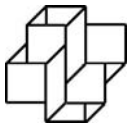

Laboratório Nacional de Computação Científica

Programa de Pós Graduação em Modelagem Computacional

CNViewer: Aplicativo Baseado em Navegador Web para Análise de Variações de Número de Cópias (CNV) do Genoma Humano

Por

Cintia Cristina Palu

PETRÓPOLIS, RJ – BRASIL

JULHO DE 2010

CNVIEWER: APLICATIVO BASEADO EM NAVEGADOR WEB PARA ANÁLISE DE VARIAÇÕES DE NÚMERO DE CÓPIAS (CNV) DO GENOMA HUMANO

Cintia Cristina Palu

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO LABORATÓRIO NACIONAL DE COMPUTAÇÃO CIENTÍFICA COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM MODELAGEM COMPUTACIONAL

Aprovada por:

_____

Prof.ª Ana Tereza Ribeiro Vasconcelos. D.Sc.
(Presidente)

_____

Prof. Jonas S. Almeida, Ph.D.

_____

Prof. Milton Ozório Moraes, D.Sc.

_____

Prof. Jauvane Cavalcante de Oliveira, Ph.D.

PETRÓPOLIS, RJ - BRASIL
JULHO DE 2010

Palu, Cintia Cristina

XXXX CNViewer: Aplicativo Baseado em Navegador Web para Análise de Variações de Número de Cópias (CNV) do Genoma Humano / Cintia Cristina Palu. Petropólis, RJ. : Laboratório Nacional de Computação Científica, 2010.

xx, yy p. : il.; 29 cm

Orientadore(s): Ana Tereza Ribeiro Vasconcelos e Jonas S. Almeida

Dissertação (M.Sc.) – Laboratório Nacional de Computação Científica, 2010.

1. ASSUNTO. 2. palavra chave. 3. palavra chave. 4. palava chave. I. II. LNCC/MCT. III. Título.

CDD XXX.XXX

epigrafe

dedicatória

Agradecimentos

Resumo da Dissertação apresentada ao LNCC/MCT como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

CNVIEWER: APLICATIVO BASEADO EM NAVEGADOR WEB PARA ANÁLISE DE VARIAÇÕES DE NÚMERO DE CÓPIAS (CNV) DO GENOMA HUMANO

Cintia Cristina Palu

Julho, 2010

Orientadora: Ana Tereza Ribeiro Vasconcelos

Co-orientador: Jonas S. Almeida

As variações de número de cópias (CNVs) caracterizam regiões genômicas cujo número de cópias pode ser diferente de dois, podendo ter como principal consequência alterações da expressão gênica. Apenas recentemente, o surgimento de ensaios moleculares em larga escala, permitiu o melhor entendimento da distribuição e ocorrência das CNVs, ampliando o interesse em seu estudo e reforçando o seu uso potencial em diagnósticos e prognósticos.

Nesta dissertação é apresentado um aplicativo para análise exploratória de CNV, capaz de comparar perfis moleculares, além de representar graficamente diversas amostras simultaneamente. Através de uma interface dinâmica, o usuário pode delimitar quaisquer regiões genômicas para a análise, e pode também exibir dados complementares às CNVs. O uso de ferramentas de análise e visualização de dados é essencial para a pesquisa, porém nem sempre está ao alcance de todo o meio científico devido a restrições de acesso ou por requerer conhecimento avançado de informática.

Portanto o desenvolvimento de interfaces amigáveis e acessíveis é essencial para a pesquisa.

Utilizando somente os recursos oferecidos pelos navegadores Web (JavaScript e HTML), o CNViewer é capaz de processar dados e executar tarefas rapidamente, com independência de servidor, pois mantêm os dados em memória durante seu uso, aperfeiçoando a interatividade com o usuário. Além disso, foi criado um módulo de exportação, que permite ao usuário salvar e recuperar suas análises, servindo também para compartilhamento de dados. O CNViewer é um aplicativo que supera os limites dos programas baseados em Web clássicos, comportando-se como um aplicativo *desktop*, com a vantagem de ser acessado diretamente, sem requerer instalação ou atualização.

Abstract of Thesis presented to LNCC/MCT as a partial fulfillment of the requirements for the degree of Master of Sciences (M.Sc.)

CNVIEWER: WEB BROWSER BASED APPLICATION FOR COPY NUMBER VARIATION ANALYSIS (CNV) OF HUMAN GENOME

Cintia Cristina Palu

Julho, 2010

Advisor: Ana Tereza Ribeiro Vasconcelos

Co-advisor: Jonas S. Almeida

The copy number variations (CNVs) characterize genomic regions whose copy number may differ from two, which could causes changes in gene expression as main consequence. Recently, the emergence of high-throughput techniques has allowed a better understanding of the distribution and occurrence of CNVs, increasing interest in their study and enhancing its potential use in diagnosis and prognosis.

This dissertation presentes an application for exploratory analysis of CNV that is able to compare molecular profiles, and graphically represent several samples simultaneously. Through a dynamic interface, user can define any genomic regions for analysis, and can also display additional data on CNVs. The use of tools for analysis and visualization of data is essential for the research, but they are not always accessible to the entire scientific world due to access restrictions or because they require advanced knowledge of computers. Therefore the development of accessible and friendly interfaces is essential for research.

Using only the features native to the Web browsers (JavaScript and HTML), the CNViewer is able to process data and perform tasks quickly regardless of server,

because it manages data in memory during its use, enhancing the user interactivity. In addition, an exportation module was created, which allows users to save and retrieve their analysis, also serving for data sharing. The CNViewer is an application that overcomes the limitations of traditional Web-based programs, behaving like a desktop application, with the advantage of being accessed directly, without requiring any installation or upgrade.

Sumário

1 Introdução 1

1.1 Objetivos 4

1.1.1 Gerais 4

1.1.2 Específicos 4

1.2 Organização da Dissertação 4

2 Biologia das CNVs 6

2.1 Ocorrência 7

2.2 Importância 9

2.3 Tipos de CNV 11

2.4 Associação Entre CNV e Doenças 12

2.4.1 CNV em câncer 15

2.5 Outras Perspectivas 18

3 Detecção de CNVs 20

3.1 Técnicas de Acesso à CNV 20

3.1.1 Contexto atual 22

3.2 Identificação de Regiões com Mesmo Número de Cópias 28

3.3 Aplicativos para Exploração 30

4 Web e Recursos Oferecidos 37

4.1 Tecnologias para Desenvolvimento Web 37

4.2 Estrutura Web 40

4.3 *Simple Sloppy Semantic Database* - S3DB 40

5 Metodologia 42

5.1 Implementação 42

5.1.1 Arquitetura do sistema 42

5.1.2 Representação de cromossomos 45

5.1.3 Acesso ao *UCSC Genome Browser* 46

5.2 Entrada de Dados 46

5.2.1 Dados-modelo 47

5.2.2 Dados mantidos com o usuário 50
5.2.3 Dados em repositório S3DB 51
5.3 Métricas de Dissimilaridade 54
5.3.1 Conceito de métrica 54
5.3.2 Métrica selecionada 55
5.4 Representação por Mapa de Calor 59
5.5 Exportação e Recuperação de Análise 60
5.6 Teste de desempenho 60
6 Resultados e Discussão 62
6.1 Fluxo de Trabalho 62
6.2 Dados de Entrada 64
6.2.1 Dados mantidos com o usuário 64
6.2.2 Dados em repositório S3DB 66
6.3 Representação dos CNVs 68
6.4 Regiões Genômicas 71
6.4.1 Explorando *UCSC Genome Browser* 72
6.5 Métricas de Dissimilaridade 74
6.6 Explorando Informações Complementares 81
6.6.1 Métricas de Dissimilaridade e Dados Complementares 82
6.7 Exportação e Recuperação da Análise 85
6.8 Segurança 86
6.9 Desempenho 88
6.9.1 Execução em Navegadores 88
6.9.2 CNViewer versus Cancer Genome Browser 91
6.10 Modelo de Implementação 93
7 Conclusões e Perspectivas Futuras 96

Lista de Figuras

Figura

2.1: Cobertura de CNV por cromossomo 8

2.2: Relação entre ocorrência de genes e CNVs no cromossomo 16 9

2.3: Exemplos de formas de CNV 11

2.4: Formas pelas quais CNV pode ocasionar doenças 12

2.5: Quantidade média de CNV somático em 27 tipos de cânceres 16

3.1: Acesso à CNVs por aCGH 23

3.2: Comparação entre três plataformas para acesso de CNV 27

5.1: Modelo de dados de CNV 50

5.2: Metodologias para acessar o conteúdo de arquivo mantido pelo usuário ... 51

5.3: Estrutura de dados de repositório S3DB compatível com CNViewer 52

5.4: Modelo de dados de CNV para repositório S3DB 53

5.5: Locais utilizados nas medidas de dissimilaridade 57

5.6: Correspondência de cores do mapa de calor 59

6.1: Diagrama de fluxo simplificado do CNViewer 63

6.2: Inserção de dados mantidos com o usuário 65

6.3: Inserção de dados em repositório S3DB 67

6.4: Representação gráfica 68

6.5: Configuração do gráfico 70

6.6: Seleção de regiões-alvo para análise 71

6.7: Consulta ao *UCSC Genome Browser* 73

6.8: Distância Euclidiana em pontos de quebra 76

6.9: Distância Euclidiana ponderada por tamanho 79

6.10: Distância Euclidiana ponderada por sondas 80

6.11: Identificação dos dados Complementares 81

6.12: Módulo de Exportação 86

6.13: Tempo médio de plotagem do Cancer Genome Browser e CNViewer..... 92

6.14: Modelo de transferência de dados para aplicativos baseados em navegador Web. 94

Lista de Tabelas

Tabela

2.1: Fenótipos ilustrativos de patologias associadas à CNV 13

3.1: Aplicativos para Visualização e Análise de CNV 32

6.1:Tempo médio de execução 89

Lista de Siglas e Abreviaturas

- aCGH - *Comparative Genomic Hybridization*
- API – *Application Programming Interface*
- AIDS – *Acquired Immunodeficiency Syndrome*
- BAC – *Bacterial Artificial Chromosome*
- CBG – *Cancer Genome Browser*
- CBS – *Circular Binary Segmentation*
- CGH –*Comparative Genomic Hybridization*
- CNV – *Copy Number Variation*
- CGI – *Common Gateway Interface*
- CSS – *Cascading Style Sheets*
- DE – Distância Euclidiana
- DNA – *DeoxyriboNucleic acid*
- DOM – *Document Object Model*
- EUA – Estados Unidos da América
- FISH – *Fluorescence In Situ Hybridization*
- HIV – *Human Immunodeficiency Virus*
- HTML – *HyperText Markup Language*
- HTTP – *HyperText Transfer Protocol*
- JMV – *Java Virtual Machine*
- kb – quilobase (1.000 pares de base)
- Mb – Megabase (1.000.000 pares de base)
- MPCBS – *Multi-Platform Circular Binary Segmentation*

- mRNA – *messenger RiboNucleic Acid*
- NC – Número de Cópias
- NIH – *National Institutes of Health*
- pb – pares de bases
- PCR – *Polymerase Chain Reaction*
- RNA – *RiboNucleic Acid*
- S3DB – *Simple Sloppy Semantic Database*
- SNP – *Single Nucleotide Polymorphism*
- STR – *Short Tandem Repeat*
- TLS – *Transport Layer Security*
- URI – *Uniform Resource Identifier*
- URL – *Uniform Resource Locator*
- VNTR – *Variable Number of Tandem Repeats*

Capítulo 1

Introdução

As pesquisas na área biológica estão fortemente interligadas à tecnologia, principalmente nas análises médicas e moleculares. Os avanços alcançados permitem cada vez mais obter dados com maior velocidade e precisão, ampliando as possibilidades de pesquisa. O volume de dados gerados trás como consequência a necessidade de novos métodos para análise e interpretação, sendo necessário otimizar a interpretação dos dados biológicos. Deste modo, há uma demanda constante de desenvolvimento de novas metodologias para integração de dados e sistematização das análises, bem como aplicativos que permitam a manipulação, processamento e representação deste conjunto de informações.

Foi o aprimoramento das pesquisas genômicas que levou a detecção de variações do número de cópias (CNV), alterações que influenciam a variabilidade humana e de outras espécies. CNVs são regiões genômicas com número de cópias (NC) variável, que consistem de segmentos que podem estar repetidos ou até mesmo ausentes. Acreditava-se que organismos diplóides normais portassem apenas duas cópias das sequências de DNA, e que alterações no NC estariam associadas a patologias, tais como síndromes e desenvolvimento de câncer. Nos últimos anos foram criados recursos capazes de analisar o genoma com melhor precisão e foi descoberto que as CNVs ocorrem com maior frequência do que se esperava e que, na verdade, são uma variação presente em todos os indivíduos.

Essas descobertas ampliaram o campo de análises de CNV, buscando compreender suas origens e seu papel biológico. Em consequência, pesquisas em diferentes domínios vêm sendo realizadas, tais como estudos de associação com doenças e perfis clínicos, além do estudo em indivíduos normais. A CNV possui um grande potencial para ser utilizada em diagnósticos e prognósticos, especialmente com o avanço da medicina personalizada.

A pesquisa das CNVs só é possível com os recursos da bioinformática e sendo um alvo de estudos relativamente recentes, muito há para ser explorado no contexto de ferramentas para análise de CNV. Inúmeros recursos já foram desenvolvidos, mas ainda existem características a serem aprimoradas. As áreas a serem desenvolvidas são várias, desde as técnicas para detecção de variantes até metodologias para análise dos resultados finais. O presente trabalho explorou o ambiente de navegadores Web para desenvolver um novo tipo de ferramenta, buscando solucionar problemas de acessibilidade, portabilidade e visualização.

Os principais pontos trabalhados nesta dissertação foram desenvolver uma ferramenta de acesso irrestrito, que também permitisse o compartilhamento de dados e principalmente, fornecesse recursos para a exploração simultânea do conjunto de dados. O problema de acessibilidade foi abordado desenvolvendo um aplicativo Web livre, o qual é acessado via URL através de qualquer máquina com acesso à internet. A segunda etapa, a portabilidade, envolveu o uso de repositório Web de arquivos e a criação de um módulo de exportação, oferecendo duas formas distintas de armazenar dados e inclusive, salvar as análises. A visualização de toda a população de amostras foi atingida com a representação dos segmentos e NC por mapa de calor. Foram adicionadas a este gráfico, ferramentas que permitem comparações entre amostras e combinações ilimitadas de regiões genômicas. Não obstante, considerando a

importância de estudos associados, foi criado um gráfico contendo dados complementares, paralelos às CNVs.

É importante ressaltar que muitos dos aplicativos não permitem a visualização do conjunto de dados, gerando gráficos individuais, ou perfis do conjunto de amostras. Além disso, em grande parte, estes programas não exibem dados complementares relativos aos espécimes. Um dos desafios de permitir a visualização de conjuntos de CNVs está justamente no volume de dados a ser veiculado simultaneamente, exigindo o processamento imediato de todas as amostras. Ainda, enquanto um aplicativo *desktop* (programas executados localmente) costuma processar os dados no mesmo local onde é executado (máquina do usuário), um aplicativo Web por padrão usa um servidor remoto para processá-los, o que representa um consumo de tempo maior devido à transferência de dados. Assim, visando evitar transferências de dados repetitivas, o CNViewer foi implementado de tal forma a ser processado inteiramente na máquina cliente, isto é, suas funções são executadas no navegador, independentes de servidor.

Até recentemente os navegadores Web não poderiam suportar tais funcionalidades, com o desenvolvimento da Web 2.0, passou a ser possível o processar os dados no navegador. A internet, bem como as demais tecnologias relacionadas à informática, sofre um desenvolvimento acelerado e existe um grande investimento em seu aperfeiçoamento. Assim, existem navegadores Web capazes de suportar o armazenamento e processamento de dados do porte das CNVs, capacidades que certamente serão expandidas.

1.1 Objetivos

1.1.1 Gerais

Um requerimento básico para pesquisas genômicas em larga escala é a visualização de dados, a qual facilita a análise, interpretação estatística e identificação de padrões de interesse biológicos. Considerando estes aspectos, o presente trabalho tem como objetivo explorar os recursos Web para produzir um aplicativo de livre acesso, para análise exploratória de CNV, capaz de integrar dados clínicos e que seja acessível a pessoas sem conhecimento avançado de informática.

1.1.2 Específicos

Desenvolver um programa:

- voltado a usuários pouco familiarizados com bioinformática;
- para visualização interativa de CNV
- que possibilite inserção de dados complementares;
- que seja flexível quanto à metodologia de obtenção de dados brutos de CNV;
- capaz de armazenar análises;
- de fácil acesso
- de fácil manutenção.

1.2 Organização da Dissertação

O Capítulo 2 apresenta uma revisão sobre os principais aspectos da biologia das CNVs, levando em consideração as recentes descobertas sobre o assunto. Como o CNViewer oferece suporte para o processo de análise das CNVs, é importante

compreender como são obtidos os dados de CNVs e quais são as ferramentas de análise já existentes, assuntos abordados pelo Capítulo 3.

O aplicativo é desenvolvido utilizando recursos Web, os quais são descritos brevemente no Capítulo 4. Os detalhes específicos de sua implementação são descritos no Capítulo 5, seguido pelos resultados e discussão no Capítulo 6 e conclusões e perspectivas futuras, no Capítulo 7.

Capítulo 2

Biologia das CNVs

A detecção de ganhos e perdas de segmentos genômicos em tumores é possível desde 1992, porém somente em 1998 a técnica aCGH (do inglês *array comparative genomic hybridization*) permitiu acessar estes dados em larga escala e com melhor resolução. Ainda assim, foi em 2004 que a perspectiva do estudo dos CNVs começou a ser alterada, passando a serem reconhecidos como parte da variabilidade genética normal do ser humano. Nos últimos cinco anos muitos avanços abrangendo este tópico foram alcançados (Lee e Scherer, 2010), mas há um longo caminho a ser cursado até se conhecer os mecanismos que envolvem as CNV e suas consequências.

As CNVs foram inicialmente caracterizadas como segmentos genômicos com número de cópias variável, que possuíam tamanho igual ou maior que 1 quilobase (kb) (Feuk et al., 2006). Com mudanças na precisão das técnicas de detecção de CNV, foi possível observar eventos de variações de NC em regiões mais curtas e assim muitos estudos passaram a incluir em suas análises de CNV sequência menores que 1 kb (Conrad et al., 2010).

As CNVs podem ser herdadas e assim estarem presentes em praticamente todas as células do organismo, ou podem ter origem somática, variações adquiridas que vão estar limitadas a um conjunto específico de células (Lee e Scherer, 2010). O termo CNV pode ser usado de forma abrangente, porém alguns autores preferem denominar as CNV adquiridas “alterações do número de cópias”, ou mesmo “variações do número de

cópias somáticas". O presente trabalho refere-se às CNVs de humanos, incluindo alterações adquiridas e variações herdadas. Os diferentes NC de uma mesma região são denominados "variantes".

Catalogar a natureza e o padrão da variação do genoma da população em geral é fundamental para o entendimento da diversidade fenotípica humana (Pinto et al., 2007). CNVs são comuns no genoma humano (Carter, 2007, Stranger et al., 2007) e podem ter consequências drásticas no fenótipo (Stranger et al. 2007) motivando ainda mais sua pesquisa.

Este capítulo traz uma visão geral das variações do número de cópias gênicas, revisando as mais recentes e relevantes descobertas. Visa apresentar as principais perspectivas e importância do estudo das CNVs. São apresentas evidências da relação das CNVs com doenças de forma resumida, permitindo contextualizar posteriormente a utilização do CNViewer, o qual provê ferramentas para a exploração combinada de fenótipos e CNVs.

2.1 Ocorrência

As CNVs ocorrem em uma fração significativa do genoma humano (Figura 2.1), apesar de representarem uma porção pequena do conteúdo genético individual (Conrad et al., 2006, Cooper et al., 2007, Wong et al., 2007). Elas são comuns e de distribuição heterogênea no genoma (Redon et al., 2006). As variantes costumam ter frequência baixa, elas são raras em sua maioria, mas a cobertura de parte significativa do genoma é garantida pela diversidade de variantes (de Smith et al., 2007).

É importante considerar que, as estatísticas referentes à presença de CNV não são precisas devido a limitações das técnicas (Cooper et al., 2007). O tamanho e distribuição das CNVs encontradas em cada estudo dependem da tecnologia utilizada para analisá-las e sua variedade é limitada não apenas ao número de indivíduos

acessados, mas também a diversidade de origens étnicas (de Smith et al., 2007). Em um estudo envolvendo 2.026 indivíduos norte americanos saudáveis (65,2% caucasóides e 34,2% afro-descendentes), Shaikh et al., (2009) encontraram 3.272 regiões com CNV recorrente, compreendendo 5,9% do genoma. Esta porcentagem tende a aumentar se incluídos mais indivíduos e um maior número de etnias.

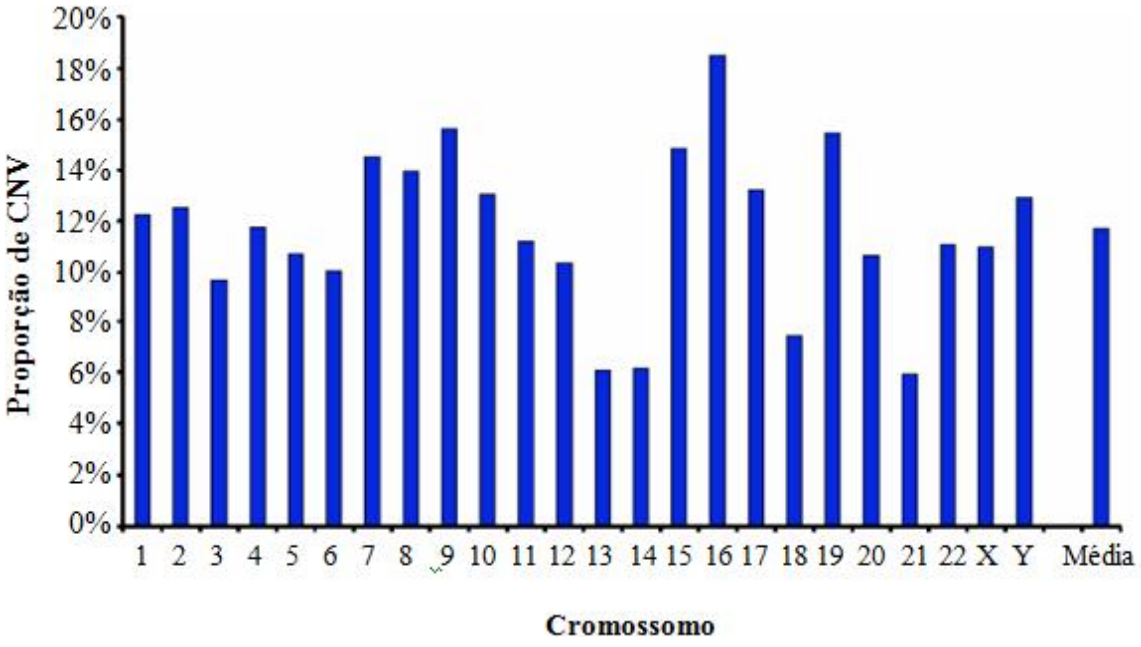

Figura 2.1: Cobertura de CNV por cromossomo (adaptado de Redon et al., 2006).
Dados do total de CNV encontrados em estudo de quatro populações, num total de 270 amostras.

As regiões portadoras de CNVs coincidem com a presença de genes (Figura 2.2), estima-se que 2,7% delas contêm éxons, em comparação com uma frequência média de 2,1% no genoma inteiro (Cooper et al., 2007, Wong et al., 2007). Em estudo mais recente, usando técnicas com melhor precisão para de identificação de CNVs, Conrad et al. (2010) encontraram 38,8% das CNVs coincidindo com 13,4% dos genes registrados no RefSeq, porém neste estudo foram contabilizados também segmentos maiores que 443 pares de base.

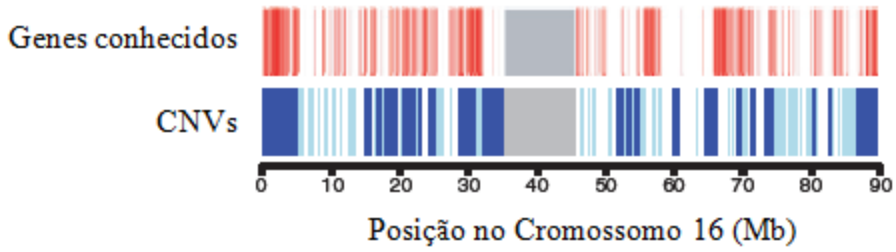

Figura 2.2: Relação entre ocorrência de genes e CNVs no cromossomo 16 (adaptado de Cooper et al., 2007). O eixo *x* corresponde às coordenadas da *hg17 assembly*; a heterocromatina está representada em cor cinza. A primeira linha contém em vermelho os éxons de todos os genes conhecidos em 2007. A segunda linha apresenta os CNVs provenientes do levantamento realizado no referente trabalho, sendo que em azul escuro estão CNVs com frequência maior que 3%, e em azul claro os demais.

As CNVs podem conter partes de genes, muitos genes e elementos regulatórios ou mesmo não conter elementos conhecidos. Assim, sua ocorrência pode interromper regiões funcionais do genoma, ou gerar novas combinações de sequências, produzindo novas moléculas (Lee e Scherer, 2010).

2.2 Importância

As CNVs alteram os transcritos gênicos e sequências codificadoras dos mRNAs (Conrad et al., 2010). Estima-se que as CNVs contribuam em 17,7% para as alterações de expressão gênica ocasionadas por variações genéticas – a maior contribuição é feita pelos polimorfismos de nucleotídeo único (SNPs), contabilizando 83,6% (Stranger et al., 2007).

Considerando-se que existem CNVs frequentes sobrepondo regiões de transcrição de microRNAs, bem como genes relacionados ao sistema sensorial, sistema imune, a doenças, adesão celular e proteínas estruturais, este tipo de variação deve estar envolvida com determinação de fenótipo, predisposição a doenças e etiologia de patologias (Freeman et al., 2006, Cooper et al., 2007, de Smith et al., 2007, Wong et al.,

2007). Dentre os genes associados a doenças destacam-se oncogenes, genes supressores de tumor entre outros envolvidos na susceptibilidade e progressão do câncer (Wong et al., 2007). Existem também genes associados à diabetes mellitus, distrofia muscular, atrofia muscular espinhal, esquizofrenia, Alzheimer, Parkinson, dependência de nicotina, obesidade e doenças coronárias que se encontram em regiões de CNV (Wong et al., 2007). A seção 2.4 trata dos mecanismos e patologias conhecidas relacionadas a esta variação.

As CNVs são responsáveis não apenas pelas variações interindividuais, mas contribuem também para as diferenças entre os tecidos do mesmo organismo. Chaignat et al. (2009) comprovaram que, em camundongos (gênero *Mus*), os genes com maior variação de expressão entre tecidos e entre espécies ocorriam em regiões de CNV. Os genes contidos por essas regiões eram expressos em menos tecidos – neste estudo foram comparadas a expressão de genes autossômicos do cérebro, pulmões, coração, fígado, rins e testículos. Os autores evidenciaram também que a influência das CNVs não é restrita aos elementos de seu conteúdo: genes distantes até 450 kb também apresentam maior variação de expressão. Assim CNVs afetam a transcrição alterando a dosagem dos genes e causando efeitos de longa distancia nas regiões vizinhas (Henrichsen et al., 2009).

Portanto, interferindo de tantas formas no genoma, as CNVs podem contribuir substancialmente para a diversidade humana – normal ou patológica, bem como divergências entre espécies.

Estudos genéticos comumente analisam os perfis de expressão gênica, porém a instabilidade dos RNAs restringe a pesquisa em alguns tipos de amostra. A análise de CNV permite o estudo de um maior número de amostras pois o DNA é uma molécula mais fácil de ser isolada e mais estável (Lockwood et al., 2006).

2.3 Tipos de CNV

As formas mais simples de CNVs são as deleções (Figura 2.3 A), duplicações consecutivas (Figura 2.3 B) e inserções de duplicações (Figura 2.3 C). As variações mais complexas são ocasionadas por combinações destes eventos ou até mesmo translocações (Figura 2.3 E). Em uma população, pode ser encontrada mais de uma variante, isto é, o número de segmentos repetidos pode variar entre indivíduos (Figura 2.3 D) e por isso, em alguns contextos são denominadas "alelos" (Lee e Scherer, 2010).

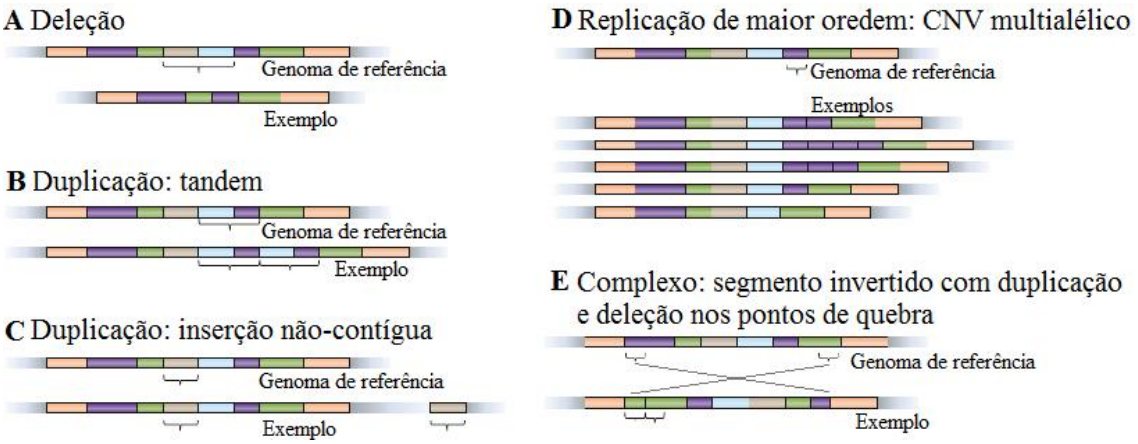

Figura 2.3: Exemplos de formas de CNV (adaptado de Lee e Scherer, 2010). Um trecho do genoma de referência é apresentado para identificação dos variantes. As diferentes cores simbolizam segmentos de DNA com sequências idênticas. Pode ocorrer (A) perda de segmentos (cinza e azul), duplicação de trechos (em marrom) em tandem (B) ou em regiões distintas do genoma (C). Os segmentos (roxo) também podem ocorrer em quantidades diferentes em genomas distintos ou ainda as CNVs podem ser decorrentes de uma série de eventos (E). O exemplo mostra uma inversão, somada à duplicação (parte do segmento verde) e deleção (segmento lilás).

A maioria das CNVs é decorrente de uma a três mutações. A ocorrência de regiões próximas contendo CNVs, leva em muitos casos a interpretá-las como um único CNV, aparentando ser uma variação com origem complexa (McCarroll et al., 2008). Estima-se que a taxa de mutação dos CNVs seja aproximadamente 10^8, semelhante à dos SNPs (Pinto et al., 2007).

2.4 Associação Entre CNV e Doenças

As primeiras observações de CNV foram realizadas há 50 anos, quando houve as primeiras descobertas de aneuploidias ocasionando síndromes. As células aneuplóides são caracterizadas por apresentarem um número de cromossomo diferente

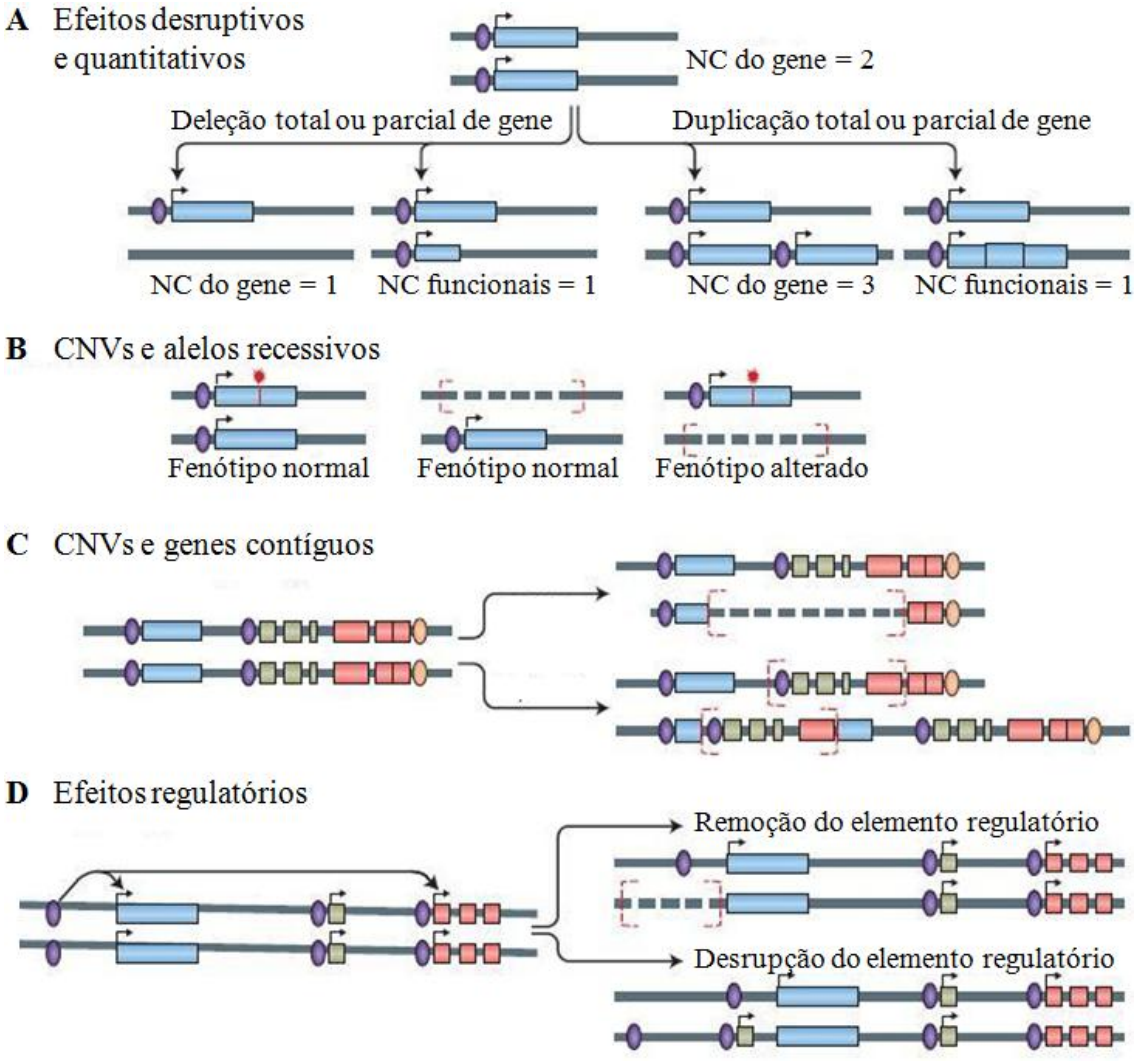

Figura 2.4: Formas pelas quais CNV pode ocasionar doenças (adaptado de Lee e Scherer, 2010). (A) As CNVs podem alterar o número de cópias funcionais dos genes por duplicações/deleções totais os parciais dos genes. (B) Um alelo recessivo (marca vermelha) pode passar a ser expresso caso seu homólogo seja deletado. (C) A deleção de genes contínuos pode eliminar (verde) ou interromper (azul e vermelho) genes funcionais, enquanto duplica (verde) outro gene. (D) A regulação gênica pode ser alterada pela deleção de regiões regulatórias (roxo) ou pela sua interrupção.

do normal da espécie. Apenas recentemente verificou-se que a presença de CNVs não necessariamente estava relacionada a um perfil patológico e que podiam ocorrer em sequências genômicas menores (Lee e Scherer, 2010).

Tabela 2.1: Fenótipos ilustrativos de patologias associadas à CNV (adaptado de Lee e Scherer, 2010). Estão citadas algumas das patologias e respectivas regiões genômicas nas quais a presença de CNV provoca alterações desencadeando ou participando da ocorrência da patologia. (*) Os números citados como referência correspondem à bibliografia do artigo do qual esta tabela foi extraída.

(continua)

Característica genotípica da CNV	**Patologia**	**Referências***
Apenas um gene ocasionando fenótipo (nome do gene)	• Distrofia muscular de Duchenne/Becker (*DMD*)	173, 174
	• Neurofibromatose 1 (*NF1*)	175 - 177
	• Esclerose tuberosa (*TSC1* ou *TSC2*)	87, 133
	• Síndrome de Soto (*NSD1*)	35, 178, 179
	• Síndrome de CHARGE (*CHD7*)	143
	• Doença de Pelizaeus-Merzbacher (*PLP1*)	28, 180
	• Início precoce do Mal de Alzheimer (*APP*)	181 - 183
	• Autismo 22q (*SHANK3*)	146, 184 - 186
	• Psoríase (*LCE3C/LCE3B*)	58, 59
	• Aniridia (*PAX6*)	88, 187
	• Síndrome da plolisindactelia trifalangel (*ZRS*)	188, 189
	• Doença de Crohn (*IRGM*)	126
	• Predisposição à doença de Crohn (*DEFB4*)	190, 191
	• Predisposição à autoimunidade sistêmica (*FCGR3B*)	124, 192, 193 194
	• Mal de Parkinson (*SNCA*)	
	• Susceptibilidade a HIV/AIDS (*CCL3L1*)	127, 128, 130
	• Susceptibilidade a doença de Kawasaki (*CCL3L1*)	195
	• Predisposição a artrite reumatóide (*CCL3L1*)	130

Tabela 2.1: Fenótipos ilustrativos de patologias associadas à CNV (adaptado de Lee e Scherer, 2010).

(conclusão)

Característica genotípica da CNV	**Fenótipo Ilustrativo**	**Referências***
Múltiplos genes potencialmente envolvidos no fenótipo (Coordenadas cromossômicas)	• Síndrome velocardiofacial / Síndrome de DiGeorge (deleção de 22q11,2)	74, 75, 196, 197
	• Síndrome de Williams-Beuren (duplicação de 7q11,23)	36, 37, 67, 96, 98, 100, 198, 199
	• Síndrome de microdeleção/microduplicação de 17q21,3	32-34, 200, 201
	• Neuroblastoma (1q21,1)	61
	• Síndrome da microdeleção de 15q13,3	105-109
	• Síndrome da microdeleção de 15q41q42	202
	• Síndrome da microdeleção de 16p11,2-p12,2	203
	• Síndrome(s) da microdeleção/microduplicação de 16p11,2	145-149, 204
	• Síndrome(s) da microdeleção/microduplicação de 3q29	205, 206
	• Síndrome de Potocki-Lupski (duplicação de 17p11,2)	207
	• Síndrome de WAGR (11p13)	88
	• Autismo	47, 118, 144-147, 208-212
	• Desordem Bipolar	213-215
	• Esquizofrenia	47, 89, 106, 113, 149, 214-218
	• Degeneração macular relacionada à idade	219, 220
	• Teratologia de Fallot	90, 221
	• Síndrome da ausência do rádio trombocitopenia (1q21,1)	77, 78
	• Atrofia muscular espinhal (*SMN1/SMN2*)	79, 222, 223
	• Síndrome de Silver-Russell (duplicação de 11p15)	80
	• Síndrome de Rubinstein-Taybi (*CREBBP, EP300*)	84-86
	• Esclerose tuberosa (*TSC1, TSC2*)	87, 133
	• Aniridia (*11p13*)	88
	• Síndrome de Li-Fraumeni	60, 133
	• Neurofibromatose	224

As CNVs influenciam na propensão a doenças de diferentes formas, podendo interferir de forma quantitativa ou disruptiva (Figura 2.4 A). Genes cuja expressão é determinada pelo número de cópias (dependentes de dosagem) terão sua expressão diretamente alterada pelas CNVs, sendo que cópias completas extras aumentam a transcrição, enquanto a inserção de segmento no gene e deleções completas ou parciais levam à inativação da cópia do gene afetado (Figura 2.4 A). A deleção ou inativação de um gene pode também levar à expressão de um homólogo mutante (Figura 2.4 B), o qual pode gerar moléculas com funções alteradas ou não funcionais. A predisposição a doenças pode ser determinada também quando deleções ou duplicações alteram genes (Figura 2.4 C) e as regiões de regulação gênica (Figura 2.4 D) (Lee e Scherer, 2010).

2.4.1 CNV em câncer

A ocorrência do câncer está relacionada com alterações genéticas da linhagem germinativa que, assim como mutações somáticas, podem influenciar na propensão a doenças (Futreal et al., 2004, Stuart e Sellers, 2009). Durante seu desenvolvimento uma série de novas alterações genéticas é acumulada, as quais podem modificar os níveis de expressão dos genes envolvidos. Dentre as alterações desencadeadas pela progressão do tumor encontram-se CNVs e outras deleções/amplificações de segmentos genéticos menores, bem como mutações de ponto e translocações (Stuart e Sellers, 2009).

Os perfis de ganhos e perdas de segmentos genômicos são particulares de cada tipo de câncer (Figura 2.5), porém algumas alterações são comuns entre eles, indicando que diferentes tipos de tumores podem ter alguns mecanismos moleculares semelhantes envolvidos no seu desenvolvimento (Mertens et al., 1997). Distinguir entre mutações causadoras de câncer daquelas que são consequência de sua progressão é um desafio complexo (Beroukhim et al., 2010), mas alterações em regiões específicas já permitiram a descoberta de oncogenes, sugerindo inclusive técnicas terapêuticas (Stuart e Sellers,

2009). Em estudo de câncer de ovário, observou-se correlação entre o NC e nível de expressão gênica, sendo que os genes de loci com alteração de NC frequente apresentavam maior correlação com sobrevivência que quaisquer outros (Kingsley et al. 2007).

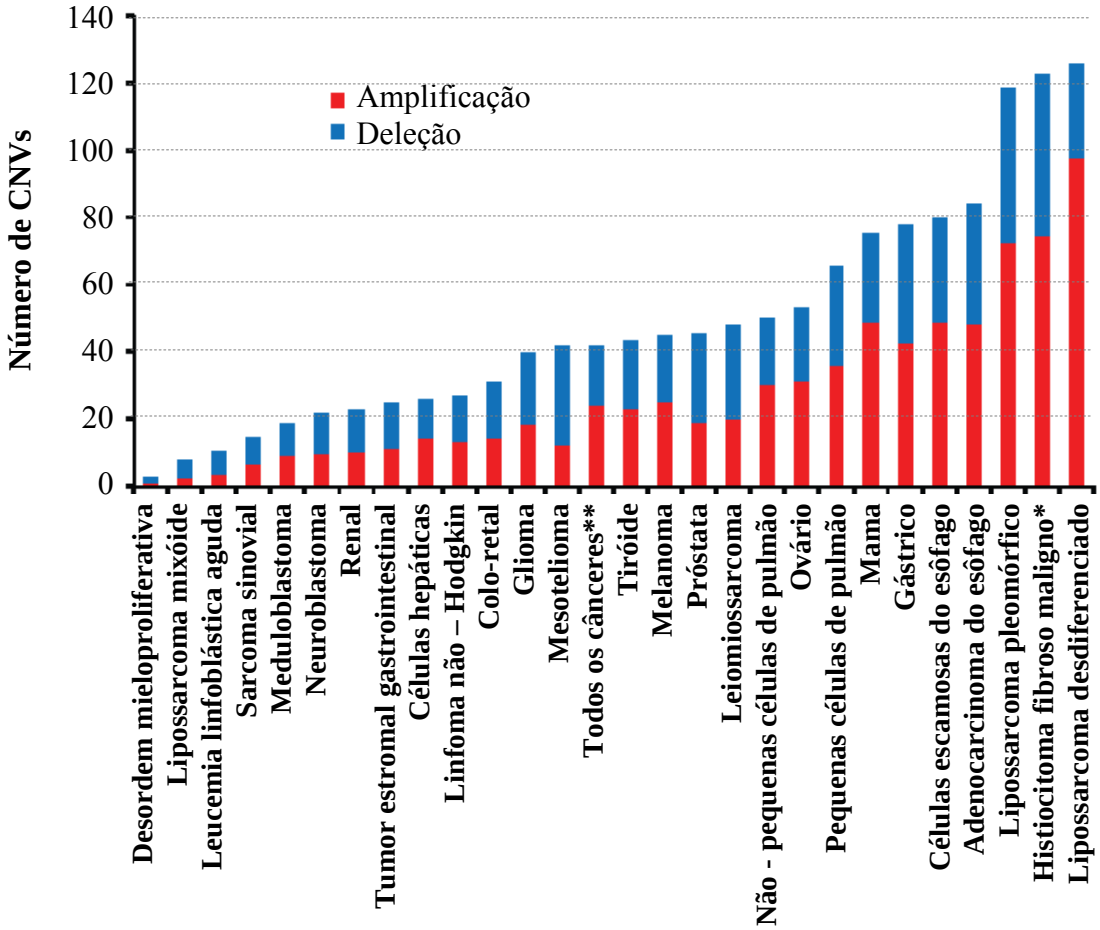

Figura 2.5: Quantidade média de CNV somático em 27 tipos de cânceres (adaptado de Beroukhim et al., 2010). Os tumores representados individualmente são aqueles para os quais havia mais de 20 amostras no estudo, exceto pelo histocarcinoma fibroso maligno (*), com apenas quatro amostras. (**) Além dos tumores representados no gráfico, foram contabilizados dados de outras 161 amostras, as quais incluíam outros tipos de tumores hematopoiéticos, sarcomas, epiteliais e neurais, totalizando a média de CNV somático de 3.131 amostras de câncer.

Outro aspecto do câncer é o desenvolvimento de novos tumores após o aparecimento de um primeiro. Muitas vezes o tumor adquire a capacidade de colonizar

regiões distantes no corpo do paciente, gerando tumores secundários, evento conhecido como metástase, em outros casos, tumores novos (primários) passam a se desenvolver. A distinção da origem de um câncer é parte importante para determinar a abordagem de tratamento, estabelecendo qual metodologia é mais apropriada, tratamento local (cirurgia) ou sistêmico por exemplo. A investigação da origem pode ser feita comparando o perfil das mutações entre dois tumores do mesmo paciente, inclusive avaliando-se exclusivamente o padrão das CNVs (Ostrovnaya et al., 2010).

A CNV tem papel importante no surgimento e progressão de neoplasias (Heidenblad et al. 2005), em média 16% do genoma de uma amostra de tumor sofre deleção e em torno de 17% duplicação, enquanto a média de um tecido normal é respectivamente 0,35% e 0,1% (Beroukhim et al. 2010). Regiões com alto número de cópias estão correlacionadas com a superexpressão gênica em grande parte dos casos (Heidenblad et al. 2005), assim despertam interesse não apenas para compreender a tumorigênese, mas também como possíveis alvos para terapias com drogas (Stuart e Sellers 2009). A inspiração para o desenvolvimento de terapia baseia-se não apenas em genes super expressos, mas também na observação de que a baixa expressão de certas proteínas influência a susceptibilidade ou resistência a tratamentos (Stuart e Sellers 2009).

PERFIL DE DISTRIBUIÇÃO GENÔMICA DE CNV EM CÂNCER

Em um estudo referência de CNV em câncer, Beroukhim et al. (2010) analisaram 3.131 amostras cancerosas e 26 tipos de tumor, caracterizando o perfil comum entre eles. Em sua maioria, as CNVs são focais (segmentos curtos) ou tão grandes quanto o tamanho do braço do cromossomo em que se encontram, cobrindo 10 e 25% do genoma respectivamente, com apenas 2% de sobreposição. As CNVs focais

têm em média 1,8 megabases (variando entre 0,5 kb a 85 Mb), e sua frequência é inversamente proporcional ao seu tamanho.

Notoriamente, os braços cromossômicos têm diferentes tendências à ocorrência de deleções e duplicações, porém esta característica varia conforme o tipo de câncer, possuindo um padrão mais parecido em cânceres de linhagens semelhantes.

Em busca dos genes atuantes no processo desencadeador da neoplasia, as variações focais são os alvos mais adequados. Beroukhim et al. (2010) selecionaram 158 regiões independentes com CNVs significativos, das quais 76 eram amplificações e 82 deleções. Em 25 das 76 regiões foram encontrados oncogenes comprovadamente ativados por amplificação. As deleções continham menos genes, das quais apenas nove possuíam genes supressores de tumor e duas eram relacionadas à translocações que geram oncogenes.

2.5 Outras Perspectivas

O padrão de variação entre indivíduos oferece uma extensa fonte de informações para entender a história de uma espécie, permitindo a inferência de migrações e interpretação das adaptações ocorridos ao longo do tempo (Jakobsson et al. 2008). Estes aspectos, tratados pela genética de populações, ainda estão sendo explorados sob a perspectiva das CNVs (de Smith et al. 2007) e somam-se aos diversos fatores que colocaram as CNVs em evidência.

Além do aspecto histórico, é essencial caracterizar e catalogar as CNVs em indivíduos saudáveis, para haver uma referência e para permitir a posterior identificação de aberrações envolvidas em distúrbios genéticos, bem como reconhecimento das variáveis neutras (de Smith et al. 2007), Shaikh et al., 2009). Conhecendo as regiões específicas envolvidas em determinado fenótipo, será possível criar plataformas de acesso a CNVs com alvos específicos (Lockwood et al. 2006).

Os polimorfismos genéticos são a chave do estudo de genética de populações, pois pressupõem que os padrões genéticos de um indivíduo serão perpetuadas em parte para os seus descendentes. Assim estudando as aquisições de mutações, pode-se reconstruir a história da espécie. Conhecendo a variabilidade genética e relacionando-a com aspectos fenotípicos, pode-se descobrir tendências e predisposições a doença que um indivíduo possui, fatores que em muitos casos afetam inclusive o tratamento.

O polimorfismo genético mais estudado em genética de populações é o SNP. Estudos da população mundial sob a perspectiva de CNV, com análise paralela de SNP (Jakobsson et al. 2008), mostraram similaridade parcial entre a estrutura populacional inferida por CNVs em relação à inferida por SNPs. Assim, as características da distribuição destas variações são distintas quando se trata de populações evolutivamente próximas.

Capítulo 3

Detecção de CNVs

3.1 Técnicas de Acesso à CNV

Alterações cromossômicas são observadas desde o início da citogenética humana (Carter, 2007, Lee e Scherer, 2010), tradicionalmente utilizando a técnica de coloração de cromossomos. Em algumas metodologias, como o bandeamento-G, o padrão de bandas coradas no cariótipo é usado para a identificação dos cromossomos e suas partes, permitindo a visualização de grandes alterações genômicas (Monni e Hautaniemi, 2009). A hibridização fluorescente *in situ* (FISH) é uma técnica de coloração surgida em 1980 (Lee e Scherer, 2010), na qual sondas com marcadores fluorescentes ligam-se às sequências complementares dos cromossomos (hibridização), permitindo a detecção de alterações em nível gênico. Nesta técnica, é necessária a escolha de regiões específicas para serem estudadas, pois poucas regiões podem ser analisadas simultaneamente (Dhawan e Padh, 2009).

As técnicas de coloração utilizam células em processo de divisão celular (mitose), pois os cromossomos estão mais condensados e, para serem coradas na fase correta da mitose, as células precisam ser cultivadas *in vitro*. Esta característica é um fator limitante, pois alguns tumores, principalmente os sólidos, são de difícil cultivo (Dhawan e Padh, 2009, Monni e Hautaniemi, 2009).

No início dos anos 90 a técnica CGH revolucionou a pesquisa de câncer, por não necessitar de cultivo celular, permitindo o acesso às alterações de NC em uma maior

gama de tumores – inclusive os sólidos, nos quais as alterações de NC são mais comuns. A CGH tem como princípio a marcação molecular diferencial dos DNAs da amostra teste e de uma referência, seguida de hibridização em cromossomos normais (por exemplo, a amostra teste é corada com vermelho fluorescente e a referência em verde fluorescente). A hibridização consiste da ligação das sequências do DNA corado com suas complementares pré-fixadas em uma lâmina. Os DNAs teste e referência estarão competindo pela ligação e as variações de cópias serão identificadas pelas cores, no exemplo sugerido, os segmentos do DNA teste que estiverem amplificados apresentarão maior intensidade de cor vermelha, enquanto que as deleções estarão evidenciadas em verde (Monni e Hautaniemi, 2009).

A resolução oferecida pela CGH está entorno de 2 a 10 Mb, identificando apenas grandes alterações no genoma e foi muito utilizada para acessar genes candidatos a estarem envolvidos na tumorigênese. Estes estudos partem do princípio que deleções podem conter genes envolvidos na manutenção celular, sinalização de dano, apoptose e regulação do ciclo celular, enquanto os genes amplificados estariam relacionados ao crescimento celular e inibição de apoptose (Monni e Hautaniemi 2009). Anos mais tarde, durante seu desenvolvimento, o Projeto do Genoma Humano disponibilizou bibliotecas genômicas, permitindo aprimorar a técnica de CGH, iniciando a pesquisa em larga escala de CNV (Carter 2007).

O surgimento de técnicas do tipo arranjo, em especial aCGH abriu as portas para o estudo das CNVs ao longo do genoma, pois permitiu a quantificação da presença de segmentos genômicos menores. Os primeiros arranjos utilizavam como sondas segmentos clonados em cromossomos artificiais bacterianos (BACs), os quais eram relativamente grandes. Com o tempo, o tamanho dos segmentos diminuiu e o número de sondas aumentou, permitindo maior cobertura do genoma com mais precisão. Arranjos

inicialmente desenhados para detecção de SNPs também passaram a ser utilizados, quantificando trechos genômicos com base na intensidade de sinal das sondas (Lee e Scherer 2010).

As sondas contidas nos arranjos resultam em uma coleção de pontos com intensidade de sinais diversa. Programas específicos para cada plataforma convertem estes sinais em NC e posteriormente mapeiam no genoma. O reconhecimento de variações no NC é possível após análise estatística (Lockwood et al., 2006).

O sequenciamento completo também permite a identificação de CNVs, através da comparação de resultados contra um genoma de referência. Porém seus custos são mais elevados, além de consumir maior tempo, impedindo sua utilização em larga escala (Lockwood et al., 2006, Lee e Scherer, 2010). O consumo de tempo e gastos, também impedem o uso em larga escala de PCR quantitativa, hibridização de sondas amplificáveis multiplex e hibridização alelo-específica dinâmica (Carter 2007).

3.1.1 Contexto atual

Com o Projeto Genoma Humano, foi possível criar sondas de sequências cuja posição genômica era conhecida e ligá-las em lâminas de vidro (microarranjos) para fazer a hibridização das amostras teste e referência, devidamente marcadas, através do mesmo princípio utilizado na CGH. A este ensaio denomina-se aCGH ou microarranjo CGH. Sua sensibilidade é maior que a da CGH, variando de acordo com o tamanho e número de sondas contidas no microarranjo (Carter 2007). Com intuito de atingir melhor precisão, replicas de uma mesma sonda estão presente no microarranjo. Após hibridização, a lâmina é escaneada e são medidas as fluorescências relativas de cada sonda, média ou mediana entre réplicas e em seguida elas são mapeadas no genoma, permitindo uma quantificação dos segmentos ao longo do genoma (Figura 3.1) (Carter, 2007, Chari et al., 2007). Em 1997 foi publicado o primeiro artigo no qual a aCGH foi

aplicada. Devido as suas qualidades e eficiência, esta técnica se popularizou rapidamente, abrindo as portas para a análise de CNVs em larga-escala (Carter 2007).

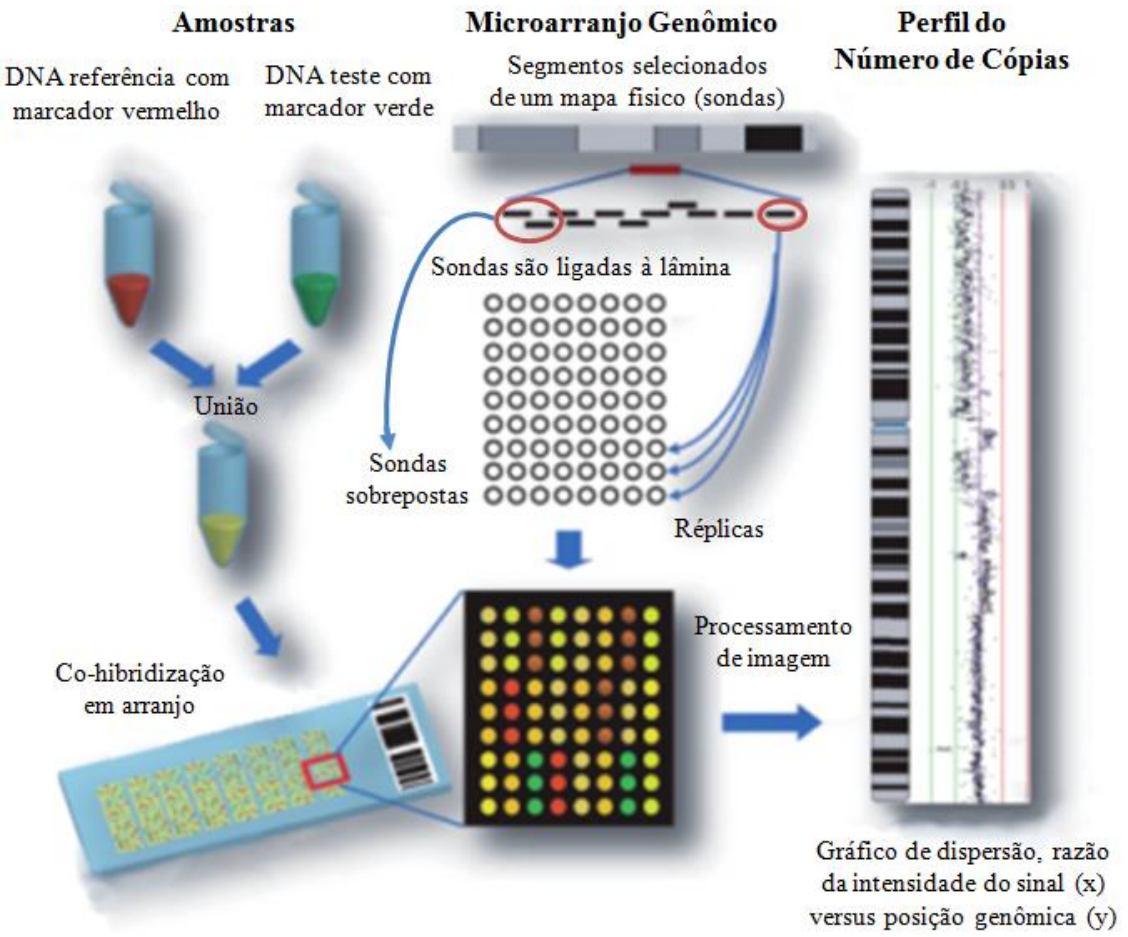

Figura 3.1: Acesso à CNVs por aCGH (adaptado de Chari et al., 2007). Utilizam-se a amostra teste e uma amostra referência, cada uma com um marcador fluorescente distinto. As amostras são misturadas e co-hibridizadas em um microarranjo, no qual irão competir pela ligação com sondas complementares. Estas sondas provem de sequências cuja posição genômica é previamente conhecida, costuma-se utilizar replicas da mesma sonda e/ou sondas sobrepostas para aumentar a precisão do resultado. A lâmina é escaneada para acessar a intensidade do sinal de cada sonda e estimar o número de cópias relativo, baseado na razão dos sinais dos marcadores fluorescente. Com a posição e sinal (representado em razão de $\log_2$) de cada sonda determinados, é possível identificar regiões de perdas e ganhos de sequência, gerando um perfil.

Os segmentos ligados à lamina de vidro ou outra superfície rígida, podem ter origens distintas, variando a sensibilidade do teste. Os arranjos podem ser construídos

com clones grandes (40-200 kb), clones pequenos (1,5-4,5 kb), clones de cDNA (0,5-2 kb), produtos de PCR genômica (100 pares de base – 1,5 kb) e oligonucleotídeos (25-80 pares de base – pb) (Carter, 2007, Monni e Hautaniemi, 2009). A variação entre tamanho das sequências utilizadas determina o tamanho mínimo das regiões de CNVs que podem ser detectadas, e cada técnica tem uma razão entre qualidade de sinal e ruído distinta. Ainda não existe uma técnica que favoreça todas as características necessárias, oferecendo boa resolução, com razão sinal/ruído alta e orçamento razoável. Segmentos provenientes de clones de cromossomos artificiais de bactéria (BAC) oferecem a melhor cobertura e razão sinal/ruído, porém são longos, detectando CNV de no mínimo 50 kb (Lockwood et al., 2006, Carter, 2007).

Diferentes técnicas visando aprimoramento da aCGH são aplicadas. Comumente se realiza uma segunda hibridização, porém trocando os marcadores das amostras teste e referência, para detecção de sinais falsos (Dhawan e Padh 2009). Outro procedimento é a análise de microarranjo de oligonucleotídeos representacional (ROMA), que através da digestão do DNA genômico com enzimas de restrição seguida de PCR, seleciona segmentos de até 1,2 kb. Esta estratégia diminui a complexidade do DNA, otimizando a hibridização de modo a diminuir o ruído – porém metodologias que utilizam segmentos curtos estão sujeitas a erro, pois a probabilidade de ocorrência de sequências semelhantes em diferentes regiões genômicas é maior, e neste caso, podem ser interpretadas como CNVs. O uso de enzimas de restrição para a digestão do DNA também pode gerar erros, pois cada indivíduo possui um padrão diferente de digestão, devido às variações genéticas específicas de cada um. Com o objetivo de precisar NC, muitos arranjos estrategicamente usam segmentos sobrepostos (Figura 3.1) – espera-se que sondas que não possuam sequências semelhantes no genoma e estejam sobrepostas tenham mesma intensidade de sinal (Lockwood et al., 2006, Carter, 2007).

A tecnologia dos arranjos é também utilizada para a genotipagem de SNPs. Neste método, as lâminas contêm sondas de poucas bases às quais se hibridizam somente o DNA teste, pré-digerido. Estas sondas são específicas para cada um dos alelos dos SNPs. O número de cópias é identificado com base na intensidade do sinal de cada sonda, em comparação a outras amostras de referência (Carter, 2007, Dhawan e Padh, 2009).

Como foi desenvolvida inicialmente para a identificação de SNPs, a cobertura do genoma dos primeiros ensaios não era uniforme. Os arranjos atuais oferecem sondas além daquelas usadas para genotipar SNPs, permitindo a identificação de mais CNVs (Lockwood et al., 2006, Carter, 2007). Além de permitir a análise simultânea de SNPs e CNVs, as variações genéticas mais presentes no genoma humano, o arranjo de SNPs oferece vantagens semelhantes à aCGH, pois ambas utilizam tecnologias dos arranjos, possuindo uma grande cobertura a custos razoáveis (Winchester et al., 2009).

Curtis et al. (2009) realizaram um amplo estudo comparativo de quatro plataformas líderes em ensaios de CNV em larga escala – os arranjos de SNP Affymetrix Genome-Wide Human SNP Array 5.0 e Illumina HumanCNV370-Duo DNA Analysis BeadChip; e os aCGH Agilent High-Density CGH Human 244A array e Nimblegen 385 K oligonucleotide array. A metodologia utilizada por estes arranjos são bem distintas e conforme explicado anteriormente, elas variam quanto ao tamanho e distribuição das sondas, além de outros fatores mais específicos, como quantidade de DNA necessária. Apesar das diferenças, todas foram capazes de identificar grandes aberrações, porém nem sempre alterações menores foram encontradas por todas. Foram também observadas variações no sinal das sondas, o que influencia na estimativa da magnitude da variação (estimativa do NC).

Devido às variações das medições das técnicas de hibridização (aCGH e arranjo de SNP) falsos positivos sempre são gerados. Visando evitar a inclusão destes dados errôneos, muitos estudos aplicam uma segunda técnica para validar as CNVs encontradas (Carter 2007), por exemplo o uso de arranjo SNP para identificação de regiões candidatas a portarem CNVs, seguido por confirmação com aCGH, com sondas específicas para as candidatas. Outra estratégia é o uso de mais de uma plataforma (multiplataforma), com o objetivo delas se complementarem, porém a variação da distribuição das sondas e a magnitude das alterações de NC que ocorre entre ensaios de fabricantes distintos (Figura 3.2) dificultam atingir um consenso entre os dados gerados (Zhang et al., 2010).

As características das amostras também possuem fatores que influenciam na precisão do resultado. Especificamente estudos de câncer, cujo alvo é alterações somáticas, as amostras costumam conter células normais, interferindo na identificação das alterações (van de Wiel et al., 2010).

Assim, apesar dos grandes avanços nos métodos, ainda é necessário o aperfeiçoamento das técnicas, para melhor delimitar os segmentos, bem como determinar a magnitude dos CNVs. Considerando as limitações das técnicas utilizadas, principalmente no início das pesquisas com foco em CNV, Lee e Scherer, 2010 sugeriram que a definição do termo seja revisada, de forma a englobar segmentos com NC variável de tamanho menor que 1 kb. Neste contexto, seriam contabilizadas mais de 11.700 CNVs conhecidas, maiores que 443 pares de base (Conrad et al. 2010).

É necessário precisar qual é a ocorrência de CNVs em populações saudáveis, para se estabelecer uma referência, permitindo inferir com mais precisão quais são as variações candidatas a associação com patologias (Lee e Scherer, 2010). Existem

diversos trabalhos neste sentido, mais ainda a muito a ser desvendado (McCarroll 2010).

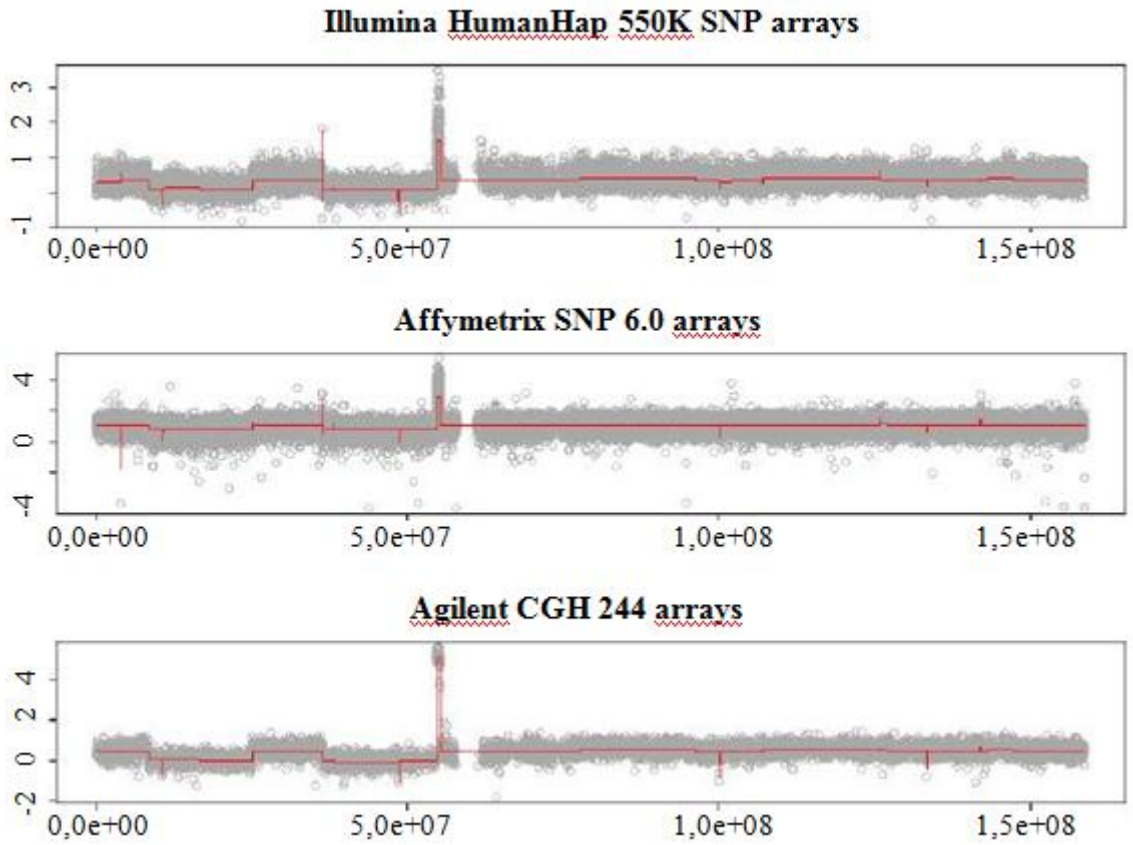

Figura 3.2: Comparação entre três plataformas para acesso de CNV (adaptado de Zhang et al., 2010). Resultado da análise de uma mesma amostra, utilizando três plataformas de fabricantes distintos. Os círculos cinza indicam o NC medido em cada sonda, a magnitude está representada em razão de log2 (eixo y). Os resultados apresentados se referem ao cromossomo 7, cuja posição está indicada em pb (eixo x). As linhas vermelhas representam os segmentos com mesmo NC, identificados utilizando a segmentação circular binária (CBS). A amostra não identificada provém do projeto *The Cancer Genome Atlas* (TCGA) que utilizou um método multiplataforma para identificar CNV em pacientes com câncer.

Ainda não existe ensaio em larga escala capaz de identificar haplótipos, apenas alguns algoritmos para inferência (Su et al. 2010). Em estudos de associação, a análise

haplotípica é mais eficiente do que o estudo de um único marcador molecular – isto pois o fenótipo é resultado da combinação da expressão de mais de um gene ou de diversas variações do mesmo gene. Os haplótipos também contribuem para a análise da história evolutiva dos CNVs, porém foram pouco explorados.(Su et al. 2010).

3.2 Identificação de Regiões com Mesmo Número de Cópias

Os resultados gerados pelos arranjos são valores de intensidade particulares de cada sonda, por isso são chamados de dados brutos. A interpretação destes resultados é possível apenas após processamento, pois o sinal das sondas é variável, sendo necessária a utilização de algoritmos estatísticos para inferir quais são os trechos genômicos com variação no NC (Lockwood et al., 2006, Carter, 2007, van de Wiel et al., 2010).

Estes algoritmos geralmente utilizam a estratégia de segmentação, a qual divide o genoma em regiões de mesmo NC. Assim, estes métodos visam delimitar a magnitude e o limite de cada segmento – denominado ponto de quebra (Carter, 2007, van de Wiel et al., 2010). Os gráficos da Figura 3.2 permitem uma melhor compreensão dos resultados obtidos: podem ser observados os dados brutos (círculos cinza), cujo sinal é bem variável, e os dados segmentados (linhas vermelhas) para os quais há um NC preciso.

O desenvolvimento de algoritmos de segmentação em sua maioria assume que a presença de CNV acarretará uma mudança brusca de sinal das sondas. Lai et al. (2005) e Willenbrock e Fridlyand (2005) analisaram o desempenho dos diversos algoritmos para análise de aCGH desenvolvidos até então e identificaram a segmentação binária circular (CBS) como a mais eficiente. Várias abordagens estatísticas são usadas para identificar os pontos de quebra, mas a CBS demonstrou ser a metodologia mais sensível e com menor índice de erro, apesar de demorar e não detectar variações pequenas

(segmentos com cobertura de uma única sonda) (Lai et al., 2005, Willenbrock e Fridlyand, 2005, Hofmann et al., 2009).

Desde sua divulgação (Olshen et al., 2004) a CBS é a escolha metodológica de inúmeros estudos e em 2007, foi apresentada uma versão mais veloz da CBS, superando um dos seus pontos falhos (Venkatraman e Olshen, 2007). Desde a análise comparativa realizada por Lai et al. (2005) e Willenbrock e Fridlyand (2005) mais de 30 novos métodos foram criados e uma nova comparação conjunta dos mesmos ainda não foi feita (van de Wiel et al. 2010).

Existem algoritmos específicos para arranjos SNP ou aCGH, alguns foram adaptados para serem aplicados a ambos, como o algoritmo para detecção de alteração genômica (GADA) e o CBS. Os fabricantes oferecem também aplicativos para detecção de alterações de NC (Winchester et al. 2009). Em estudo comparativo de técnicas para arranjos de SNP Winchester et al. (2009) recomendam o uso de dois algoritmos distintos num mesmo conjunto de dados para uma análise com maior confiança, sendo indicado que um deles seja o oferecido pelo fornecedor do microarranjo.

Conforme explicado na seção 3.1.1, uma estratégia para maior cobertura e precisão de resultados, são as análises multiplataforma. Apenas recentemente Zhang et al. (2010) desenvolveram um método robusto para estes casos, a Segmentação Circular Binária Multiplataforma (MPCBS), a qual normaliza os resultados das diferentes plataformas, gerando resultados mais precisos, com o poder de detecção da CBS.

Usuários com interesse em explorar diferentes métodos de segmentação, podem a ferramenta CGHweb. Com intuito de auxiliar na escolha dos algoritmos de segmentação, Lee e Kim (2009) criaram uma aplicativo Web (http://compbio.med.harvard.edu/CGHweb), na qual o usuário submete seus dados brutos para comparar os resultados de até dez métodos de segmentação. Além da

visualização comum sob forma de gráfico (posição genômica x NC), mapas de calor das amplificações e deleções detectadas por cada um dos softwares são gerados. A interface é amigável, porém é necessário aguardar alguns minutos para acessar o resultado, pois são processados em servidor remoto. Um endereço Web é destinado ao resultado gerado, e o usuário recebe um hyperlink via e-mail, podendo acessar o resultado por sete dias.

3.3 Aplicativos para Exploração

A técnicas baseadas em microarranjo são as mais populares para a identificação de CNVs, assim a grande maioria dos softwares e algoritmos para análise são voltados para os ensaios aCGH e arranjo de SNP. Portanto esta seção traça um perfil geral dos aplicativos capazes de processar dados provenientes de microarranjos.

O primeiro processamento computacional ao qual o microarranjo é submetido é o escaneamento para capturar o sinal dos marcadores e alinhá-los no genoma. Posteriormente, seguem as etapas de normalização, detecção de alterações de NC, determinação do NC e demais análises (para extração de informações e elaboração/teste de hipóteses), as quais são realizadas também com suporte de aplicativos (Chari et al., 2007, Kingsley et al., 2007, Hofmann et al., 2009).

Os programas que realizam os processamentos iniciais dos dados costumam ser plataforma-específicos, atendendo as particularidades de cada ensaio (Chari et al., 2007). Assim, softwares que oferecem recursos para o processamento e análise completos de CNV podem ter seu uso restringindo pelo formato dos dados de entrada (Shankar et al., 2006). Foram desenvolvidos alguns aplicativos capazes de aceitar dados de plataformas distintas, mas geralmente as diferentes etapas da análise se encontram em programas separados.

Existe um grande número de programas para visualização e análise estatística sendo disponibilizado. Suas funcionalidades e requerimentos são bastante diversificados, variando quanto ao ambiente de execução, arquivos de entrada, algoritmos disponíveis, custo e modo de visualização (Lockwood et al., 2006, Chari et al., 2007). A Tabela 3.1 compara diversos programas envolvidos nas análises de dados pós normalização de sinal, os quais foram avaliados com base na literatura disponível, sendo acessados principalmente softwares acadêmicos com acesso livre.

São comuns os softwares que oferecem algoritmos para segmentação e outras análises de dados (ver Tabela 3.1 para referências), porém sua utilização costuma requerer um bom conhecimento de bioinformática e programação (Chari et al., 2007, Kotliarov et al., 2010). Um exemplo são os pacotes do projeto Bioconductor, que reúne uma coleção de métodos e algoritmos para análise de dados genômicos, inclusive diversos métodos para cada um dos passos da análise de CNV. Estes recursos são executáveis em ambiente R, uma plataforma estatística de acesso livre, *cross*-plataforma (Hofmann et al. 2009).

Dentre os trinta softwares citados, dezenove permitem a visualização de múltiplas amostras, porém três não puderam ser encontrados on-line e alguns dos outros oferecem exibição simultânea de um número limitado de amostras. O SnoopCGH apresenta as amostras no mesmo gráfico de dispersão (posição genômica x NC), sobrepondo os pontos, de forma pouco esclarecedora, os demais representam resultados individuais ou consenso entre o conjunto de amostras.

Em sua revisão de aplicativos para análise de CNVs obtidas por aCGH, Chari et al., 2007 concluem que são necessários programas capazes de realizar meta-análise entre plataformas, principalmente para estudo conjunto de outros tipos de informação, como dados epigenéticos e transcriptomas. Afirmam também que pesquisadores com

Tabela 3.1: Aplicativos para Visualização e Análise de CNV

. (*) L- Linux, M – Mac OS, U – Unix, W – Microsoft Windows. (**) Análises de dados pós normalização ou segmentação. (-) Informação não disponibilizada

(continua)

Aplicativo	Sistema operacional*	Requerimento	Disponi-bilidade	Linguagem	Detecção de pontos de quebra	Visualiza-ção	Análises**	Página Web	Referência
aCGH-Smooth	W	Excel 97	Livre	C++	indisponível	individual	disponível	www.few.vu.nl /~vumarray/	Jong et al., Lockwood et al., 2006, Chari et al., 2007
aCGHViewer / Overlay Tool	W	J2SE 1.3, Microsoft .NET Framework 2.0	Livre	Java	indisponível	individual	disponível	falcon.roswellpark .org/aCGHview/	(Shankar et al. 2006), Lo et al., 2007
ArrayCyGHt	Baseado em Web	-	Livre	R, Perl, PHP	indisponível	individual	indisponível	genomics.catholic .ac.kr/arrayCGH/	Kim et al., 2005
Asterias	Baseado em Web	compilador C/C++, MPI, PHP, Python, R, Web server	Livre	C/C++, JavaScript, MPI, Perl, PHP, Python, R, shell	disponível	individual	integrada	www.bioinformatics .org/asterias/wiki/	Díaz-Uriarte et al., 2007
Cancer Genome Browser	L, M, W	módulo MATLAB	Livre	MATLAB, s3ql	indisponível	múltipla	disponível		Freire et al., 2008
Caryoscope	U, W	J2SE 1.4.2_03	Livre	Java	disponível	individual	integrada	caryoscope. stanford.edu/	Awad et al, 2004, Lockwood et al., 2006, Chari et al., 2007
CGH Analytics	L, M, W	nenhum	Pago	-	disponível	múltipla	integrada	www.chem.agilent.com/en-US/products/software/lifesciencesinformatics/pages/gp35719.aspx	Agilent Technologies, Lockwood et al., 2006, Chari et al., 2007, Chi et al., 2008

Tabela 3.2: Aplicativos para Visualização e Análise de CNV

(continuação)

Aplicativo	Sistema operacional*	Requerimento	Disponi-bilidade	Linguagem	Detecção de pontos de quebra	Visuali-zação	Análises**	Página Web	Referência
CGHAnalyzer	U, W	JRE 1.4.1	Livre	Java	indisponível	múltipla	disponível	não encontrado	Margolin et al., 2005, Lockwood et al., 2006, Chari et al., 2007
CGH-Explorer	L, M, W	JRE 1.4	Livre	Java	disponível	múltipla	disponível	www.ifi.uio.no/bioinf/Papers/CGH/	Lingjærde et al., 2005, Lockwood et al., 2006, Chari et al., 2007, Chi et al., 2008
CGH-Miner	U, W	Excel 2000, R	Livre	Excel, R	indisponível	múltipla	indisponível	peiwang.fhcrc.org/internal/temp/CGH-Miner-1.0.exe/view	Wang et al., 2005, Lockwood et al., 2006, Chari et al., 2007
CGHPRO	L, W	JRE 1.4, MySQL, R	Livre	Java, SQL, R	disponível	múltipla	indisponível	www.diss.fu-berlin.de/diss/servlets/MCRFileNodeServlet/FUDISS_derivate_000000002359/	Chen et al., 2005
ChARM	L, M, W	JRE 1.4.2	Sob requisição	-	disponível	múltiplo	-	function.princeton.edu/ChARM/	Myers et al., 2004
CHESS	Baseado em Web, W, L	JRE 1.6	Livre	Java	disponível	multipla	integrada	biostone.khu.ac.kr/CHESS/	Lee e Kim, 2009
CNAG	W	-	Livre	C++	disponível	individual	disponível	www.genome.umin.jp	Nannya et al., 2005
CNAReporter	L, M, S, U, W	MATLAB 2007b, Perl 5 e GenePattern 3.0	-	MATLAB, Perl	disponível	individual	disponível	gforge.nci.nih.gov/projects/cnareport	Kotliarov et al., 2010
dCHIP	W	-	Livre	C++	disponível	multipla	disponível	www.dchip.org	Li, 2008
DIGMAP Viewer	Independente	JRE 1.3	Livre	Java	indisponível	múltipla	integrada	geneexplorer.mc.vanderbilt.edu/digmap	Yi et al., 2005
ECN-tool/ CGH-Plotter	L, M, W	MATLAB 6.1	Livre	MATLAB	indisponível	múltipla	integrada	não encontrado	Autio et al., 2009

Tabela 3.3: Aplicativos para Visualização e Análise de CNV

(conclusão)

Aplicativo	Sistema operacional*	Requerimento	Disponibilidade	Linguagem	Detecção de pontos de quebra	Visualização	Análises**	Página Web	Referência
ISACGH	Baseado em Web	-	Livre	-	disponível	múltipla	integrada	não encontrado	Chari et al., 2007, Conde et al., 2007, Chi et al., 2008, Lee e Kim, 2009
Magellan	Baseado em Web, W	-	-	C, Java, R, SQL	-	individual, mapa consenso	integrada	não encontrado	Kingsley et al., 2007
M-CGH	Independente	MATLAB 6.1	Livre	MATLAB	disponível	individual	disponível	folk.uio.no/junbaiw/mcgh/	Wang et al., 2004, Lockwood et al., 2006, Chari et al., 2007, Chi et al., 2008
MD-SeeGH	W	MySQL	Livre	C++, SQL	disponível	múltipla	integrada	www.arraycgh.ca/	Chi et al., 2008
Nexus	L, M, W	-	Pago	-	disponível	múltipla	integrada	www.biodiscovery.com/index/nexus	BioDiscovery
Nimblegen SignalMap	W	nenhum	Pago	-	indisponível	múltipla	disponível	www.Nimblegen.com/products/software/signalmap.html	NimbleGen Systems, 2006, Chari et al., 2007
SIGMA2	W	JRE 1.6, R, MySQL	Livre	Java, R, SQL	disponível	individual	integrada	não encontrado	Chari et al., 2008
SnoopCGH	Independente	Java	Livre	Java	disponível	múltipla	anotação	snoopcgh.sourceforge.net/	Almagro-Garcia et al., 2009
SNPexpress	Independente	Java 1.5	Livre	Java	disponível	múltipla	integrada	people.genome.duke.edu/~dg48/SNPExpress/	Sanders et al., 2008
SNPscan	Baseado em Web (Independente)	nenhum	Livre	HTML, Perl, R	indisponível	múltipla	integrada	pevsnerlab.kennedykrieger.org/snpscan.htm	Ting et al., 2006
Spectral Ware	Baseado em Web	JRE 1.4.2_05 plugin	Pago	-	disponível	individual	-	www.spectralgenomics.com/spectralware.	Lockwood et al., 2006, Chari et al., 2007
VAMP	U, W	Java 1.4.2	Sob requisição	Java, PHP	disponível	múltipla	integrada	bioinfo.curie.fr/vamp	La Rosa et al., 2006

conhecimento computacional limitado não podem usufruir prontamente destas análises. Apontam que o desenvolvimento de programas Web com interfaces amigáveis que permitam integração de base de dados multidisciplinares permitirão a mineração de dados genômicos e correlação com dados clínicos. Além disso, prevêem que com o avanço dos estudos de associação de aCGH com dados clínicos, surgiriam problemas relacionados com segurança e automação, entre outros.

Desde então, alguns programas para análise integrada foram criados (Tabela 3.1, categoria "integrada" da coluna "Análises"). Os softwares CGHviewer/Overlay Tool, Asterias, Caryoscope, CHESS, CGH Analytics, DIGMAP Viewer, ISACGH, Magellan, MD-SeeGH, Nexus, SIGMA^2, SNPexpress, VAMP permitem análise conjunta com expressão gênica, além de outras funções, como por exemplo suporte para estudo caso-controle (CHESS, ISACGH, SNPscan) ou meta-análise de dados clínicos (Asterias, Magellan).

A grande maioria dos aplicativos ainda não é voltada para Web, necessitam de instalação na máquina do usuário. As linguagens de implementação mais populares são R, MATLAB, C e Java (Lockwood et al., 2006). Uma das vantagens do desenvolvimento de softwares baseados em Web está na flexibilidade de execução entre sistemas operacionais e a potencial independência de instalação. Ainda assim, programas Web como o Caryoscope, CHESS, Magellan e Spectral Ware, requerem carregamento de *applets*, por serem implementados em Java e são restritos a plataformas computacionais específicas. Outros programas, como o Asterias fornecem o software baseado em Web, porém este deve ser instalado e mantido pelo usuário, o que requer a instalação de diversos componentes para suporte.

CHESS é versátil, pois pode ser executado localmente, como os programas *desktop*, pois sua implementação é independente de servidor e outros recursos de rede, necessita somente do aplicativo Java (Lee e Kim 2009).

Capítulo 4

Web e Recursos Oferecidos

A World Wide Web surgiu como uma tecnologia para compartilhar texto e imagem estáticos através da Internet, mas suas funções foram estendidas muito além do objetivo inicial. Ela tornou-se um meio para conectar usuários remotos e aplicativos, bem como permitiu integrar programas pela Internet, sendo atualmente também uma plataforma para criação de sistemas operacionais interativos (Alonso et al., 1998, Goodman e Morrison, 2007, O'Reilly, 2007). A Web tem seu núcleo no protocolo de transferência de hipertexto (HTTP), linguagem de marcação de hipertexto (HTML), servidores Web e navegadores Web. A informação é trocada sob o HTTP em forma de documentos identificados pelo identificador-padrão de recursos (URI). O conteúdo destes documentos pode ser estático ou dinâmico – isto é, gerado ao ser requisitado. Todo recurso disponível na Web pode ser acessado através de seu localizador padrão de recursos (URL) (Alonso et al., 1998).

4.1 Tecnologias para Desenvolvimento Web

A princípio os navegadores Web foram desenvolvidos apenas para exibição de documentos estáticos, provenientes do servidor e enviados ao usuário via HTTP, não sendo possível construir programas sofisticados. A primeira solução foi literalmente embutir no navegador aplicativos Java (máquina virtual Java – JMV), os quais ficaram conhecidos como *"applets"*, gerando interfaces dinâmicas. O próximo passo foi o surgimento do *Common Gateway Interface* (CGI), tecnologia que permite gerar páginas

dinamicamente, pois torna possível que o navegador passe parâmetros para um *script* (conjunto de passos pré-programados) alojado num servidor web, que os interpreta e gera uma página após processamento (Alonso et al., 1998).

O uso de JMV era mais popular no final da década de 90, e diminui por razões técnicas e corporativas. Os demais aplicativos para navegador, como o popular Flash, exigem a instalação da versão compatível, bem como certo tempo para carregar os arquivos do próprio programa, demandando por vezes mais tempo do que o usuário pode aguardar (Goodman e Morrison 2007).

O desempenho de programas CGI é dispendioso, pois geralmente é criado um processo separado para cada solicitação submetida. Este processo requer certo tempo para ser executado e em caso de múltiplas requisições, os processos competirão por recursos e memória do servidor, limitando a escalabilidade do sistema. Assim, foram criados Servlets, programas Java que oferecem os mesmos recursos do CGI, sem gerar processos paralelos. Outras tecnologias surgiram neste sentido, porém mantendo o paradigma de gerar novas páginas num servidor remoto, em resposta a solicitações do cliente (Alonso et al., 1998). Dentre as plataformas de desenvolvimento em servidor mais utilizadas estão PHP, ASP, .Net, JSP e ColdFusion, sendo também muito comum as linguagens Perl, Python, Java, C++, C# e VisualBasic (Goodman e Morrison, 2007).

Por mais robusta e útil que possa ser a execução de *scripts* no servidor, sua eficiência é pouca quando se trata de interatividade com usuário, pois qualquer mudança a ser feita na página precisa ser processada no servidor. A real interatividade com o usuário foi atingida com a linguagem JavaScript em 1995, que ao ser incorporada no navegador, possibilitou o processamento de requisições no próprio computador do usuário, sem necessidade de submissão às bases remotas. A inovação consiste na

programação no navegador Web, isto é, com o JavaScript parte da execução de tarefas pôde ser deslocada para a máquina do usuário (Goodman e Morrison, 2007).

Além de desempenho, outro aspecto a ser considerado ao trabalhar com recursos Web é a segurança. A defesa básica do JavaScript contra códigos maliciosos é sua incapacidade de leitura/escrita de arquivos no servidor ou máquina do usuário. Dessa forma, não é possível implantar vírus ou apagar os dados. O JavaScript também não permite o estabelecimento do contato com servidores diferente do seu próprio e a manipulação de conteúdo está restrita aos elementos gerados pelo código a que pertence. Isto é, a linguagem não pode ler ou alterar propriedades e elementos de páginas cuja origem difere da sua. Assim, as limitações embutidas na linguagem provêm o mecanismo básico de segurança contra os principais tipos de ataque (Flanagan 1998).

Por outro lado, a leitura/escrita de arquivos é uma função necessária. O Ajax (*Asynchronous JavaScript e XML*) é um recurso que permite a comunicação com o servidor no *background*, portanto com seu uso, a troca de informações com o servidor pode ser feita sem a necessidade de alterar a interface com o usuário (Goodman e Morrison 2007). Desta forma, os dados XML (*eXtensible Markup Language*, método estruturado para descrição de conteúdo) obtidos do servidor são examinados pelo JavaScript, o qual pode então alterar porções específicas da página Web (Macedo, 2004, Goodman e Morrison, 2007). Esta é a metodologia empregada por muitos programas, principalmente em interfaces de correio eletrônico e na exploração de mapas do Google Maps (http://maps.google.com) (Goodman e Morrison 2007).

Trabalhando em conjunto com as demais tecnologias Web, o JavaScript possibilitou a criação de páginas dinâmicas e com vários recursos, pois utilizando o *Document Object Model* (DOM) é capaz de manipular os elementos HTML e definir a

aparência da interface (Flanagan, 1998, Goodman e Morrison, 2007). Assim, as páginas passaram a parecer cada vez mais com aplicativos *desktop*, surgindo o conceito da Web 2.0, na qual além de processar dados no navegador, o usuário pode editar documentos e realizar alterações nas páginas, como observado nos populares blogs e wikis (O'Reilly 2007).

4.2 Estrutura Web

Os serviços Web são uma forma de sistema de informação distribuído e sua estrutura de implementação clássica pode ser dividida em três camadas: apresentação, lógica da aplicação e gerenciamento de recursos. A camada de apresentação é a responsável pela interação com o cliente. O termo cliente refere-se às entidades que utilizam os serviços oferecidos pela aplicação, especificamente, o navegador Web (Alonso et al., 1998).

O cliente utiliza a camada de apresentação para solicitar informações que são processadas pela camada de lógica da aplicação. Esta camada é a responsável pelo processamento de dados e execução de tarefas. Normalmente, as páginas Web mantêm a camada lógica num servidor remoto, executando apenas pequenos processamentos localmente. Os dados por sua vez são administrados pela camada de gerenciamento de recursos, a qual define também a fonte de dados e sua organização. Em programas baseados em Web, os dados geralmente provêm de arquivos na máquina cliente ou bancos de dados que podem estar em servidores locais (máquina do usuário) ou remotos (dados transmitidos via Web) (Alonso et al., 1998).

4.3 *Simple Sloppy Semantic Database* - S3DB

Tim Berners-Lee, considerado o criador da Internet, aposta que o próximo grande passo da Web será permitir que os computadores, não apenas apresentem

informações, mas interpretem e processem as informações contidas nos bilhões de páginas disponíveis – o que atualmente só pode ser executado por humanos. A essa nova Internet denominou-se "Web Semântica" ou Web 3.0. Seu princípio baseia-se em atribuir informação aos dados (metadado) contidos nas páginas de forma a permitir extração de conhecimento automaticamente (Breitman, 2005).

Motivado por esta proposta foi criado o S3DB, um repositório no qual os dados são representados de forma flexível, em oposição às bases de dados clássicas, com objetivo de extrair informações com mais facilidade. No S3DB a representação dos dados utiliza triplas sujeito-verbo-objeto, enquanto as bases relacionais armazenam os dados como tabelas. O S3DB é um software livre e pode ser instalado localmente ou em servidores remotos (Almeida et al., 2006, Deus et al., 2008).

Entre os recursos oferecidos pelo S3DB, destacam-se a capacidade de controle de acesso às informações, possibilidade de consulta e alteração da base utilizando comandos JavaScript em combinação com s3ql (a linguagem de manipulação de bancos S3DB). O controle de acesso é feito por identificação/senha e as autorizações de uso das informações (leitura, alteração e escrita) podem ser controladas em diferentes instâncias do banco de dados, isto é, cada usuário possui um conjunto de permissões para as subdivisões do banco S3DB (Deus, Deus et al., 2008).

Ao acessar uma base de dados, o usuário obtém uma palavra-chave (*key*) e com ela pode trabalhar com o banco S3DB de acordo com suas autorizações. Utilizando a Interface de Programação de Aplicativos (API) do S3DB estas tarefas podem ser executadas através de programas Web. Por comandos JavaScript, podem ser enviadas solicitações ao banco de dados remoto e assim, realizar leitura, alteração, criação e deleção de informações (Deus).

Capítulo 5

Metodologia

Inspirada nos recursos oferecidos pela Web 2.0 e no rápido desenvolvimento dos navegadores, foi proposta a implementação de uma ferramenta que explorasse ao máximo os recursos oferecidos pelo ambiente de navegadores de Internet. A ferramenta em questão seria empregada na análise exploratória de CNV humano.

Dados genômicos humanos são caracteristicamente volumosos, consequentemente costumam gerar arquivos grandes, exigindo certo tempo para serem processados. O CNViewer teve o desafio de representar não apenas dados de uma amostra, mas sim de um conjunto de indivíduos simultaneamente. O modelo de representação utilizado baseou-se em uma das ferramentas do programa *Cancer Genome Browser* (*CGB*), a qual foram acrescentadas novas características, oferecendo maior flexibilidade para o usuário e aprimorando as formas de exploração.

O desenvolvimento do CNViewer foi feito em parceria com a *University of Texas – M. D. Anderson Cancer Center*.

5.1 Implementação

5.1.1 Arquitetura do sistema

Visando usufruir da praticidade oferecida pelos programas baseados em Web, com as vantagens dos aplicativos *desktop*, o CNViewer foi desenvolvido no ambiente de navegador Web, utilizando sua linguagem de programação nativa, o JavaScript, para

torná-lo independente de carregamento de aplicativos suporte (como ocorrem com programas Java) e evitando processamentos externos. Infelizmente, algumas opções do CNViewer ainda são dependentes do servidor, porém foram criadas soluções opcionais para realizar as mesmas tarefas localmente.

Os navegadores Web naturalmente unem o cliente e a camada de apresentação do programa Web (Alonso et al., 1998). No caso do CNViewer, a camada lógica também foi deslocada inteiramente para o navegador, bem como a estrutura de dados a ser utilizada, evitando assim problemas de latência e performance associados com a transferência de dados. Uma das prioridades no desenvolvimento do CNViewer foi criar um aplicativo de fácil acesso e utilização. O desenvolvimento baseado em Web potencialmente torna o software disponível para qualquer plataforma computacional sem necessidade de instalação, facilmente acessível via URL e com a versão mais recente pronta para uso.

Com a Web 2.0 é possível também criar interfaces dinâmicas e amigáveis. Por outro lado, o modelo de análise exploratório que deveria ser suportado pelo CNViewer exige o acesso frequente aos mesmo dados e portanto, se utilizado o padrão comum de aplicativos Web, o tempo consumido em cada processo de análise e representação gráfica seria inviável devido ao volume de dados. Isto ocorre pois os programas Web costumam acessar dados em servidores remotos a cada requisição. Portanto, apesar de os dados serem provenientes de diferentes fontes (seção 5.2), durante a execução do programa uma cópia deles estará armazenada temporariamente no navegador e o CNViewer se comportará de forma semelhante a um aplicativo *desktop*.

O aCGHViewer é um exemplo de programa que foi implementado como um aplicativo *desktop* para evitar problemas de performance e latência, porém é restrito a Windows e exige instalação (Shankar et al., 2006). Desde 2006 as próprias companhias

aprimoraram o desempenho de seus navegadores e assim dependendo da forma de implementação, o uso de navegadores pode ser mais vantajoso em relação aos programas *desktop*.

O JavaScript é linguagem de programação leve com capacidade para orientação a objeto (Flanagan 1998). A implementação fez uso de bibliotecas Dojo (The Dojo Foundation, 2009) para a realização de algumas funções. O CNViewer possui três formas de inclusão de dados para análise, a primeira, acessa um repositório de arquivos semântico, do tipo S3DB (ver seção 4.3), através de comandos s3ql . A segunda é mais simples, pois o usuário insere manualmente os dados, que são imediatamente interpretados no navegador. O terceiro modo de entrada de dados depende de um *script* PHP para poder ler arquivos da máquina do usuário, devido às restrições impostas ao JavaScript abordadas na seção 4.1. Pelas mesmas razões, este *script* deve estar localizado junto ao arquivo HTML para poder acessar as propriedades do DOM da interface.

Um dos problemas observados em softwares acadêmicos é seu curto prazo de disponibilidade, conforme ocorrido com alguns dos aplicativos citados na Tabela 3.1. Geralmente o programa fica disponível na página de uma instituição enquanto o seu desenvolvedor é vinculado a ela e depois perde-se o acesso. Existem algumas páginas de internet que oferecem hospedagem gratuita de programas de distribuição livre, sendo possível em alguns casos como o Google code, realizar controle das versões submetidas, permitindo o acesso de versões anteriores caso haja interesse.(Google Project Hosting 2010).

Desta forma o código do CNViewer foi submetido ao Google code, passando a estar disponível indeterminadamente para uso. Os serviços oferecidos pelo Google code não abrangem o uso de servidores para execução de tarefas e processamentos em geral.

Assim os arquivos PHP e consequentemente uma versão do HTML, estão localizados num servidor remoto (http://odin.mdacc.tmc.edu/~ccpalu/CNViewer/, pertencente ao *M. D. Anderson Cancer Center*) enquanto os *scripts* em JavaScript e imagens estão no Google code (http://www.googlecode.com/). Apesar de o Dojo utilizar a linguagem JavaScript, seu método de localização de bibliotecas se baseia no endereço da URL da página de internet e portanto todos os arquivos Dojo também foram armazenados no servidor do *M. D. Anderson Cancer Center*. Uma segunda versão do CNViewer, sem as opções dependentes de PHP, foi disponibilizada também no Google code (em http://www.cnviewer.googlecode.com/hg/cnviewer.html).

5.1.2 Representação de cromossomos

Um dos recursos oferecidos pelo CNViewer é a escolha de regiões genômicas, as quais devem ser delimitadas nos cromossomos de interesse. Para tanto, foi implementado um idiograma no qual o usuário especifica as regiões a serem analisadas com marcadores. Estes marcadores móveis foram feitos utilizando a biblioteca Dojo "*Movable*" com alterações para atender as necessidades do CNViewer, tais como capturar a posição final do marcador e movimentar-se apenas horizontalmente. Posteriormente, quando o resultado da análise é exibido, as regiões escolhidas são destacadas nas imagens dos cromossomos, identificando-as.

Os cromossomos estão ilustrados de acordo com o padrão de bandas-G e foram criados através do software Idiographica 2.0 (Kin e Ono, 2007). As bandas-G são o padrão de coloração assumido pelos cromossomos ao serem corados com Giemsa, corante que se liga às regiões ricas em adenina e timina gerando um perfil de bandas comum entre indivíduos de uma mesma espécie.

As regiões genômicas são estimadas com base no *Human Genome Assembly Information* montagem 36. Através da razão entre a imagem do cromossomo 1 e seu

tamanho em pares de base, calculam-se as posições no genoma a serem utilizadas na análise, bem como as marca no gráfico com resultados.

5.1.3 Acesso ao *UCSC Genome Browser*

Outra característica implementada, é o acesso ao *University of California Santa* Cruz *Genome Browser* (*UCSC Genome Browser*), um portal para acesso aos dados contidos na *UCSC Genome Browser Database*,uma fonte de dados de sequencias genômicas de diversas espécies e respectivas anotações. A interface do *UCSC Genome Browser* oferece recursos como alinhamentos, predição de genes, elementos regulatórios, dados de expressão gênica, repetições no genoma, SNPs e outras variações, além de diversas ferramentas de análise (Kuhn et al. 2009)

O *UCSC Genome Browser* aceita submissão de requisições via URL, permitindo o envio de pesquisas a partir de outros programas. Assim é possível ao usuário do CNViewer consultar informações sobre regiões de interesse automaticamente. O CNViewer acessa a posição do trecho genômico selecionado e submete ao *UCSC Genome Browser*, informando também a montagem do genoma e a espécie, gerando uma consulta. É então criada uma janela para cada região genômica submetida.

O acesso desta forma ao *UCSC Genome Browser* usa uma chave alfanumérica, gerada através de requisição à página. Devido às limitações de segurança do JavaScript, esta requisição é feita por um *script* PHP, armazenado no servidor http://odin.mdacc.tmc.edu/~ccpalu/CNViewer/.

5.2 Entrada de Dados

O conceito fundamental do CNViewer é oferecer ferramentas para os dados fornecidos para o usuário. Conforme explicado previamente (seção 3.2), os dados brutos costumam ser segmentados antes de serem interpretados, assim os dados brutos são

utilizados poucas vezes e as diversas análises são feitas com base nos resultados da segmentação. Os arquivos brutos são grandes, pois possuem os resultados do conjunto de sondas utilizadas no microarranjo, já o resultado da segmentação possui tamanho mais compacto e consequentemente otimiza processamentos posteriores. Outra motivação para utilizar estes dados está nos valores de NC. Uma vez realizada a segmentação, espera-se que os ruídos da técnica empregada sejam suavizados, permitindo a obtenção de um NC mais próximo ao real. Portanto, o CNViewer interpreta dados segmentados, sob formato texto delimitado por tabulação. Além de tornar o formato de entrada independente da plataforma utilizada para acessar CNV, o usuário tem a flexibilidade de escolher entre um maior número de algoritmos de segmentação. Procedimento semelhante ao utilizado no desenvolvimento do aCGHviewer (Shankar et al. 2006). Assim, é possível analisar dados provenientes de plataforma distintas, com o intuito inclusive de comparar as diferenças entre as técnicas ou tendo o cuidado e conhecimento das variações normais geradas pelas características de cada plataforma.

A CBS é um algoritmo popular, que pode ser utilizado para segmentação de dados provenientes tanto de plataformas aCGH quanto de SNParray (ver Seção 3.2). Portanto o CNViewer foi desenvolvido com base nos arquivos gerados por CBS, através do programa estatístico R, mas atentando para não restringir a um modelo padrão de entrada.

5.2.1 Dados-modelo

Conforme mencionado anteriormente, o CNViewer é baseado em uma ferramenta do CGB, aplicativo desenvolvido pelo mesmo grupo de pesquisa, porém feito em MATLAB. O CGB acessa principalmente dados brutos e segmentado do projeto TCGA, alvo de pesquisas realizadas no *M. D. Anderson Cancer Center*,

instituição parceira no desenvolvimento de ambos os aplicativos. Por essas razões, foi utilizado um conjunto de dados segmentados previamente, para uso no CGB.

O conjunto de dados usados para modelar o CNViewer foram de Glioblastoma multiforme primário, constando de 210 amostras, provenientes da *Harvard Medical School* e do *Brigham and Women's Hospital*, obtidas no portal do projeto TCGA (http://tcga-data.nci.nih.gov/) e posteriormente filtradas utilizando a CBS. O algoritmo para esta segmentação está implementado em linguagem R, no pacote DNAcopy; foram utilizados os parâmetros padrão. Regiões com ocorrência normal de CNV foram eliminadas para não desorientar a interpretação, de acordo com a versão 18v1 da *Database of Genomic Variants* (http://projects.tcag.ca/variation/) e das anotações de variantes estruturais do *UCSC Genome Browser*.

Estes dados são os mesmos acessados pelo programa CGB, exceto pelas amostras TCGA-06-0208-01B e TCGA-06-0211-01B que foram incluídas no CNViewer. Este conjunto de dados foi armazenado em um repositório S3DB (http://ibl.mdanderson.org/CNViewer), com estrutura distinta à acessada pelo CGB. Os arquivos de dados segmentados estão disponíveis publicamente utilizando a palavra "*public*" como usuário e senha.

No portal TCGA estão disponíveis também dados clínicos, complementares aos dados brutos de CNV; ambos os tipos dados foram utilizados para ilustrar o uso do CNViewer neste trabalho. Todas estas informações provêm do banco de dados público do TCGA e estão sujeitas às políticas de privacidade de doadores e confidencialidade de dados. Estas políticas permitem acesso mínimo aos dados clínicos, oferecendo um perfil de informações único para cada um dos indivíduos, porém contendo poucas informações para evitar re-identificação do paciente.

O PROJETO *THE CANCER GENOME ATLAS* (TCGA)

Este é um projeto dos Institutos Nacionais de Saúde (*National Institutes of Health* – NIH) do governo dos Estados Unidos da América (EUA), iniciado em dezembro de 2005, que visa compreender a base molecular do câncer, através de análises genômicas, para no futuro aperfeiçoar os métodos de diagnóstico, tratamento e prevenção de câncer.

Os procedimentos incluídos no projeto iniciam-se desde a coleta e processamento de amostras de tecidos cancerosos e controle até o sequenciamento e outras caracterizações genômicas, bem como a análise de resultados. As principais características estudadas incluem CNV, aberrações cromossomais, perda de heterozigosidade, alterações epigenéticas,mudanças na expressão gênica e de micro RNAs, além da análise de mutações por sequenciamento.

O TCGA financia também o aprimoramento ou desenvolvimento de novas técnicas de larga escala e metodologias com melhor relação custo-benefício. Objetivam também o desenvolvimento de técnicas eficientes para correlação entre o perfil molecular e clínico, além de criar interfaces amigáveis para divulgação de dados.

O projeto pretende analisar mais de vinte tipos de tumor e inicialmente, para a realização de um projeto piloto, foram escolhidos o glioblastoma multiforme, câncer de ovário e câncer de células escamosas de pulmão. Esta escolha foi feita com base em uma pesquisa realizada, que avaliou as coleções de espécimes quanto a sua qualidade e quantidade. Destes três o Glioblastoma Multiforme foi o primeiro a ser pesquisado e os primeiros resultados foram divulgados em setembro de 2008, mas a pesquisa deste tumor prossegue até hoje (National Cancer Institute e National Human Genome Research Institute n.d.)

5.2.2 Dados mantidos com o usuário

Praticamente todos os programas de segmentação são capazes de gerar arquivos texto delimitados por tabulação. O CNViewer é flexível quanto ao formato de entrada, requerendo apenas que o arquivo contenha dados de um segmento em cada linha, com as informações sobre a amostra a que ele pertence, cromossomo, NC, posição genômica de início e fim (Figura 5.1). Estes arquivos podem conter inúmeras amostras.

```
     1  sample.ID       chromosome      segment.start   segment.end     segment.mean
     2  exemplel        1       1018584 74674800        2.3068
     3  exemplel        1       75010081        123000895       1.7863
     4  exemplel        1       123002455       184638762       2.0785
     5  exemplel        10      216098  133432567       2.3024
        ...
 45161  exemple231      1       1018584         5478935 2.8759
```

Figura 5.1: Modelo de dados de CNV

Ao carregar o arquivo, é solicitado ao usuário que seja feita a correspondência entre as colunas identificadoras de seus dados e as informações exigidas, tornando assim o CNViewer independente de um formato padrão, palavra-chave ou ordem no arquivo de saída. Não há impedimento de leitura de arquivos que possuam colunas com outros, elas são simplesmente ignoradas. O arquivo pode ser carregado via formulário, para ser lido com auxílio de *scripts* PHP ou seu conteúdo pode ser colado em uma área de texto e lido diretamente pelo programa (Figura 5.2).

Ao carregar dados via formulário, o usuário informa a localização de seu arquivo, a qual é submetida para o *script* PHP que lê o conteúdo e remete a uma estrutura do DOM, para então ser acessível ao JavaScript, que continua automaticamente com o processamento. Este processo é otimizado, quando o usuário realiza o procedimento de inserir o conteúdo de seu arquivo numa estrutura DOM, copiando e colando os dados numa caixa de texto.

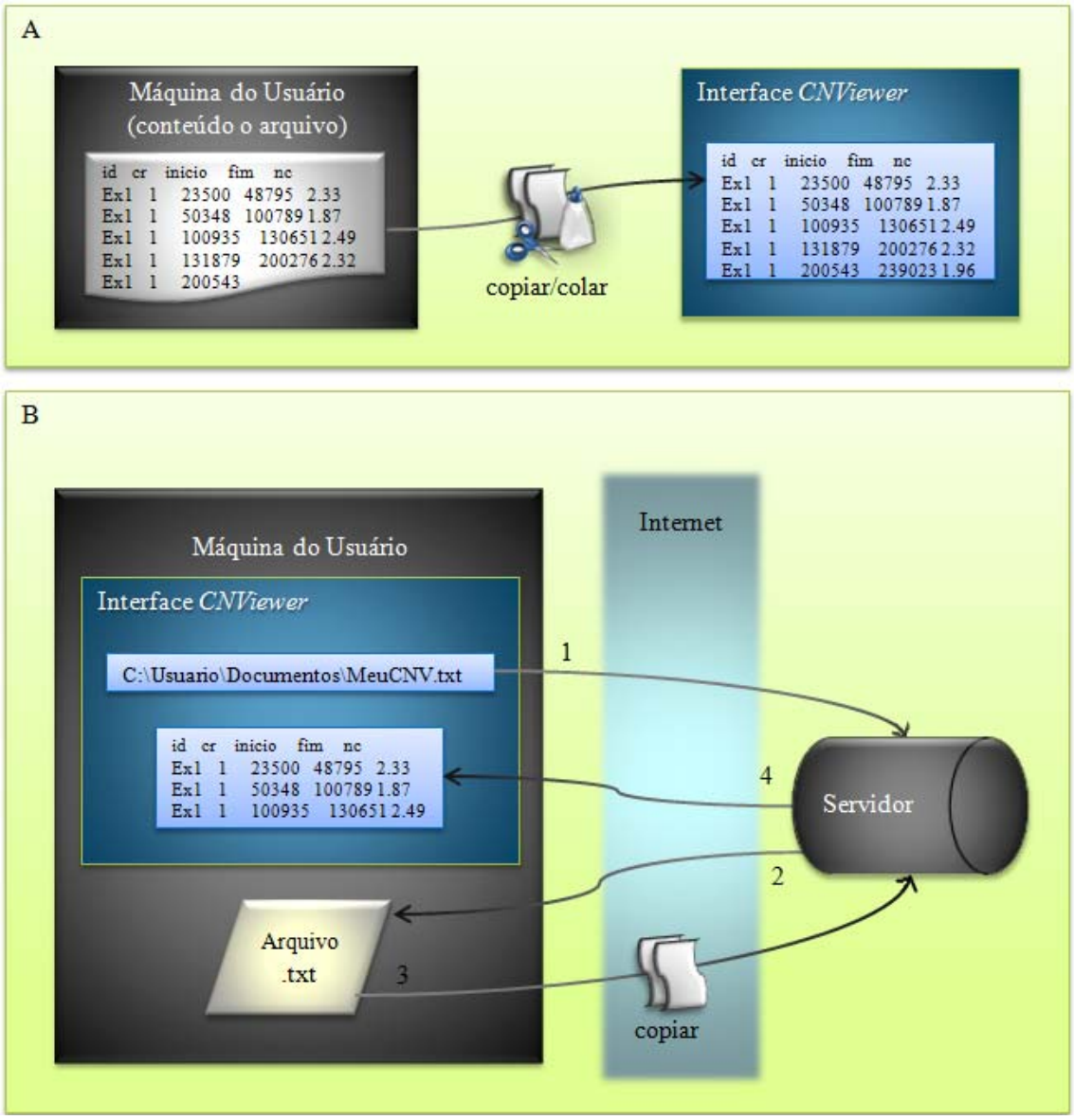

Figura 5.2: Metodologias para acessar o conteúdo de arquivo mantido pelo usuário. O método A consiste de um procedimento simples, no qual o usuário esta encarregado de executar a tarefa de inserir no programa o conteúdo do arquivo. O método B, exige que o usuário localize o arquivo através da interface do CNViewer, para então este endereço ser enviado (1) a um *script* no servidor, que então acessa (2) o arquivo e faz uma cópia (3) no servidor, que envia (4) o conteúdo a um elemento da interface.

5.2.3 Dados em repositório S3DB

Manter dados em simples arquivos texto é mais simples, porém não necessariamente mais prático. Frequentemente ocorre o armazenamento desnecessário

de dados, devido à criação de múltiplas cópias, ou mesmo perda ou alteração de informações. Uma maneira segura e prática de manter estes dados é utilizar um banco de dados. Assim foi implementado no CNViewer métodos para acesso de dados em bases do tipo S3DB.

Este tipo de banco oferece diversas vantagens explanadas na seção 4.3 e é usado em vários projetos do *M. D. Anderson Cancer Center*, inclusive é a única forma de incluir dados no CGB. Assim, a implementação de um banco de dados emprega certo esforço, mas oferece vastos benefícios, inclusive compartilhamento de dados.

Os dados mantidos no S3DB estão sob forma de arquivos individualizados por amostra. Neste modelo o administrador do banco de dados pode oferecer senhas de acesso aos dados a quaisquer usuários e inclusive conceder acesso público.

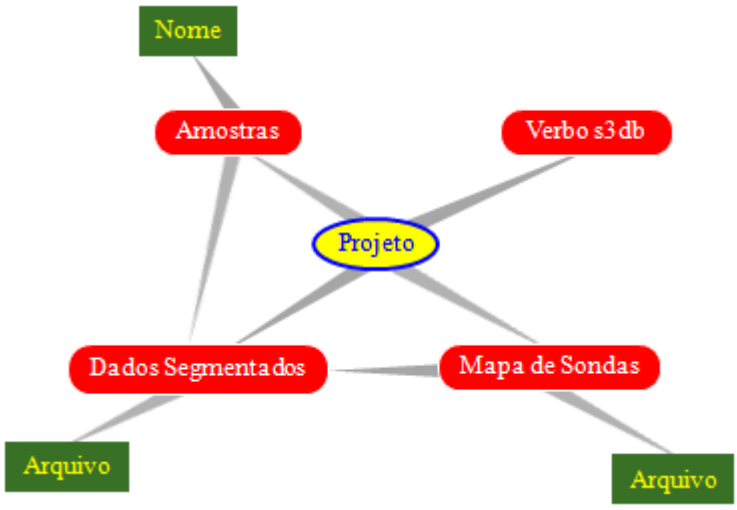

Figura 5.3: Estrutura de dados de repositório S3DB compatível com CNViewer. Mapa adaptado com base na representação gerada pela interface do S3DB. Os elementos do banco são vinculados por verbos, este relacionamento está representado em cinza, sendo que a extremidade espessa da linha indica o sujeito da relação e a outra ponta, o objeto. Um Projeto possui diversas coleções (caixas vermelhas), as quais estão armazenando os dados (caixas verdes) sob forma de texto ou arquivo.

CNViewer oferece uma interface gráfica para entrada de usuário, senha e endereço do banco de dados. Quando o usuário é reconhecido, uma palavra-chave é gerada e é utilizada para carregamento automático das informações contidas no repositório. O CNViewer reconhece os verbos de ligação entre as coleções de informações no S3DB, identificando os tipos de dados, assim após o carregamento, o usuário já pode explorar os dados.

As informações contidas num repositório S3DB têm formato restrito, representado pela Figura 5.3. Todas as informações estão distribuídas em coleções, separadas em nome, dados de CNV e mapa de sondas. Uma quarta coleção contém os verbos utilizados para relacionamento dos elementos do banco. Cada amostra possui uma identificação na coleção "Amostras" e um arquivo texto em "Dados Segmentados", com seus CNVs, seguindo o modelo da Figura 5.4. Não possuindo cabeçalho, cada linha deste arquivo deve conter o cromossomo, posição genômica de início e fim do segmento e seu NC, sempre nesta ordem. Este formato deve ser seguido para garantir que os dados possam ser interpretados corretamente pelo CNViewer.

```
1   1       1018591 58086531        2.04968185977168
2   1       58093497        74674749        2.13006875655443
3   1       74693681        74877559        1.55606249133981
4   1       74885033        74952090        1.24444286960328
5   1       74961547        75110280        1.59825765081566
6   1       75148431        109266600       2.13657505819862
7   1       109467543       184039686       2.06322334190196
    ...
118 9       28145093        136964428       2.17015787514697
```

Figura 5.4: Modelo de dados de CNV para repositório S3DB

5.3 Métricas de Dissimilaridade

Os estudos frequentemente buscam descobrir subtipos de um grupo de amostras, tendo como princípio que a semelhança entre um conjunto de fatores específicos possa estar refletindo em outras características. Os perfis de CNV são um dos critérios que podem ser utilizados para estabelecimento de subtipos, aplicação bem sucedida na identificação de subtipos moleculares de tumores de um mesmo tecido (van de Wiel et al., 2010).

5.3.1 Conceito de métrica

Uma métrica em Álgebra Linear é definida como uma função (d: M x $M \rightarrow \mathbb{R}$) que associa a cada par ordenado de elementos (p, q) pertencentes a um conjunto (M) um número real (produto interno de p e q), chamado de distância de p a q, de modo a satisfazer as seguintes condições para quaisquer p, q, $r \in M$ (Lima, 1975, Rolewicz, 1985):

a) $d(p, p) = 0$;

b) Se $p \neq q$ então $d(p, q) > 0$;

c) $d(p, q) = d(q, p)$;

d) $d(p, r) \leq d(p, q) + d(q, r)$.

Os postulados **a** e **b** dizem que $d(p, q) \geq 0$ e que $d(p, q) = 0$, se, e somente se, $p = q$. O postulado **c** afirma que a distância $d(p, q)$ é uma função simétrica das variáveis p, q. A condição **d**, chamada "desigualdade do triângulo", baseado no fato de que, em um plano euclidiano, o comprimento de um dos lados de um triângulo não excede a soma dos outros dois (Lima, 1975), portanto $d(p, r) = d(p, q) + d(q, r)$, se , e somente se, q estiver contido na reta que intercede p e r.

A distância entre dois pontos geométricos, pertencente ao conjunto $\mathbb{R}$ dos reais é dada por $d\ (p,\ r) = d\ |p,\ q|$, chamada de "métrica usual" da reta. Há três maneiras naturais de se medir a distância entre dois elementos – elementos podem ser pontos, conjuntos, funções, etc (Lima, 1975). Dados $p= (p_1, ..., p_n)$ e $q= (q_1, ..., q_n)$:

(5.1)

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + \ldots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i} \cdot d(p,q) =$$

$$\sqrt{(p_1 - q_1)^2 + \ldots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}\,,$$

$$d'(p,q) = |p_1 - q_1| + \ldots + |p_n - q_n| = \sum_{i=1}^{n}|p_i -$$

$q_i|\,,$ $\qquad d'(p,q) = |p_1 - q_1| + \ldots + |p_n - q_n| = \sum_{i=1}^{n}|p$ (5.2)

$$d''(p,q) = \max\{|p_1 - q_1|, \ldots, |p_n - q_n|\} =$$

$\max\limits_{1\le i\le n}|p_i - q_i|\,.$ $\qquad d''(p,q) = \max\{|p_1 - q_1|, \ldots, |p_n - q_n|\} = \max\limits_{1\le i\le n}|p$ (5.3)

As três funções acima são métricas, sendo que a métrica dada pela Equação 5.1 é chamada de "euclidiana". Ela provém da fórmula para menor distância entre dois pontos em um plano cartesiano, a qual é provada por Teorema de Pitágoras. A métrica euclidiana é considerada uma métrica natural pois fornece a distância da Geometria Euclidiana. As outras duas métricas são "equivalentes" à euclidiana, porém formalmente mais simples (Lima 1975).

5.3.2 Métrica selecionada

O estudo de perfis genômicos ou moleculares envolve a análise de fatores complexos e necessitam de estudo a parte para o desenvolvimento de métodos

específicos. Nesta proposta inicial de análise comparativa exploratória entre perfis de CNV, optou-se por utilizar a métrica euclidiana, por ser uma métrica natural

O objetivo de inserir a DE é dar suporte à análise exploratória da relação de uma amostra de CNV e um conjunto. Considerando cada amostra um elemento e cada uma dos *n* segmentos um componente deste elemento, a DE pode ser utilizada para comparar quais amostras são mais próximas e portanto mais similares. Por exemplo, as amostras *p* e *q* compostas respectivamente pelos segmentos (p_1, ..., p_n) e (q_1, ..., q_n) cujos valores correspondem ao NC. Se $p = q$ a DE calculada é 0 (postulado **a**, seção anterior), portanto não há dissimilaridade – elas são idênticas. Caso $p \neq q$, o valor da DE calculada será tão grande quanto for a diferença de cada par (p_i, q_i), medindo a dissimilaridade entre os NC das amostras.

Os pontos utilizados na métrica para dissimilaridade de CNV são os pontos de quebra, pois representam exatamente as regiões nas quais ocorrem as mudanças de NC (Figura 5.5 A). Cada amostra possui um padrão diferente de segmentos assim, antes da realização do cálculo intercalam-se os pontos de quebra do par de amostras (Figura 5.5 B) e posteriormente é obtida a DE comparando os valores de CNV em cada uma destas regiões.

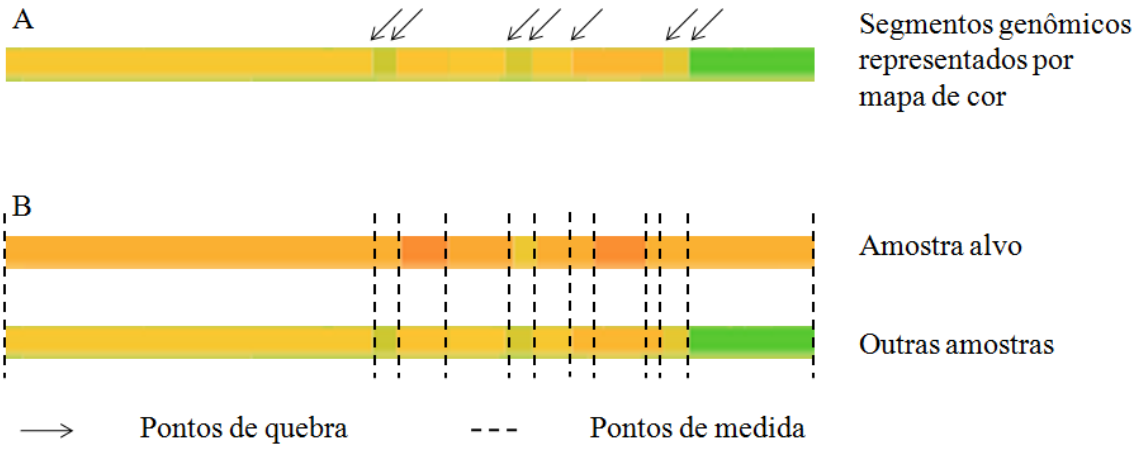

Figura 5.5: Locais utilizados nas medidas de dissimilaridade . (A) Exemplo de segmento genômico, no qual diferentes cores representam valores distintos de NC. As regiões delimitantes dos segmentos são os pontos de quebra. (B) A sobreposição destes pontos é feita antes do cálculo da DE e a medição é realizada com base nos segmentos gerados por essa sobreposição.

O cálculo da DE realizado desta forma não inclui a região genômica selecionada de forma contínua, pois os intervalos entre as variações não são contabilizados. Assim, é uma medida cumulativa das diferenças entre pontos do genoma, chamada de distância euclidiana (DE) em pontos de quebra (*Breakpoint Euclidean Distance*). Considerando-se estes trechos que não foram medidos, foram criadas duas metodologias de DE ponderada.

Antes de detalhar como o cálculo ponderado é realizado, é importante notar a metodologia de cálculo em relação à sua implementação. Os dados de CNV são armazenados sob forma de segmentos, para os quais se tem o valor de NC e posição das sondas que delimitam seu início e fim; portanto o acesso aos pontos de quebra é favorecido pela arquitetura dos dados. Ainda, para as medidas baseadas em sondas, o cálculo da distância sonda a sonda seria redundante, podendo ser substituído pelo cálculo nos pontos de quebra multiplicado pelo número de sondas, o qual consome menos tempo e recursos computacionais.

A ponderação (Equação 5.4) multiplica o quadrado da diferença entre dois pontos (p e q) por um peso (w), resultando numa dissimilaridade entre os segmentos. Cada segmento de uma amostra é obtido através da normalização de um conjunto de sondas. Portanto, se w é o número de sondas do segmento, multiplicar por w, é equivalente a somar o quadrado das diferenças obtidas para cada uma das sondas de um segmento.

$$\sqrt{\sum_{i=1}^{n} w_i (p_i - q_i)^2} \tag{5.4}$$

A DE ponderada por sondas (*Probe Weighted Euclidean distance*) utiliza este princípio. Porém, como o número de sondas é muito grande, o w utilizado é um valor relativo, entre o segmento e o cromossomo (Equação 5.5). É calculado através da contabilização do total de sondas no intervalo em questão (s) dividido pelo valor total de sondas (S) do cromossomo (c). (5.5)

$$w_i = {}^{s_i}/_{S_c}$$

A implementação desta medida, depende da disponibilização de um mapa de sondas, que infelizmente nem sempre está disponível. Foi desenvolvida assim a DE ponderada por tamanho (*Size Weighted Euclidean distance*) na qual o w é obtido com base no tamanho dos segmentos, medido em pb. O comprimento do i-ézimo segmento é obtido pela subtração entre os pontos de quebra que o delimitam, isto é, a diferença entre as posições genômicas das sondas final e inicial. Posteriormente este valor (Equação 5.6 variável l) é dividido pelo tamanho total (L) do cromossomo (c), medido em pares de base de acordo com a montagem 36 do *Human Genome Assembly Information*.

$$w_i = {}^{l_i}/_{L_c} \tag{5.6}$$

5.4 Representação por Mapa de Calor

O mapa de calor permite a visualização das CNVs de múltiplas amostras, pois diferentes perfis podem ser exibidos lado a lado, com seus NC representados em cores. Esta estratégia é utilizada nos softwares CGH Analyzer, CGB, Magellan, MD-SeeGH e VAMP.

Foi implementado no CNViewer um mapa de calor com intervalo fixo de valores, com gradiente de cores variando uniformemente entre 0 e 4 (Figura 5.6), destacando com clareza deleções (azul escuro em caso de deleção de duas cópias, azul claro para uma cópia) e duplicações (amarelo para NC entorno de três e vermelho para mais cópias). No caso de valores superiores a 4, apenas a intensidade do vermelho é aumentada, até 100. Valores superiores recebem a cor equivalente a 100.

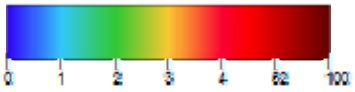

Figura 5.6: Correspondência de cores do mapa de calor

Ao carregar o arquivo com dados segmentados, o programa calcula a cor correspondente aos NC de cada amostra e armazena junto com as demais informações, para otimizar o processamento das futuras análises.

5.5 Exportação e Recuperação de Análise

Uma característica importante do CNViewer é a exportação da análise, de forma a incluir não apenas o resultado, mas também permitir que novas explorações de dados sejam realizadas. Uma propriedade intrínseca dos navegadores é a opção de salvar uma página, porém são armazenados os diversos arquivos relacionado ao site, mas o conteúdo que consta em memória destinada ao navegador (como os dados acessados pelo CNViewer) não são salvos. Além disso, os arquivos armazenados ocupariam um espaço desnecessário na memória e a cada novo arquivo salvo, haveria uma série de arquivos repetidos. Assim foi criado um módulo de exportação, que armazena apenas os dados necessários para recuperação da análise.

O módulo de exportação consiste de uma nova página gerada pelo CNViewer, a qual é povoada com os dados de CNV e complementares. A interface do módulo contém informações relevantes sobre a última análise realizada pelo usuário e um *link*, apontando para a URL do aplicativo, acrescida de uma estrutura contendo todos os dados utilizados na análise. Estes dados contidos na URL estão no formato de notação de objetos do JavaScript (*JavaScript Object Notation* – JSON), uma forma leve para transferência de dados, facilmente interpretada.

Os dados são submetidos via URL pois é atualmente a única forma de envio de informações a uma página, utilizando somente JavaScript. O CNViewer verifica o conteúdo da URL primeiramente e em seguida gera a interface utilizando os dados submetidos, e recriando a análise salva.

5.6 Teste de desempenho

A forma de implementação utilizada requer que os navegadores possuam uma boa capacidade de armazenamento. Foram realizados testes do desempenho do CNViewer em quatro navegadores populares (Google Chrome, Mozilla Firefox 3.5.9,

Opera 9.80 e Apple Safari 4.0.3), excluiu-se o Internet Explorer por não oferecer as funcionalidades padrão, impedindo o desenvolvimento de interface com qualidade. Sete rotinas foram testadas: carregamento da interface; carregamento dos dados modelo armazenados no S3DB; plotagem da análise utilizando as três métricas e os 22 cromossomos; carregamento do módulo de exportação; e recuperação da análise.

Com o objetivo de comparar o desempenho de uma implementação em navegador com uma *stand-alone*, mediu-se também o desempenho do CNViewer em relação ao CGB, por ser o aplicativo a oferecer a funcionalidade mais próxima a do CNViewer, além de ter acesso aos mesmos dados utilizados. A estrutura do CGB é distinta, assim comparou-se especificamente a plotagem do conjunto de amostras sob a forma de mapa de calor. No CGB a visualização é feita cromossomo a cromossomo e consequentemente foram feitas medições para cada um dos autossomos. Apenas o CNViewer solicita a escolha de uma amostra alvo, assim, os testes foram feitos utilizando a amostra TCGA-02-0001-01C como alvo por ela ser a primeira na listagem de amostras. Os testes foram realizados no Mozilla Firefox 3.5.9, devido ao fato de ser o ambiente de desenvolvimento do CNViewer e para execução do CGB versão 12, foi utilizado o MATLAB 7.4.0 (R2007a).

O objetivo destes testes foi fazer uma comparação qualitativa, para estimar a eficiência deste modelo de implementação. Assim, foi realizado um teste empírico com cinco tomadas de tempo para cada condição, no qual as mediadas foram feitas de forma intercalada, uma por ambiente, para criar condições de teste mais próximas possíveis. O tempo foi medido com cronometro simples. Os testes foram realizados em sistema operacional Microsoft Windows 6.0 (Home Premium) 32 bit, com processador Intel Core 2 Duo 1.50GHz e memória RAM com 2038 MB.

Capítulo 6

Resultados e Discussão

Diferentemente dos programas baseados em Web, o CNViewer executa as solicitações do usuário no próprio navegador, enquanto os outros processam muitas das funções no servidor externo. Devido a esta característica descrevemos o CNViewer como um aplicativo baseado em navegador.

Essa abordagem de implementação é interessante pois os dados biológicos costumam ser volumosos, assim a troca de informações via rede consome tempo e o processamento em servidores externos, demanda maior manutenção e suporte.

Mover programas de análise para o navegador torna a integração de dados mais fácil, e as vantagens da acessibilidade de programas online possui diversas ramificações. Por exemplo, o compartilhamento/divulgação de resultados geralmente é feito através de gráficos estáticos, mas com a criação do módulo de exportação, há uma mudança nesta perspectiva: o envio do módulo ou publicação de um *link*, levam a uma interface dinâmica, ampliando as possibilidades de interpretação e permitindo a realização de mais análises.

6.1 Fluxo de Trabalho

O Capítulo anterior descreveu individualmente as características implementadas no CNViewer, mas para compreender o fluxo de trabalho do aplicativo, os principais passos de execução estão representados na Figura 6.1.

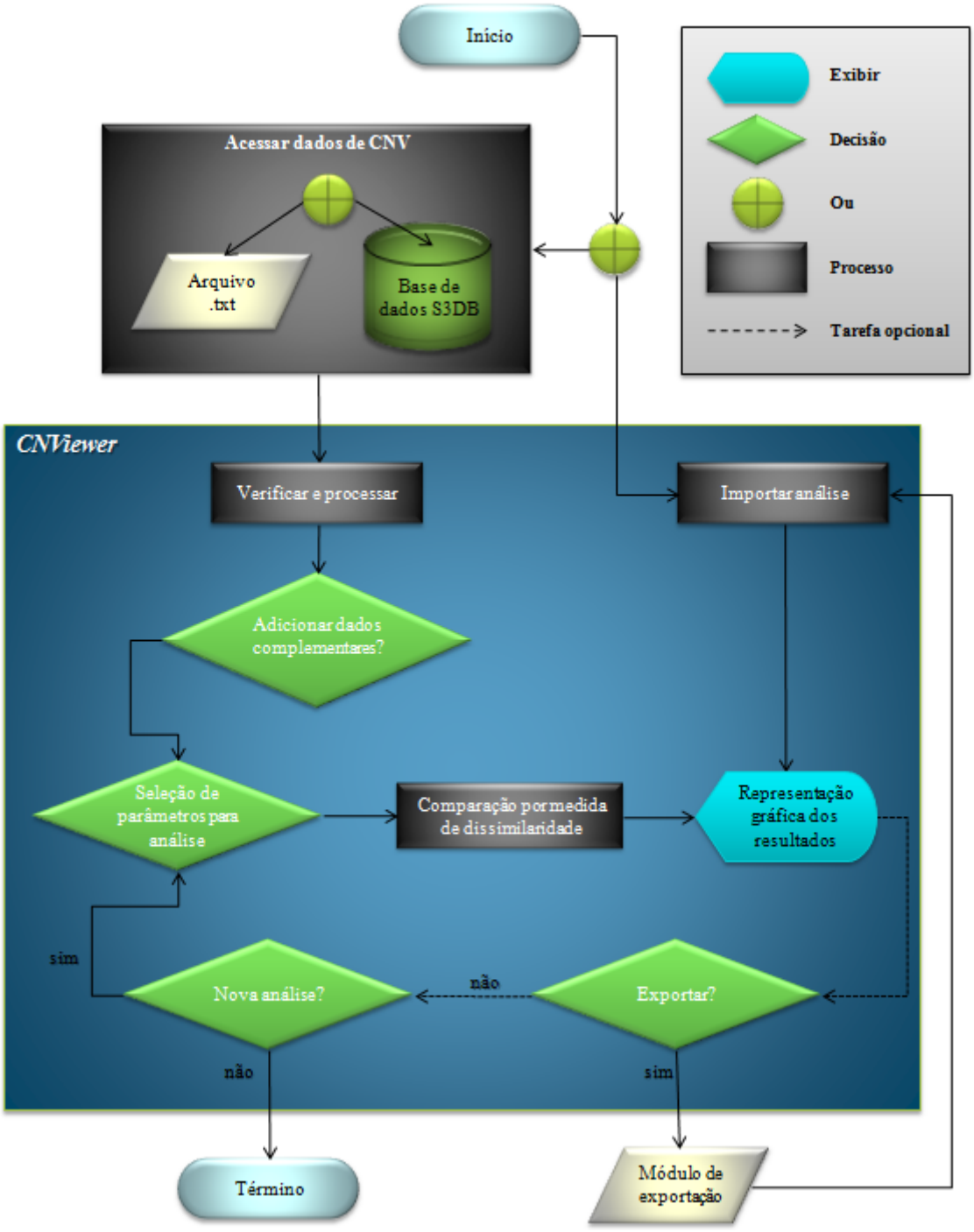

Figura 6.1: Diagrama de fluxo simplificado do CNViewer

O acesso direto do CNViewer (sem envolver o módulo de exportação) gera uma interface com todas as ferramentas, mas sem dados, requerendo a entrada de dados.

Estes dados devem estar no formato de texto limitado por tabulação, em posse do usuário ou num banco de dados (local ou remoto). Uma vez verificado o formato dos dados, eles são processados, e armazenados em uma estrutura JavaScript do navegador. Posteriormente é oferecido ao usuário a possibilidade de inserir dados complementares, ou mais dados de CNV.

Após estes procedimentos, o CNViewer está pronto para realizar a análise, sendo necessário apenas que o usuário escolha os parâmetros a serem utilizados. O aplicativo então executa a comparação entre as amostras, gerando um gráfico com os resultados obtidos. Caso haja interesse, um módulo de exportação é gerado e independente disso, o usuário pode iniciar uma nova análise alterando os parâmetros ou simplesmente finalizar o uso do programa.

Com o módulo de exportação disponível, o usuário sempre poderá enviar automaticamente para o CNViewer os dados necessários para recuperação da análise. Quando a recuperação é solicitada, o aplicativo primeiro recupera os dados submetidos, para então gerar a interface e os gráficos para visualização. Novas análises continuam podendo ser feitas e salvas.

6.2 Dados de Entrada

6.2.1 Dados mantidos com o usuário

Assim que o CNViewer é iniciado, é exibida uma janela com as opções de carregamento (Figura 6.2 A), levando à caixa de texto para inserção manual de dados (Figura 6.2 B) ou ao campo para carregamento de arquivos (Figura 6.2 C). Em ambos os casos o formato dos dados submetidos é verificado e, se compatível, uma nova interface é apresentada.

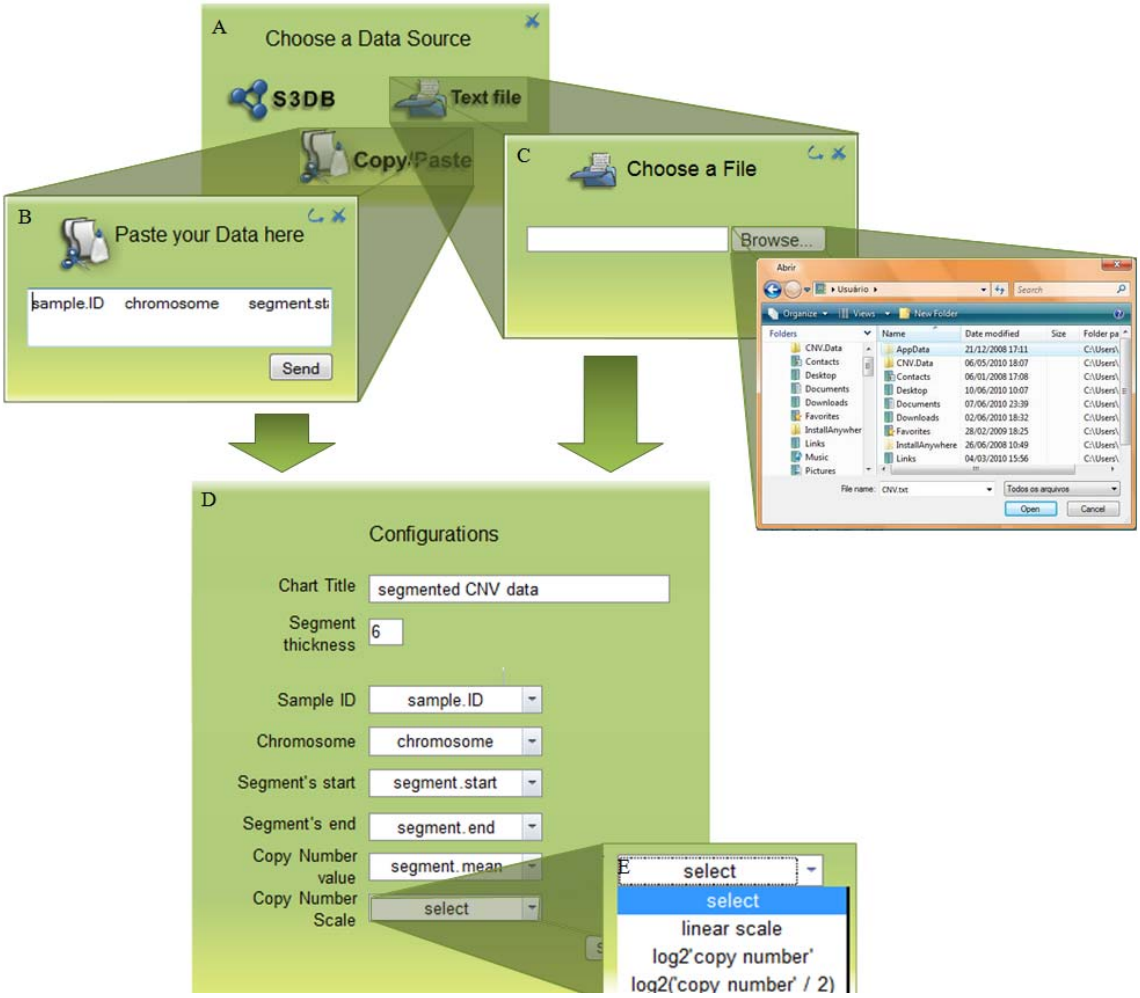

Figura 6.2: Inserção de dados mantidos com o usuário. Três opções de inserção de dados são oferecidas ao usuário (A), duas delas permitem que sejam carregados dados mantidos em arquivos texto. A primeira opção, copiar e colar (B), exibe uma caixa de texto para colocar o conteúdo do arquivo. A segunda opção (C) permite o carregamento do arquivo. Uma vez que os dados estão disponíveis no CNViewer, uma nova interface surge (D), listando opções do gráfico (título e espessura de segmento) e também solicitando ao usuário que identifique quais são as colunas que correspondem aos campos necessários para uso do programa. O usuário deve também informar a escala que o valores de NC estão representados.

Esta nova janela, possui uma lista com as informações requisitadas pelo programa: identificação da amostra, número do cromossomo, posição de início e fim do segmento em pares de base e o NC (Figura 6.2 D). Ao lado de cada um destes campos, são listados os identificadores das colunas dos dados submetidos. O usuário deve correlacioná-los, e informar a notação utilizada para representação dos NC (Figura 6.2

E). O CNViewer verifica se não há repetição de colunas selecionadas para então o processamento de dados ocorrer. Em caso de erro, o usuário é solicitado a corrigir a correspondência de colunas e campos ou a inserir um novo arquivo.

O procedimento de inserção de dados pode ser repetido inúmeras vezes e, ao final deste processo, os dados complementares podem ser adicionados. O carregamento de dados complementares utiliza a mesma metodologia, porém não existem campos obrigatórios, basta apenas que os tipos de dados estejam identificados na primeira linha do arquivo e que as demais linhas correspondam apenas a uma amostra. É imperativo que os dados complementares estejam ordenados com correspondência aos dados de CNV.

Os dados segmentados podem ser obtidos de inúmeros programas, cujo arquivo gerado pode ter variações no posicionamento de colunas e escala dos valores de NC. A possibilidade de indicar a correspondência das colunas com as informações requisitadas, assim como a escolha da escala dos dados, amplia a quantidade de formatos de arquivos compatíveis com o uso do CNViewer.

Apesar disso, alguns detalhes do formato do arquivo devem ser considerados com cuidado, por exemplo, os números decimais devem ser demarcados com pontos e não vírgulas, pois a maioria do algoritmos de análise de CNV utilizam esta notação. O usuário deve evitar o uso de espaço e anotar os cromossomos omitindo o caractere zero a esquerda. Uma particularidade observada em alguns tipos de arquivos de saída, é o uso de número neperiano para denotação dos NC, neste caso o CNViewer esta apto a processar este tipo de entrada automaticamente.

6.2.2 Dados em repositório S3DB

Os repositórios S3DB ao serem criados recebem um endereço que localiza o banco de dados e o servidor no qual estão instalados. Independente de estar num

servidor local ou remoto, este endereço é uma URL e deve ser indicada todas as vezes que for solicitado o acesso ao banco. Deste modo, ao escolher a opção “S3DB” na interface inicial do CNViewer (Figura 6.3 A), o usuário deve preencher um pequeno formulário de acesso (Figura 6.3 B), no qual insere sua identificação, senha e domínio do banco, que é exatamente a URL. O par usuário/senha pode estar sob autoridade do S3DB ou do Google, assim também deve ser informada qual destas duas autoridades validará esses dados.

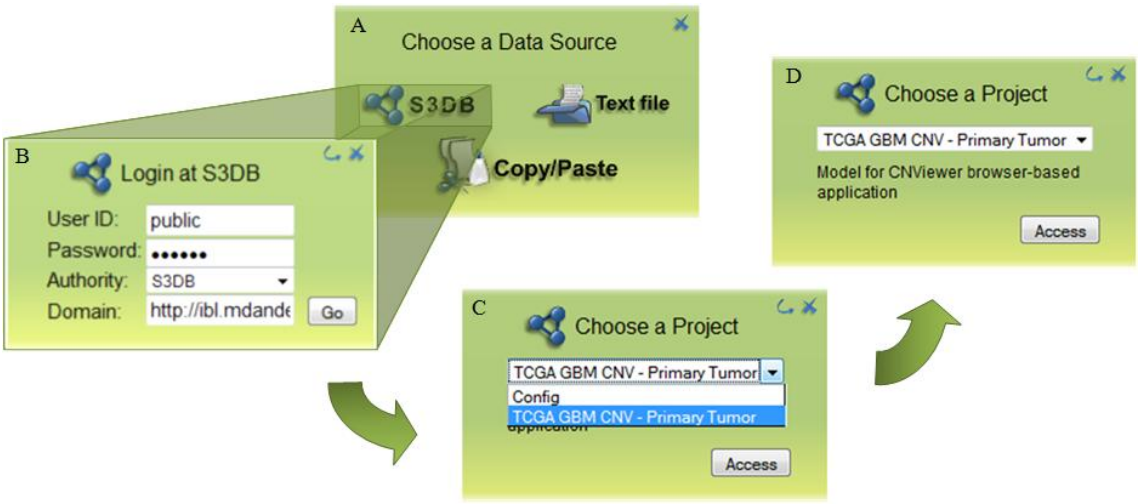

Figura 6.3: Inserção de dados em repositório S3DB. A primeira opção da interface inicial (A) leva o usuário ao formulário (B) para conexão com um repositório do tipo S3DB. Depois de validados os dados enviados, o CNViewer recebe uma lista dos projetos vinculados àquela conta (C). É exibida uma breve descrição do projeto selecionado (D) e uma vez escolhido um projeto, o usuário solicita o acesso.

Os dados são enviados ao domínio informado e, se reconhecidos, são listados todos os projetos a que o usuário tem acesso. Apenas um projeto deve ser escolhido e se ele tiver o formato compatível ao CNViewer, o carregamento de dados é iniciado, levando alguns minutos para finalmente o usuário poder iniciar suas análises.

6.3 Representação dos CNVs

Uma vez estando com os dados para análise, o usuário precisa apenas determinar os parâmetros que serão utilizados, escolhendo a métrica, a amostra alvo e as regiões genômicas. A análise é rapidamente executada e exibida em forma de gráfico. Os CNVs são retratados em um mapa de calor (Figura 6.4 E) no qual o eixo *x* possui a posição genômica e o eixo *y*, as amostras. Em cada linha horizontal consta uma amostra, a qual é identificada por pontos do lado esquerdo do gráfico (Figura 6.4 C) – ao colocar o cursor sobre estes pontos, o nome da amostra é exibido. Os segmentos das amostras estão presentes ao longo do eixo *y* e os NC são representado em cores, assim, cada mudança de cor observada, indica um segmento diferente (Figura 6.4 B). Regiões para as quais não há dados, tais como trechos genômicos sem cobertura de sondas ou como aquelas regiões excluídas da análise(ver seção 5.2.1) aparecem em branco.

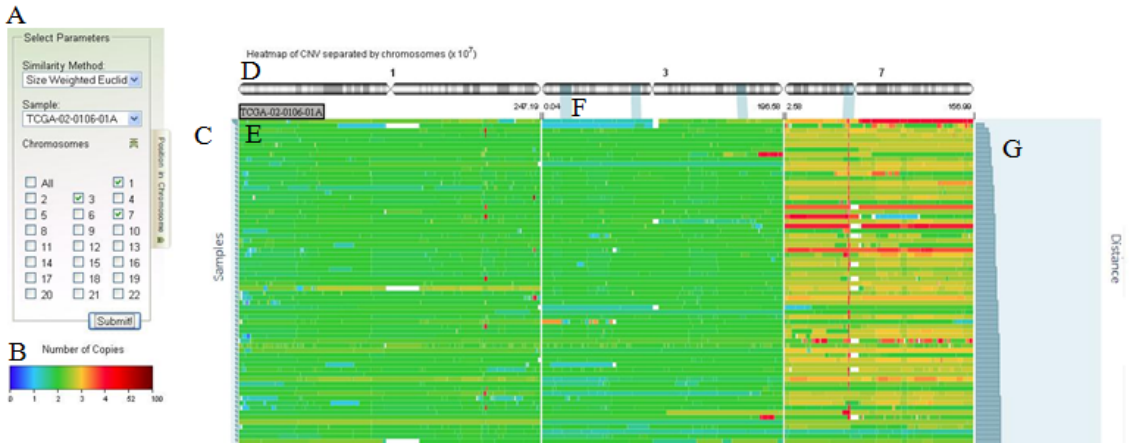

Figura 6.4: Representação gráfica

. (A) Parâmetros da análise; (B) escala de cores e correspondência com o valor dos NC. (C) Eixo contendo as amostras (representadas pelos pontos azuis), em detalhe a identificação da amostra. (D) Regiões genômicas separadas por cromossomo – apenas os cromossomos selecionados em (A) são mostrados. (E) Segmentos exibidos através de mapa de calor; (F) as regiões genômicas analisadas (em azul) relacionando a posição no cromossomo com os segmentos plotados. (G) Gráfico com a distância calculada.

Apenas as regiões genômicas dos cromossomos selecionados para análise são plotadas. O gráfico é separado em autossomos (Figura 6.4 D), nos quais são demarcados os intervalos analisados do genoma (Figura 6.4 F). Decidiu-se exibir todo o conjunto de segmentos de um cromossomo e apenas indicar a região analisada (Figura 6.4 F), pois assim, o usuário sempre visualizará o cromossomo completo e caso deseje observar com mais detalhe uma determinada região, pode utilizar o *zoom* do próprio navegador (teclas de atalho "ctrl" + "+", *zoom out* "ctrl" + "-"). É possível também plotar outros cromossomos sem incluí-los na análise, por exemplo o cromossomo 1 na Figura 6.4, para tanto basta eliminar os marcadores de regiões cromossômicas (abordado na seção 6.4).

O ordenamento dos segmentos é feito de acordo com a dissimilaridade calculada. A primeira amostra listada é a selecionada como alvo (Figura 6.4 A, amostra TCGA-02-0106-01A) e as conseguintes são ordenadas crescentemente, de acordo com a dissimilaridade calculada. Em caso de empate, a ordem é estabelecida pela posição na lista de amostras (Figura 6.4 A). É possível alterar a espessura dos segmentos, permitindo flexibilidade na sua visualização, além de poder modificar a qualquer momento, o título do gráfico, basta pressionar o botão "*Chart Options*", conforme demonstrado na Figura 6.5.

O CNViewer oferece uma forma prática para análise exploratória de populações, pois além de possibilitar a visualização do conjunto completo de amostras, oferece métodos de comparação entre elas (abordado na seção 6.5). A visão global dos dados permite a busca por padrões característicos de uma amostra, indicando regiões candidatas à análise detalhada.

Com o auxílio destas ferramentas, usuários também podem buscar alvos específicos, selecionando regiões genômicas. Finalmente, podem-se entrar dados de

uma nova amostra para ser comparada com um conjunto pré-existente. A analise do perfil novo no contexto de um conjunto já conhecido, tem como objetivo a inferência de características desconhecidas na amostra nova, com base no conhecimento prévio das características das amostras com perfil molecular semelhante.

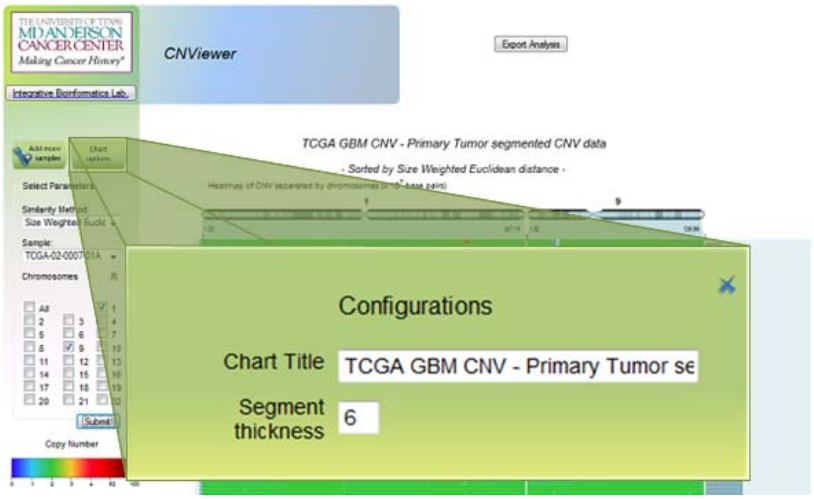

Figura 6.5: Configuração do gráfico . Durante o uso do CNViewer, o título e espessura de segmentos podem ser alterados. A interface é acessível pelo botão "*Chart Options*".

Um dos principais aspectos do CNViewer é a visualização de múltiplas amostras, mas não há impedimento para a exibição de uma única amostra. A visualização de um perfil de único também é importante para o diagnóstico e tratamento de algumas doenças. Um exemplo é a distinção entre os gliomas do tipo oligodendroglioma maligno e o glioblastoma. A identificação deles é essencial para o tratamento, pois o primeiro é menos agressivo e pode ser tratado apenas com drogas na maioria das vezes, já o glioblastoma precisa de quimioterapia e radioterapia.

Infelizmente a abordagem diagnóstica padrão, baseada em histologia, nem sempre permite a distinção entre estes tumores. Por outro lado, eles são facilmente identificados por perfil genético: o oligodendroglioma caracteristicamente possui

deleção dos braços cromossômicos 1p e 19q e o glioblastoma costuma apresentar amplificações no cromossomo 7, principalmente na região do *EGFR*, deleção no 10 e dupla-deleção no 9 (Kotliarov et al. 2010). Portanto, o CNViewer pode ser utilizado também para análise de perfis individuais, funcionalidade que possui inclusive aplicação em diagnóstico, além da comparação de perfis ter potencial para sugestão de outros diagnósticos baseado em tratamentos bem sucedidos com padrão de alterações de NC similar.

6.4 Regiões Genômicas

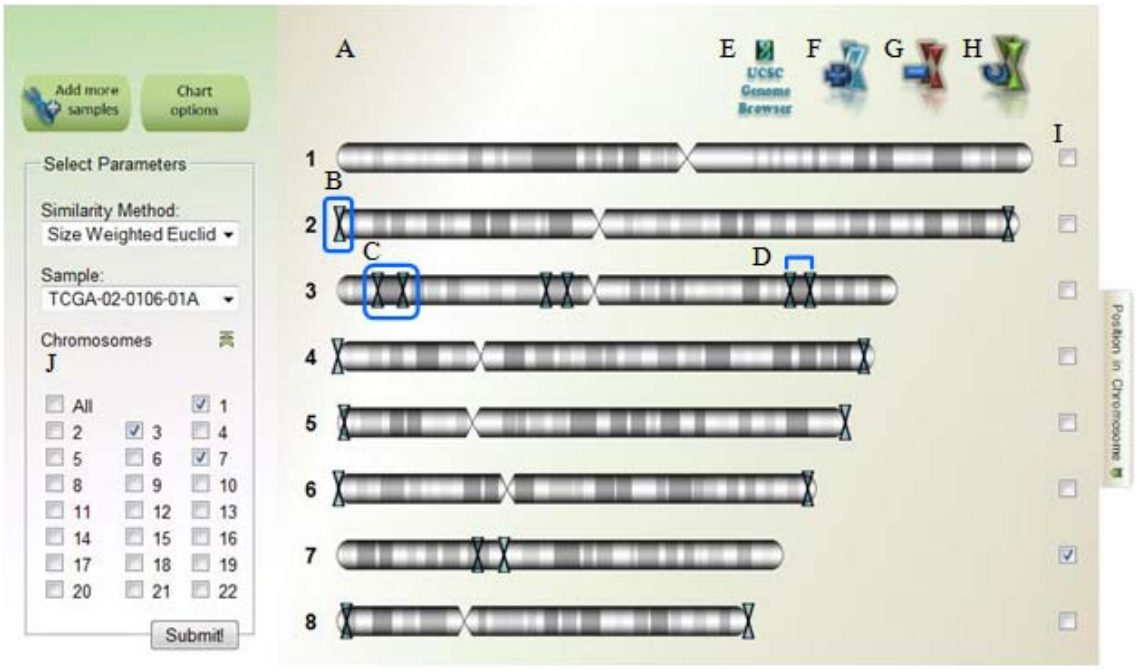

Figura 6.6: Seleção de regiões-alvo para análise. (A) Menu de seleção de regiões autossômicas, os 22 autossomos são representados com o padrão de bandas G. (B) Marcadores móveis, são usados aos pares (C) para delimitar regiões genômicas (D). Delimitam-se estas regiões para serem analisadas no CNViewer e também exploradas no *UCSC Genome Browser* (E). (F) Podem ser adicionados mais pares de marcadores; (E) removidos ou (F) resetados. Ao resetar ou adicionar, o novo par de marcadores é adicionado nas extremidades do autossomo. (I) Área de seleção de cromossomos para os quais as ações dos botões (E-H) serão executadas. (J) A análise será executada apenas nas regiões dos cromossomos selecionados neste menu.

A importância da escolha de regiões genômicas é ressaltada por diversas vezes nesta dissertação. O mapeamento das posições genômicas está de acordo com o *Human Genome Assembly Information,* montagem 36 e é medido em pares de bases. O CNViewer permite um número ilimitado de combinações de regiões do genoma para análise, pois podem ser escolhidos quaisquer autossomos, nos quais diferentes trechos podem ser selecionados (Figura 6.6).

A flexibilidade de regiões a serem analisadas é essencial para os usuários com um alvo específico no genoma e pode ser também aplicada para correlacionar dois ou mais trechos de interesse. Durante a exploração de resultados, quer seja uma análise do genoma global ou não, o usuário pode explorar qualquer região que lhe chame atenção usando o menu de seleção para especificá-la (Figura 6.6) e acessar o *UCSC Genome Browser* (Figura 6.6 E) em paralelo à utilização do CNViewer.

6.4.1 Explorando *UCSC Genome Browser*

O *UCSC Genome Browser* permite explorar de forma prática dados de vários genomas. Apresenta anotações de genes conhecidos e preditos, SNPs, além de comparar dados de outras espécies com o trecho genômico escolhido (Figura 6.7).

A análise apresentada na Figura 6.7 usou os recursos e seleção demonstrados na Figura 6.6. As coordenadas genômicas marcadas no cromossomo 7 foram exportadas para o *UCSC Genome Browser* – a área do genoma escolhida pode ser visualizada em ambas as imagens. A vinculação do CNViewer a esta aplicação enriquece a análise, provendo dados informativos com precisão.

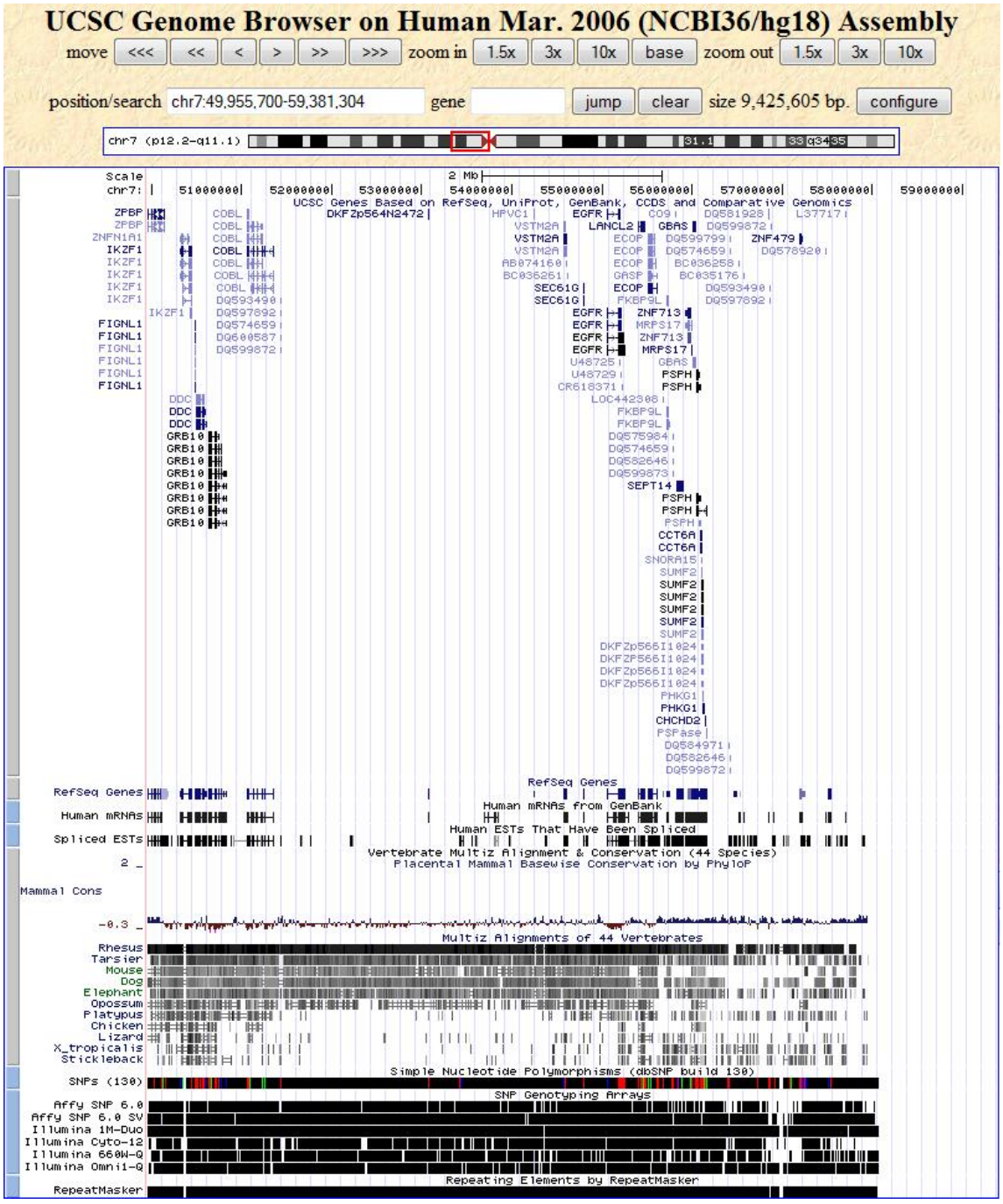

Figura 6.7: Consulta ao *UCSC Genome Browser*. Imagem da análise no *UCSC Genome Browser* submetida através do CNViewer. A região apresentada corresponde à seleção mostrada na Figura 6.6. Na região superior, o ideograma indica a região cujas anotações estão sendo exibidas. Estão indicadas na imagem os diversos genes contidos neste intervalo genômico, o alinhamento com alguns animais (em preto as regiões similares) e regiões de SNP.

6.5 Métricas de Dissimilaridade

A comparação entre amostras é uma das características mais importantes do CNViewer. A (DE) foi disponibilizada com diferentes métodos de ponderação, fornecendo três formas de explorar as relações de uma amostra e seu conjunto. CNVs são potenciais classificadores genômicos para uso em diagnósticos e prognósticos. Atualmente são utilizados com sucesso classificadores com base em mRNA, a perspectiva de classificadores baseado em CNV é vantajosa pois a molécula de DNA é mais estável que a de RNA, além de não estar sujeita a variações ambientais e circadianas (van de Wiel et al. 2010).

A comparação dos perfis de CNV oferecida pelo CNViewer somada a flexibilidade de escolha de regiões genômicas, oferece um ambiente excelente para aplicação de classificadores. É possível armazenar uma coleção de perfis moleculares, para os quais o quadro clínico já está estabelecido, e comparar com a CNV de um paciente, permitindo estabelecer um prognóstico/diagnóstico com mais segurança e eficiência.

Poucas estratégias para análise de predição utilizando CNV foram desenvolvidas, basicamente aplicando classificadores já conhecidos, mas ainda não está estabelecida uma metodologia satisfatória de comparação entre perfis de CNV (van de Wiel et al. 2010). Este fator foi considerado durante a implementação do aplicativo, assim futuramente, quando forem criadas melhores técnicas comparativas para este caso, novas métricas de dissimilaridade poderão ser implementadas de forma prática.

Conforme descrito na seção 5.3, o CNViewer realiza as comparações com base nos pontos de mudança de NC. Duas métricas ponderam a dissimilaridade, uma de acordo com o tamanho dos segmentos entre estes pontos e outra, conforme o número de sondas contidas neles. A terceira métrica não é ponderada e apenas soma os valores das

medidas nos pontos – DE em pontos de quebra. Com objetivo de ilustrar os diferentes resultados obtidos com as três métricas, serão apresentados a seguir imagens dos gráficos gerados pelas análises que utilizam a amostra TCGA-02-0001-01C como alvo (escolhida simplesmente por ser a primeira da lista de amostras) e nas quais são comparados todos os segmentos dos cromossomos 7, 9 e 10. Utilizou-s estes três cromossomos por serem os que apresentam as mudanças mais características do glioblastoma multiforme.

AS CNVs da amostra-alvo estão ilustradas na primeira linha dos gráficos das Figuras Figura 6.8, Figura 6.9 e Figura 6.10. Seu perfil é caracterizado por amplificações ao longo de todo cromossomo 7 (entre 3 e 4 repetições, denotadas pelas cores amarelo e laranja) e deleção numa região do braço *p* do cromossomo 9 (em azul claro, indicando a existência de uma cópias apenas). O cromossomo 10 aparentemente possui NC estável, porém a coloração do segmento é verde levemente azulado, denotando que provavelmente, algumas células do tumor sofreram deleção de parte do cromossomo e portanto seu NC está entre 1, 5 e 2.

A DE de pontos de quebra, por consequência da ausência de ponderação, tende a calcular distâncias maiores para segmentos com mais pontos de quebra, pois são feitos um maior número de medições. Assim serão classificadas como similares primeiramente as amostras com pontos de quebra coincidentes com os do alvo, ou com poucos pontos de quebra. O resultado pode ser observado na Figura 6.8, na qual a amostra mais semelhante possui um NC menor para os cromossomo 7 e 10, além de não possuir nenhum ponto de quebra. Observando o gráfico como todo, pode ser notado que todas as amostras na região superior – portanto mais similares ao alvo – tem poucos pontos de quebra, enquanto na região inferior estão presentes amostras com mais variação de NC ou com regiões com NC muito altos.

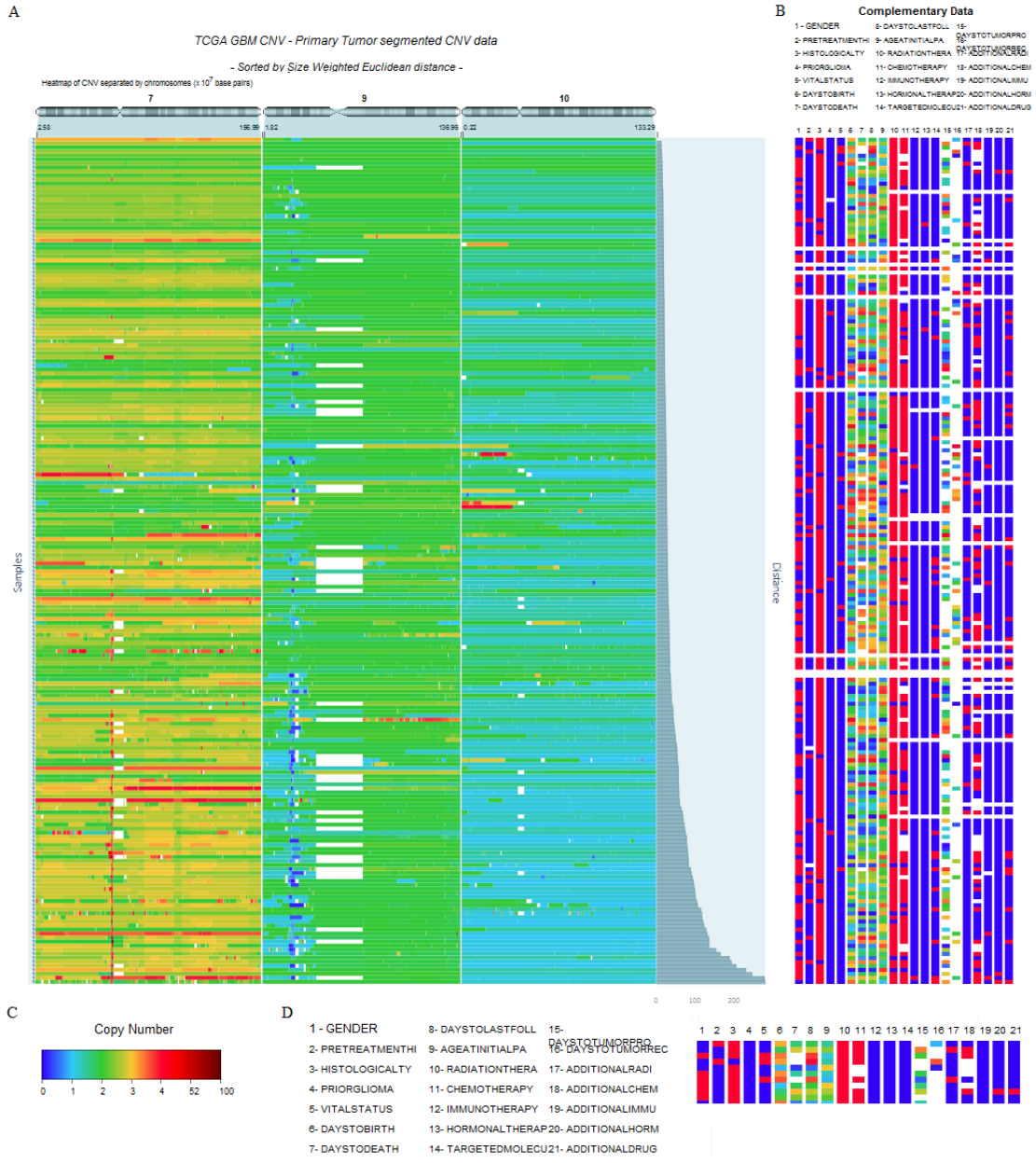

Figura 6.8: Distância Euclidiana em pontos de quebra. Resultado da análise dos dados modelos, tendo como alvo os cromossomos 7, 9 10 e amostra TCGA-02-0001-01C, representada na primeira linha dos gráficos.(A) Gráfico com segmentos ordenados pela DE de pontos de quebra, com valores de NC representados e cores correspondentes com a legenda (C). (B) Dados complementares e (D) detalhe da legendas dos dados complementares e dos dados da amostra alvo e dos nove perfis indicados como mais próximos

Considerando-se estas características, a DE em pontos de quebra é uma métrica para ser aplicada em casos cujo interesse esteja na quantidade ou nas regiões de pontos de mudança de NC. Em oposição, nas distâncias ponderadas (Figura 6.9 e Figura 6.10) o número de segmentos contido em cada amostra não influi no resultado, é a diferença entre cada região do genoma que determina a similaridade. O NC de cada segmento é a média dos valores das sondas de sua região, assim, nesta aplicação, ponderar a distância DE seria refletir a medida obtida no ponto de quebra, em toda região por ele representado.

A distância ponderada por tamanho torna a dissimilaridade calculada em um ponto proporcional ao tamanho do segmento. Essa estratégia tem como consequência suavizar os efeitos de grandes alterações de cópias em pequenos pontos e valorizar a semelhança de NC, independente do número de segmentos contidos nas amostras.

O resultado na Figura 6.9 A ilustra bem estas características, pois as duas primeiras amostras mais próximas tem amplificação do cromossomo 7, com NC por volta de três, sendo que a primeira amostra mais semelhantes tem pontos de quebra coincidentes neste mesmo cromossomo e NC parecido também para o cromossomo 10. Apesar desta primeira amostra não ter deleção no cromossomo 9, a segunda amostra mais parecida possui deleção em região semelhante. Assim a categorização da similaridade por DE ponderada por tamanho possui correspondência visual. Outra característica perceptível é que os segmentos estão distribuídos com maior independência do seu número de pontos de quebra. Ainda assim, a DE ponderada por sondas aparentemente consegue fazer uma melhor comparação de valores, devido a aspecto particulares da análise de CNV.

As sondas não cobrem todo o genoma, mas estão posicionadas de forma a representá-lo de maneira eficiente, porém certas regiões não são acessadas durante a

obtenção de dados, além daquelas que eventualmente são removidas propositalmente da análise. A CBS usada para filtrar dados de CNV ignora estas regiões, assim se os segmentos vizinhos às regiões sem sondas possuírem mesmo NC, o resultado do CBS irá considerar um único segmento a região de NC desconhecido e suas vizinhas. Caso haja diferença nos valores, esta região não é representada por valor algum, ilustrada no CNViewer como trechos em branco. Este aspecto não é tratado pela DE ponderada por tamanho, nem pela DE em pontos de quebra. Portanto, as medidas obtidas em locais sem sondas alteram o resultado, pois considera-se que seu NC é igual a zero.

A DE ponderada por sondas soluciona o impasse de maneira simples, tornando as medidas nos pontos de mudança de NC proporcionais às sondas contidas no segmento. Os resultados das métricas ponderadas diferem levemente, no exemplo das Figura 6.9 e Figura 6.10, as duas amostras mais semelhantes são as mesmas. A terceira amostra mais próxima a alvo sugerida pela DE ponderada por sondas, é visualmente mais parecida, que a sugerida pela DE ponderada por tamanho. Mas a maior diferença é vista na distribuição das amostras com grandes trechos em branco nas proximidades dos centrômero do cromossomo 9, com a nova ponderação (por sondas) estes perfis deixam de estar muito concentrados na região inferior do gráfico. Portanto a ausência de dados para estas regiões não influenciou na classificação das amostras, anteriormente, como a amostra alvo não tinha intervalo nesta região, a ausência de dados aumentava o valor da DE calculada. Portanto a ponderada por sondas possui as mesmas características descritas para a ponderada por tamanho, apenas os resultados passam a ser obtidos com maior precisão.

A consequência do uso das sondas para ponderação esta no desempenho do programa. Um mapa contendo todas as posições das sondas deve ser submetido e o

cálculo demora aproximadamente o dobro das demais métricas, como discorrido na seção 6.9.

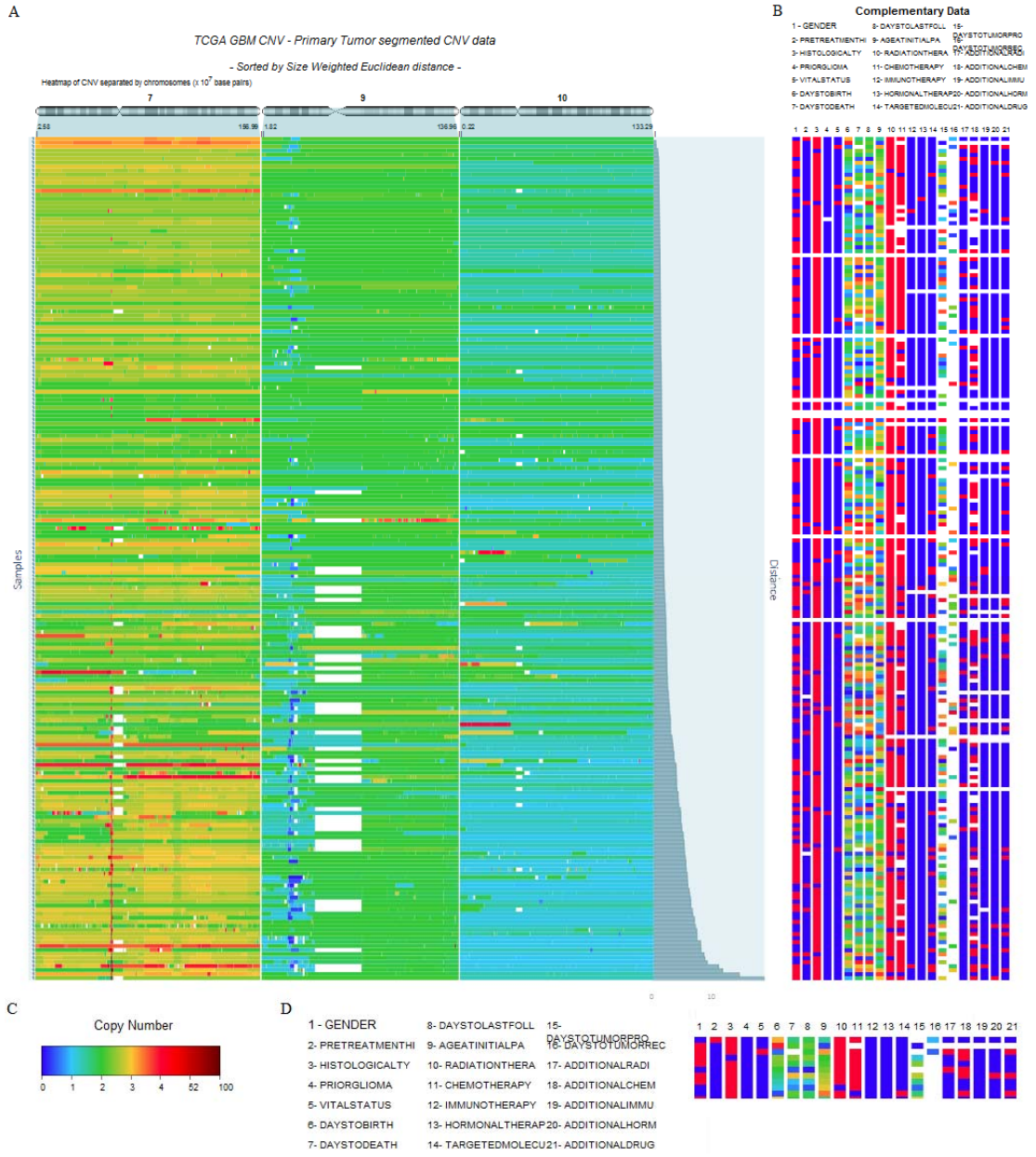

Figura 6.9: Distância Euclidiana ponderada por tamanho. Resultado da análise dos dados modelos, tendo como alvo os cromossomos 7, 9 10 e amostra TCGA-02-0001-01C, representada na primeira linha dos gráficos.(A) Gráfico com segmentos ordenados pela DE ponderada por tamanho, com valores de NC representados e cores correspondentes com a legenda (C). (B) Dados complementares e (D) detalhe da legendas dos dados complementares e dos dados da amostra alvo e dos nove perfis indicados como mais próximos

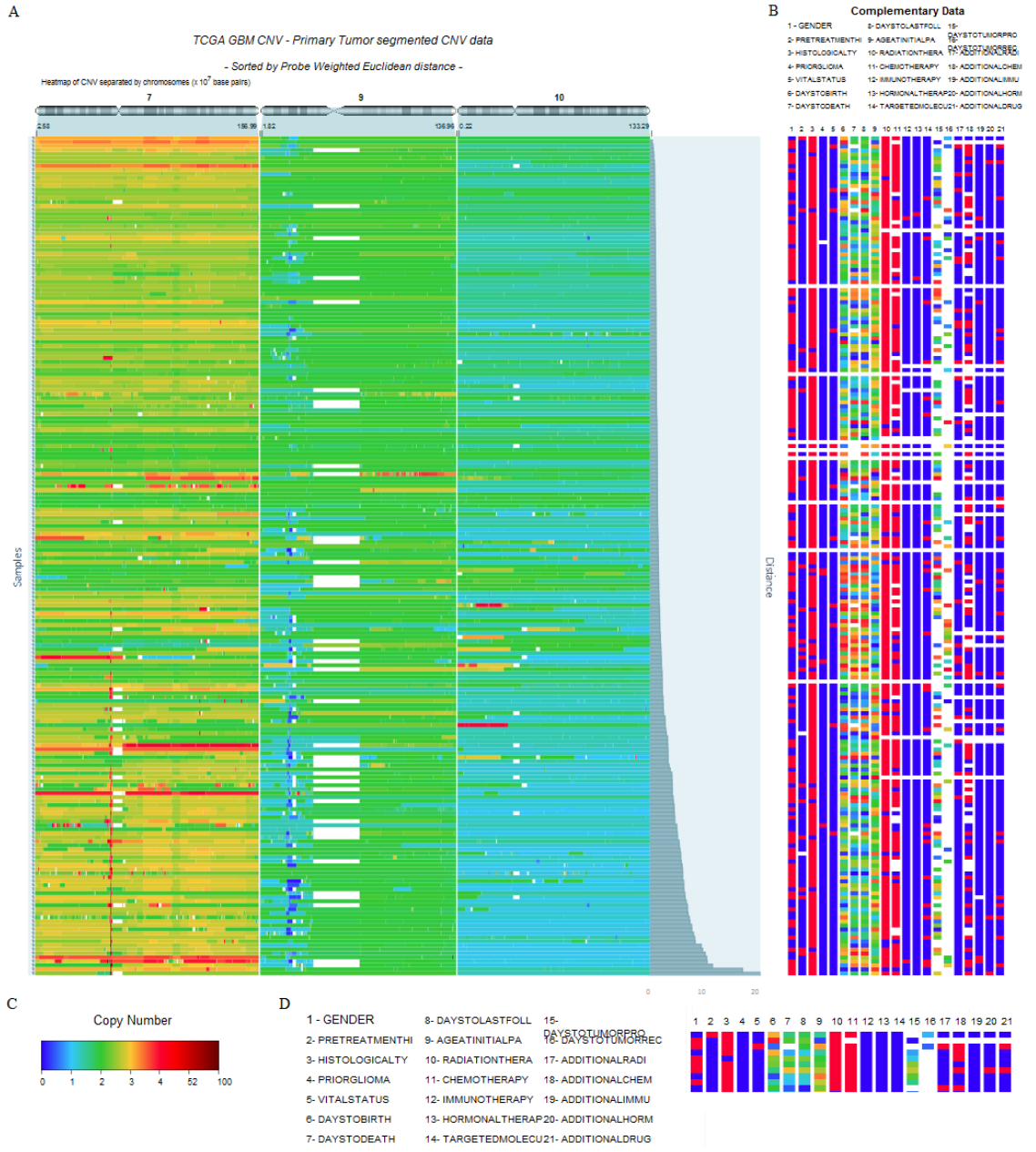

Figura 6.10: Distância Euclidiana ponderada por sondas. Resultado da análise dos dados modelos, tendo como alvo os cromossomos 7, 9 10 e amostra TCGA-02-0001-01C, representada na primeira linha dos gráficos.(A) Gráfico com segmentos ordenados pela DE ponderada por sondas, com valores de NC representados e cores correspondentes com a legenda (C). (B) Dados complementares e (D) detalhe da legendas dos dados complementares e dos dados da amostra alvo e dos nove perfis indicados como mais próximos.

6.6 Explorando Informações Complementares

As informações complementares são quaisquer tipos de dados referentes às amostras (Figura 6.11 A). Esta opção foi adicionada ao CNViewer com intuito de ser utilizada para correlação ou caracterização das CNVs.

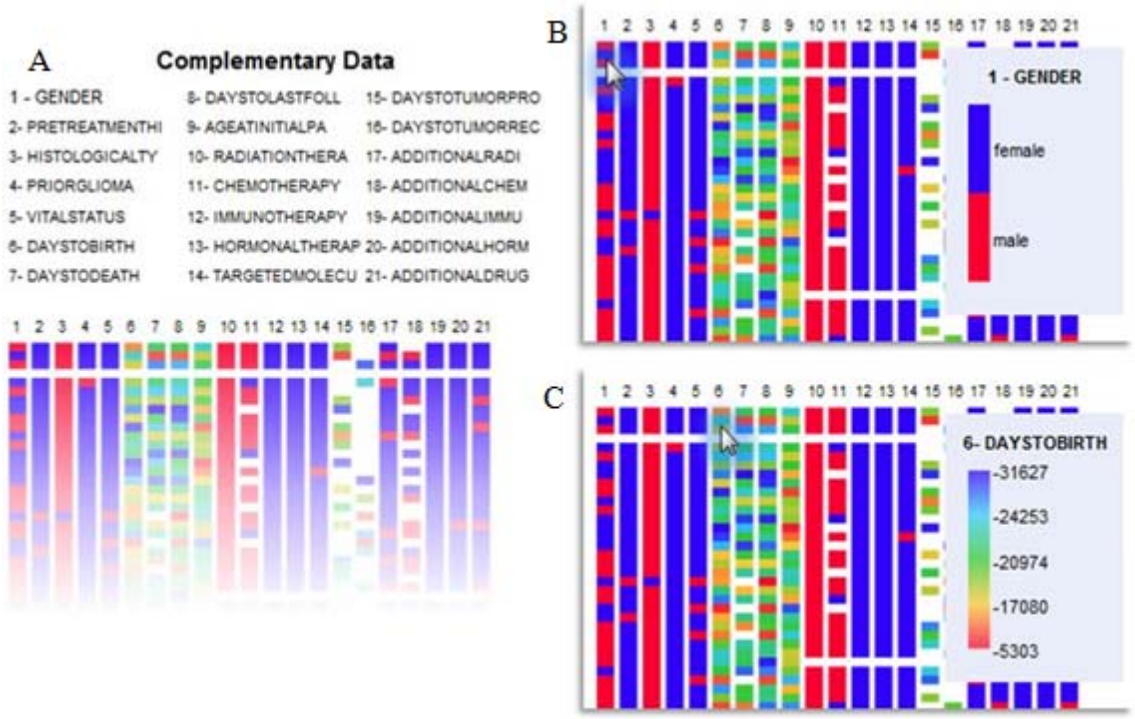

Figura 6.11: Identificação dos dados Complementares. Imagens parciais do resultado de análise realizada com os dados-modelos; gráfico plotado ao lado do gráfico com os dados de CNV. (A) Lista de categorias presentes e seus respectivos dados representados em cores. As linhas correspondem às amostras de mesma posição no gráfico de CNVs. Categorias: 1- sexo; 2- histórico de pré-tratamento; 3- tipo histológico; 4- glioma prévio; 5- estado vital; 6- dias para o nascimento; 7- dias para a morte; 8- dias para a última consulta; 9 – idade ao primeiro diagnóstico patológico; 10 – terapia de radiação; 11-quimioterapia; 12- imunoterapia; 13- hormonioterapia; 14- terapia molecular marcada; 15- dias para progressão tumoral; 16- dias para recorrência do tumor; 17- radioterapia adicional; 18- quimioterapia adicional; 19- imunoterapia adicional; 20- hormonioterapia adicional; 21- terapia adicional com drogas. (B, C) Legenda de cores, correspondente à coluna na qual o cursor está sobre.

Comumente os dados disponibilizados sobre as amostras são demográficos e fenotípicos. Estes e outros dados informativos podem ser visualizados junto com a

CNV. Outra possibilidade é a indicação de categorias, como diferentes tipos de tumor e níveis de progressão, as quais seriam facilmente traduzidas em cores no CNViewer, identificando com clareza cada amostra.

Esta funcionalidade permite a inclusão de dados categóricos e numéricos. Assim dados como sexo (Figura 6.11, categoria 1) são representados com cores específicas (Figura 6.11 B) e informações quantitativas, como idade (ou em nosso exemplo, "dias para o nascimento" - Figura 6.11, categoria 6), podem ser visualizadas em forma de um gradiente de cores (Figura 6.11 C). A interpretação desses dados deve ser feita em conjunto com as métricas de dissimilaridade e é abordada na seção 6.5.

Considerando-se o número ilimitado de informações que podem ser adicionadas, optou-se por apenas identificar o tipo de dados em uma legenda simples (Figura 6.11 A) e a legenda de cores é mostrada ao cursor ser sobreposto sobre os dados (Figura 6.11 B, C).

Conforme desenvolvido na seção anterior, existe uma busca por classificadores moleculares com base em CNV e o CNViewer oferece um ambiente apropriado para futura aplicação dos mesmo. Com a apresentação de dados extras, o CNViewer complementa esta funcionalidade e cria também um ambiente de suporte para a identificação destes classificadores, pois perfis podem ser comparados em conjunto aos dados clínicos, indicando e testando regiões candidatas.

6.6.1 Métricas de Dissimilaridade e Dados Complementares

A interpretação dos dados complementares é feita com base nas amostras mais similares, pois se espera que a semelhança entre NC se reflita nos dados complementares. Portanto deve-se selecionar para análise as amostras com as menores distâncias calculadas e preferencialmente visualmente distinguíveis da maioria.

Deve ser lembrado que, conforme as restrições explicadas na seção 5.2, os dados complementares apresentados neste trabalho não são os mais adequados por serem pouco informativos. A análise realizada é empírica, utilizada para rápida visualização de um conjunto amostral complexo, com o objetivo de sugerir potenciais focos para análise aprofundada. É importante observar que os dados do tipo 2-5, 10-14, 17, 19-21 nas Figuras Figura 6.8, Figura 6.9 e Figura 6.10 possuem um resultado predominante, uma característica presente em quase todos os pacientes. Portanto a coincidência entre estes dados ao examinar amostras semelhantes, não deve ser tratada como um resultado em potencial, mas a correspondência entre dados menos comuns apontaria para uma possível característica peculiar àquele padrão de NC. Outro fator importante para a análise é a existência de dados correlacionados, gerando resultados semelhantes (5, 7 e 8; 6 e 9) ou excludentes (15 e 16).

O gráfico da Figura 6.8 apresenta pouca diferença entre as distâncias das amostras mais similares. Comparando os dados complementares do alvo com as das 5 amostras mais parecidas (a amostras mais próxima será denotada apor A1 e as demais serão denominadas sucessivamente), poucas semelhanças são observadas. Desconsiderando os campos com predominância de um tipo de característica, observa-se que três amostras tiveram correspondência entre sexo (tipo 1) e que quatro pacientes também receberam radioterapia adicional (tipo 17). Os demais dados clínicos de A1 e A4 são bem distintos, porém A5 possui grandes semelhanças quanto as campos 6, 7, 8 e 9, que estão relacionados ao tempo de vida, idade de desenvolvimento e diagnóstico do tumor. A2 e A3, tem correspondência apenas com o tipo de dado 6 e assim, de forma geral os resultados desta análise foram pouco informativos.

A DEs ponderadas agrupam com mais eficiência amostras com perfis de dados complementares semelhantes. As cinco primeiras amostras denotadas como mais

semelhantes foram as mesmas, assim os resultados das duas métricas serão avaliados em conjunto. Os dados do tipo 7 e 8 destacam-se por sua correspondência entre quatro e cinco amostras, respectivamente. Por outro lado, não foi observado coincidências entre os dados tipo 1, 6 e 9 em nenhuma das cinco amostras mais similares. Além disso, apenas A2 é semelhante quanto aos dados 16 e 18, caracterizando pacientes com recorrência de tumor e uso adicional de quimioterapia.

Portanto, a comparação da amostra TCGA-02-0001-01C com este conjunto com a DE ponderada por tamanho, obteve sucesso em agrupar amostras com sobrevivência parecida. A previsão de tempo de vida é uma das informações mais importantes de um prognóstico, porém esta é uma análise preliminar, não se pode afirmar que este perfil molecular está relacionado necessariamente um período de vida determinado (por volta de 422 dias).

A similaridade de perfis moleculares é aparentemente calculada com eficiência pelas distâncias ponderadas, pois a representação gráfica dos resultados ilustra claramente sua capacidade de agrupamento de amostras com CNVs parecidas. Porém a relação entre a biologia molecular e fenótipos na grande maioria das vezes não é linear. Uma série de fatores (aspectos genéticos, fisiológicos, ambientais, entre outros) atuam no resultado final, que no caso do glioblastoma multiforme é o tempo de vida.

Além destes aspectos, outras questões devem ser exploradas antes mesmo de se desenvolver um método para predição com base em CNV. Deve ser estabelecido qual é o alvo ideal da análise: variações pontuais ou o genoma global. Outro ponto discutido é como ponderar as relações de perdas e ganhos genômicos, pois deleções variam entre 0 e 2 e amplificações tem teto indeterminado. Provavelmente cada um destes aspecto irá variar conforme a característica que se deseja prever.

6.7 Exportação e Recuperação da Análise

A exportação é uma propriedade do CNViewer muito importante pois permite guardar os dados e a análise. Assim, além de não ser necessário ao usuário recarregar dados na interface, ele pode salvar seus resultados sob a forma gráfica. Portanto é possível o compartilhamento de análises e dados, garantido reprodutibilidade dos resultados.

A disponibilização de tal ferramenta abrange também a divulgação dos dados, já que resulta em um arquivo HTM que pode ser vinculado em uma página de internet como hyperlink. Deve ser observado que esse arquivo contém todos os dados carregados no CNViewer e sua disponibilização sob qualquer forma, torna-os acessíveis em forma de texto também.

O arquivo gerado pelo módulo de exportação tem tamanho correspondente ao dos arquivos submetido ao CNViewer. Estes dados são recarregados no aplicativo através de submissão por URI, e a transmissão desses dados é proporcional ao volume de informação enviada. Assim, o tempo de recuperação de uma análise depende da capacidade do serviço de internet utilizado e da quantidade de dados contidos por ela, consequentemente a memória virtual dedicada a este serviço pelo navegador de internet é também um limitante.

Considerando o volume gerado de dados biológicos, a recuperação de dados pode nem sempre ser tão eficiente quanto o desejado, como os serviços de internet e produtos relacionados estão sendo aprimorados rapidamente, brevemente a capacidade de processamento de dados será expandido. Enquanto isso, CNViewer demonstra que já é possível armazenar os dados e análise, recuperando com sucesso à informação de 210 amostras de CNV, com tamanho aproximado de 3,5 MB.

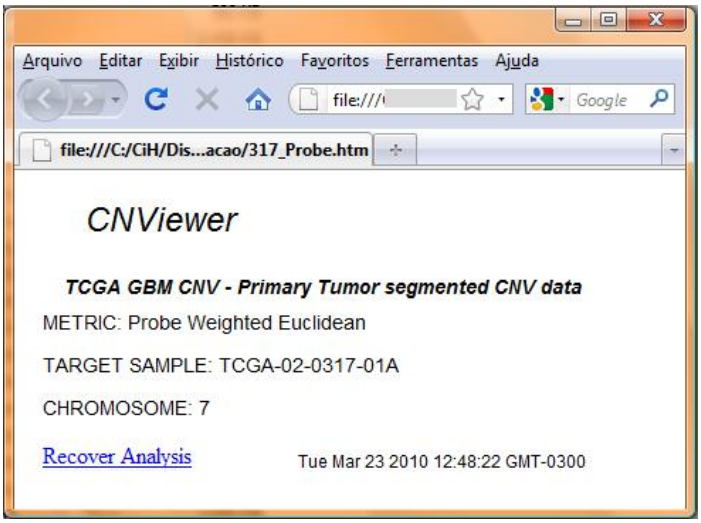

Figura 6.12: Módulo de Exportação

A criação do módulo de exportação se destaca por conter basicamente os dados carregados. As ferramentas do CNViewer não estão presentes no módulo (Figura 6.12), tornando-o o mais leve. O arquivo gerado é uma HTM simples, que não utiliza o DOM, e contem informações sobre os dados e a análise. Quem acessá-lo saberá a data e hora de geração, título e parâmetros da análise. Ao clicar no link "*Recover Analysis*", o usuário estará submetendo os dados ao CNViewer que irá interpretá-los e re-apresentar a análise.

Existem outras formas e salvar e recuperar dados, a opção de "Salvar como" intrínseca dos navegadores ou o armazenamento em arquivos do tipo *cookie*. A geração de um módulo cria um arquivo mais leve do que salvar diretamente a página Web, pois as ferramentas não são salvas junto e também permite o compartilhamento e transporte de dados, enquanto um cookie fica vinculado à maquina que o criou.

6.8 Segurança

A segurança de dados é uma das maiores preocupações de usuários de softwares, devido à privacidade de pacientes ou confidencialidade da pesquisa, no caso específico

de dados biológicos. Pensando no caso específico do CNViewer e também no conceito de utilização do navegador Web como um ambiente de análise de dados, os limites e soluções de segurança também foram explorados.

Um dos passos mais sensíveis é o carregamento de dados, para tanto foram oferecidas três alternativas, duas das quais são seguras. A segurança dos dados em serviços do tipo S3DB já está implementada no próprio banco de dados, com o controle por senha. Essa segurança pode ser melhorada futuramente com a implementação de TLS (*Transport Layer Security*) um protocolo de comunicação criptográfico, garantindo a segurança na transferência de informações.

O carregamento via copiar e colar não envolve transferência de dados, portanto a segurança depende exclusivamente do ambiente da máquina do usuário. O carregamento direto de arquivos precisa de suporte do servidor, com a implementação futura de TLS passaria a ser seguro também. Apesar desta opção ainda não oferecer garantia de segurança, o copiar e colar supre essa função caso o usuário queira garantir a proteção de seus dados.

Conforme descrito anteriormente, os processos realizados pelo CNViewer e a geração do módulo de exportação ocorrem inteiramente no navegador, sem comunicação com um serviço externo, assim a segurança se restringe ao ambiente onde o aplicativo está sendo executado.

A única etapa para a qual a segurança não foi oferecida, é a recuperação da análise. Como é necessário submeter os dados via URI eles são enviados a um serviço externo, apesar de só serem lidos no navegador do usuário. Este envio não oferece proteção. Uma estratégia para oferecer segurança, seria gerar um arquivo texto cujo conteúdo deveria ser copiado em uma caixa de texto, e o CNViewer interpretaria as informações da mesma forma que é feito o carregamento via copiar e colar.

6.9 Desempenho

6.9.1 Execução em Navegadores

Os navegadores possuem características heterogêneas assim, o comportamento do aplicativo em cada um deles é distinto, conforme demonstrado na Tabela 6.1. A compatibilidade entre Navegadores Web é um problema de longa data e, apesar de padrões terem sido estabelecidos, o problema persiste. Primeiramente devido ao fato de nem todos os navegadores seguirem estes padrões, como o caso do Internet Explorer, mas também porque a rapidez do surgimento de novos recursos não demanda apenas a incorporação dos mesmo em todos os navegadores, mas exige também que os usuários atualizem seus navegadores – procedimento comumente adiado (Goodman e Morrison, 2007).

A análise de desempenho do CNViewer permitiu observar também outro aspecto que torna os navegadores ainda mais distintos: a capacidade para armazenamento e processamento de dados. Conforme discutido posteriormente nesta seção, para muitos navegadores há um limite para os dados inseridos em áreas de texto, e ultrapassar este limite pode levar a fim da execução do programa em alguns casos.

Era esperado também que houvesse diferenças entre os navegadores quanto ao desempenho durante a execução de tarefas, mas foi possível observar que eles oferecem a estrutura necessária para a computação satisfatória de um grande volume de dados. Certamente, o campo de desenvolvimento de navegadores é vasto, existindo um vasto campo a ser expandindo, permitindo a criação de navegadores ainda mais eficientes e com características novas, expandindo a interação com o usuário.

Tabela 6.1: Tempo médio de execução. Valor médio de cinco tomadas de tempo, medido em segundos, com base nos dados-modelo (210 amostras).

Navegador	Chrome	Firefox	Opera	Safari
Carregar CNViewer	47.4	26.3	52.8	28.4
Carregar Dados no S3DB	48.1	35.7	54.2	37.6
DE ponderada por sonda	3.5	12.9	17.8	8.2
DE ponderada por tamanho	2.0	6.6	6.2	3.9
DE em pontos de quebra	1.8	6.8	28.0	3.9
Carregar módulo de exportação	2.3	2.1	10.0	2.0
Recuperar análise	37.4	33.2	49.3	42.9

CARREGAR CNVIEWER

O programa carrega a grande maioria dos *scripts* ao ser iniciado, visando poupar tempo ao longo de sua execução. A primeira vez que é acessado, demora em torno de 30 s para ser carregado, independentemente do navegador. Em acessos futuros, os *scripts* estarão em memória cachê, diminuindo este tempo para aproximadamente 10 s.

CARREGAR DADOS DO S3DB

A obtenção de dados do S3DB não é imediata, pois deve ser aguardado o *download* dos arquivos. Esta tarefa é executada com mais eficiência pelo Google Chrome e Apple Safari, enquanto o maior tempo é despendido pelo Opera.

DISTÂNCIAS EUCLIDIANAS

Um aspecto interessante é o processamento gráfico. Ao ser medido o tempo de cálculo das distâncias Euclidianas, estão envolvidos o processamento numérico e gráfico. Até recentemente seria inviável processar no ambiente do navegador tais gráficos, esta tarefa teria que ser feita em um servidor, gerando um novo código HTML

para ser apresentado ao usuário – estratégia ainda usada por inúmeras páginas de internet.

O processamento é um dos aspectos mais heterogêneo dos navegadores. Google Chrome executa o calculo das distâncias e plotagem gráfica com eficiência superior aos demais, sendo aproximadamente duas vezes mais rápido que o Apple Safari e entre três a quatro vezes mais rápido que o Mozilla Firefox e Ópera.

O navegador escolhido para modelar o CNViewer, Mozilla Firefox, ainda não oferece um processamento ideal, mas seu desempenho é satisfatório, plotando 210 amostras em até 16 s. Observando o tempo de execução no Google Chrome, fica claro este fator pode ser aprimorado e que os navegadores têm potencial para realizar tais processos com maior eficiência.

É interessante notar que a DE ponderada por sonda consome em média, o dobro do tempo que as demais medidas, e seu cálculo difere somente em um aspecto principal – a busca de do valor da posição do ponto de quebra num vetor contendo a localização das sondas. Portanto esse procedimento é provavelmente o que esta levando a um maior tempo de execução.

MÓDULO DE EXPORTAÇÃO

Ao módulo de exportação ser carregado, todos os dados são exportados para uma nova página, fator que pode causar certo atraso na reposta. No caso de estudo, com 210 amostras, apenas o Ópera demorou a executar a tarefa, sendo quatro a cinco vezes mais lento que os demais navegadores.

RECUPERAÇÃO DA ANÁLISE

Recuperar a análise infelizmente é uma tarefa limitada devido a restrições no processamento de muitos dados na URL. A recuperação de dados é lenta, entorno de 40

s, mas a facilidade proveniente de seu uso deve superar este fator. Por exemplo, o tempo de recuperação é menor que a soma do tempo de carregamento da página e de dados do S3DB, e ainda traz por completo a análise salva.

As características dos navegadores Google Chrome e Apple Safari não permitem utilizar a opção "Salvar como", para armazenar o módulo de exportação gerados nestes aplicativos deve-se copiar o endereço do link "*Recover Analysis*" e colar em um arquivo vazio, salvando o como HTM. A cópia do endereço é feita clicando-se no link com o botão direito do mouse, e selecionando a opção correspondente no menu.

COPIAR E COLAR

Outro aspecto limitante dos navegadores está no método de copiar e colar. Apenas o Mozilla Firefox suporta grande quantidade de caracteres, Opera permite aproximadamente quatro milhões de caracteres, o que seria entorno de 4 MB de dados tabulados e Google Chrome e Safari, suportam 525 mil caracteres (525 KB).

6.9.2 CNViewer versus Cancer Genome Browser

A comparação do desempenho do CNViewer e aplicativos com outras formas de implementação (dependentes de servidores de rede ou executados localmente) deveria ser feita, idealmente, com um programa que oferecesse as mesmas funcionalidades. Como não há um aplicativo idêntico ao CNViewer, a comparação foi feita com o CGB, pois tem acesso aos mesmos dados que o CNViewer, além de oferecer representação de CNV por mapa de calor. Apesar da semelhança entre os gráficos,a implementação dele sé distinta e o CNViewer realiza adicionalmente a comparação entre amostras.

Os resultados obtidos deixam claro que o desempenho do CNViewer é o mais eficiente (Figura 6.13), pois na maioria dos casos a velocidade de exibição do gráfico é maior e porque o tempo possui pouca oscilação. O processamento do CNViewer é

influenciado pelo número de segmentos das amostras, principalmente da amostra alvo, mas as diferenças entre os cromossomos são refletidas minimamente no tempo total de execução, que se manteve por volta dos 4 segundos. O CGB tem o tempo de plotagem variando proporcionalmente ao tamanho dos cromossomos, exceto pelos picos observados nos cromossomos 7 e 12, sugerindo que quanto maior o número de segmentos presentes nas amostras, maior tempo de processamento necessário à plotagem.

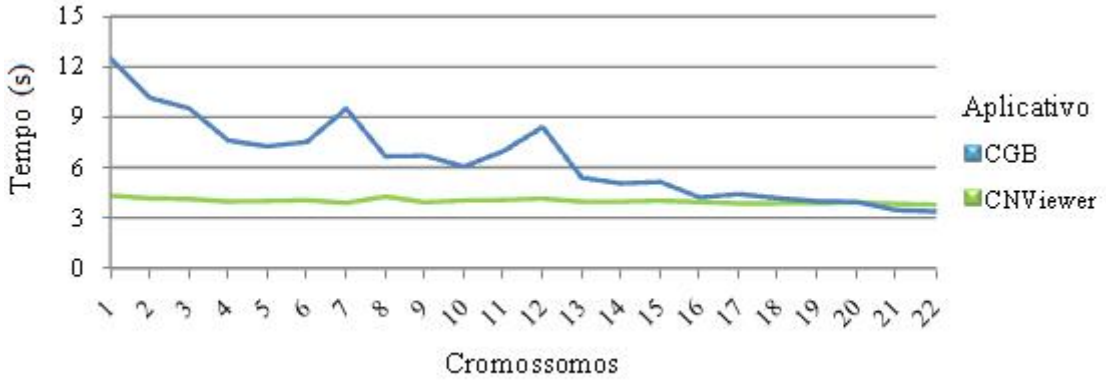

Figura 6.13: Tempo médio de plotagem do Cancer Genome Browser e CNViewer. Valor médio de cinco tomadas de tempo referentes à plotagem de mapas de calor para cada cromossomo, medido em segundos.

Esta comparação demonstra que o uso do navegador como ambiente de desenvolvimento de softwares para análise de dados pode superar aplicativos locais. Os cromossomos de maior tamanho (1-13) ao serem plotados individualmente no CGB (Tabela 6.1) consomem mais tempo do que o processamento da dissimilaridade e plotagem de todos os autossomos no CNViewer (Tabela 6.1), tornando evidente esse fato. Certamente, programas escritos em outras linguagens ou utilizando outras formas de implementação podem ser mais eficientes que o CNViewer, mas com este teste foi

possível comprovar a aptidão dos navegadores para processamento de um grande volume de dados e representação gráfica.

6.10 Modelo de Implementação

Até onde conhecemos, este é o primeiro trabalho que reúne no navegador a interface com o usuário, o processamento e análise de dados, com geração de gráficos e interface dinâmica, além de permitir o armazenamento/recuperação de análise. Esta implementação possui diversas vantagens discutidas ao longo desta dissertação e pode beneficiar também o desenvolvimento de outros aplicativos.

O modelo desenvolvido está ilustrado na Figura 6.14. Como é característico de uma página Web, os *scripts* e imagens do aplicativo estão armazenados em um servidor remoto e são carregados no navegador (Figura 6.14 A) quando o usuário acessa a página. O modelo propõe que todas as funcionalidades estejam implementadas em JavaScript para serem executadas inteiramente no navegador e sem requerer carregamento de aplicativos à parte, unindo o cliente com a camada de apresentação e a camada lógica do aplicativo. Desta forma, o programa fará comunicação com recursos remotos, somente sob solicitação do usuário, para realização de alguns tipos de inserção de dados ou mesmo interação com outras páginas Web.

Este modelo tem como principal diferencial o processamento de tarefas localmente, sendo possível também o processamento local de dados (Figura 6.14 B), com colaboração direta o usuário que deve inserir o conteúdo desejado diretamente no navegador. A forma comum de carregamento de dados é a que utiliza *scripts* localizados num servidor remoto para leitura de arquivos locais. Ela também pode ser implementada (Figura 6.14 C) e neste caso, uma vez que o arquivo seja lido remotamente, os dados obtidos são remetidos ao navegador que então pode processá-los localmente durante todo o uso do programa. Finalmente, uma terceira forma de

importação de dados é oferecida, resgatando todas as informações de interesse contidas num banco de dados remoto (Figura 6.14 A) ou local, que semelhante as formas anteriores, serão mantidas no navegador durante o uso do programa.

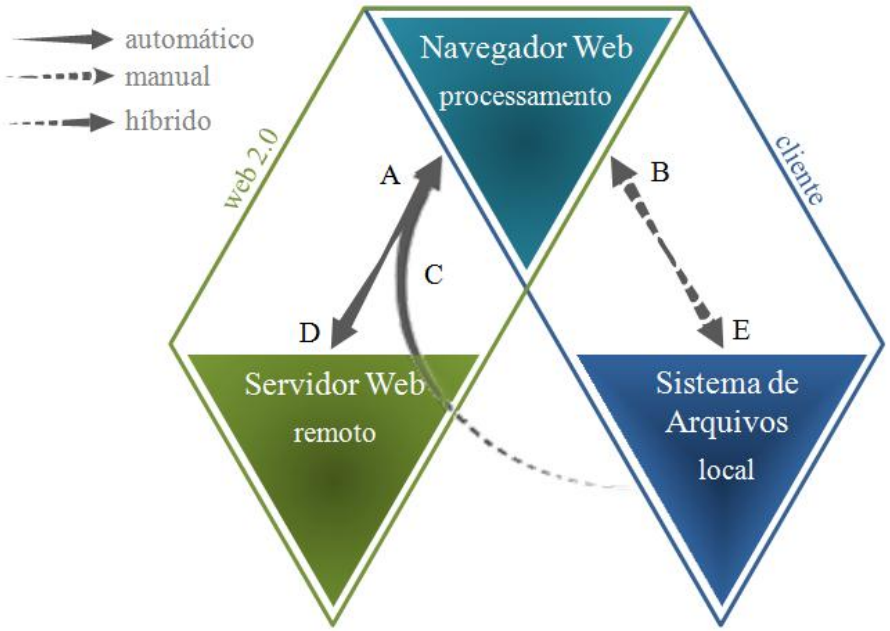

Figura 6.14: Modelo de transferência de dados para aplicativos baseados em navegador Web.

Esses dois recursos descritos, oferecem a possibilidade única de inclusão de informações, característica que pode ser suficiente para aplicativos como o CNViewer, cuja funcionalidades se assemelham a um programa *desktop* pois ele oferece ferramentas para manipulação de dados oferecidos pelo usuário, como os pacotes para análise de dados, editores de texto e visualizadores de imagem. Porém a Web é utilizada para a criação de programas que vão além disso, como os sites de compra e venda de produtos, e-mail e de relacionamento. Nestes casos, o usuário obrigatoriamente deve enviar informações suas para um servidor que armazena dados e além disso, é inviável

alojar no navegador toda a gama de informações a que o usuário pode ter acesso. Atualmente, sites como Amazon, Gmail e JBrowse utilizam Ajax para acessar os dados conforme o necessário e processá-los localmente. Em nosso modelo o mesmo pode ser feito (Figura 6.14 A e D): um banco de dados pode ser criado para uso particular, isto é, manutenção de uma coleção de informações (localização local ou remota), ou para prestação de serviços (localização remota). Com o uso de Ajax ou banco como o S3DB, o acesso e submissão de dados (Figura 6.14 A e D respectivamente) pode ser feito no *background*, com processamento no navegador. Certamente, para estes sites que manipulam um vasto número de informações e que já utilizam a Web 2.0, o uso do modelo proposto acarretaria em pouca mudança.

O ultimo aspecto abordado é a gravação manual de informações na máquina do usuário (Figura 6.14 E). Propõem-se a geração de um módulo de exportação contendo os dados e informações relevantes sobre a atividade recente executada no aplicativo – este módulo é uma página HTM, que pode ser salva pelo usuário usando as propriedades do navegador. A recuperação da atividade é realizada através de um link para a página original acrescida dos dados necessário para recuperar a atividade salva. Atualmente, a única forma de interpretar dados submetidos a uma página utilizando comandos JavaScript é enviá-los via URL. É importante notar que, apesar de os dados serem submetidos à rede pela URL, sua interpretação será feita no navegador, quando a página for carregada.

Um aplicativo que utiliza inserção de dados com interação direta do usuário (Figura 6.14 B) ou que acesse uma base de dados num servidor local, oferece a possibilidade de execução *stand-alone*, isto é, o usuário pode manter uma cópia do código do programa em seu computador e executá-lo sem necessidade de conectar-se à Internet.

Capítulo 7

Conclusões e Perspectivas Futuras

O desenvolvimento de programas baseado em Web pode significar a necessidade de inúmeras linguagens de programação. A implementação por JavaScript permite o uso eficiente dos navegadores Web, sem necessitar explorar outras linguagens e ainda criando um aplicativo *desktop* e independente de sistema operacional. Tem como vantagem também, poder depositar o código em repositórios *on-line*, tornando-o utilizável e com manutenção dispensável.

Apesar de o CNViewer não ser um aplicativo *desktop*, sua implementação demonstrou que é possível usar o ambiente do navegador Web para criar aplicativos que possam ser executados localmente e satisfazendo as funcionalidades mais importantes de um programa de desenvolvimento/análise: inserção de grande volume de dados, processamento e análise matemática destes dados, representação gráfica, geração de arquivo para armazenamento de analise e dados, bem como compartilhamento de dados/resultado. Além disso, esta forma de implementação praticamente independe de manutenção de servidores, tornando a manutenção do programa mais fácil e barata. Com base no acelerado desenvolvimento de navegadores e a própria competição de mercado, é esperado que em breve a execução de aplicativos como o CNViewer torne-se ainda mais eficiente.

O desenvolvimento do CNViewer visou abranger o maior número de usuários possível, o que foi buscado através da versatilização nos formatos de entrada de dados e

análises. Assim foi possível atingir usuários com domínio restrito de informática e aqueles com maior conhecimento, bem como usuários interessados em análises genômicas globais ou em trechos específicos, oferecendo um conjunto de recursos distintos dos apresentados pelos outros programas para análise de CNV.

A implementação do CNViewer foi feita com os olhos no futuro, pensando na medicina personalizada ao implementar métodos comparativos e representação de dados complementares. O método estatístico não é robusto o suficiente, para uso em prognósticos, mas o desenvolvimento de um algoritmo para correlação entre CNVs e dados clínicos é um desafio enfrentado por toda a comunidade científica. Enquanto isso, a visualização dos perfis de CNV já serve de apoio para o uso em diagnósticos.

Futuramente, novas características podem ser adicionadas ao CNViewer, melhorando as formas de análise e fortalecendo a segurança de seu uso com a implementação de protocolo *TSL*. Pretendemos futuramente inserir opções de escolha de genes para análise, além de novas métricas.

Com o desenvolvimento de aplicativos com cada vez mais funcionalidades executadas em navegador, é esperado que as companhias desenvolvam navegadores cada vez mais aptos para processamento de um grande volume de dados e que o uso deles para a implementação de programas se torne cada vez mais popular.

Referências Bibliográficas

Agilent Technologies. (n.d.). "CGH Analytics Software Showcase."

Almagro-Garcia, J., Manske, M., Carret, C., Campino, S., Auburn, S., Macinnis, B. L., Maslen, G., Pain, A., Newbold, C. I., Kwiatkowski, D. P., e Clark, T. G. (2009). "SnoopCGH: software for visualizing comparative genomic hybridization data." *Bioinformatics (Oxford, England)*, 25(20), 2732-3.

Almeida, J. S., Chen, C., Gorlitsky, R., Stanislaus, R., Aires-de-Sousa, M., Eleutério, P., Carriço, J., Maretzek, A., Bohn, A., Chang, A., Zhang, F., Mitra, R., Mills, G. B., Wang, X., e Deus, H. F. (2006). "Data integration gets ‘ Sloppy ’." *Nature Biotechnology*, 24(9), 1070-1071.

Alonso, G., Casati, F., Kuno, H., e Machiraju, V. (1998). *Web Services Concepts, Architectures and Applications*. Springer, Berlin, 354.

Autio, R., Saarela, M., Järvinen, A., Hautaniemi, S., e Astola, J. (2009). "Advanced analysis and visualization of gene copy number and expression data." *BMC bioinformatics*, 10 Suppl 1, S70.

Awad, I. A., Rees, C. A., Hernandez-Boussard, T., Ball, C. A., e Sherlock, G. (2004). "Caryoscope: an Open Source Java application for viewing microarray data in a genomic context.." *BMC bioinformatics*, Stanford University, 5, 151.

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J. S., Dobson, J., Urashima, M., Henry, K. T., Pinchback, R. M., Ligon, A. H., Cho, Y., Haery, L., Greulich, H., Reich, M., Winckler, W.,

Lawrence, M. S., Weir, B. A., Tanaka, K. E., Chiang, D. Y., Bass, A. J., Loo, A., Hoffman, C., Prensner, J., Liefeld, T., Gao, Q., Yecies, D., Signoretti, S., Maher, E., Kaye, F. J., Sasaki, H., Tepper, J. E., Fletcher, J. A., Tsao, M., Demichelis, F., Rubin, M. A., Janne, P. A., Tabernero, J., Daly, M. J., Nucera, C., Levine, R. L., Ebert, B. L., Gabriel, S., Rustgi, A. K., Antonescu, C. R., Ladanyi, M., Letai, A., Garraway, L. A., Loda, M., Beer, D. G., True, L. D., Okamoto, A., Pomeroy, S. L., Singer, S., Golub, T. R., Lander, E. S., Getz, G., Sellers, W. R., e Meyerson, M. (2010). "The landscape of somatic copy-number alteration across human cancers." *Nature*, 463(February), 899-905.

BioDiscovery. (n.d.). "Nexus Copy Number - Accelerating Copy Number Analysis Research."

Breitman, K. (2005). "Web Semântica: a Internet do futuro." *Rio de Janeiro: LTC*, LTC, Rio de Janeiro, 1-13.

Carter, N. P. (2007). "Methods and strategies for analyzing copy number variation using DNA microarrays." *Nature genetics*, 39(7 Suppl), S16-21.

Chari, R., Coe, B. P., Wedseltoft, C., Benetti, M., Wilson, I. M., Vucic, E. A., MacAulay, C., Ng, R. T., e Lam, W. L. (2008). "SIGMA2: a system for the integrative genomic multi-dimensional analysis of cancer genomes, epigenomes, and transcriptomes." *BMC bioinformatics*, 9, 422.

Chari, R., Lockwood, W. W., e Lam, W. L. (2007). "Computational methods for the analysis of array comparative genomic hybridization." *Cancer informatics*, 2, 48-58.

Chen, W., Erdogan, F., Ropers, H., Lenzner, S., e Ullmann, R. (2005). "CGHPRO -- a comprehensive data analysis tool for array CGH.." *BMC bioinformatics*, 6, 85.

Chi, B., DeLeeuw, R. J., Coe, B. P., Ng, R. T., MacAulay, C., e Lam, W. L. (2008). "MD-SeeGH: a platform for integrative analysis of multi-dimensional genomic data." *BMC bioinformatics*, 9, 243.

Conde, L., Montaner, D., Burguet-Castell, J., Tárraga, J., Medina, I., Al-Shahrour, F., e Dopazo, J. (2007). "ISACGH: a web-based environment for the analysis of Array CGH and gene expression which includes functional profiling." *Nucleic acids research*, 35(Web Server issue), W81-5.

Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E., e Pritchard, J. K. (2006). "A high-resolution survey of deletion polymorphism in the human genome." *Nature genetics*, 38(1), 75-81.

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T. D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C. H., Kristiansson, K., Macarthur, D. G., Macdonald, J. R., Onyiah, I., Wing, A., Pang, C., Robson, S., Stirrups, K., Valsesia, A., Walter, K., Wei, J., Trust, W., Control, C., Tyler-smith, C., Carter, N. P., Lee, C., Scherer, S. W., e Hurles, M. E. (2010). "Origins and functional impact of copy number variation in the human genome." *Nature*, Nature Publishing Group, 464(7289), 704-712.

Cooper, G. M., Nickerson, D. A., e Eichler, E. E. (2007). "Mutational and selective effects on copy-number variants in the human genome." *Nature genetics*, 39(7 Suppl), S22-9.

Curtis, C., Lynch, A. G., Dunning, M. J., Spiteri, I., Marioni, J. C., Hadfield, J., Chin, S., Brenton, J. D., Tavaré, S., e Caldas, C. (2009). "The pitfalls of platform

comparison: DNA copy number array technologies assessed." *BMC genomics*, 10, 588.

Deus, H. F. (n.d.). "S3DB API Basics."

Deus, H. F., Stanislaus, R., Veiga, D. F., Behrens, C., Wistuba, I. I., Minna, J. D., Garner, H. R., Swisher, S. G., Roth, J. A., Correa, A. M., Broom, B., Coombes, K., Chang, A., Vogel, L. H., e Almeida, J. S. (2008). "A Semantic Web management model for integrative biomedical informatics." *PloS one*, 3(8), e2946.

Dhawan, D., e Padh, H. (2009). "Pharmacogenetics: technologies to detect copy number variations." *Current opinion in molecular therapeutics*, 11(6), 670-80.

Díaz-Uriarte, R., Alibés, A., Morrissey, E. R., Cañada, A., Rueda, O. M., e Neves, M. L. (2007). "Asterias: integrated analysis of expression and aCGH data using an open-source, web-based, parallelized software suite." *Nucleic acids research*, 35(Web Server issue), W75-80.

Feuk, L., Carson, A. R., e Scherer, S. W. (2006). "Structural variation in the human genome." *Nature reviews. Genetics*, 7(2), 85-97.

Flanagan, D. (1998). *JavaScript The Definitive Guide*. (P. Ferguson), O'Reilly, Sebastopol.

Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., Aburatani, H., Jones, K. W., Tyler-Smith, C., Hurles, M. E., Carter, N. P., Scherer, S. W., e Lee, C. (2006). "Copy number variation: new insights in genome diversity." *Genome research*, 16(8), 949-61.

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., e Stratton, M. R. (2004). "A census of human cancer genes." *Nature reviews. Cancer*, Nature Publishing Group, 4(3), 177-83.

Goodman, D., e Morrison, M. (2007). *Javascript® Bible*. *Annals of Physics*, Wiley Publishing, Inc., Indianapolis, 1173.

Google Project Hosting. (2010). "Project Hosting on Google Code."

Heidenblad, M., Lindgren, D., Veltman, J. A., Jonson, T., Mahlamäki, E. H., Gorunova, L., van Kessel, A. G., Schoenmakers, E. F., e Höglund, M. (2005). "Microarray analyses reveal strong influence of DNA copy number alterations on the transcriptional patterns in pancreatic cancer: implications for the interpretation of genomic amplifications." *Oncogene*, 24(10), 1794-801.

Henrichsen, C. N., Vinckenbosch, N., Zöllner, S., Chaignat, E., Pradervand, S., Schütz, F., Ruedi, M., Kaessmann, H., e Reymond, A. (2009). "Segmental copy number variation shapes tissue transcriptomes." *Nature genetics*, 41(4), 424-9.

Hofmann, W. a., Weigmann, A., Tauscher, M., Skawran, B., Focken, T., Buurman, R., Wingen, L. U., Schlegelberger, B., e Steinemann, D. (2009). "Analysis of Array-CGH Data Using the R and Bioconductor Software Suite." *Comparative and functional genomics*, 2009, 201325.

Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., Fung, H., Szpiech, Z. A., Degnan, J. H., Wang, K., Guerreiro, R., Bras, J. M., Schymick, J. C., Hernandez, D. G., Traynor, B. J., Simon-Sanchez, J., Matarin, M., Britton, A., van De Leemput, J., Rafferty, I., Bucan, M., Cann, H. M., Hardy, J. A., Rosenberg, N.

A., e Singleton, A. B. (2008). "Genotype, haplotype and copy-number variation in worldwide human populations." *Nature*, 451(7181), 998-1003.

Jong, K., Marchiori, E., van Der Vaart, A., e Ylstra, B. (n.d.). "VU Micro-Array Data Analysis."

Kim, S. Y., Nam, S. W., Lee, S. H., Park, W. S., Yoo, N. J., Lee, J. Y., e Chung, Y. (2005). "ArrayCyGHt: a web application for analysis and visualization of array-CGH data." *Bioinformatics (Oxford, England)*, 21(10), 2554-5.

Kin, T., e Ono, Y. (2007). "Idiographica: a general-purpose web application to build idiograms on-demand for human, mouse and rat." *Bioinformatics (Oxford, England)*, 23(21), 2945-6.

Kingsley, C. B., Kuo, W., Polikoff, D., Berchuck, A., Gray, J. W., e Jain, A. N. (2007). "Magellan: a web based system for the integrated analysis of heterogeneous biological data and annotations; application to DNA copy number and expression data in ovarian cancer." *Cancer informatics*, 2(415), 10-21.

Kotliarov, Y., Bozdag, S., Cheng, H., Wuchty, S., Zenklusen, J., e Fine, H. A. (2010). "CNAReporter: a GenePattern pipeline for the generation of clinical reports of genomic alterations." *BMC medical genomics*, 3(1), 11.

Kuhn, R. M., Karolchik, D., Zweig, a. S., Wang, T., Smith, K. E., Rosenbloom, K. R., Rhead, B., Raney, B. J., Pohl, a., Pheasant, M., Meyer, L., Hsu, F., Hinrichs, a. S., Harte, R. a., Giardine, B., Fujita, P., Diekhans, M., Dreszer, T., Clawson, H., Barber, G. P., Haussler, D., e Kent, W. J. (2009). "The UCSC Genome Browser Database: update 2009.." *Nucleic acids research*, 37(Database issue), D755-61.

La Rosa, P., Viara, E., Hupé, P., Pierron, G., Liva, S., Neuvial, P., Brito, I., Lair, S., Servant, N., Robine, N., Manié, E., Brennetot, C., Janoueix-Lerosey, I., Raynal, V., Gruel, N., Rouveirol, C., Stransky, N., Stern, M., Delattre, O., Aurias, A., Radvanyi, F., e Barillot, E. (2006). "VAMP: visualization and analysis of array-CGH, transcriptome and other molecular profiles." *Bioinformatics (Oxford, England)*, 22(17), 2066-73.

Lai, W. R., Johnson, M. D., Kucherlapati, R., e Park, P. J. (2005). "Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data." *Bioinformatics (Oxford, England)*, 21(19), 3763-70.

Lee, C., e Scherer, S. W. (2010). "The clinical context of copy number variation in the human genome." *Expert reviews in molecular medicine*, 12(March), e8.

Lee, M., e Kim, Y. (2009). "CHESS (CgHExpreSS): a comprehensive analysis tool for the analysis of genomic alterations and their effects on the expression profile of the genome." *BMC bioinformatics*, 10, 424.

Li, C. (2008). "Automating dChip: toward reproducible sharing of microarray data analysis." *BMC bioinformatics*, 9, 231.

Lima, E. L. (1975). "Espaços Métricos." *Espaços Métricos*, Instituto de Matemática Pura e Aplicada do CNPq, Rio de Janeiro, 1-51.

Lingjærde, O. C., Baumbusch, L. O., Liestøl, K., Glad, I. K., e Børresen-Dale, A. (2005). "CGH-Explorer: a program for analysis of array-CGH data." *Bioinformatics (Oxford, England)*, 21(6), 821-2.

Lo, K. C., Shankar, G., Turpaz, Y., Bailey, D., Rossi, M. R., Burkhardt, T., Liang, P., e Cowell, J. K. (2007). "Overlay Tool for aCGHViewer: An Analysis Module Built for aCGHViewer used to Perform Comparisons of Data Derived from Different Microarray Platforms." *Cancer informatics*, 3(716), 307-19.

Lockwood, W. W., Chari, R., Chi, B., e Lam, W. L. (2006). "Recent advances in array comparative genomic hybridization technologies and their applications in human genetics." *European journal of human genetics*, 14(2), 139-48.

Macedo, M. D. (2004). *Construindo Sites Adotando os Padrões da Web*. (P. A. Marques), Ciência Moderna, Rio de Janeiro, 249.

Margolin, A. A., Greshock, J., Naylor, T. L., Mosse, Y., Maris, J. M., Bignell, G., Saeed, A. I., Quackenbush, J., e Weber, B. L. (2005). "CGHAnalyzer: a stand-alone software package for cancer genome analysis using array-based DNA copy number data." *Bioinformatics (Oxford, England)*, 21(15), 3308-11.

McCarroll, S. A. (2010). "Copy number variation and human genome maps." *Nature genetics*, 42(5), 365-6.

McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M. H., de Bakker, P. I., Maller, J. B., Kirby, A., Elliott, A. L., Parkin, M., Hubbell, E., Webster, T., Mei, R., Veitch, J., Collins, P. J., Handsaker, R., Lincoln, S., Nizzari, M., Blume, J., Jones, K. W., Rava, R., Daly, M. J., Gabriel, S. B., e Altshuler, D. (2008). "Integrated detection and population-genetic analysis of SNPs and copy number variation." *Nature genetics*, 40(10), 1166-74.

Mertens, F., Johansson, B., Höglund, M., e Mitelman, F. (1997). "Chromosomal imbalance maps of malignant solid tumors: a cytogenetic survey of 3185 neoplasms.." *Cancer research*, 57(13), 2765-80.

Monni, O., e Hautaniemi, S. (2009). "Bioinformatics of gene expression and copy number data integration." *Statisticsand informatics in molecular cancer research*, C. Wiuf e C. L. Andersen, Oxford University Press, New York, 78-81.

Myers, C. L., Dunham, M. J., Kung, S. Y., e Troyanskaya, O. G. (2004). "Accurate detection of aneuploidies in array CGH and gene expression microarray data." *Bioinformatics (Oxford, England)*, 20(18), 3533-43.

Nannya, Y., Sanada, M., Nakazaki, K., Hosoya, N., Wang, L., Hangaishi, A., Kurokawa, M., Chiba, S., Bailey, D. K., Kennedy, G. C., e Ogawa, S. (2005). "A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays." *Cancer research*, 65(14), 6071-9.

National Cancer Institute, e National Human Genome Research Institute. (n.d.). "The Cancer Genome Atlas."

NimbleGen Systems, I. (2006). *SignalMap 1.9 User's Guide*.

O'Reilly, T. (2007). "What is Web 2.0: Design patterns e business models for the next generation of software." *Communications & Strategies*, (65), 17-37.

Olshen, A. B., Venkatraman, E. S., Lucito, R., e Wigler, M. (2004). "Circular binary segmentation for the analysis of array-based DNA copy number data." *Biostatistics (Oxford, England)*, 5(4), 557-72.

Ostrovnaya, I., Olshen, A. B., Seshan, V. E., Orlow, I., Albertson, D. G., e Begg, C. B. (2010). "A metastasis or a second independent cancer? Evaluating the clonal origin of tumors using array copy number data." *Statistics in medicine*.

Pinto, D., Marshall, C., Feuk, L., e Scherer, S. W. (2007). "Copy-number variation in control population cohorts." *Human molecular genetics*, 16 Spec No(2), R168-73.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., González, J. R., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W., e Hurles, M. E. (2006). "Global variation in copy number in the human genome." *Nature*, 444(7118), 444-54.

Rolewicz, S. (1985). "Definition of metric linear spaces and the theorem on the invariant norm." *Metric Linear Spaces*, D. Reidel Publishing Company, Dordrecht, 1-6.

Sanders, M. A., Verhaak, R. G., Geertsma-Kleinekoort, W. M., Abbas, S., Horsman, S., van Der Spek, P. J., Löwenberg, B., e Valk, P. J. (2008). "SNPExpress: integrated visualization of genome-wide genotypes, copy numbers and gene expression levels." *BMC genomics*, 9, 41.

Shaikh, T. H., Gai, X., Perin, J. C., Glessner, J. T., Xie, H., Murphy, K., O'Hara, R., Casalunovo, T., Conlin, L. K., D'Arcy, M., Frackelton, E. C., Geiger, E. A.,

Haldeman-Englert, C., Imielinski, M., Kim, C. E., Medne, L., Annaiah, K., Bradfield, J. P., Dabaghyan, E., Eckert, A., Onyiah, C. C., Ostapenko, S., Otieno, F. G., Santa, E., Shaner, J. L., Skraban, R., Smith, R. M., Elia, J., Goldmuntz, E., Spinner, N. B., Zackai, E. H., Chiavacci, R. M., Grundmeier, R., Rappaport, E. F., Grant, S. F., White, P. S., e Hakonarson, H. (2009). "High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications." *Genome research*, 19(9), 1682-90.

Shankar, G., Rossi, M. R., McQuaid, D. E., Conroy, J. M., Gaile, D. G., Cowell, J. K., Nowak, N. J., e Liang, P. (2006). "aCGHViewer: A Generic Visualization Tool For aCGH data." *Cancer informatics*, 2, 36-43.

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon, R., Bird, C. P., de Grassi, A., Lee, C., Tyler-Smith, C., Carter, N., Scherer, S. W., Tavaré, S., Deloukas, P., Hurles, M. E., e Dermitzakis, E. T. (2007). "Relative impact of nucleotide and copy number variation on gene expression phenotypes." *Science (New York, N.Y.)*, 315(5813), 848-53.

Stuart, D., e Sellers, W. R. (2009). "Linking somatic genetic alterations in cancer to therapeutics." *Current opinion in cell biology*, 21(2), 304-10.

Su, S., Asher, J. E., Jarvelin, M., Froguel, P., Blakemore, A. I., Balding, D. J., e Coin, L. J. (2010). "Inferring combined CNV/SNP haplotypes from genotype data." *Bioinformatics (Oxford, England)*, 1-10.

The Dojo Foundation. (2009). "Dojo toolkit - Unbeatable JavaScript Tools."

Ting, J. C., Ye, Y., Thomas, G. H., Ruczinski, I., e Pevsner, J. (2006). "Analysis and visualization of chromosomal abnormalities in SNP data with SNPscan." *BMC bioinformatics*, 7, 25.

Venkatraman, E. S., e Olshen, A. B. (2007). "A faster circular binary segmentation algorithm for the analysis of array CGH data." *Bioinformatics (Oxford, England)*, 23(6), 657-63.

Wang, J., Meza-Zepeda, L. A., Kresse, S. H., e Myklebost, O. (2004). "M-CGH: analysing microarray-based CGH experiments." *BMC bioinformatics*, 5, 74.

Wang, P., Kim, Y., Pollack, J., Narasimhan, B., e Tibshirani, R. (2005). "A method for calling gains and losses in array CGH data.." *Biostatistics (Oxford, England)*, 6(1), 45-58.

Willenbrock, H., e Fridlyand, J. (2005). "A comparison study: applying segmentation to array CGH data for downstream analyses." *Bioinformatics (Oxford, England)*, 21(22), 4084-91.

Winchester, L., Yau, C., e Ragoussis, J. (2009). "Comparing CNV detection methods for SNP arrays." *Briefings in functional genomics & proteomics*, 8(5), 353-66.

Wong, K. K., DeLeeuw, R. J., Dosanjh, N. S., Kimm, L. R., Cheng, Z., Horsman, D. E., MacAulay, C., Ng, R. T., Brown, C. J., Eichler, E. E., e Lam, W. L. (2007). "A comprehensive analysis of common copy-number variations in the human genome." *American journal of human genetics*, 80(1), 91-104.

Yi, Y., Mirosevich, J., Shyr, Y., Matusik, R., e George, A. L. (2005). "Coupled analysis of gene expression and chromosomal location.." *Genomics*, 85(3), 401-12.

Zhang, N. R., Senbabaoglu, Y., e Li, J. Z. (2010). "Joint estimation of DNA copy number from multiple platforms." *Bioinformatics (Oxford, England)*, 26(2), 153-60.

de Smith, A. J., Tsalenko, A., Sampas, N., Scheffer, A., Yamada, N. A., Tsang, P., Ben-Dor, A., Yakhini, Z., Ellis, R. J., Bruhn, L., Laderman, S., Froguel, P., e Blakemore, A. I. (2007). "Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases." *Human molecular genetics*, 16(23), 2783-94.

van de Wiel, M. A., Picard, F., van Wieringen, W. N., e Ylstra, B. (2010). "Preprocessing and downstream analysis of microarray DNA copy number profiles." *Briefings in bioinformatics*.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Administração
Baixar livros de Agronomia
Baixar livros de Arquitetura
Baixar livros de Artes
Baixar livros de Astronomia
Baixar livros de Biologia Geral
Baixar livros de Ciência da Computação
Baixar livros de Ciência da Informação
Baixar livros de Ciência Política
Baixar livros de Ciências da Saúde
Baixar livros de Comunicação
Baixar livros do Conselho Nacional de Educação - CNE
Baixar livros de Defesa civil
Baixar livros de Direito
Baixar livros de Direitos humanos
Baixar livros de Economia
Baixar livros de Economia Doméstica
Baixar livros de Educação
Baixar livros de Educação - Trânsito
Baixar livros de Educação Física
Baixar livros de Engenharia Aeroespacial
Baixar livros de Farmácia
Baixar livros de Filosofia
Baixar livros de Física
Baixar livros de Geociências
Baixar livros de Geografia
Baixar livros de História
Baixar livros de Línguas

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo