UNIVERSIDADE CATÓLICA DE BRASÍLIA

PROGRAMA DE PÓS-GRADUAÇÃO
STRICTO SENSU EM CIÊNCIAS GENÔMICAS E BIOTECNOLOGIA

Mestrado

ANÁLISE COMPARATIVA DE ALGORITMOS DE AGRUPAMENTO DE ESTs ("Expressed Sequence Tags")

Autor: Alexandre Peixoto Figueira

Orientador: Georgios Joannis Pappas Júnior

BRASÍLIA **2006**

Alexandre Peixoto Figueira

ANÁLISE COMPARATIVA DE ALGORITMOS DE AGRUPAMENTO DE ESTs ("Expressed Sequence Tags")

Dissertação apresentada ao Programa de Pós-Graduação "Stricto Sensu" em Ciências Genômicas e Biotecnologia da Universidade Católica de Brasília, como requisito para a obtenção do Título de Mestre em Ciências Genômicas e Biotecnologia.

Orientador: Georgios Joannis Pappas Júnior.

Brasília
2006

TERMO DE APROVAÇÃO

Dissertação defendida e aprovada como requisito parcial para a obtenção do Título de Mestre em Ciências Genômicas e Biotecnologia, em 31 de agosto de 2006, pela banca examinadora constituída por:

_____

Georgios Joannis Pappas Júnior

_____

Marcos Mota do Carmo Costa

_____

Natália Florêncio Martins

Brasília
UCB

Ao Prof. Dr. Georgios,

Meus agradecimentos pela paciência e compreensão.

ÍNDICE DE TABELAS

ÍNDICE DE FIGURAS

RESUMO

As ESTs ("Expressed Sequence Tags") são seqüências curtas de DNA geradas a partir do seqüenciamento aleatório de uma biblioteca de cDNA. Por conter informações sobre os genes expressos na célula, encontram aplicação em vários tipos de pesquisa, principalmente para a descoberta de genes e avaliação de seu nível de expressão em diferentes tipos celulares. Devido a sua redundância inerente, faz-se necessária a organização das ESTs em grupos que contenham mensagens do mesmo transcrito. Este processo é denominado de agrupamento de ESTs e resulta em uma redução da complexidade dos dados e ao mesmo tempo provê estimativas da abundância dos mRNAs correspondentes. Existem diversas ferramentas especializadas nesta tarefa descritas na literatura, e o principal objetivo deste estudo é realizar a primeira comparação objetiva de acuidade entre cinco destas ferramentas (CAP3, d2_cluster, ESTate, TGICL e XSACT), utilizando um agrupamento de referência construído com o auxílio da seqüência completa do genoma humano. Diversas análises comparativas foram realizadas demonstrando que de maneira geral as ferramentas geram agrupamentos com boa qualidade em relação ao agrupamento de referência, e que estas apresentam resultados semelhantes entre si. No entanto, em alguns casos, os resultados das ferramentas podem ser drasticamente afetados pela biblioteca de ESTs. À luz dos critérios utilizados na avaliação das ferramentas, a ferramenta XSACT foi a que apresentou os melhores resultados, mas não existe uma diferença significativa que impeça a utilização das outras ferramentas estudadas.

PALAVRAS-CHAVE: ESTs, agrupamento de ESTs, genômica, análise de seqüências biológicas.

ABSTRACT

Expressed sequence tags (ESTs) are short single pass sequences generated by random sequencing selected clones of a cDNA library. Since they contain information about expressed genes in the cell, they are suited for several applications, mainly for gene discovery and expression profiling. Due to their intrinsic redundancy, it is important to classify ESTs in groups containing all the messages of the same transcript. This process is called ESTs clustering and leads to a data complexity reduction and provides estimates of mRNA abundance. Several specialized tools are available in the literature for this task, and the main objective of this study was to provide the first objective comparison regarding the accuracy of five such tools (CAP3, d2_cluster, ESTate, TGICL and XSACT), based on a reference clustering constructed based on the information of the complete human genome sequence. Several comparative analyses were conducted and they showed that the clustering tools display good agreement with standard clustering, and that they produce similar results. However, in some cases the results of the tools are affected by the cDNA library itself. Based on several criteria, XSACT displayed more consistent results; nevertheless there is no significant difference that points to the utilization of a specific EST clustering tool.

KEYWORDS: ESTs, EST clustering, genomics, biological sequence analysis.

SUMÁRIO

1 INTRODUÇÃO

1.1 A Era Genômica

Na década de 80, com os aperfeiçoamentos no método de seqüenciamento de DNA e o surgimento de seqüenciadores automáticos houve um aumento expressivo na capacidade de produção de seqüências. Em 1995, a primeira publicação do genoma completo de um organismo de vida livre (Fleischmann et al., 1995), a bactéria *Haemophilus influenzae*, marcou o início da Era Genômica. A partir daí, vários genomas já foram seqüenciados, como pode se constatar no GOLD ("Genomes On Line Database") (Liolios et al., 2006), recurso que disponibiliza informações sobre projetos de seqüenciamento de genomas completos e em andamento. A publicação do primeiro esboço do genoma humano em 2001 (Lander et al., 2001; Venter et al., 2001) e a publicação de edições especiais das revistas *Science* e *Nature* em abril de 2003 sobre a finalização do genoma humano consolidaram em definitivo a Era Genômica.

Inicialmente, o principal interesse da Bioinformática era criar bancos de dados de informações biológicas e prover acesso a essas informações, principalmente seqüências de DNA e proteína. Atualmente, a Bioinformática busca relacionar vários tipos de dados, através da integração de diferentes bancos de dados biológicos e desenvolvimento de novos algoritmos, de modo a fornecer um panorama abrangente sobre as atividades celulares.

A Bioinformática permite que muito tempo de pesquisa seja poupado ao fazer inferências que podem dispensar experimentos *in vivo* ou *in vitro*, sendo estes necessários somente em uma fase tardia de confirmação das inferências. Entretanto, a utilização de ferramentas computacionais não dispensa o julgamento humano e nem sempre provê respostas definitivas para os problemas com os quais os biólogos se deparam, por isso, não pode ser encarada como uma panacéia e tem que ser utilizada com prudência.

1.2 Estratégias de Seqüenciamento de Genomas

Existem duas estratégias principais para o seqüenciamento de genomas. A primeira é denominada "hierarchical shotgun sequencing" - também conhecida pelas denominações "map-based sequencing" ou "BAC-based sequencing" - e a segunda é denominada "whole-genome shotgun sequencing". O genoma humano, por exemplo, foi seqüenciado paralelamente e independentemente por duas iniciativas, cada qual baseada em uma das estratégias descritas acima (Lander et al., 2001; Venter et al., 2001).

O método de seqüenciamento "shotgun" (Gardner et al., 1981; Anderson, 1981; Sanger et al., 1982), primeiramente descrito no início dos anos 80, consiste em: fragmentar um pedaço grande de DNA em pedaços menores; gerar quantidades redundantes de seqüências a partir dos fragmentos aleatórios; e, por fim, utilizar o computador para juntar as seqüências e revelar o DNA inicial. As ferramentas computacionais utilizam as regiões de sobreposição entre as seqüências para chegar ao DNA original, portanto, a redundância dos dados de seqüências, também chamada de cobertura, é primordial para que não ocorram lacunas que impeçam que o DNA inicial seja montado por completo. Tipicamente, para produzir seqüências de alta qualidade, a cobertura deve ser de 8 a 10 vezes o tamanho do DNA original. Significa dizer que, para obter uma cobertura de 10 vezes o trecho inicial, assumindo um tamanho médio de seqüência de 500 pares de bases, aproximadamente 3.000 seqüências devem ser geradas para possibilitar a montagem de um BAC[1] de 150.000 pares de bases (500 x 3.000 é igual a 150.000 x 10). Um BAC é um vetor de clonagem que pode receber trechos de DNA de até 150 Kb aproximadamente. As duas estratégias de seqüenciamento de genomas que serão descritas a seguir utilizam o método "shotgun", porém em momentos diferentes.

[1] Bacterial Artificial Chromosome.

1.2.1 Hierarchical Shotgun Sequencing

A Figura 1a apresenta uma visão geral esquemática da estratégia hierárquica. Por analogia o genoma é representado como uma enciclopédia, onde cada volume corresponde a um cromossomo individual. Nesta estratégia procede-se inicialmente à construção de mapas físicos baseados em clones de BAC, produzindo uma série destes que se sobrepõem, formando uma região contígua do genoma original. Cada clone pode ser imaginado como contendo o DNA representado por uma página de um volume. Para aplicar o método de seqüenciamento "shotgun", cada clone de BAC selecionado para seqüenciamento a partir do mapa de clones é sub-clonado em bibliotecas de insertos menores, que são seqüenciados aleatoriamente. Após a montagem dos BACs individuais, procede-se à montagem completa do genoma de forma hierárquica baseando-se no mapa físico.

1.2.2 Whole-Genome Shotgun Sequencing

A Figura 1b apresenta uma visão geral esquemática da estratégia "whole-genome shotgun sequencing". Diferentemente da estratégia hierárquica, nessa não há preocupação em construir um mapa físico dos BACs. O genoma inteiro é quebrado aleatoriamente em fragmentos pequenos que são clonados em plasmídios e seqüenciados nos dois sentidos. Uma vez que as seqüências são obtidas, as ferramentas computacionais montam os fragmentos em regiões contíguas através da busca por sobreposições. Essa estratégia de seqüenciamento se mostra mais frágil do que a hierárquica no caso de genomas que são ricos em seqüências repetitivas, pois estas podem levar à montagem de fragmentos espúrios.

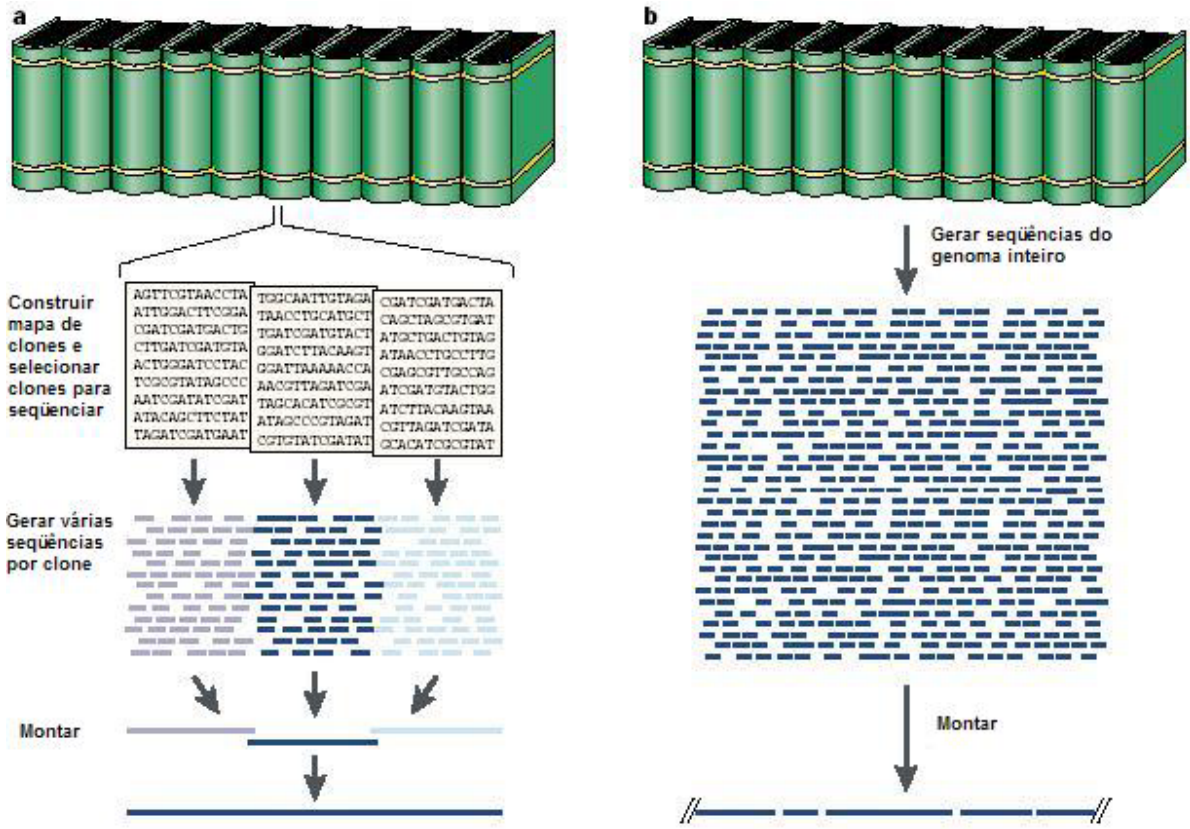

Figura 1 - As duas principais estratégias de seqüenciamento de genomas. a) Hierarchical Shotgun Sequencing. b) Whole Genome Shotgun Sequencing. Adaptado de (Green, 2001).

1.3 Genômica Funcional

As estratégias descritas anteriormente são utilizadas quando se pretende caracterizar a estrutura do genoma inteiro. Em espécies com genomas pequenos com até 10 Mb, como as bactérias, seqüenciar o genoma completo de um organismo não representa atualmente um problema, porém, em espécies com genomas maiores, como no caso dos eucariotos (30 Mb até dezenas Gb), os custos representam um grande obstáculo em função do tamanho do esforço, que é considerável. Assim, nem sempre é viável seqüenciar o genoma completo de determinado organismo se este possui um genoma muito extenso. Uma alternativa a esta dificuldade é o seqüenciamento de uma parcela do genoma, especificamente das moléculas de mRNA transcritas de genes codificadores de proteínas da célula. A este subconjunto de seqüências dá-se o nome de

transcritoma. Essas moléculas de mRNA direcionam a síntese do produto final resultante da expressão do genoma, as proteínas. Para que seja possível seqüenciar os mRNAs, inicialmente são construídas bibliotecas de DNA complementar (cDNA) a partir dos mRNAs isolados de um tecido ou tipo celular específico (Alberts et al., 2002). O seqüenciamento do transcritoma exige um esforço bem menor que o do genoma, além de fornecer informações valiosas por se tratar das porções gênicas do genoma. Por esse motivo, também é utilizado complementarmente em projetos de seqüenciamento de genomas completos para ajudar na identificação de genes, na elucidação da estrutura de genes, na identificação de polimorfismos entre indivíduos e análise de expressão celular ou tecido-específica.

1.3.1 Expressed Sequence Tags

No contexto da transcritômica, as bibliotecas de cDNA são produzidas e os insertos são seqüenciados. Caso este seqüenciamento se dê por uma única leitura dos clones, gerando entre 300 a 500 bases, estas seqüências são denominadas ESTs ("Expressed Sequence Tags"). ESTs representam somente uma porção do gene e inicialmente foram vislumbradas como uma maneira de identificar genes expressos. As primeiras 609 ESTs foram descritas por Adams e colaboradores em 1991 (Adams et al., 1991).

Apesar de sua natureza fragmentária e de ser relativamente inexata em virtude de ser seqüenciada em uma única passada, as ESTs se mostraram um recurso valioso na descoberta de novos genes (Sikela e Auffray, 1993; Boguski et al., 1994). Após a demonstração inicial de sua utilidade e efetividade, vários projetos de descoberta de genes foram estabelecidos baseados nessa tecnologia. Adicionalmente, projetos de seqüenciamento de ESTs em larga escala foram iniciados para vários organismos de interesse experimental. Em 1992, um banco de dados público chamado dbEST ("database of Expressed Sequence Tags") (Boguski et al., 1993) foi criado para

armazenar as ESTs. O dbEST representa a divisão do GenBank (Benson et al., 2002) com o maior número de submissões. Em sua versão de junho de 2006 ("dbEST release 060206") apresentava mais de 36 milhões de seqüências armazenadas (36.750.628). As ESTs também já provaram sua utilidade no mapeamento genético (Khan et al., 1992), na anotação genômica, na descoberta de SNPs[2] (Hu et al., 2002; Picoult-Newberg et al., 1999) e na detecção de junções alternativas ("alternative splicing") de mRNAs (Lee, 2003; Heber et al., 2002; Xu et al., 2002; Modrek e Lee, 2002; Modrek et al., 2001).

1.3.2 Produção de ESTs

A produção de ESTs envolve a construção de uma biblioteca de cDNA e o posterior seqüenciamento dos clones de cDNA da biblioteca. A Figura 2 mostra uma representação esquemática geral deste processo. O primeiro passo na construção da biblioteca de cDNA consiste em isolar a população de mRNAs do tecido ou tipo celular de interesse. As moléculas de mRNA são então utilizadas como moldes para a síntese de moléculas de DNA por uma enzima denominada transcriptase reversa. Essas moléculas de DNA transcritas a partir de mRNAs são denominadas cDNAs, as quais são inseridas em vetores de clonagem. O conjunto de clones de cDNA derivado da população de mRNAs constitui a biblioteca de cDNA. Os clones da biblioteca de cDNA são então seqüenciados para produzir as ESTs.

É importante ressaltar que a biblioteca de cDNA representa uma população de moléculas de mRNA que depende do tecido e do momento em que foi construída, portanto, uma alteração em qualquer uma dessas variáveis pode levar a um perfil diferente das populações de mRNAs.

[2] Single Nucleotide Polymorphisms.

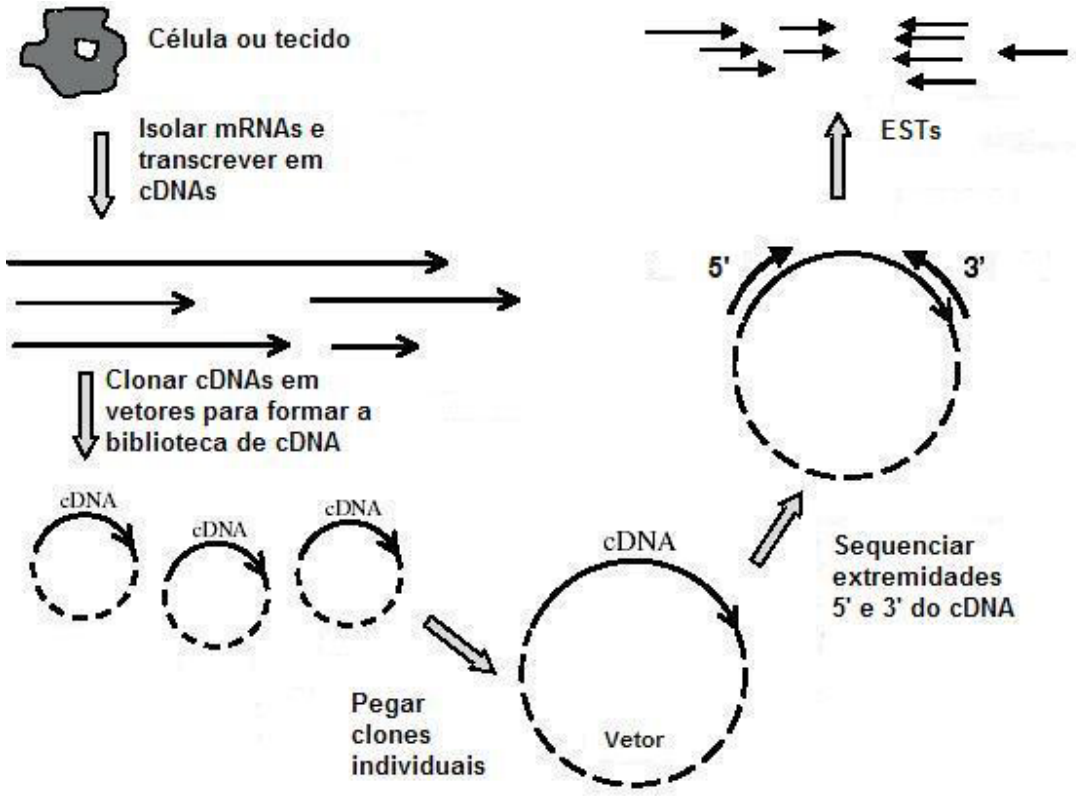

Figura 2 – Fabricação de ESTs. Adaptado de (Wolfsberg e Landsman, 2001).

1.4 Agrupamento de ESTs ("EST clustering")

Em uma biblioteca de ESTs, um gene pode ser representado por várias ESTs, o que freqüentemente dificulta o uso efetivo das mesmas. Em função da redundância inerente se faz necessária uma análise especial de forma a identificar a real representatividade das seqüências. Esta é denominada agrupamento de ESTs ("EST clustering") e tem por objetivo separar seqüências similares em grupos homogêneos e distintos. Como conseqüência, além de se possuir um perfil das seqüências únicas existentes, pode-se também obter estimativas sobre os níveis relativos de expressão do gene no contexto fisiológico da biblioteca correspondente. Uma vez organizadas em grupos, a análise das ESTs pode ser racionalizada, tornando-as valiosas para iniciativas como descoberta de novos genes e mapeamento de genomas.

Existem vários projetos que buscam produzir índices de genes para vários organismos. Um índice de genes nada mais é que um catálogo dos genes de um determinado organismo

gerado a partir da organização de ESTs e/ou mRNAs em grupos, sendo que cada grupo representa um gene. Um mesmo gene pode gerar transcritos diferentes em função de um evento de junção alternativa. Neste evento éxons são removidos juntamente com íntrons gerando transcritos com diferentes configurações de éxons. Assim, há duas formas de se organizar os transcritos nesses projetos. A primeira consiste em agrupar os transcritos por gene. Isso significa que todos os transcritos de um gene, incluindo aqueles que sofreram junção alternativa, estarão no mesmo grupo. A segunda consiste em agrupar as ESTs por transcrito. Isso significa que, mesmo sendo originados do mesmo gene, transcritos diferentes serão colocados em grupos distintos. A seguir são descritos os três principais projetos de índices de genes.

1.4.1 Índices de Genes

1.4.1.1 UniGene

O UniGene (Pontius et al., 2003), desenvolvido e mantido pelo NCBI ("National Center for Biotechnology Information"), é um sistema de organização automática das seqüências do GenBank em um conjunto não redundante de grupos orientados a gene. Cada grupo contém transcritos, incluindo aqueles que sofreram junção alternativa de éxons, que representam um gene único, bem como informações sobre tipos de tecido no qual o gene foi expresso e a localização de mapa. Os recursos do UniGene podem ser acessados através do endereço http://www.ncbi.nlm.nih.gov/UniGene/.

1.4.1.2 TGI

O "TIGR Gene Indices" (TGI) (Quackenbush et al., 2000) é uma iniciativa que busca consolidar ESTs e outras seqüências de genes anotadas. Uma diferença significante entre o TGI e o UniGene é que o primeiro consiste em montagens de ESTs e outras seqüências de genes ao invés de grupos. O TGI trata as ESTs como elementos de um projeto de seqüenciamento *shotgun*

do transcriptoma, onde primeiramente estas são agrupadas e elementos de um grupo são montados para produzir uma seqüência consenso de alta qualidade. Critérios estringentes de sobreposição são utilizados para montar as seqüências consenso tentativas (TC – "Tentative Consensus"). As seqüências consenso tendem a representar um transcrito, dessa forma, produtos de junção alternativa são agrupados separadamente. A geração de seqüências consenso apresenta algumas vantagens: a seqüência consenso produzida normalmente é maior que as ESTs individuais que a compõem, provendo um recurso que pode ser usado mais efetivamente para anotação funcional; as seqüências consenso podem ser melhor utilizadas que as ESTs individuais na anotação de seqüências genômicas.

O TIGR mantém índices de genes para animais, plantas, protistas e fungos. Entre os vários organismos estão humano, rato, camundongo, *Drosophila*, *Arabidopsis*, e outros. Os recursos do TGI podem ser acessados através do endereço http://www.tigr.org/tdb/tgi/.

1.4.1.3 STACK

O projeto STACK ("Sequence Tag Alignment and Consensus Knowledge Base") (Christoffels et al., 2001), mantido pelo "South African National Bioinformatics Institute" (SANBI), tem por objetivo gerar uma representação abrangente de cada um dos genes expressos no genoma humano através do processamento extensivo de fragmentos de genes para produzir alinhamentos acurados, destacar a diversidade e prover um conjunto de seqüências consenso cuidadosamente agrupado para cada gene. O projeto STACK é composto pelo índice de genes humanos STACKdb™, um banco de dados de transcritos humanos virtuais, e pelo stackPACK™, que é o pacote que contém as ferramentas utilizadas na criação do STACKdb™. Diferentemente do UniGene e do TGI, as ESTs no STACK são separadas por tipo de tecido antes de serem

agrupadas. Os recursos do STACK podem ser acessados através do endereço http://www.sanbi.ac.za/Dbases.html.

1.4.2 Problemas dos dados de ESTs

De forma geral, a obtenção de ESTs envolve duas fases: a construção da biblioteca de cDNA e o seqüenciamento dos clones de cDNA. Devido a fatores relacionados ao processo experimental e à própria natureza das ESTs, surgem problemas que influenciam na tarefa de organizar as ESTs em grupos.

Pelo fato de serem geradas em uma única leitura, ESTs têm uma taxa de erro da ordem de 3% maior que seqüências que são verificadas por múltiplas reações de seqüenciamento (Boguski et al., 1993). Durante a construção da biblioteca de cDNA as ESTs podem ser contaminadas por seqüências de bactéria, mitocôndria, vetor de clonagem, tRNA e rRNA. Além disso, podem conter substituições, deleções e inserções em comparação à seqüência de mRNA que as originou. Outro artefato são as seqüências quiméricas resultantes da ligação entre cDNAs distintos, que forma uma molécula híbrida.

Em virtude dos problemas citados anteriormente, o agrupamento de ESTs demanda um pré-processamento que busca eliminar prováveis fontes de erros. Seqüências contaminantes e repetitivas devem ser identificadas e marcadas com caracteres especiais, em um processo chamado mascaramento. Apesar de ser um dado intrínseco da seqüência, a seqüência repetitiva também pode levar a erro, pois contribui para a formação de sobreposições falsas de ESTs não relacionadas, o que resulta em grupos com tamanhos superdimensionados. Assim, deve ser mascarada juntamente com a seqüência contaminante. Outro aspecto importante diz respeito à qualidade da leitura de cada base a partir do arquivo bruto resultante do seqüenciamento, o cromatograma. O trabalho de traduzir os sinais contidos no cromatograma em uma seqüência de

bases é denominado nomeação de bases (“base-calling”). O programa Phred (Ewing et al., 1998; Ewing e Green, 1998) é largamente utilizado pela comunidade científica para realizar esta tarefa. Para cada base lida pelo programa é associado um valor de qualidade. Este é calculado a partir da probabilidade da base estar incorreta, sendo que quanto maior a qualidade menor a probabilidade de erro. Assim, se juntamente com o arquivo de seqüência é fornecido também o arquivo de qualidade, pode-se remover trechos - ou seqüências inteiras - da análise segundo critérios mínimos de qualidade definidos pelo pesquisador. Na prática, a maioria das ESTs armazenadas em repositórios públicos não possui valores de qualidade para as bases.

1.5 Análise Estatística de Agrupamento

No contexto da Estatística, a análise de agrupamento (“clustering”) é um processo de divisão de um conjunto de objetos em subconjuntos menores, que são uniformes sob um determinado aspecto. O número de subconjuntos pode ou não ser conhecido no início do processo (Ptitsyn, 2000). Para formalizar o problema da análise de agrupamento, os objetos são representados como pontos em um espaço correspondente. Nesse espaço, os objetos que pertencem a um grupo estão situados a uma distância relativamente pequena entre si. A questão fundamental da análise de agrupamento é a escolha da métrica de similaridade ou dissimilaridade entre os objetos, a qual define o espaço.

Diferentemente de outros procedimentos estatísticos, os métodos de análise de agrupamento, na sua maioria, são utilizados quando não existe uma hipótese a priori, mas a pesquisa está na fase exploratória. De certa forma, a análise de agrupamento busca encontrar a solução mais significante possível. Por conseguinte, outro ponto crucial na análise de agrupamento é a escolha de uma medida de qualidade do agrupamento. Essa métrica é otimizada durante o procedimento de agrupamento e freqüentemente é chamada de regra de ligação. Na

maioria das vezes essa métrica é construída a partir de alguma medida de distância entre os grupos formados pelo processo de agrupamento. Se a medida é selecionada corretamente para os dados, então quanto mais distante os grupos resultantes estão um do outro, mais satisfatória será a solução.

O agrupamento de ESTs visto sob a ótica da bioinformática não é exatamente a análise de grupos clássica como entendida pelos estatísticos. Normalmente, o problema de agrupamento de ESTs não é expresso nos termos estatísticos corretos, definindo os objetos, o espaço, as métricas para as distâncias entre objetos e entre grupos e a medida de qualidade de grupo. Os algoritmos atualmente utilizados para ESTs são heurísticos. Assim, do ponto de vista estatístico, o agrupamento de ESTs utiliza uma versão simplificada da análise de agrupamento.

1.6 Comparação de Seqüências Biológicas

A comparação de seqüências biológicas (DNA, RNA ou seqüências de aminoácidos) se tornou essencial na biologia molecular. São vários os motivos para comparar seqüências: encontrar uma seqüência em um banco de dados; deduzir relações evolutivas entre seqüências; inferir função e estrutura. No agrupamento de ESTs, mesmo que informações extrínsecas à seqüência também sejam utilizadas, como é caso do STACK, que separa as ESTs por tecido antes de agrupá-las, a cadeia de nucleotídeos é a informação primordial para a realização desta tarefa. Portanto, no agrupamento de ESTs a comparação das seqüências é o passo mais importante, bem como o mais crítico, por consumir recursos computacionais que podem ser proibitivos frente ao grande volume de dados atual. Os métodos de comparação se dividem em duas classes principais: aqueles que utilizam alinhamento de seqüências e aqueles que não utilizam alinhamento de seqüências (Vinga e Almeida, 2003).

1.6.1 Métodos baseados em alinhamento de seqüência

Os métodos de comparação de seqüências biológicas baseados em alinhamento associam pontos para inserções, deleções e substituições e computam um alinhamento entre duas seqüências que corresponde ao conjunto das mutações com o menor custo. Tal alinhamento pode ser visto como minimizando a distância evolucionária ou maximizando a similaridade entre as duas seqüências comparadas. Em ambos os casos, o custo do alinhamento é uma medida de similaridade. Baseado na pontuação o algoritmo garante que o alinhamento é ótimo.

Os algoritmos que produzem alinhamentos ótimos utilizam programação dinâmica e são de dois tipos, global e local. Na literatura biológica, o alinhamento global freqüentemente é mencionado como alinhamento Needleman-Wunsch (Needleman e Wunsch, 1970) por terem sido os autores os primeiros a discutir similaridade global entre seqüências. Analogamente, o alinhamento local freqüentemente é mencionado como alinhamento Smith-Waterman (Smith e Waterman, 1981). Algoritmos de similaridade global otimizam o alinhamento geral entre duas seqüências, ou seja, alinha as seqüências em toda a sua extensão, podendo incluir longos trechos de baixa similaridade. Algoritmos de similaridade local buscam somente subseqüências relativamente conservadas, e uma única comparação pode produzir vários alinhamentos (subseqüências) distintos. Embora estas soluções algorítmicas sejam satisfatórias, consomem grandes recursos computacionais, tornando-as inapropriadas para determinados problemas, como, por exemplo, pesquisas em grandes bancos de dados.

Depois dos algoritmos baseados em programação dinâmica, surgiram algumas abordagens heurísticas baseadas no reconhecimento de “sementes” de alinhamentos, sendo as principais o FASTA (Pearson e Lipman, 1988) e o BLAST(Altschul et al., 1997). Estas abordagens buscam produzir resultados próximos aos que seriam obtidos com algoritmos que usam programação dinâmica. De maneira geral, primeiro identificam palavras curtas (“sementes”) comuns entre as

seqüências e depois estendem alinhamentos a partir delas. A grande vantagem do BLAST é a busca rápida e acurada contra um grande número de seqüências. Logo depois do seu surgimento, o BLAST se tornou o mecanismo de pesquisa dominante para bancos de dados de seqüências biológicas. As razões para tanto foram a velocidade, o fato de apresentar mais de uma solução na saída do programa, e o fato de fornecer uma estimativa da significância estatística de cada resultado.

1.6.2 Métodos não baseados em alinhamento de seqüência

A necessidade de velocidades cada vez maiores levou ao desenvolvimento de outros métodos para comparação de seqüências biológicas, muitos deles utilizando estratégias alternativas em detrimento do alinhamento de seqüências. Entre esses métodos as estratégias mais utilizadas são aquelas baseadas em tabela hash, freqüências de palavras e árvores de sufixos.

O algoritmo de pesquisa de seqüências em bancos de dados SSAHA ("Sequence Search and Alignment by Hashing Algorithm") (Ning et al., 2001), por exemplo, divide o banco de dados em *k*-tuplas de *k* bases contíguas e usa uma tabela hash para armazenar a posição de cada ocorrência de cada *k*-tupla. A pesquisa de uma seqüência no banco de dados é realizada obtendo da tabela hash uma lista de posições para cada *k*-tupla da seqüência consultada e ordenando os resultados. Outros métodos baseados em tabela hash foram descritos por Waterman (Waterman, 1995) e Miller e colaboradores (Miller et al., 1999).

O algoritmo de pesquisa d2 (Hide et al., 1994), baseado na distância d2 (Torney et al., 1990), compara seqüências de DNA usando a multiplicidade de palavras como uma medida simples de dissimilaridade. Hide e colaboradores (1994) investigaram a habilidade do algoritmo em detectar resultados biologicamente significantes entre uma seqüência e grandes conjuntos de seqüências de DNA variando parâmetros específicos como o comprimento da palavra ("word

length") e o tamanho da janela ("window size"). No exemplo particular discutido no trabalho (a pesquisa de lípases em um banco de dados genômico) um comprimento de palavra igual a 8 alcançou resultados similares aos obtidos com o FASTA. O algoritmo de pesquisa d2 também foi empregado com sucesso no agrupamento de ESTs (Hide et al., 1997; Burke et al., 1999).

A ferramenta ESTmapper (Wu et al., 2005) utiliza árvores de sufixos para alinhar ESTs ao genoma. O programa constrói uma árvore de sufixos especial ("WOTD - write-only, top-down") (Giergerih et al., 2003) para o genoma, que é utilizada para detectar segmentos comuns entre a EST e o genoma acima de um comprimento mínimo especificado pelo usuário. Esses segmentos são utilizados como ponto de partida para o algoritmo que computa o alinhamento entre a EST e o genoma. Árvores de sufixos e métodos derivados delas, como os arranjos de sufixos (Manber e Myers, 1993), também foram utilizados no agrupamento de ESTs (Kalyanaraman et al., 2003; Malde et al., 2003).

1.7 Ferramentas de Agrupamento de ESTs

Devido ao papel de destaque que as ESTs assumiram em vários tipos de pesquisa, a tarefa de agrupar ESTs ganhou enorme relevância, pois esta é primordial para que as ESTs possam ser utilizadas. Assim, as ferramentas de agrupamento são de fundamental importância para qualquer um que necessite trabalhar com ESTs. Dentre as várias ferramentas existentes, foram selecionadas para avaliação por este trabalho as cinco que serão descritas a seguir.

1.7.1 CAP3

O CAP3 (Huang e Madan, 1999) é a terceira geração do programa de montagem de seqüências de DNA CAP ("Contig Assembly Program") (Huang, 1992). O CAP3 realiza a detecção de sobreposições entre as seqüências, de forma a uni-las e criar regiões contíguas, os contigs. O programa aplica um filtro para eliminar pares de fragmentos que possivelmente não

teriam sobreposições. Um par de fragmentos passa pelo filtro somente se houver um alinhamento sem lacuna (“gap”) de comprimento 20 entre os fragmentos tal que o alinhamento contenha pelo menos 9 bases iguais consecutivas e no máximo duas bases diferentes. O programa examina cada ocorrência de uma palavra comum de comprimento 9 nos dois fragmentos para ver se esta pode ser estendida para um alinhamento sem lacuna de comprimento 20 com no máximo duas bases diferentes. Se o par passar pelo filtro então a sobreposição entre os fragmentos é computada por um algoritmo de programação dinâmica. Após isso, as seqüências são agrupadas para formar contigs em ordem decrescente de escore de sobreposição. Por último, um alinhamento múltiplo das seqüências é construído e uma seqüência consenso é gerada para cada contig. A despeito de ter sido projetado para montagem de seqüências de DNA genômico, o CAP3 é largamente utilizado como ferramenta de agrupamento de ESTs, seguindo o pressuposto de que as ESTs do mesmo transcrito formam um único contig.

1.7.2 d2_cluster

O d2_cluster é um método de agrupamento aglomerativo, ou seja, cada seqüência começa no seu próprio grupo e o agrupamento final é construído através de uma série de uniões (Burke et al., 1999). O d2_cluster utiliza fechamento transitivo, isso significa que duas seqüências A e B estarão no mesmo grupo mesmo que não compartilhem similaridade, desde que exista uma seqüência C que compartilhe similaridade com ambas as seqüências A e B. O único critério utilizado para agrupar as seqüências é a sobreposição das mesmas, outras informações, como anotação, não são usadas. Para detectar a sobreposição das seqüências o d2_cluster utiliza o algoritmo d2 (Hide et al., 1994). Esse algoritmo usa a distância d2 (Torney et al., 1990), que se baseia em freqüências de palavras. O d2_cluster integra o pacote de “software” stackPACK™, utilizado no projeto STACK.

1.7.3 ESTate

O sistema de agrupamento de ESTs ESTate foi desenvolvido por Guy Slater como parte de seu projeto de pesquisa de doutorado (Guy St.C.Slater, 2000). O sistema inclui uma série de ferramentas para análise de ESTs, duas das quais desenvolvidas especificamente para a tarefa de agrupar ESTs. O *precluster* é o programa que calcula a pontuação das palavras antes do agrupamento das ESTs, ele emprega Máquinas de Estado Finito Virtuais e um algoritmo de comparação de palavras eficiente para permitir a comparação todos-contra-todos de um grupo de seqüências em tempo sub-quadrático. As pontuações geradas são simplesmente o número de palavras iguais entre as seqüências. Os resultados são escritos em um formato binário compactado, e posteriormente podem ser utilizados pelo programa *estcluster* para gerar o agrupamento de ESTs. O *estcluster* utiliza teoria de grafos para permitir a geração de grupos de ESTs. As pontuações de palavras fornecidas pelo *precluster* são usadas na priorização dos alinhamentos para a detecção de sobreposições significativas.

1.7.4 TGICL

O TGICL ("TIGR Gene Indices clustering tools") (Pertea et al., 2003) é um protocolo para análise de ESTs no qual as seqüências primeiro são agrupadas com base em comparações par-a-par, e então montadas em grupos maiores para produzir seqüências consenso mais completas. O algoritmo é dividido em duas fases, sendo que a primeira consiste em um filtro global onde se realizam comparações de similaridade de seqüências usando o aplicativo mgblast, que é uma versão modificada do megablast (Zhang et al., 2000). Na segunda fase o CAP3 é utilizado para montar os grupos de seqüências construídos na fase anterior, gerando os grupos finais e seus respectivos consensos. O TGICL é usado na geração do TIGR Gene Indices.

1.7.5 XSACT

O XSACT (Malde et al., 2003) é uma ferramenta para agrupamento de ESTs que utiliza um algoritmo de complexidade sub-quadrática baseado em arranjos de sufixos. De maneira geral o algoritmo é divido em três partes. A primeira parte do algoritmo identifica os pares de seqüências com blocos comuns, de tamanho *k*. A segunda parte usa a informação gerada a partir deste processo para calcular um escore para os pares de seqüências. Por fim, os escores são utilizados para agrupar as seqüências hierarquicamente.

1.7.6 Outras ferramentas de agrupamento de ESTs

Além das ferramentas descritas anteriormente, existem outras. A ferramenta PaCE (Kalyanaraman et al., 2003), por exemplo, enfocou a questão do desempenho como sendo um gargalo no agrupamento das ESTs, e foi projetada para executar em computadores paralelos. O Phrap, apesar de ser um programa desenvolvido para montagem de fragmentos de DNA genômico, também já foi utilizado para agrupar ESTs quando ainda não havia ferramentas específicas de agrupamento. A ferramenta miraEST (Chevreux et al., 2004) é um montador que provê várias funcionalidades, entre elas a detecção de SNPs enquanto agrupa as ESTs.

1.8 Avaliação de ferramentas de agrupamento de ESTs

Liang et al. (2000) avaliaram quatro programas de montagem de seqüência para determinar qual deles reproduz com maior fidelidade as seqüências de transcritos a partir de dados de ESTs. Foram escolhidos os programas CAP3, Phrap e duas versões do TIGR Assembler (Sutton et al., 1995), sendo uma versão otimizada para a montagem de ESTs, denominada TA-EST, e a outra modificada para a montagem de seqüências genômicas, denominada TIGR Assembler. O CAP3 superou as outras ferramentas, produzindo seqüências consenso de alta fidelidade e mantendo um alto nível de sensibilidade para membros de famílias gênicas sem

deixar de lidar efetivamente com erros de seqüenciamento. Baseado nesta análise o CAP3 foi selecionado como ferramenta de montagem do TIGR Gene Indices.

Wang e colaboradores (Wang et al., 2004) investigaram a estrutura do erro de agrupamento de ESTs, a relação entre critério (comprimento de sobreposição e percentual de identidade) e erro de agrupamento, e possíveis métodos de correção de erro. Considerando-se o resultado do agrupamento, os erros foram classificados em dois tipos, chamados de Tipo I e Tipo II, em analogia à terminologia introduzida por Burke et al. (1999) para testar hipóteses estatísticas. O erro Tipo I ocorre quando ESTs do mesmo gene, que deveriam estar no mesmo grupo, são colocadas em grupos diferentes, enquanto que o erro Tipo II ocorre quando ESTs de genes distintos, que deveriam estar em grupos diferentes, são colocadas no mesmo grupo. Duas bibliotecas de ESTs da planta *Aradidopsis thaliana* foram utilizadas. Uma contendo somente ESTs seqüenciadas a partir da extremidade 5' e outra contendo somente ESTs seqüenciadas a partir da extremidade 3'. Cada biblioteca foi agrupada com a ferramenta CAP3, variando-se os parâmetros *comprimento da sobreposição* (O = 25, 30, 35, 40, 45) e *percentual de identidade* (P = 75, 80, 85, 90, 95, 97.5). O *comprimento da sobreposição* não se mostrou sensível dentro da faixa O = 25-45 e por isso foi adotado O = 40 em todos os experimentos. A identificação e quantificação dos erros Tipo I e II foram realizadas a partir da comparação dos agrupamentos produzidos pelo CAP3 com um agrupamento de referência, considerado o perfil de grupos de genes verdadeiro. O agrupamento de referência foi construído utilizando-se o BLASTN para alinhar as ESTs ao genoma anotado da *Aradidopsis thaliana*. Enquanto o percentual de erro Tipo II foi menor que 1,5% para ambos os agrupamentos de ESTs 3' e 5', o erro Tipo I foi 10 vezes maior no agrupamento de ESTs 5' que no 3' (30% versus 3%). Verificou-se que um *percentual de identidade* $P \geq 95\%$ pode inflar o erro Tipo I em ambos os casos, e demonstrou-se que aproximadamente 80% do erro Tipo I ocorre em virtude do erro de sobreposição insuficiente em

ESTs presentes no agrupamento 5'. Foi proposta uma nova abordagem estatística para corrigir o erro de sobreposição insuficiente com o objetivo de prover estimativas mais acuradas do perfil de grupos de genes verdadeiro.

Dada a importância de se organizar as ESTs em grupos, avaliar a acurácia dos agrupamentos produzidos pelas ferramentas se torna essencial. Infelizmente, cada ferramenta quando disponibilizada é avaliada de maneira diferente, o que inviabiliza uma comparação direta entre as mesmas. Em alguns casos, até existe comparação com outras ferramentas congêneres, mas, normalmente, a comparação se limita a uma ou duas ferramentas, portanto, carece de abrangência. Essa heterogeneidade torna árdua a tarefa de escolher uma ferramenta com base em dados objetivos.

2 OBJETIVO

2.1 Geral

Este trabalho tem por objetivo geral realizar uma análise comparativa entre as ferramentas de agrupamento de ESTs citadas anteriormente (CAP3, d2_cluster, ESTate, TGICL e XSACT) utilizando uma abordagem de comparação baseada em um agrupamento de referência construído com o auxílio de um genoma.

2.2 Específicos

Este trabalho tem como objetivos específicos:

- avaliar o desempenho computacional das ferramentas;

- utilizar métricas que permitam avaliar a qualidade dos agrupamentos de ESTs produzidos pelas ferramentas;

- avaliar o perfil de distribuição dos casos incorretos;

- avaliar a dispersão dos grupos em relação ao agrupamento de referência;

- analisar grupos discrepantes.

3 MATERIAIS E MÉTODOS

3.1 Terminologia

Um agrupamento de ESTs, doravante denominado somente *agrupamento*, é formado pelo conjunto de grupos e singletons produzidos por uma ferramenta de agrupamento. Um *grupo* é formado por um conjunto de ESTs. Um *singleton* é um grupo que contém somente uma EST. Um *agrupamento de referência* é um agrupamento de ESTs construído com o auxílio de um genoma, que para efeito de comparação com os agrupamentos gerados pelas ferramentas de agrupamento será considerado o agrupamento correto.

3.2 Visão Geral da Metodologia do Trabalho

Um problema recorrente quando se compara a qualidade dos agrupamentos produzidos pelas ferramentas é a incapacidade de determinar qual deles é biologicamente mais correto. Este trabalho é facilitado quando o organismo em questão possui o genoma seqüenciado. Pelo fato de ser fonte de toda a informação genética do organismo, o genoma pode ser usado para determinar a origem de cada EST. Dessa maneira, agrupar ESTs passa a ser uma questão de colocar no mesmo grupo ESTs originadas do mesmo locus. Na prática, isso se faz alinhando as ESTs ao genoma. Assim, o genoma pode auxiliar na construção de um agrupamento que pode ser utilizado como referência para avaliar a qualidade dos agrupamentos produzidos pelas ferramentas. Neste trabalho, o agrupamento supracitado foi denominado agrupamento de referência e foi considerado a solução correta para o problema de agrupar ESTs. É importante ressaltar que, a despeito de contar com o auxílio do genoma, o agrupamento de referência pode não refletir a solução biológica correta, pois há situações em que detectar a verdadeira origem das ESTs não é possível, como no caso de genes duplicados recentemente. Mesmo assim, isso não impede que o agrupamento de referência seja utilizado como parâmetro de comparação.

De maneira geral, a metodologia do trabalho consistiu em construir um agrupamento de referência, compará-lo aos agrupamentos produzidos pelas ferramentas e calcular métricas a partir dessa comparação. A Figura 3 mostra uma representação esquemática da metodologia do trabalho, que pretende fornecer uma visão geral de cada etapa e como elas se relacionam, de modo a facilitar o entendimento de cada uma quando forem descritas detalhadamente nos tópicos mais adiante.

Na etapa de pré-processamento repetições e seqüências contaminantes presentes nas ESTs são mascaradas. Depois disso as ESTs são agrupadas com o auxílio de um genoma para produzir o agrupamento de referência. Em resumo, como já dito, a construção do agrupamento de referência consiste em pesquisar cada EST no genoma e colocar no mesmo grupo aquelas que alinham no mesmo locus genômico (ESTs que não produzem nenhum alinhamento são retiradas da análise). As ESTs presentes no agrupamento de referência são então agrupadas por cada ferramenta. A seguir, a correspondência entre o agrupamento de referência e o agrupamento de cada ferramenta é determinada, e, por fim, são calculadas as métricas de comparação entre agrupamentos.

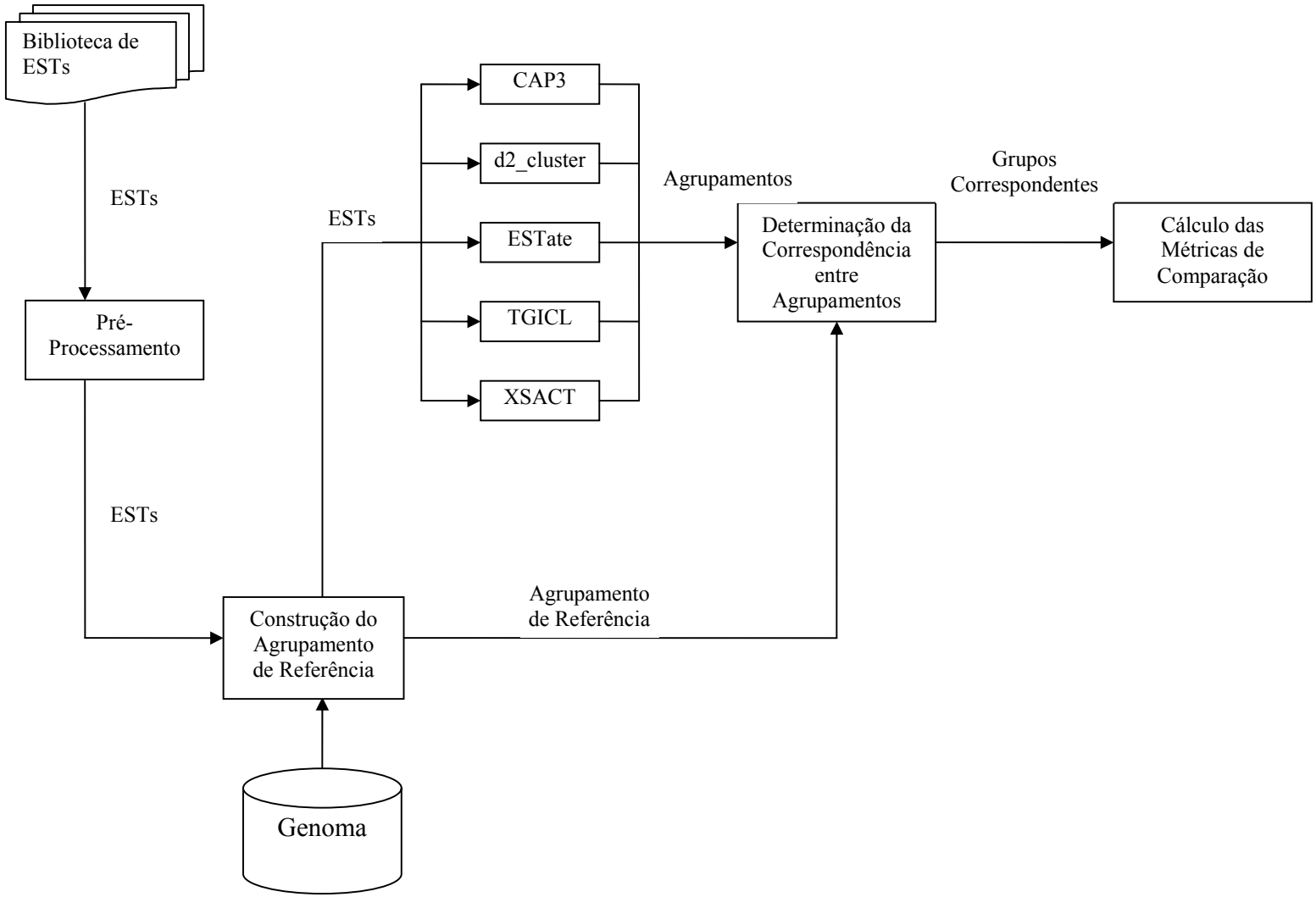

Figura 3 – Representação esquemática geral da metodologia do trabalho.

3.3 Bibliotecas de ESTs

Foram selecionadas três bibliotecas de ESTs de *Homo sapiens* no dbEST para os experimentos conduzidos neste trabalho. As bibliotecas foram identificadas pelo nome do tecido a partir do qual foi gerada a biblioteca de cDNA e pelo número aproximado de ESTs. A biblioteca FIGADO_10K é de fígado e possui 10.690 ESTs. A biblioteca CEREBRO_15K é de cérebro e possui 15.154 ESTs. A biblioteca CTRONCO_38K (Brandenberger et al., 2004) é de célula-tronco e possui 38.206 ESTs. Os identificadores das bibliotecas não possuem nenhuma relação com qualquer campo do dbEST e são significativos somente no contexto deste trabalho. As ESTs das bibliotecas podem ser recuperadas na página WEB do NCBI através dos endereços fornecidos na Tabela 1.

Tabela 1 - Endereços para a obtenção das ESTs das bibliotecas CTRONCO_38K, CEREBRO_15K e FIGADO_10K.

BIBLIOTECA	ENDEREÇO
CTRONCO_38K	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=nucest&cmd=search&term=GRN_PRENEU[Library%20Name]
CEREBRO_15K	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=nucest&cmd=search&term=NIH_MGC_56[Library%20Name]
FIGADO_10K	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=nucest&cmd=search&term="779%20(synonym:%20hncc1)"[Library%20Name]

3.4 Pré-processamento das bibliotecas de ESTs

Primeiramente as seqüências repetitivas de humanos (opção *-species* igual a *human*) foram mascaradas com o software RepeatMasker (Smit et al., 1996). O RepeatMasker utiliza como base de dados de seqüências repetitivas o Repbase (Jurka et al., 2005).

Em todos os mascaramentos posteriores foi utilizado o software cross_match (http://www.phrap.org). Tanto o RepeatMasker quanto o cross_match substituem as bases que devem ser mascaradas pela letra N. As seqüências de vetores de clonagem foram mascaradas a partir da base de dados do UniVec, que também possui seqüências de adaptadores, ligadores e iniciadores. O UniVec é mantido pelo NCBI e pode ser obtido através do endereço ftp://ftp.ncbi.nih.gov/pub/UniVec. Para as seqüências de tRNA foi utilizado O Banco de Dados de tRNA Genômico (http://lowelab.ucsc.edu/GtRNAdb/), as seqüências de tRNA humano utilizadas podem ser obtidas no endereço http://lowelab.ucsc.edu/GtRNAdb/Hsapi/Hg17-tRNAs.fa. As seqüências de mitocôndria foram mascaradas utilizando-se o banco de dados de seqüências mitocondriais disponível no NCBI, que pode ser obtido no endereço ftp://ftp.ncbi.nih.gov/blast/db/FASTA/mito.nt.gz. Para o mascaramento das seqüências de rRNA foram selecionados os seguintes GIs do Genbank: 36084, 174914, 337376 e 555853. Estas entradas contemplam todos os tipos de rRNA humanos.

Por fim, foi utilizado o software seqclean (http://www.tigr.org) para remover caudas poli A, extremidades ricas em Ns (bases não determinadas) e seqüências menores que 100 pares de bases.

3.5 Agrupamento de Referência

3.5.1 Seqüência do Genoma Humano

Neste trabalho foram utilizadas as seqüências do genoma humano disponíveis no RefSeq ("Reference Sequence") (Pruitt et al., 2005), que é um banco de dados de seqüências não-redundantes de genomas, transcritos e proteínas de vários organismos mantido pelo NCBI. Os arquivos do genoma utilizados neste trabalho são da versão liberada em 12/09/2004[3].

[3] ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/Assembled_chromosomes/

3.5.2 Fluxo de Processamento da Construção do Agrupamento de Referência

Para a produção do agrupamento de referência as ESTs foram alinhadas com o genoma e aquelas que produziram alinhamentos na mesma região, ou seja, apresentaram intersecção entre as posições dos alinhamentos no genoma, foram colocadas no mesmo grupo. Duas ferramentas foram utilizadas para determinar as posições das ESTs no genoma. Primeiramente as ESTs foram alinhadas utilizando-se o BLASTN (Altschul et al., 1990). Os alinhamentos locais produzidos pelo BLASTN não são os mais adequados para alinhar ESTs com o genoma, pois íntrons presentes no genoma que não foram transcritos em virtude de junções podem limitar o tamanho dos HSPs ("High Scoring Pairs") obtidos, prejudicando, dessa forma, a montagem dos grupos. Com o intuito de evitar este problema, foi utilizada a ferramenta est2genome (Mott, 1997), que produz alinhamentos entre seqüências com junções - ESTs e mRNAs - e seqüências de DNA genômico, inserindo os íntrons quando necessário. Somente as informações dos alinhamentos do est2genome foram utilizadas para agrupar as ESTs, mas nem por isso foi possível prescindir das informações dos alinhamentos do BLASTN. Como o tempo de processamento requerido pelo est2genome para alinhar cada EST ao genoma seria proibitivo, as ESTs foram alinhadas com uma pequena região do genoma recortada a partir das posições fornecidas pelos HSPs do BLASTN.

A Figura 4 apresenta uma representação esquemática do encadeamento dos scripts PERL utilizados na construção do agrupamento de referência. As elipses representam processos externos à construção do agrupamento de referência que fornecem ou recebem dados e os retângulos numerados simbolizam os scripts. Este fluxo de processamento foi executado para cada biblioteca. Segue abaixo a descrição do mesmo.

Script 1. Após passar pelo pré-processamento, as ESTs foram pesquisadas no genoma utilizando-se o BLASTN. Com exceção do e-value (opção –e), que foi configurado com o valor 10^{-10}, o BLASTN foi executado com os valores padrões para todas as opções.

Script 2. Foi realizado o processamento dos resultados do BLASTN e as ESTs que não produziram alinhamento foram retiradas da análise. Para as demais, foi selecionado o primeiro HSP do primeiro *hit* do resultado de cada uma, o restante dos HSPs foi desprezado.

Script 3. Foi recortada uma pequena região do genoma abrangendo 20 mil pares de bases antes do início e 20 mil depois do fim de cada HSP. Cada EST foi alinhada à respectiva região genômica utilizando-se o est2genome. O est2genome foi executado com os valores padrões para todas as opções.

Script 4. Foi realizado o processamento dos resultados do est2genome.

Script 5. As ESTs foram agrupadas utilizando-se um método aglomerativo onde cada uma começa o algoritmo no seu próprio grupo e o agrupamento final é produzido a partir de uma série de uniões dos grupos. O critério para a união dos grupos foi a intersecção entre as regiões do genoma onde as ESTs alinharam. Os grupos foram unidos empregando-se fechamento transitivo, ou seja, duas ESTs A e B que não possuem intersecção entre as regiões do genoma onde elas alinharam estarão no mesmo grupo se existir uma terceira EST que possui intersecção com as duas (A e B). Na verificação do critério de união dos grupos (intersecção) foram usados os limites extremos dos alinhamentos das ESTs, ou seja, se uma EST possui três éxons, as posições de início e fim consideradas serão respectivamente a posição inicial do primeiro éxon e a posição final do último. Assim, as ESTs serão agrupadas por gene, ou seja, transcritos de um mesmo gene que sofreram junção alternativa ficarão no mesmo grupo.

Script 6. As ESTs presentes no agrupamento de referência foram geradas para serem agrupadas pelas ferramentas.

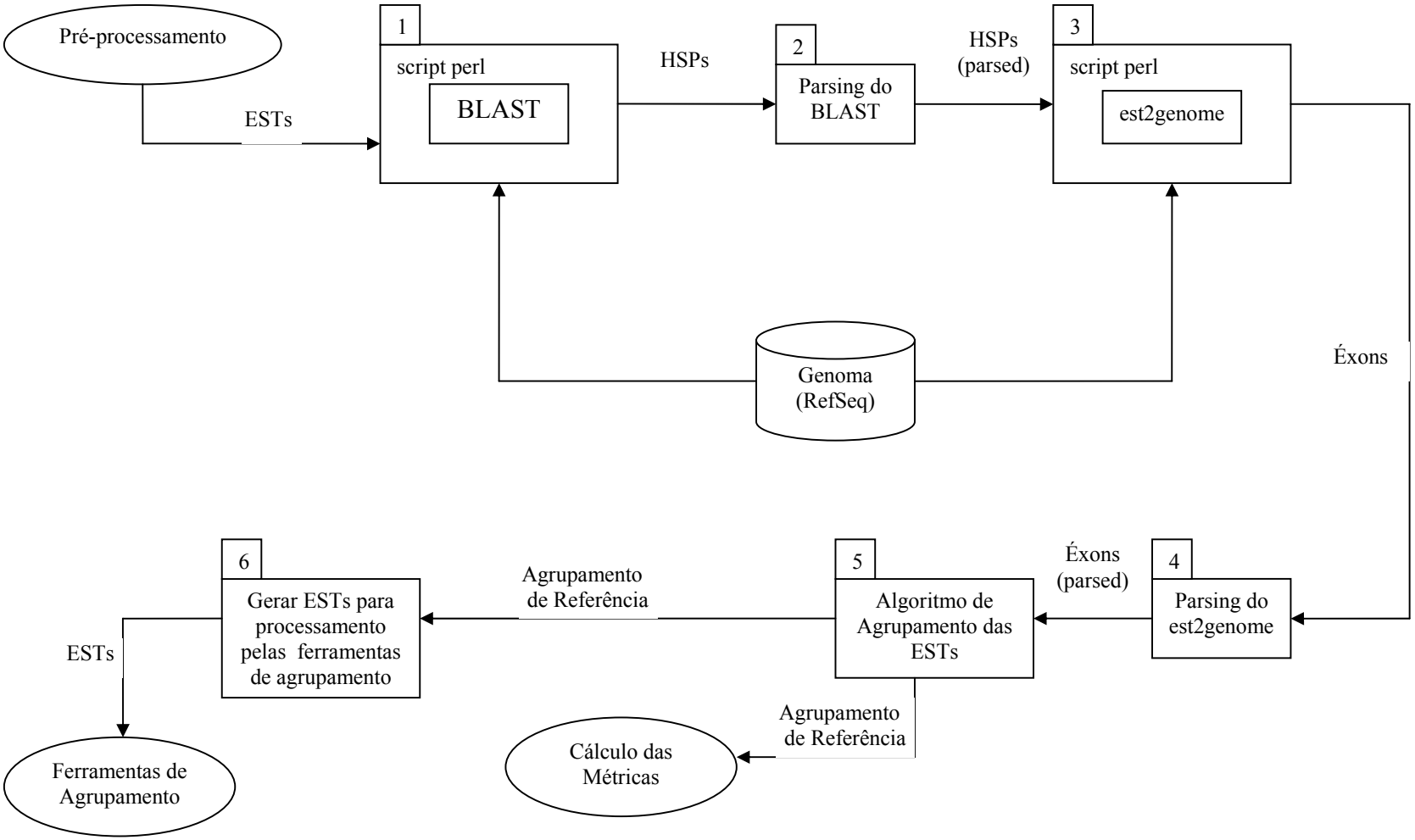

Figura 4 – Desenho esquemático representando o encadeamento dos programas utilizados na construção do agrupamento de referência.

3.5.3 Definição da ferramenta de comparação de seqüência

Além do BLASTN, a ferramenta ssahaEST também foi testada no fluxo de processamento do agrupamento de referência como alternativa para comparar as ESTs com o genoma. Esta ferramenta combina o algoritmo de pesquisa SSAHA (Ning et al., 2001) com a implementação do código do algoritmo de alinhamento de seqüência Smith-Waterman-Gotoh (Smith e Waterman, 1981; Gotoh, 1982) utilizado no programa cross_match. O programa foi executado com os valores padrões para todas as opções, com exceção da opção –best = 0, que é recomendada para ESTs, porém, não produziu resultados satisfatórios, como será descrito mais à frente.

3.6 Ferramentas de agrupamento de ESTs

A tabela abaixo lista as ferramentas de agrupamento avaliadas neste trabalho, fornecendo as referências e os endereços para obtenção das mesmas. Nos casos em que as ferramentas não estão disponíveis para download na Internet, são fornecidos endereços de e-mail que podem ser utilizados para solicitá-las.

Tabela 2 – Lista de ferramentas de agrupamento de ESTs avaliadas neste trabalho.

FERRAMENTAS	REFERÊNCIAS	ENDEREÇOS
CAP3	(Huang e Madan, 1999)	http://seq.cs.iastate.edu
d2_cluster	(Burke et al., 1999)	bsmalley@uh.edu
ESTate	(Guy St.C.Slater, 2000)	http://www.ebi.ac.uk/~guy/estate/
TGICL	(Pertea et al., 2003)	http://www.tigr.org/tdb/tgi/software/
XSACT	(Malde et al., 2003)	http://www.ii.uib.no/~ketil/bioinformatics/downloads/index.html

3.7 Execução e processamento dos resultados das ferramentas de agrupamento de ESTs

Cada ferramenta gera os seus dados de agrupamento em um formato específico. Para evitar que os programas que recebem agrupamentos como entrada tivessem que lidar com vários

formatos diferentes, foram escritos scripts na linguagem PERL para executar e converter os dados de agrupamento de cada ferramenta para um formato padrão.

As ferramentas CAP3, ESTate e TGICL foram executadas com os valores padrões para todas as opções. O d2_cluster foi executado com a seguinte configuração: <word size> = 6, <similarity cutoff> = 0.96, <min. sequence size> = 50, <window size> = 100 e <REV_COMP> = 1. O XSACT foi executado com a seguinte configuração: -k 24, -n 64 e –p 2.

3.8 Determinação da correspondência entre agrupamentos

A determinação da correspondência entre agrupamentos é fundamental para compará-los, pois é utilizada para calcular as métricas. A Figura 5 apresenta uma representação esquemática da correspondência entre dois agrupamentos **A** e **B**. Cada círculo representa um grupo. Os grupos de cada agrupamento estão separados por uma linha vertical tracejada e identificados por uma letra minúscula (que indica a qual agrupamento pertence o grupo) e um número. Assim, o primeiro grupo do agrupamento **A** é denominado **a1**, o primeiro grupo do agrupamento **B** é denominado **b1**, e assim por diante. Os números inteiros no interior dos círculos representam as ESTs.

Primeiramente, é preciso definir o termo *ESTs concordantes*. Em um par de grupos, são aquelas que são comuns aos dois grupos. Na Figura 5 as linhas arqueadas ligam as ESTs concordantes entre os grupos. Dessa forma, o par de grupos **a1-b1** possui cinco ESTs concordantes, o par **a2-b2** possui duas e o par **a2-b3** possui três. Podemos passar então à definição de *grupo correspondente*, tomando como exemplo o grupo **a2**. O grupo correspondente de **a2** no agrupamento **B** será aquele que apresentar o maior número de ESTs concordantes. Na figura, o grupo **a2** possui ESTs concordantes com os grupos **b2** e **b3**. Como **b3** apresenta o maior número de ESTs concordantes (três), este é considerado o grupo correspondente de **a2** no

agrupamento **B**. No caso de grupos com o mesmo número de ESTs concordantes a escolha é aleatória.

O par de grupos correspondentes **a1-b1** é um caso especial de correspondência. Conforme mostra a figura, todas as ESTs de **a1** são iguais às ESTs de **b1**, e vice-versa. Assim, quando os grupos são iguais, dizemos que os grupos *concordam perfeitamente*.

Uma vez introduzida a terminologia, determinar a correspondência entre dois agrupamentos **A** e **B** consiste em encontrar para cada grupo do agrupamento **A** o seu grupo correspondente no agrupamento **B**, produzindo, ao final, uma lista com os pares de grupos que se correspondem.

Alguns detalhes devem ser destacados. O primeiro é que, a não ser que os agrupamentos sejam iguais, ocorrerá o caso de alguns grupos não possuírem correspondentes no outro agrupamento. O segundo é que, por se tratar de uma comparação assimétrica, a ordem dos agrupamentos altera os resultados. Isso significa que **A** comparado a **B** é diferente de **B** comparado a **A**. Em virtude disso, em todas as comparações buscou-se os grupos correspondentes do agrupamento de referência no agrupamento da ferramenta, e nunca o inverso.

A correspondência entre o agrupamento de referência e os agrupamentos das ferramentas foi determinada para as três bibliotecas, gerando um total de 15 listas de grupos correspondentes. As listas foram utilizadas no cálculo das métricas explicadas a seguir.

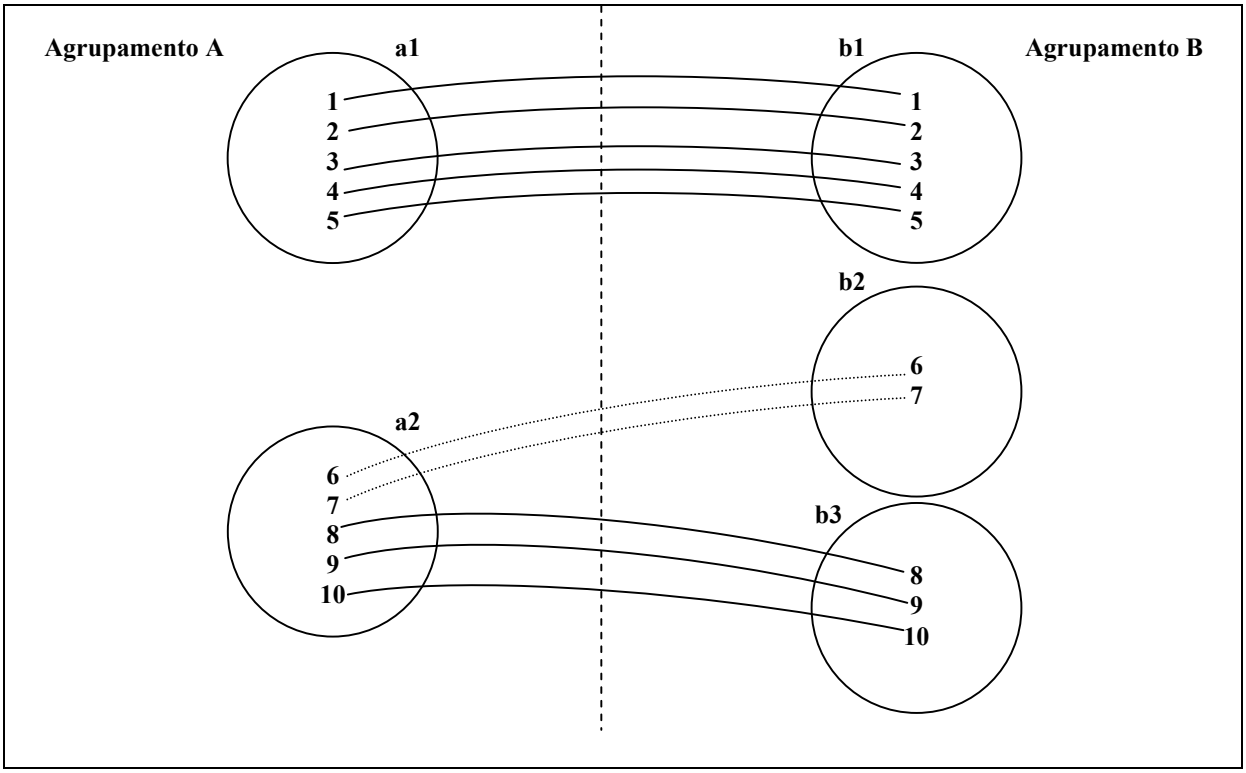

Figura 5 – Representação esquemática da correspondência entre agrupamentos.

3.9 Métricas de comparação entre agrupamentos

3.9.1 Média e Desvio Padrão do Coeficiente de Jaccard por Grupo

O coeficiente de Jaccard (Jaccard, 1908) é uma medida de similaridade de informação assimétrica. Este pode ser utilizado para medir a similaridade entre dois objetos. O coeficiente é dado pela fórmula $\mathbf{S_{ij} = p/(p + q + r)}$, onde **p** é o número de variáveis que aparece nos dois objetos, **q** é o número de variáveis que aparece somente no objeto ***i***, e **r** é o número de variáveis que aparece somente no objeto ***j***. Neste trabalho o coeficiente de Jaccard foi utilizado para medir a similaridade de grupos, sendo estes os objetos e as ESTs de cada um as variáveis.

Como exemplo, calculemos o coeficiente de Jaccard $\mathbf{S_{a2b3}}$ para o par de grupos correspondentes **a2-b3** da Figura 5. O número de ESTs iguais nos dois grupos, **p**, é igual a 3, o número de ESTs que aparece somente no grupo **a2**, **q**, é igual a 2, e o número de ESTs que

aparece somente no grupo **b3**, **r**, é igual a 0. Então, a similaridade entre os grupos **a2** e **b3** dada pelo coeficiente de Jaccard é $\mathbf{S_{a2b3}}$ = 3/(3 + 2 + 0) = **0,6**.

O cálculo da métrica para cada comparação agrupamento de referência versus agrupamento da ferramenta foi feito da seguinte forma. Dada uma lista de correspondência entre agrupamentos, calculou-se o coeficiente de Jaccard, conforme descrito anteriormente, para cada par de grupos. Para os grupos do agrupamento de referência que não possuíam grupos correspondentes atribuiu-se valor zero ao coeficiente. A média e o desvio padrão dos coeficientes foram então calculados.

3.9.2 Média e Desvio Padrão do Percentual de ESTs Concordantes por Grupo

O percentual de ESTs concordantes de um par de grupos é dado pela fórmula **P = E/T*100**, onde **E** é número de ESTs concordantes e **T** é o tamanho do grupo do agrupamento de referência que participa do par. Como exemplo, calculemos o percentual de ESTs concordantes do par de grupos **a2-b3**. Assumindo o agrupamento **A** como agrupamento de referência, **E** é igual a 3 e **T** é igual a 5. Então, o percentual entre os grupos é **P** = 3/5*100 = **60%**.

O cálculo da métrica para cada comparação agrupamento de referência versus agrupamento da ferramenta foi feito da seguinte forma. Dada uma lista de correspondência entre agrupamentos, calculou-se para cada par o percentual de ESTs concordantes, conforme descrito anteriormente. Para grupos do agrupamento de referência que não possuíam grupos correspondentes, atribuiu-se valor zero ao percentual. Após calculados os percentuais para todos os pares, foram calculados a média e o desvio padrão.

3.9.3 Percentual de Concordância Perfeita

O percentual de concordância perfeita total é dado pela fórmula **CP = P/T*100**, onde **P** é o número de pares de grupos que concordam perfeitamente e **T** é o número de grupos do

agrupamento de referência. Como exemplo, calculemos o percentual de concordância perfeita total entre o agrupamento **B** e **A** da Figura 5. Assumindo o agrupamento **B** como agrupamento de referência, **P** é igual a 1 (somente o par **b1-a1** concorda perfeitamente) e **T** é igual a 3. Então, o percentual de concordância perfeita total entre os agrupamentos **B** e **A** é CP = 1/3*100 = **33,33%**.

A variável **P** pode ser dividida em grupos singletons e não-singletons, assim, temos $\mathbf{P = P_S + P_N}$, onde $\mathbf{P_S}$ é o número de pares de grupos singletons que concordam perfeitamente e $\mathbf{P_N}$ é o número de pares de grupos não-singletons que concordam perfeitamente. Dessa forma, o percentual de concordância perfeita também pode ser calculado separadamente para grupos singletons ($\mathbf{CP_S}$) e não-singletons ($\mathbf{CP_N}$). Assim, temos $\mathbf{CP_S = P_S/T*100}$ e $\mathbf{CP_N = P_N/T*100}$.

3.10 Anotação dos grupos do agrupamento de referência

O CAP3 foi utilizado para gerar as seqüências consenso dos grupos *00248*, *00910* e *00064* do agrupamento de referência. Nos casos em que o CAP3 gerou mais de um contig por grupo todos os contigs foram consultados.

Cada contig gerado pelo CAP3 foi pesquisado na página WEB 'BLAST Human Sequences'[4] do NCBI. Para todas as consultas foram feitas as seguintes alterações nos parâmetros: no parâmetro 'Database' foi escolhida a opção 'genome (reference only)'; no parâmetro 'Program' foi escolhida a opção 'BLASTN: Compare nucleotide sequences'; e no parâmetro 'Expect' foi escolhida a opção '0.0001'. Através do NCBI Map Viewer os dois primeiros resultados de cada contig foram investigados em busca da anotação no banco de dados Entrez Gene (Maglott et al., 2005).

[4] http://www.ncbi.nlm.nih.gov/genome/seq/BlastGen/BlastGen.cgi?taxid=9606

4 RESULTADOS

4.1 Pré-processamento das bibliotecas de ESTs

Na fase de pré-processamento das ESTs as seqüências repetitivas e contaminantes foram mascaradas, extremidades ricas em bases não determinadas foram removidas e ESTs com tamanho menor que 100 pares de bases foram retiradas da análise. A Tabela 3 mostra os tamanhos das bibliotecas e os percentuais de ESTs retiradas da análise após o processamento.

Tabela 3 – Tamanhos das bibliotecas e percentuais de ESTs retiradas da análise após o pré-processamento.

	BIBLIOTECAS		
	FIGADO_10K	**CEREBRO_15K**	**CTRONCO_38K**
Tamanho original	10690	15154	38206
Tamanho após pré-processamento	9139	14687	36219
Percentual de ESTs retiradas da análise	14,50%	3,08%	5,20%

4.2 Construção do Agrupamento de Referência

Após a fase de pré-processamento das bibliotecas de ESTs foi construído o agrupamento de referência para cada uma delas. Conforme mencionado anteriormente na seção de métodos, o agrupamento de referência consiste no agrupamento de ESTs que será considerado como a solução correta para efeito de comparação com os agrupamentos produzidos pelas ferramentas. O fluxo de processamento descrito no tópico 3.5.2 (página 27) foi executado para cada biblioteca. As ESTs que não produziram alinhamentos (BLASTN) no genoma foram retiradas da análise. A Tabela 4 apresenta os tamanhos e percentuais de ESTs retiradas da análise após o fluxo de processamento do agrupamento de referência. As bibliotecas FIGADO_10K e CTRONCO_38K tiveram menos de 1% das suas ESTs retiradas da análise, ao passo que a biblioteca CEREBRO_15K apresentou um percentual bem maior, de 13,35%.

Tabela 4 – Tamanhos das bibliotecas e percentuais de ESTs retiradas da análise após o fluxo de processamento do agrupamento de referência.

	BIBLIOTECAS		
	FIGADO_10K	**CEREBRO_15K**	**CTRONCO_38K**
Tamanho após pré-processamento	9.139	14.687	36.219
Tamanho após fluxo de processamento do agrupamento de referência	9.119	12.727	36.164
Percentual de ESTs retiradas da análise	0,22%	13,35%	0,15%

4.2.1 Definição da ferramenta de comparação de seqüência

Conforme visto no tópico 3.5.3, o BLASTN foi utilizado para fazer um alinhamento prévio de cada EST com o genoma antes de realizar o alinhamento com o est2genome. O uso das informações do alinhamento do BLASTN como semente para recortar a região genômica fornecida ao est2genome buscou reduzir o tempo gasto por este último.

A ferramenta ssahaEST também foi testada no fluxo de processamento do agrupamento de referência como alternativa para comparar as ESTs com o genoma. As ESTs da biblioteca CTRONCO_38K, após a fase de pré-processamento, foram pesquisadas no genoma utilizando a ferramenta ssahaEST, que foi executada conforme descrito no tópico 3.5.3. Das 36.219 ESTs pesquisadas com o ssahaEST, somente 15.279 produziram alinhamentos, ao passo que com o BLASTN 36.164 ESTs produziram alinhamentos. Apesar ser mais rápido, o ssahaEST se mostrou muito rigoroso, penalizando um número muito grande de EST. Por isso, optou-se pelo BLASTN, que se mostrou mais permissivo, e, portanto, mais adequado aos dados de ESTs, que possuem um ruído inerente.

4.3 Execução das Ferramentas de Agrupamento de ESTs

Após a construção do agrupamento de referência, as bibliotecas de ESTs foram agrupadas por cada ferramenta conforme os parâmetros descritos no tópico 3.7. As quantidades de grupos

produzidas por cada ferramenta para cada biblioteca são apresentadas na Tabela 5. Em geral as ferramentas produziram quantidades de grupos compatíveis com as quantidades de grupos produzidas no agrupamento de referência, com exceção das ferramentas CAP3 e TGICL, que na biblioteca CEREBRO_15K apresentaram quantidades de grupos bastante superiores em virtude de uma super estimativa do número de singletons. Os dados da Tabela 5 também são apresentados em forma de gráfico nas três figuras a seguir.

Tabela 5 - Quantidades de grupos produzidas por cada ferramenta para as três bibliotecas.

Biblioteca		**Quantidade de Grupos**					
		Agrup. Ref.	**CAP3**	**d2_cluster**	**ESTate**	**TGICL**	**XSACT**
CTRONCO_38K	**Singletons**	10381	10414	10478	11593	10952	10336
	Não-Singletons	4882	4538	4415	4144	4274	4414
	Total de Grupos	**15263**	**14952**	**14893**	**15737**	**15226**	**14750**
CEREBRO_15K	**Singletons**	2411	5497	2855	2555	5595	2686
	Não-Singletons	1513	1127	1288	1325	721	1313
	Total de Grupos	**3924**	**6624**	**4143**	**3880**	**6316**	**3999**
FIGADO_10K	**Singletons**	5419	5512	5591	5728	5650	5536
	Não-Singletons	843	822	784	732	773	798
	Total de Grupos	**6262**	**6334**	**6375**	**6460**	**6423**	**6334**

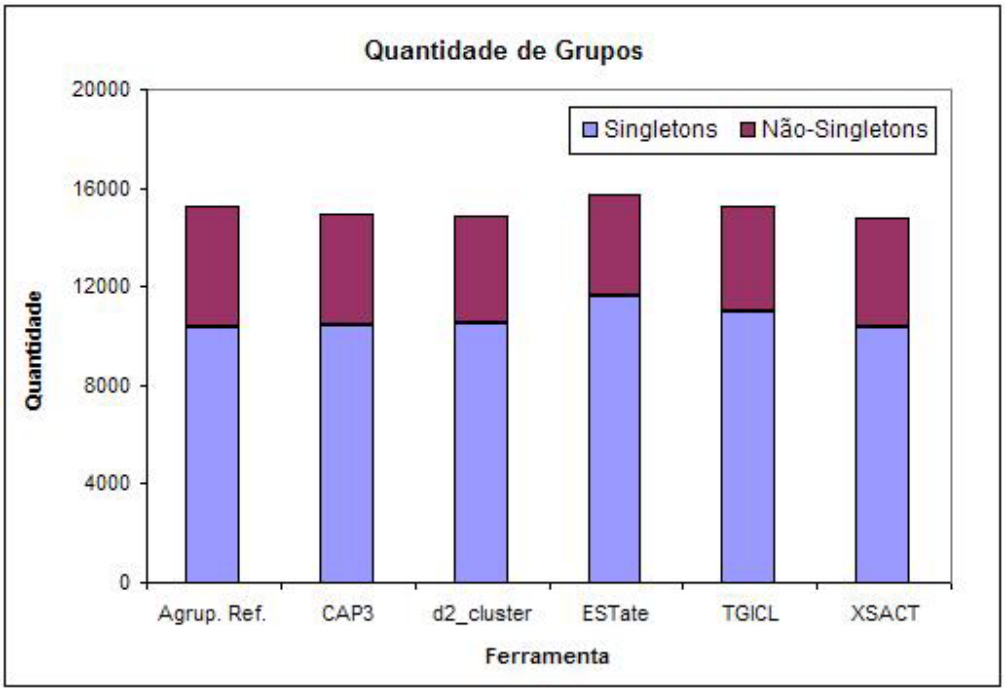

Figura 6 - Quantidade de Grupos por ferramenta para a biblioteca CTRONCO_38K.

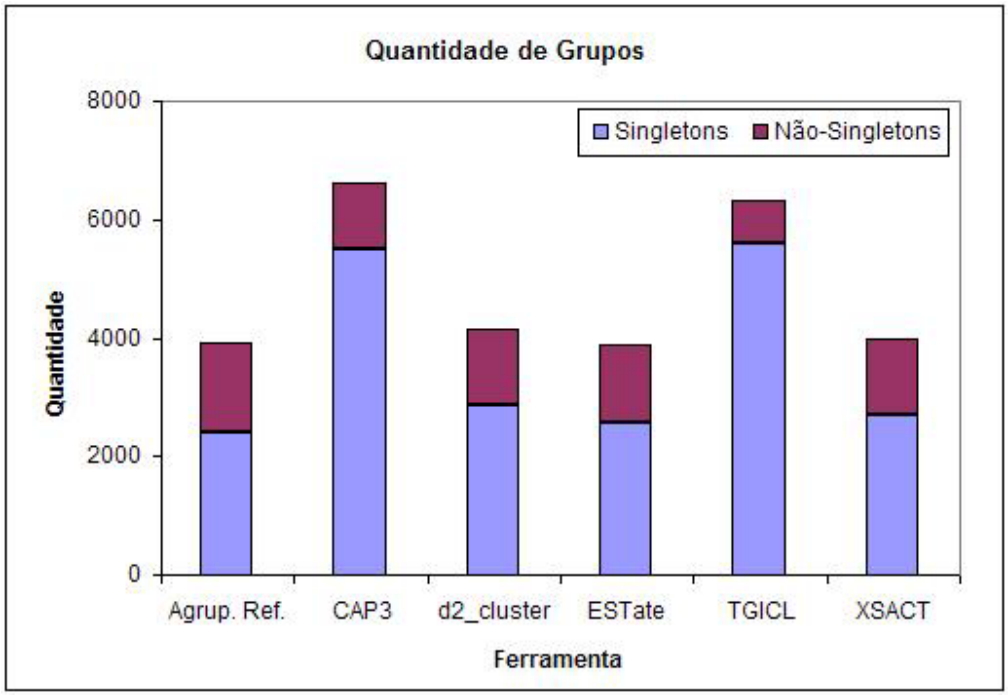

Figura 7 - Quantidade de Grupos por ferramenta para a biblioteca CEREBRO_15K.

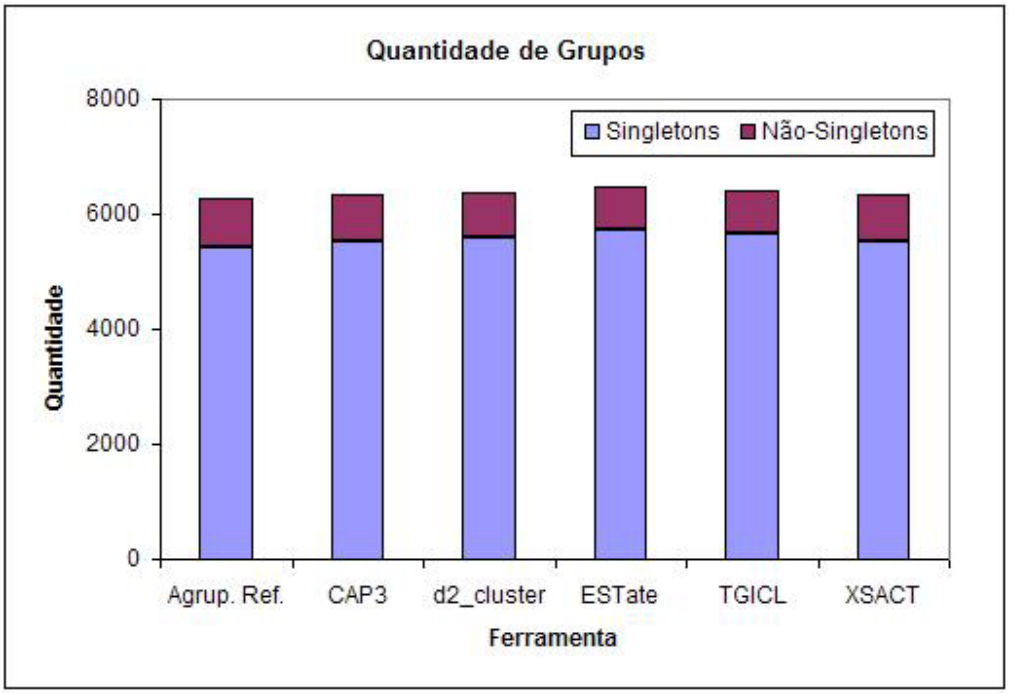

Figura 8 - Quantidade de Grupos por ferramenta para a biblioteca FIGADO_10K.

4.3.1 Avaliação do desempenho das ferramentas de agrupamento de ESTs

Com o intuito de avaliar o desempenho das ferramentas, foi realizada uma execução separada para cada uma delas onde foram coletados os tempos de execução para a biblioteca CTRONCO_38K após a fase de pré-processamento (36.219 ESTs). As execuções foram realizadas em uma máquina DELL com 4 processadores e 8 gigabytes de memória RAM rodando sistema operacional linux. Todas as ferramentas foram executadas com os parâmetros descritos no tópico 3.7. A Tabela 6 apresenta os tempos de execução por ferramenta para a biblioteca CTRONCO_38K.

Tabela 6 – Tempos de execução por ferramenta para a biblioteca CTRONCO_38K.

.	FERRAMENTAS				
	CAP3	**d2_cluster**	**ESTate**	**TGICL**	**XSACT**
Tempo de execução	01:06:01	02:43:17	00:33:21	01:22:43	00:15:56

O XSACT apresentou um desempenho bastante superior ao das outras ferramentas, sendo 2 vezes mais rápido que a segunda melhor ferramenta, o ESTate, e 10 vezes mais rápido que a

pior ferramenta, o d2_cluster. Vale ressaltar que todas as ferramentas oferecem a possibilidade de alterar os parâmetros de execução de forma a melhorar o desempenho computacional, portanto, os números acima não são indicadores absolutos do desempenho de cada ferramenta, porém, é importante destacar que alterações nos parâmetros de execução que privilegiam o desempenho sacrificam a sensibilidade das ferramentas, impactando diretamente nos resultados.

4.4 Análise dos Agrupamentos de ESTs

4.4.1 Distribuição dos grupos por tamanho

O gráfico da Figura 9 mostra a quantidade de grupos por tamanho para o agrupamento de referência e para os agrupamentos das ferramentas em relação à biblioteca CTRONCO_38K. De maneira geral o resultado das ferramentas foi similar ao do agrupamento de referência, com exceção do ESTate e do TGICL, que apresentaram um número de singletons mais elevado que as demais ferramentas. ESTate e TGICL apresentaram, respectivamente, quantidades de singletons 11,67% e 5,5% maiores que o agrupamento de referência.

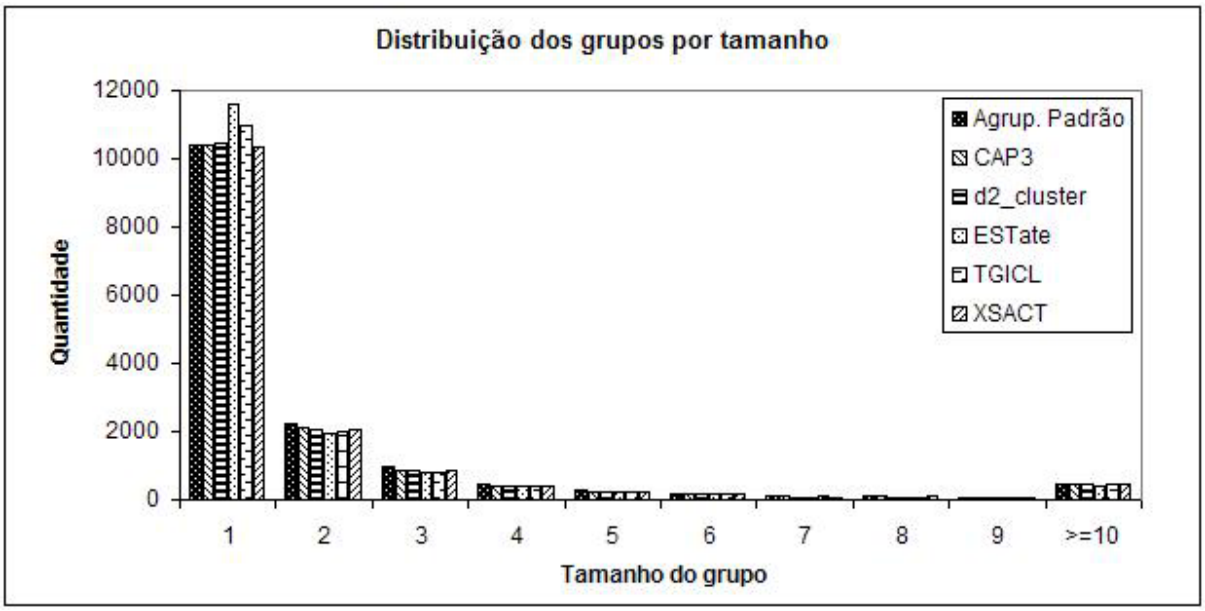

Figura 9 – Distribuição dos grupos por tamanho para a biblioteca CTRONCO_38K.

O gráfico da Figura 10 apresenta a distribuição dos grupos por tamanho para a biblioteca CEREBRO_15K. Nesta biblioteca o número de singletons apresentado pelas ferramentas divergiu bastante do número apresentado pelo agrupamento de referência. Enquanto que a ferramenta com o número de singletons mais próximo do agrupamento de referência foi a ferramenta ESTate, que apresentou um valor 5,97% maior, as duas ferramentas com os resultados mais discrepantes, TGICL e CAP3, apresentaram, respectivamente, valores 132,06% e 128% maiores que o do agrupamento de referência. As ferramentas d2_cluster e XSACT apresentaram, respectivamente, quantidades de singletons 18,42% e 11,41% maiores que o agrupamento de referência.

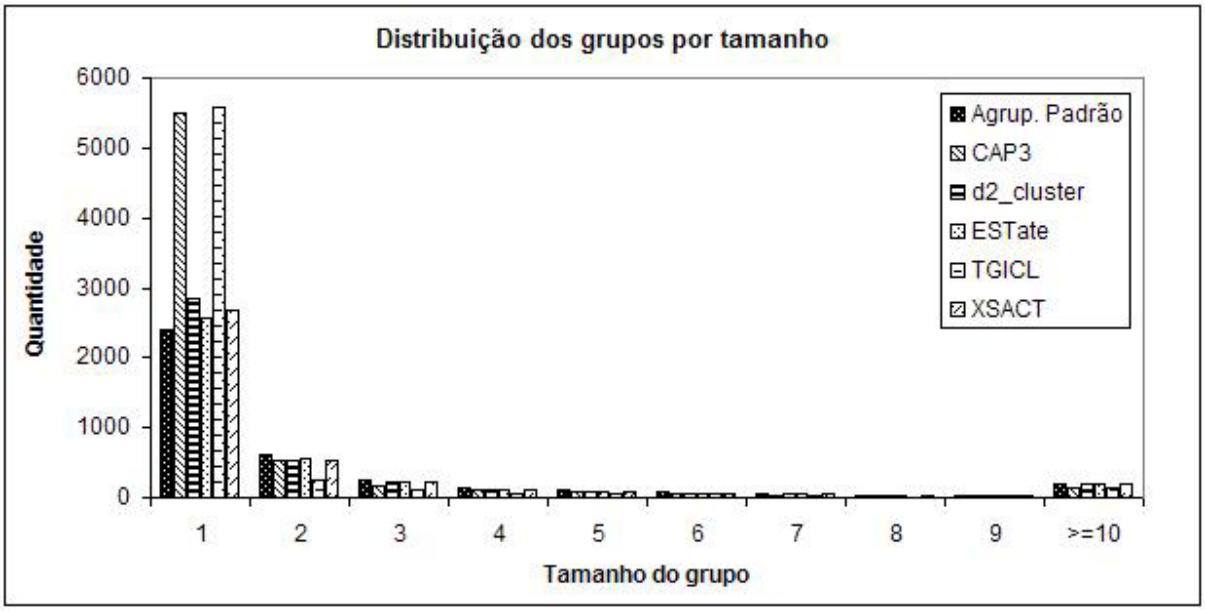

Figura 10 – Distribuição dos grupos por tamanho para a biblioteca CEREBRO_15K.

O gráfico da Figura 11 apresenta a distribuição dos grupos por tamanho para a biblioteca FIGADO_10K. De maneira geral, as ferramentas produziram resultados similares ao agrupamento de referência. Da mesma forma que nas outras duas bibliotecas, os números de singletons dos agrupamentos das ferramentas foi superior ao do agrupamento de referência. O número de singletons mais elevado foi produzido pelo ESTate, que superou o agrupamento de

referência em 5,7%. O CAP3 produziu o menor número de singletons, superando o agrupamento de referência em 1,72%.

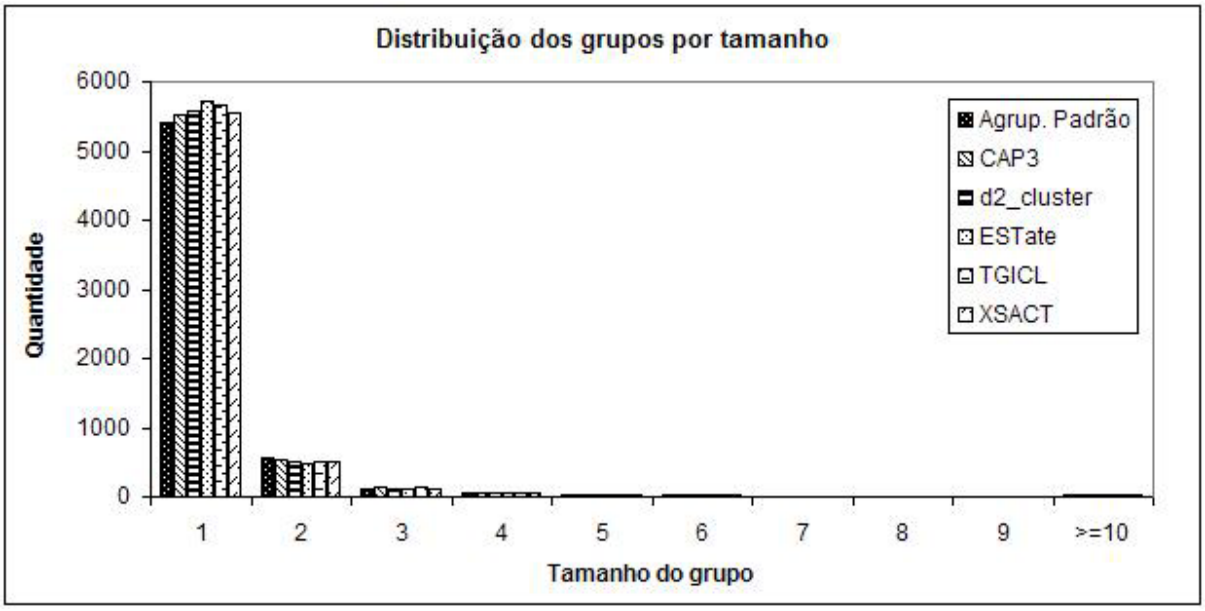

Figura 11 – Distribuição dos grupos por tamanho para a biblioteca FIGADO_10K.

4.4.2 Métricas de comparação entre agrupamentos

Com a finalidade de medir a consistência dos agrupamentos gerados pelas ferramentas em relação ao agrupamento de referência foram empregadas as métricas a seguir.

4.4.2.1 Média e Desvio Padrão do Coeficiente de Jaccard por Grupo

A média e o desvio padrão do coeficiente de Jaccard por grupo foram calculados como descrito no tópico 3.9.1. A Tabela 7 apresenta os valores por biblioteca e por ferramenta. A Figura 12 apresenta os dados da Tabela 7 em forma de gráfico, porém sem os desvios. Como demonstra o gráfico, as ferramentas mostraram desempenhos bastante semelhantes entre si nas bibliotecas CTRONCO_38K e FIGADO_10K, ficando as médias dos coeficientes entre 0,88 e 0,89 na primeira, e entre 0,96 e 0,97 na segunda. Na biblioteca CEREBRO_15K as ferramentas apresentaram um padrão um pouco diferente, com as ferramentas CAP3 e TGICL produzindo valores inferiores aos das outras ferramentas. Enquanto o CAP3 e o TGICL produziram

respectivamente os valores 0,84 e 0,82, as outras ferramentas produziram valores entre 0,87 e 0,88.

Tabela 7 – Média e Desvio Padrão do Coeficiente de Jaccard por Grupo para as três bibliotecas.

	CAP3		d2_cluster		ESTate		TGICL		XSACT	
CTRONCO_38K	0,89	±0,28	0,89	±0,28	0,88	±0,28	0,89	±0,27	0,89	±0,28
CEREBRO_15K	0,84	±0,27	0,88	±0,28	0,87	±0,30	0,82	±0,29	0,88	±0,28
FIGADO_10K	0,97	±0,14	0,97	±0,13	0,96	±0,15	0,97	±0,13	0,97	±0,14

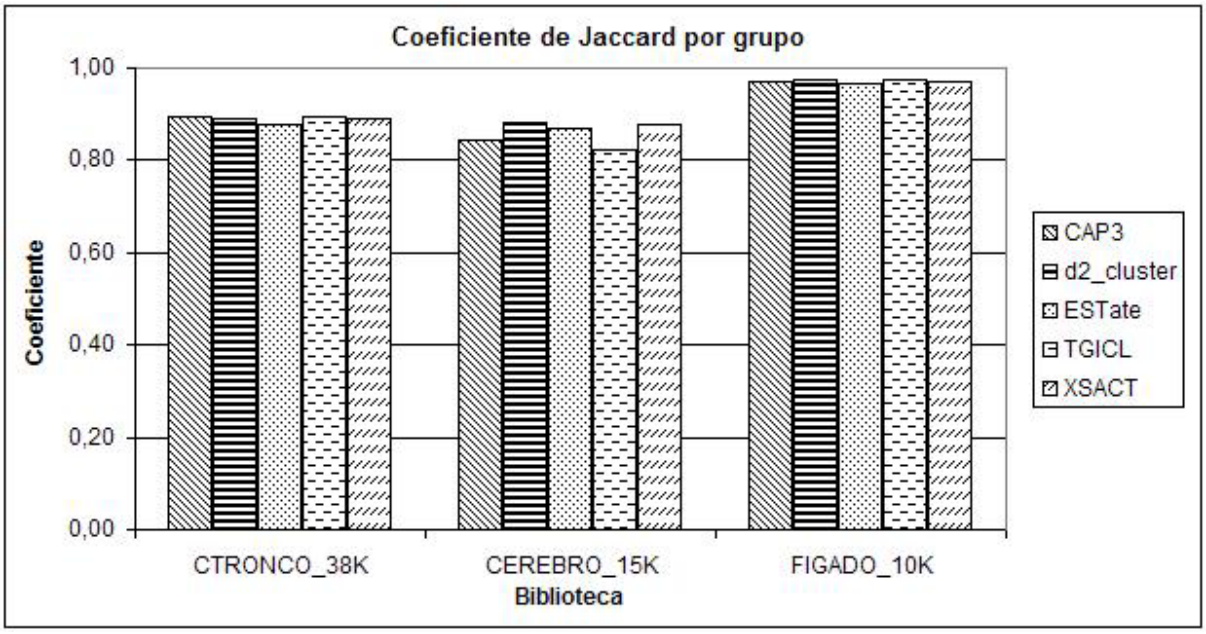

Figura 12 – Médias dos coeficientes de Jaccard por grupo para as três bibliotecas.

4.4.2.2 Média e Desvio Padrão do Percentual de ESTs Concordantes por Grupo

A média e o desvio padrão do percentual de ESTs concordantes por grupo é uma métrica menos rigorosa que a métrica anterior. A média do coeficiente de Jaccard por grupo abarca a similaridade e a diferença existente entre dois grupos, enquanto que esta só considera a similaridade entre os dois grupos em relação ao grupo do agrupamento de referência. A média e o desvio padrão do percentual de ESTs concordantes por grupo foram calculados conforme descrito no tópico 3.9.2. A Tabela 8 apresenta os percentuais por biblioteca e por ferramenta. A Figura 13

apresenta os dados da Tabela 8 em forma de gráfico. Como demonstrado no gráfico, as ferramentas praticamente repetiram o mesmo padrão de resultados da métrica anterior, com CAP3 e TGICL apresentando uma queda em relação às outras ferramentas na biblioteca CEREBRO_15K.

Tabela 8 – Média e Desvio Padrão do Percentual de ESTs Concordantes por Grupo para as três bibliotecas.

	CAP3		d2_cluster		ESTate		TGICL		XSACT	
CTRONCO_38K	91,16	±26,64	90,92	±26,95	89,40	±27,53	90,89	±26,51	90,81	±27,36
CEREBRO_15K	85,29	±26,60	89,60	±27,21	88,81	±29,31	82,67	±28,46	89,65	±27,73
FIGADO_10K	97,71	±13,34	97,77	±12,67	96,90	±14,62	97,70	±12,48	97,68	±13,33

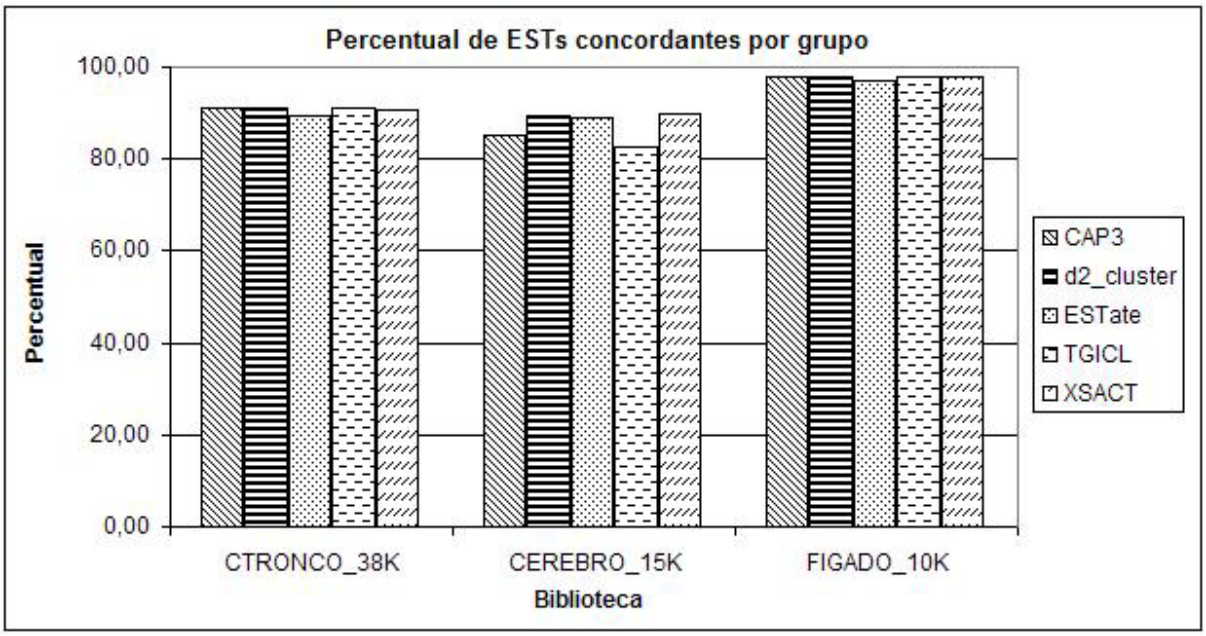

Figura 13 – Médias dos percentuais de ESTs concordantes por grupo para as três bibliotecas.

4.4.2.3 Percentual de Concordância Perfeita

De maneira simplificada, o percentual de concordância perfeita nada mais é do que o número de grupos que são exatamente iguais em dois agrupamentos em relação ao número total de grupos de um deles (neste caso, agrupamento de referência). A Tabela 9 mostra os percentuais de concordância perfeita total, por grupos singletons e por grupos não-singletons. Os gráficos das Figuras 14, 15 e 16 apresentam os dados da Tabela 9 em forma de barras empilhadas. Nas

bibliotecas CTRONCO_38K e FIGADO_10K as ferramentas apresentaram desempenhos semelhantes entre si, com exceção do ESTate que apresentou desempenho inferior. Na biblioteca CTRONCO_38K o percentual de concordância perfeita total do ESTate ficou em 79,85%, enquanto que no restante das ferramentas ficou entre 82,83% e 83,29%. Na biblioteca FIGADO_10K o percentual de concordância perfeita total do ESTate ficou em 93,69%, enquanto que no restante das ferramentas ficou entre 95,04% e 95,24%. Da mesma forma que nos resultados das métricas anteriores, na biblioteca CEREBRO_15K as ferramentas CAP3 e TGICL apresentaram desempenhos bem inferiores em comparação ao restante das ferramentas. CAP3 e TGICL produziram, respectivamente, os valores 70,13% e 66,62%, enquanto que o restante das ferramentas apresentou valores entre 78,64% e 79,46%.

Tabela 9 – Percentuais de Concordância Perfeita para as três bibliotecas.

BIBILOTECAS		**CAP3**	**d2_cluster**	**ESTate**	**TGICL**	**XSACT**
CTRONCO_38K	**Singlet.**	62,39	62,60	62,92	63,15	62,36
	Não-Singlet.	20,89	20,55	16,93	19,68	20,93
	C.P. Total	83,28	83,15	79,85	82,83	83,29
CEREBRO_15K	**Singlet.**	58,56	56,86	55,45	59,94	56,42
	Não-Singlet.	11,57	21,94	23,19	6,68	23,04
	C.P. Total	70,13	78,80	78,64	66,62	79,46
FIGADO_10K	**Singlet.**	84,67	85,05	84,80	85,20	84,75
	Não-Singlet.	10,54	10,19	84,80	9,84	10,46
	C.P. Total	95,12	95,24	93,69	95,04	95,21
MÉDIA C.P. TOTAL		82,84	85,73	84,06	81,50	85,99

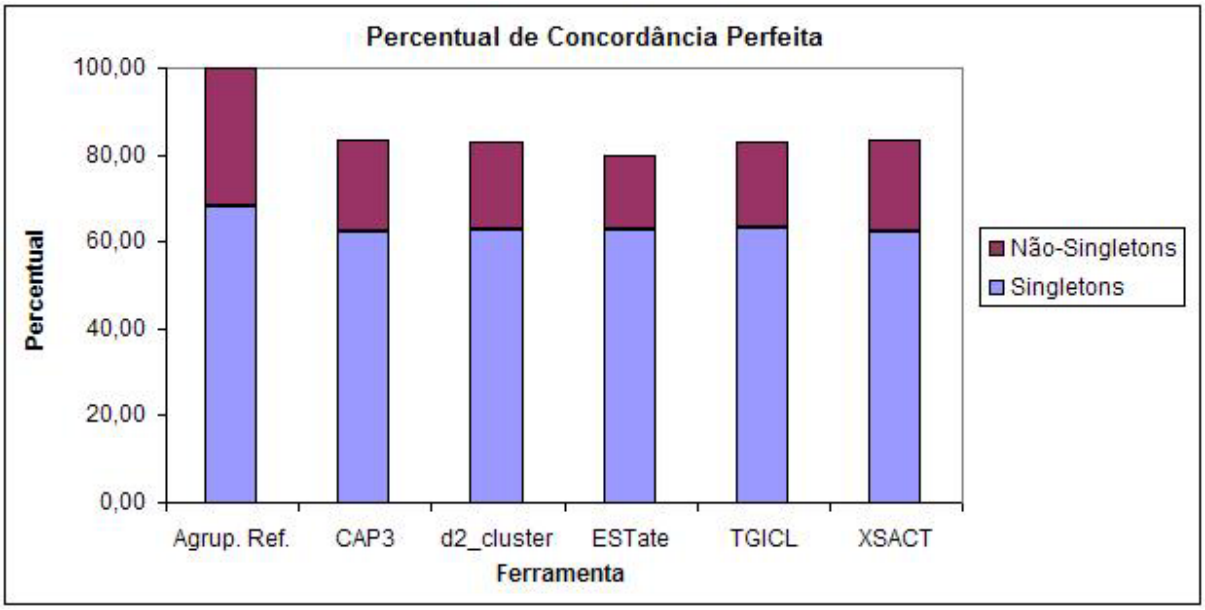

Figura 14 - Percentuais de Concordância Perfeita para a biblioteca CTRONCO_38K.

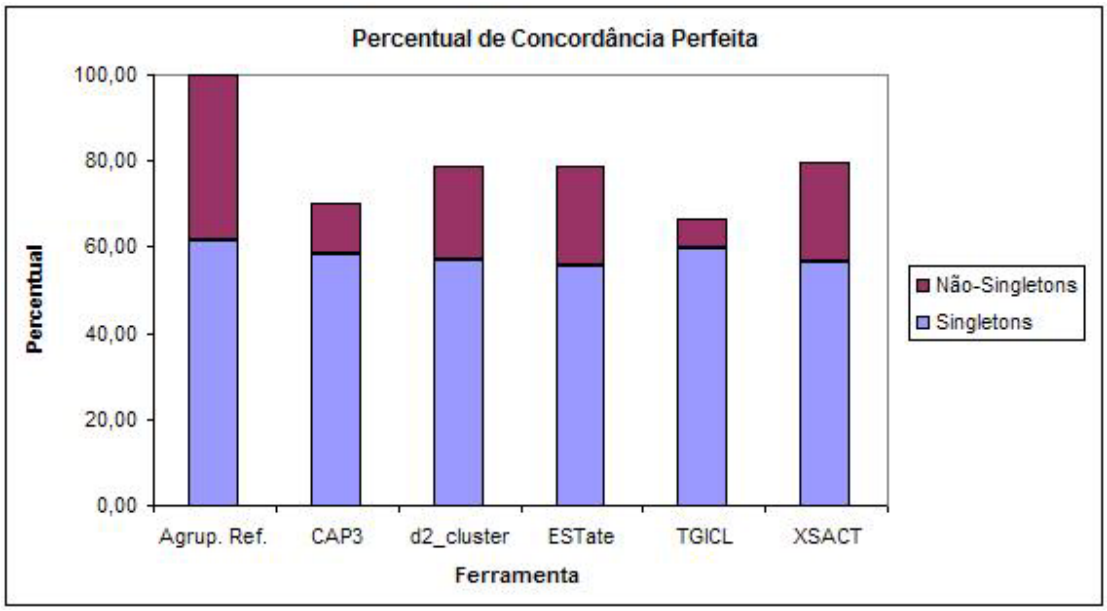

Figura 15 - Percentuais de Concordância Perfeita para a biblioteca CEREBRO_15K.

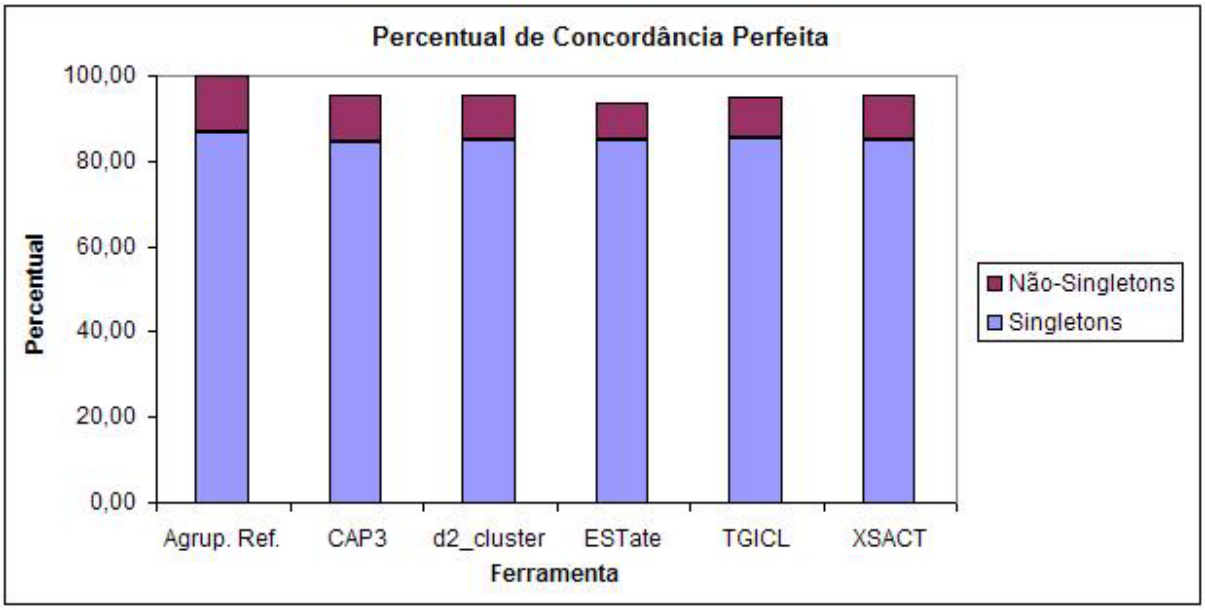

Figura 16 - Percentuais de Concordância Perfeita para a biblioteca FIGADO_10K.

4.5 Análise da super estimativa de singletons pelo CAP3 e TGICL

Conforme foi visto no gráfico da Figura 10, as ferramentas CAP3 e TGICL produziram uma quantidade de singletons bastante superior à quantidade de singletons do agrupamento de referência na biblioteca CEREBRO_15K. A super estimativa de singletons prejudicou o desempenho das ferramentas, como foi constatado nas métricas de comparação entre agrupamentos calculadas anteriormente para a biblioteca CEREBRO_15K.

Com o objetivo de investigar a super estimativa do número de singletons da biblioteca CEREBRO_15K pelo CAP3, foram realizadas novas execuções da ferramenta variando o *limite do comprimento da sobreposição* (opção –o) e o *limite do percentual de identidade da sobreposição* (opção –p), de forma a tornar os critérios de agrupamento menos rigorosos. Na primeira execução, o *limite do comprimento da sobreposição* foi modificado de 40 (valor padrão) para 25 (o menor valor permitido pelo programa é 21), mantendo-se os valores padrões para o restante das opções. Essa alteração não surtiu nenhum efeito, com o CAP3 gerando exatamente a mesma distribuição. Na segunda execução, o *limite do percentual de identidade da sobreposição*

foi modificado de 80 (valor padrão) para 70 (o menor valor permitido pelo programa é 66), mantendo-se os valores padrões para o restante das opções. Essa alteração modificou muito pouco a distribuição dos grupos. O número de singletons diminuiu de 5497 para 5472, ou seja, o CAP3 continuou produzindo um número de singletons elevado, mesmo com critérios de agrupamento mais relaxados.

Novas execuções do TGICL também foram realizadas para investigar o elevado número de singletons produzido na biblioteca CEREBRO_15K. Variou-se o *comprimento mínimo de sobreposição* (opção –l) e o *percentual de identidade mínimo para sobreposições* (opção –p). Na primeira execução, o *comprimento mínimo de sobreposição* foi modificado de 30 (valor padrão) para 20, mantendo-se os valores padrões para o restante das opções. Essa variação surtiu pouco efeito sobre o número de singletons produzido pelo TGICL. O número de singletons diminuiu de 5595 para 5582, ou seja, continuou elevado. Em seguida foi realizada uma nova execução modificando o valor da opção para 10. Novamente a alteração surtiu pouco efeito sobre o número de singletons, desta vez passando de 5595 para 5567. Para o *percentual de identidade mínimo para sobreposições* foram realizadas execuções com os valores 85, 80, 70 e 50. A Tabela 10 mostra a quantidade de singletons para cada valor de *p*. Diferentemente do CAP3, onde as alterações nos critérios de agrupamento praticamente não modificaram o número de singletons, os valores de *p* iguais a 85, 80 e 70 tiveram um efeito mais significativo sobre o número de singletons do TGICL, apesar de o número de singletons ainda ter permanecido elevado em relação ao agrupamento de referência. Curiosamente, o valor de *p* igual a 50 produziu exatamente o mesmo resultado da execução padrão do TGICL.

Tabela 10 - Quantidade de Singletons para várias execuções do TGICL na biblioteca CEREBRO_15K variando o percentual de identidade mínimo para sobreposições.

	Execuções do TGICL				
	Execução Padrão (*p* = 94)	***p* = 85**	***p* = 80**	***p* = 70**	***p* = 50**
Qtde de Singletons	5595	4744	4692	4692	5595

Com a finalidade de verificar o efeito da redução do número de singletons na qualidade dos agrupamentos, o percentual de concordância perfeita foi calculado para os agrupamentos resultantes das execuções do TGICL com valores de *p* iguais a 85 e 80. O valor de *p* igual a 70 não foi considerado por ter gerado o mesmo resultado de *p* igual a 80, e *p* igual a 50 não foi considerado por ter produzido o mesmo resultado da execução padrão. O gráfico da Figura 17 apresenta os percentuais de concordância perfeita para os agrupamentos produzidos pelo XSACT, pela execução padrão do TGICL e pelas execuções do TGICL com *p* igual a 85 e 80 em comparação ao agrupamento de referência. O agrupamento do XSACT foi incluído por ter apresentado o maior percentual de concordância perfeita dentre todas as ferramentas. Conforme demonstrado no gráfico, houve um aumento exíguo nos percentuais de concordância perfeita total para as novas execuções do TGICL em comparação à execução padrão. Assim, verificou-se que o relaxamento dos critérios de agrupamento para as ferramentas CAP3 e TGICL não foram capazes de melhorar a qualidade dos agrupamentos, que foi bastante inferior em relação às outras ferramentas na biblioteca CEREBRO_15K.

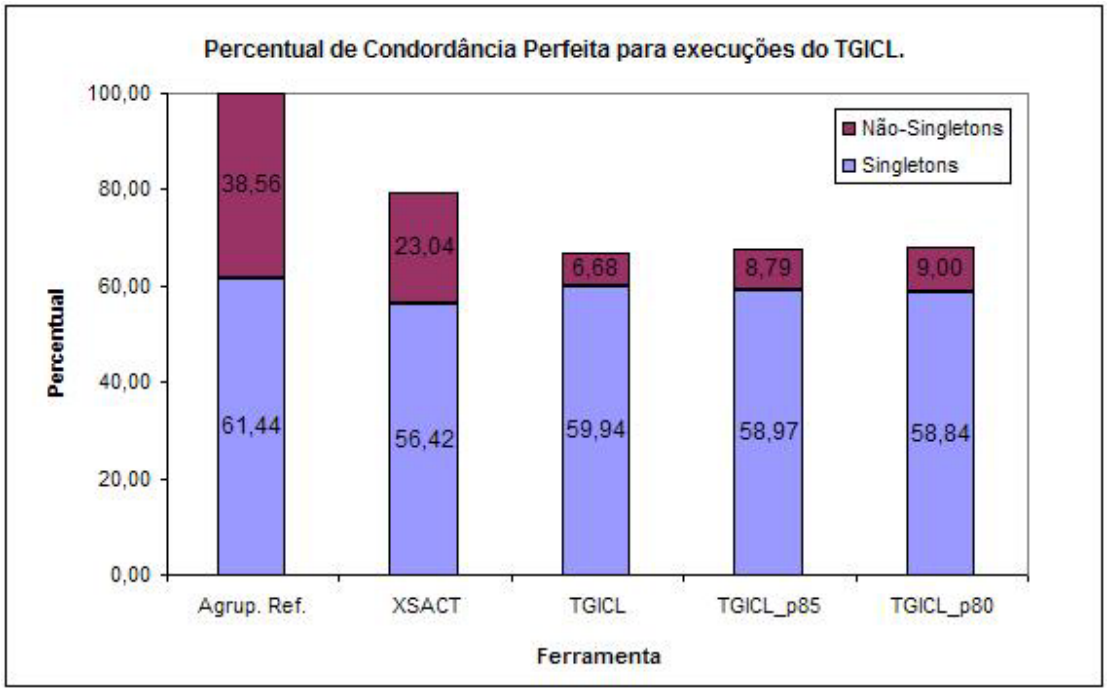

Figura 17 - Percentuais de Concordância Perfeita para execuções do TGICL variando o percentual de identidade mínimo para sobreposições na biblioteca CEREBRO_15K.

4.6 Perfil dos singletons incorretos

Com o objetivo de avaliar o perfil dos singletons incorretos gerados pelas ferramentas, buscou-se saber em que grupos estes deveriam estar distribuídos no agrupamento de referência. Nos gráficos apresentados nas Figuras 18, 19 e 20 as barras mostram o percentual de singletons incorretos por tamanho do grupo correspondente no agrupamento de referência para as bibliotecas CTRONCO_38K, CEREBRO_15K e FIGADO_10K respectivamente. Por exemplo, a primeira barra da esquerda para a direita da Figura 18 indica que aproximadamente 43% dos singletons incorretos do CAP3 deveriam fazer parte de grupos de duas ESTs no agrupamento de referência. Por uma questão de visualização nos gráficos os grupos correspondentes com tamanho de 10 a 14 foram considerados na classe ‘10-14’, aqueles com tamanho de 15 a 19 foram considerados na classe ‘15-19’, e aqueles com tamanho maior ou igual a 20 foram considerados na classe ‘>=20’.

Na biblioteca CTRONCO_38K, conforme pode ser visto no gráfico da Figura 18, as ferramentas produziram perfis semelhantes entre si. O gráfico demonstra que quase metade dos singletons incorretos produzidos pelas ferramentas deveria estar em um grupo de dois singletons, com os resultados variando de 43% (CAP3) a 51% (d2_cluster) aproximadamente. O segundo maior percentual de singletons incorretos é o daqueles que deveriam estar em grupos de tamanho 3, com os números variando de 17% (CAP3) a 20% (ESTate) aproximadamente. Esse comportamento também se repetiu para as bibliotecas CEREBRO_15K e FIGADO_10K. Assim, verifica-se que existe uma tendência dos singletons incorretos se concentrarem nos grupos de tamanho menor. De certa forma, esse comportamento já era esperado em virtude dos grupos menores serem mais numerosos, como foi visto anteriormente nos gráficos que mostram a distribuição dos grupos por tamanho, e demonstra que não houve viés na distribuição dos singletons incorretos, ou seja, uma concentração preferencial em grupos de determinado tamanho.

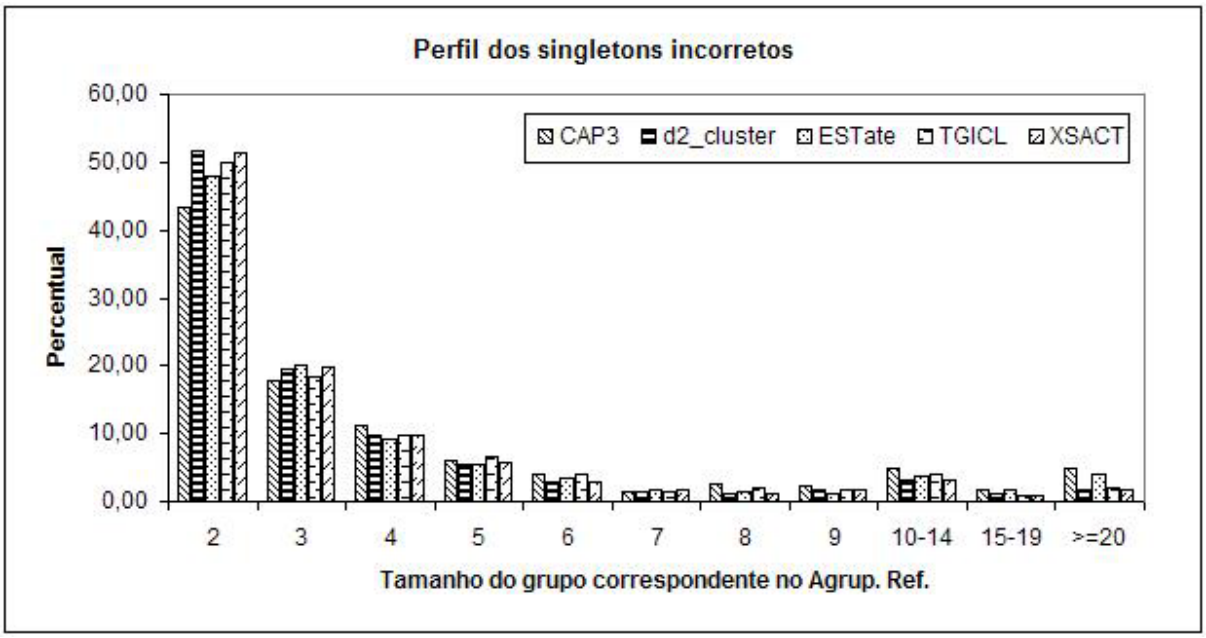

Figura 18 - Perfil dos singletons incorretos para a biblioteca CTRONCO_38K.

Na biblioteca CEREBRO_15K, o número de singletons foi super estimado pelas ferramentas CAP3 e TGICL (ver Figura 10). De maneira geral, a Figura 19 mostrou que as ferramentas CAP3 e TGICL apresentaram padrões menos similares ao restante das ferramentas no que diz respeito ao perfil de distribuição dos singletons incorretos no agrupamento de referência. Nas ferramentas d2_cluster, ESTate e XSACT houve uma concentração maior de singletons incorretos nos grupos menores do que nas ferramentas CAP3 e TGICL. Nas ferramentas d2_cluster, ESTate e XSACT as somas dos percentuais de singletons incorretos que deveriam estar em grupos de tamanho 2, 3 e 4 representaram respectivamente 61,22%, 69,39% e 66,95%, enquanto que no CAP3 e TGICL representaram 31,51% e 51,90% respectivamente. Também houve diferença nos grupos maiores. Considerando-se grupos com tamanho maior ou igual a 10 (classes ’10-14’, ’15-19’ e ‘>=20’ do gráfico) os percentuais de singletons incorretos para as ferramentas d2_cluster, ESTate e XSACT foram respectivamente 20,03%, 14,78% e 17,58%. Já para as ferramentas CAP3 e TGICL foram 46,30% e 22,88% respectivamente. Claramente, as ferramentas que super estimaram a quantidade de singletons, CAP3 e TGICL, apresentaram padrões diferentes em comparação ao restante das ferramentas, concentrando menos singletons incorretos que as outras ferramentas nos grupos menores, e mais singletons incorretos que as outras ferramentas nos grupos maiores.

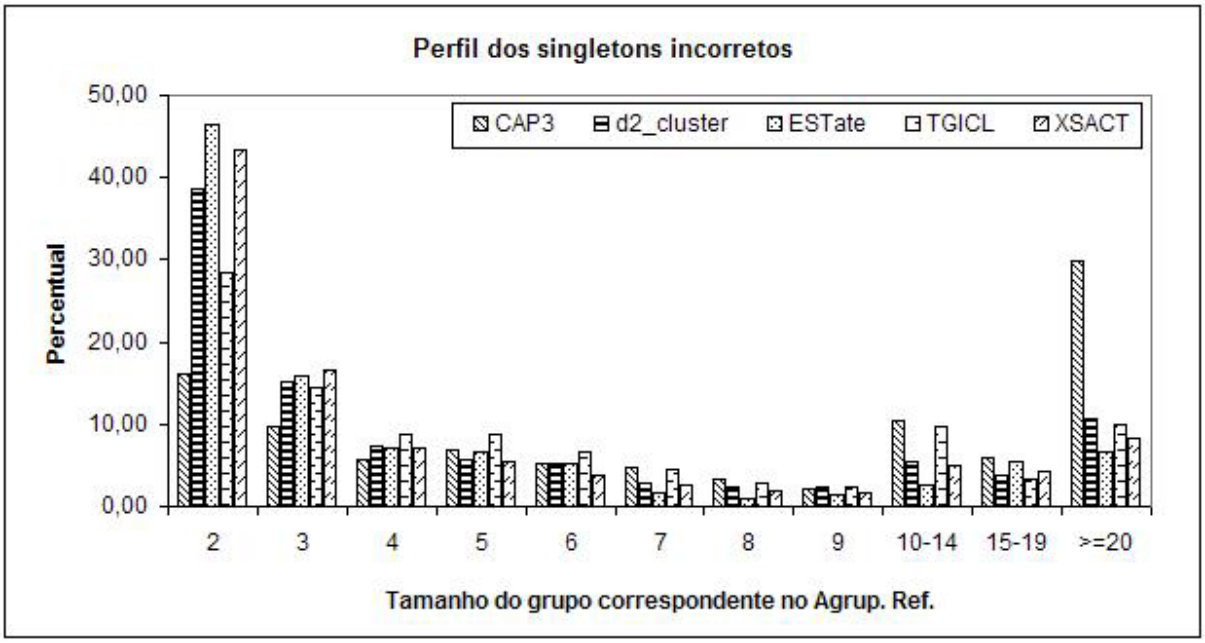

Figura 19 - Perfil dos singletons incorretos para a biblioteca CEREBRO_15K.

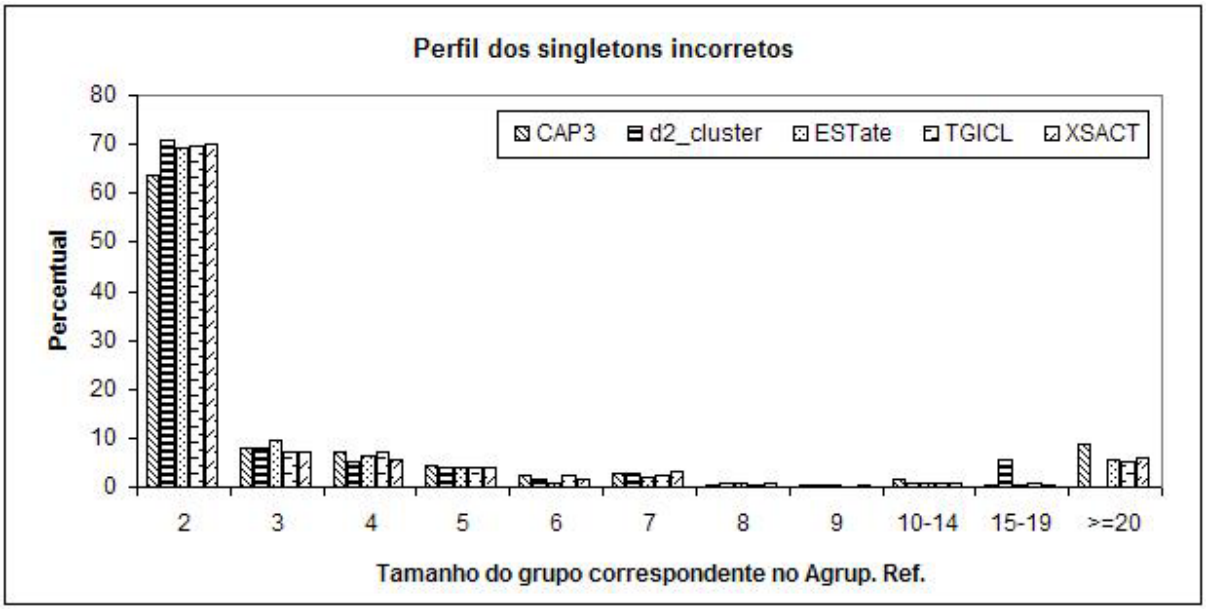

Figura 20 - Perfil dos singletons incorretos para a biblioteca FIGADO_10K.

4.7 Análise da dispersão dos grupos

No tópico 4.4 (Análise dos Agrupamentos de ESTs), o nível de acerto das ferramentas em relação ao agrupamento de referência foi avaliado, porém as medidas utilizadas não permitem visualizar como as ferramentas se comportaram nos casos de erros. Com o objetivo de ter uma

visão geral do comportamento das ferramentas, para cada lista de grupos correspondentes (agrupamento de referência versus agrupamento da ferramenta), foi utilizado um gráfico de dispersão com os tamanhos dos pares de grupos correspondentes. Cada ponto do gráfico representa um par de grupos correspondentes, sendo que o valor da coordenada *x* indica o tamanho do grupo do agrupamento de referência e o valor da coordenada *y* representa o tamanho do grupo do agrupamento encontrado por determinada ferramenta.

A Figura 21 apresenta o gráfico de dispersão entre os grupos da ferramenta XSACT e os grupos do agrupamento de referência para a biblioteca CTRONCO_38K. Pontos que aparecem acima da linha diagonal indicam que o grupo gerado pela ferramenta é maior que o seu correspondente no agrupamento de referência, pontos que aparecem abaixo da linha diagonal indicam que o grupo gerado pela ferramenta é menor que o seu correspondente no agrupamento de referência. Portanto, se o XSACT tivesse produzido um agrupamento igual ao agrupamento de referência, todos os pontos estariam em cima da linha diagonal.

De maneira geral, a dispersão dos grupos do XSACT não foi muito grande para a biblioteca CTRONCO_38K, houve uma concentração grande de pontos ao redor da diagonal. Esse fato indica que os grupos do XSACT não divergiram muito dos grupos do agrupamento de referência. Ainda assim, alguns grupos se destacaram por estarem mais dispersos que o restante, principalmente aquele localizado no canto superior esquerdo do gráfico, circulado em vermelho.

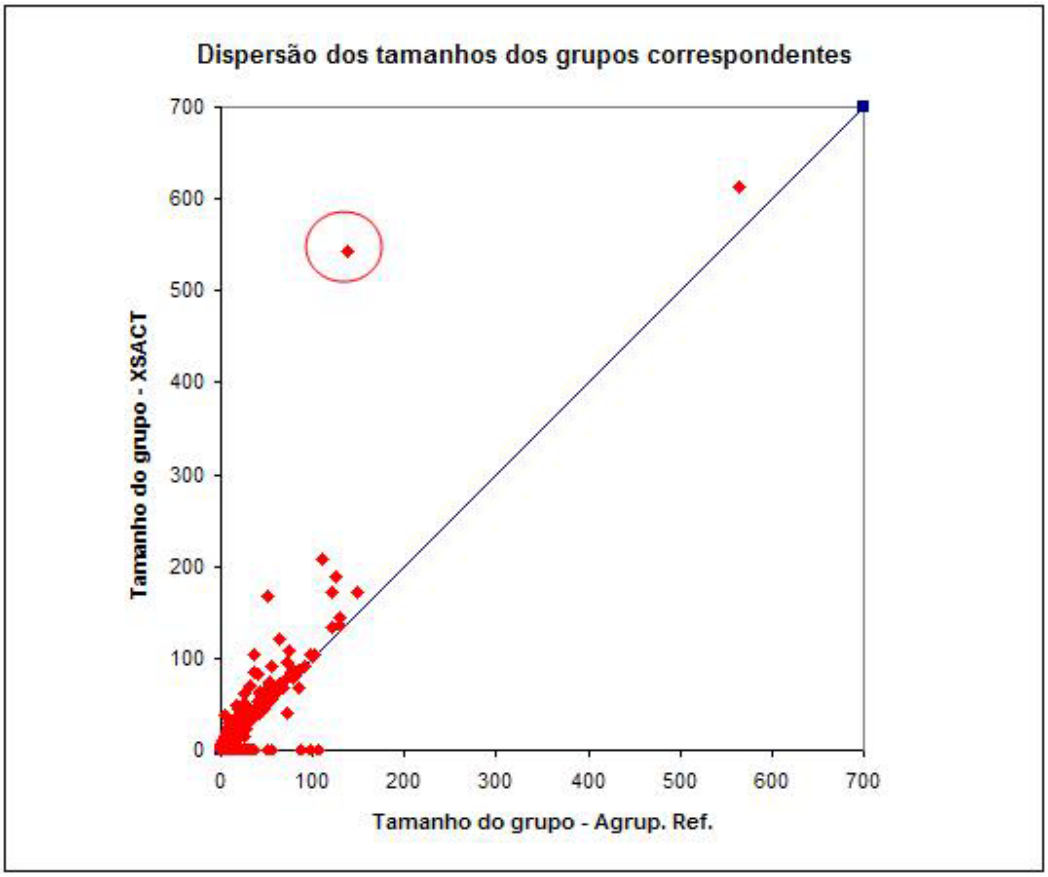

Figura 21 - Gráfico de dispersão dos tamanhos de grupos correspondentes do agrupamento de referência e XSACT para a biblioteca CTRONCO_38K.

Todas as ferramentas apresentaram gráficos de dispersão bastante similares para a biblioteca CTRONCO_38K, incluindo o grupo discrepante circulado em vermelho (dado não mostrado). O CAP3, no entanto, gerou um grupo discrepante menor, por volta de 300 ESTs, como pode ser visto no gráfico da Figura 22.

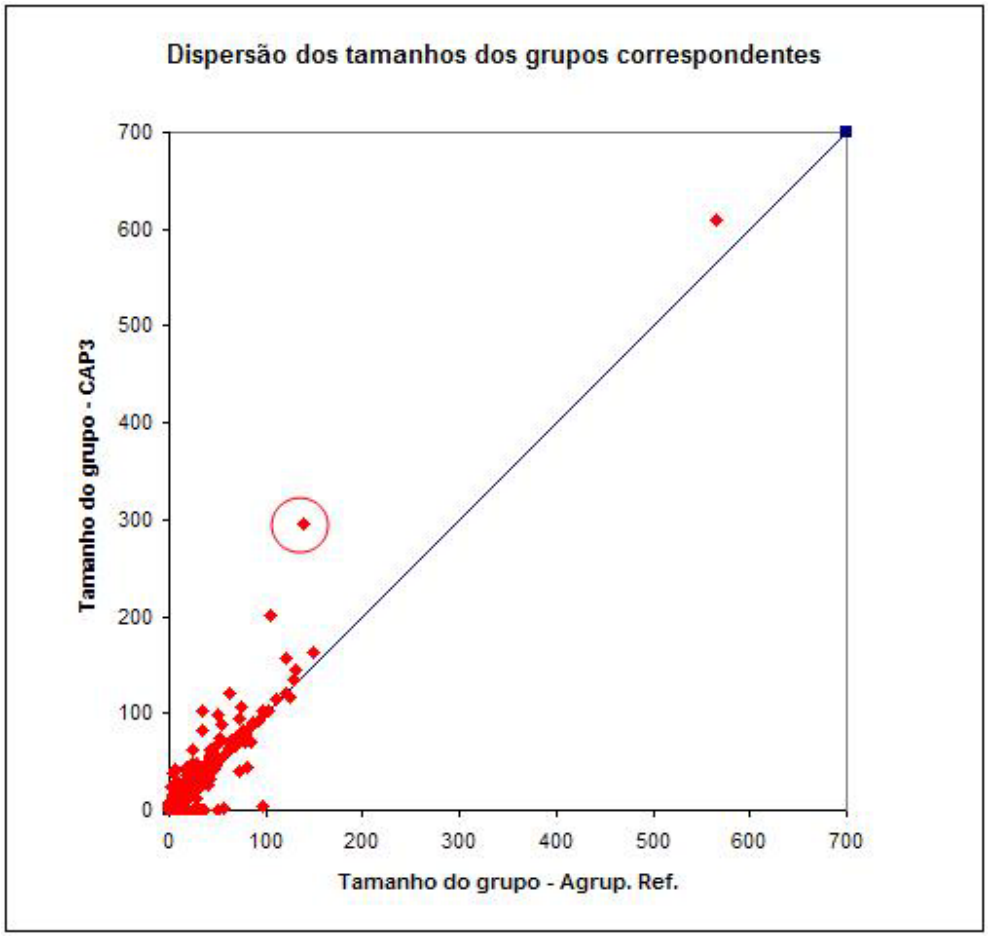

Figura 22 - Gráfico de dispersão dos tamanhos de grupos correspondentes do agrupamento de referência e CAP3 para a biblioteca CTRONCO_38K.

A Figura 23 apresenta o gráfico de dispersão do XSACT para a biblioteca CEREBRO_15K. As ferramentas d2_cluster e ESTate apresentaram gráficos de dispersão similares ao gráfico do XSACT (dado não mostrado). Assim como na biblioteca CTRONCO_38K, houve uma concentração grande de pontos próximos da linha diagonal, e também o aparecimento de alguns grupos mais dispersos que se destacaram do restante.

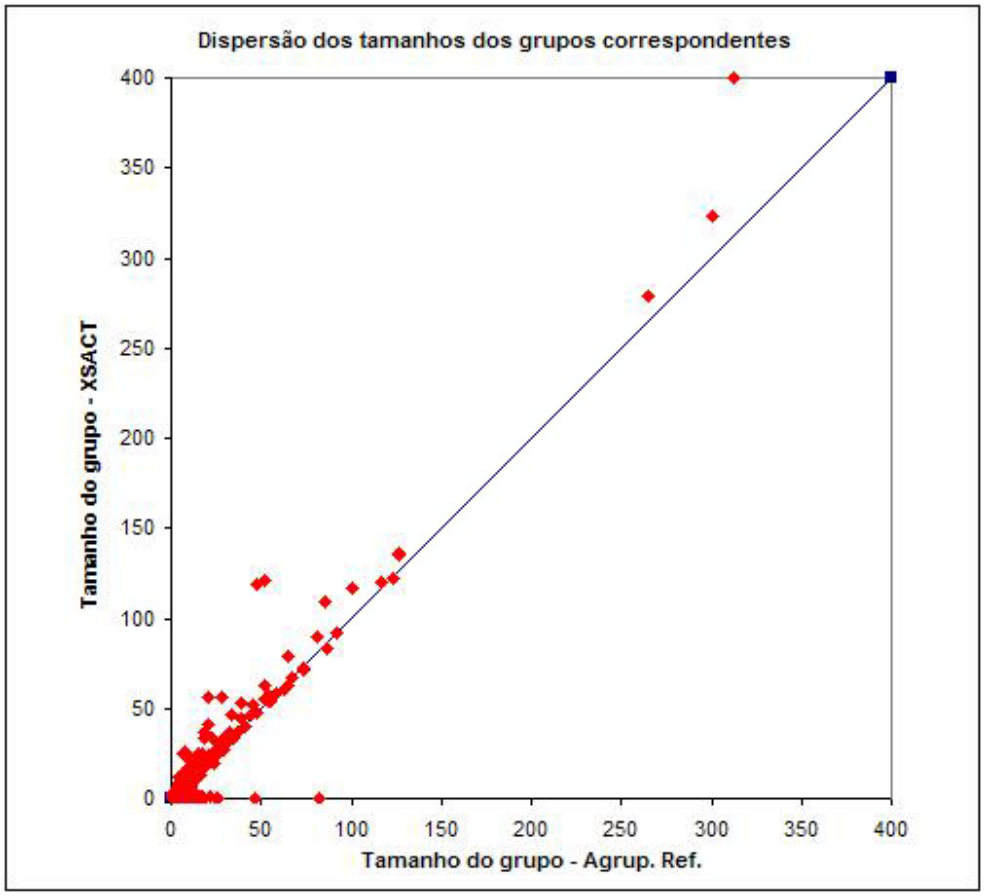

Figura 23 - Gráfico de dispersão dos tamanhos de grupos correspondentes do agrupamento de referência e XSACT para a biblioteca CEREBRO_15K.

A ferramenta CAP3 apresentou um gráfico de dispersão diferente dos gráficos das ferramentas d2_cluster, ESTate e XSACT para a biblioteca CEREBRO_15K. O gráfico da Figura 24 demonstra que boa parte dos grupos do CAP3 ficaram abaixo da linha diagonal, ao contrário do que aconteceu com as ferramentas d2_cluster, ESTate e XSACT. Esse padrão de dispersão aconteceu em virtude do alto número de singletons gerado pela ferramenta CAP3.

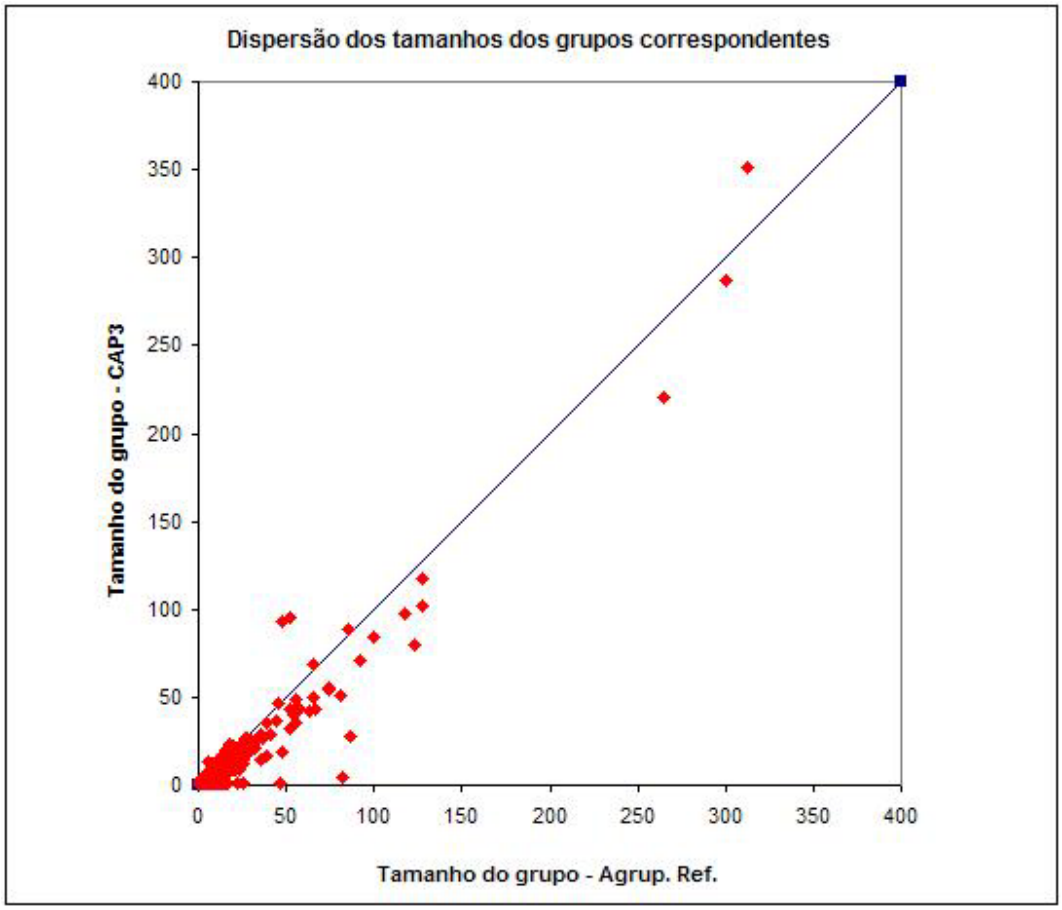

Figura 24 - Gráfico de dispersão dos tamanhos de grupos correspondentes do agrupamento de referência e CAP3 para a biblioteca CEREBRO_15K.

A ferramenta TGICL apresentou um gráfico de dispersão (Figura 25) semelhante ao gráfico do CAP3 para a biblioteca CEREBRO_15K, porém, verificou-se que os grupos do TGICL apresentaram uma dispersão menor em comparação aos grupos do CAP3. Isso se explica pela diferença entre os perfis dos singletons incorretos do CAP3 e TGICL. No CAP3 há uma concentração percentual maior de singletons incorretos em grupos maiores que no TGICL (ver Figura 19). Significa dizer que para gerar seus singletons incorretos o CAP3 retira mais singletons dos grupos maiores que o TGICL.

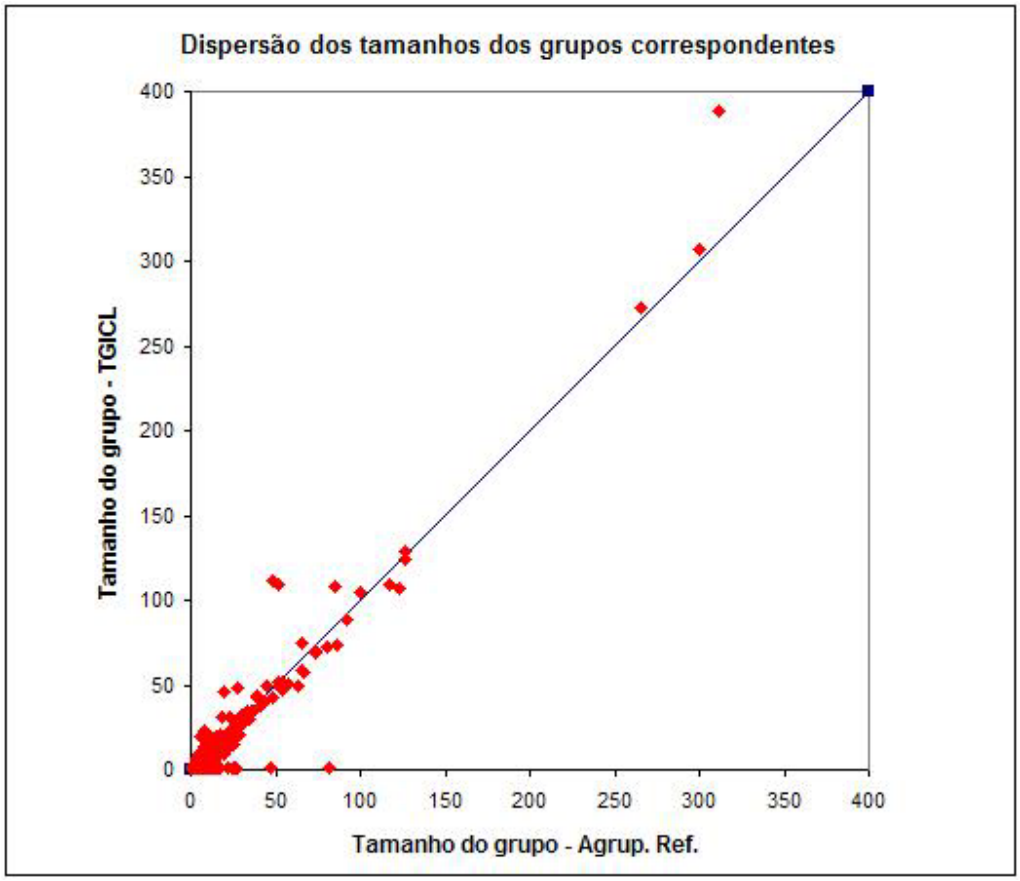

Figura 25 - Gráfico de dispersão dos tamanhos de grupos correspondentes do agrupamento de referência e TGICL para a biblioteca CEREBRO_15K.

A Figura 26 apresenta o gráfico de dispersão da ferramenta XSACT para a biblioteca FIGADO_10K. Todas as outras ferramentas apresentaram gráficos similares (dado não mostrado). Conforme pode ser visto no gráfico, os grupos divergiram muito pouco. Esse padrão está de acordo com os resultados das métricas de comparação entre agrupamentos, que indicou o excelente desempenho das ferramentas na biblioteca FIGADO_10K.

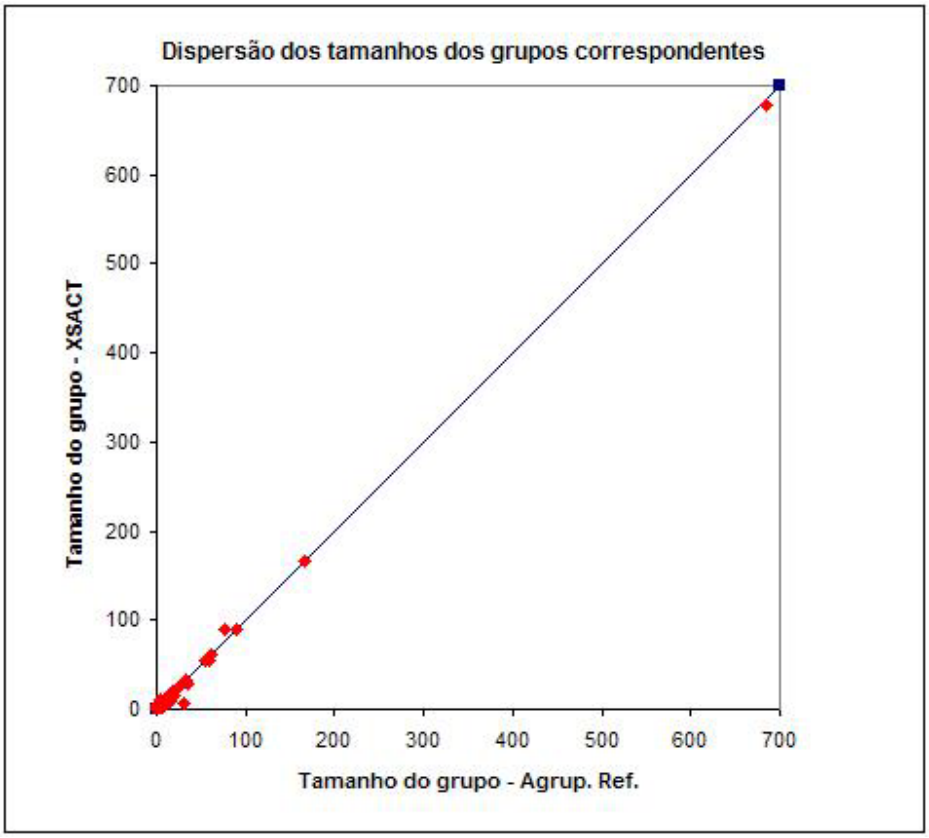

Figura 26 - Gráfico de dispersão dos tamanhos de grupos correspondentes do agrupamento de referência e XSACT para a biblioteca FIGADO_10K.

4.8 Análise de grupos discrepantes

No tópico anterior os gráficos de dispersão permitiram inspecionar visualmente como os grupos gerados pelas ferramentas divergiram dos grupos do agrupamento de referência. Em alguns casos surgiram grupos que divergiram drasticamente do seu grupo correspondente no agrupamento de referência. Apesar dos gráficos mostrarem os grupos discrepantes, eles não indicam por que o grupo gerado pela ferramenta divergiu. Assim, buscou-se investigar as causas envolvidas na geração dos grupos mais discrepantes, e para isso o grupo circulado em vermelho no gráfico da Figura 21 foi escolhido. Este grupo apareceu em todos os agrupamentos gerados pelas ferramentas para a biblioteca CTRONCO_38K. No caso do CAP3, este grupo apareceu com o tamanho menor (ver grupo circulado em vermelho no gráfico da Figura 22).

Para cada grupo do agrupamento de referência foi contabilizada a quantidade de ESTs concordantes com os grupos discrepantes gerados pelas ferramentas. Essas contagens estão reunidas na Tabela 11. As três primeiras colunas da esquerda (Id., Cr. e Tam.) indicam respectivamente o identificador do grupo do agrupamento de referência, o cromossomo onde este foi situado e o seu tamanho. Nas colunas das ferramentas, as colunas Id. e Tam. indicam respectivamente o identificador do grupo discrepante gerado pela ferramenta e o seu tamanho. Cada célula sombreada indica a quantidade de ESTs concordantes entre o grupo discrepante e o grupo do agrupamento de referência. O grupo *CL2* do TGICL, por exemplo, possui 138 ESTs concordantes com o grupo *00248* do agrupamento de referência. O grupo *Contig2641* do CAP3, por exemplo, não possui nenhuma EST concordante com o grupo *00910* do agrupamento de referência.

Tabela 11 - Quantidades de ESTs concordantes entre grupos discrepantes e grupos do agrupamento de referência (biblioteca CTRONCO_38K).

GRUPOS DO AGRUP. REF.			GRUPOS DAS FERRAMENTAS									
			CAP3		d2_cluster		ESTate		TGICL		XSACT	
			Id.	Tam.	Id.	Tam.	Id.	Tam.	Id.	Tam.	Id.	Tam.
Id.	Cr.	Tam.	*Contig 2641*	295	*20828*	538	*CL000 04144*	504	*CL2*	537	*CL26*	543
00248	5	138	138		138		138		138		138	
00910	17	105	0		105		105		105		105	
00064	1	97	0		97		97		97		97	
01476	2	56	55		56		56		56		56	
01892	7	36	36		36		0		36		36	
00635	7	33	33		33		30		33		33	
01134	24	22	0		22		22		22		22	
01705	1	13	0		13		13		13		13	
03418	1	13	13		13		13		13		13	
03008	2	11	11		11		11		11		11	
11557	1	4	1		4		3		4		4	
03250	11	3	0		1		1		1		3	
05758	2	3	3		3		3		3		3	
02819	22	2	2		2		2		2		2	
04875	3	2	0		2		2		2		2	
00859	10	3	0		0		2		0		0	

Conforme visto na Tabela 11, as ferramentas uniram praticamente os mesmos grupos do agrupamento de referência nos seus respectivos grupos discrepantes. A exceção foi o CAP3, que

deixou de unir alguns grupos, produzindo, dessa forma, um grupo discrepante menor que os demais.

4.8.1 Similaridade entre ESTs dos grupos *00248* e *CL26*

O grupo *00248* (138 ESTs) é o grupo correspondente do grupo *CL26* (543 ESTs) no agrupamento de referência por compartilhar o maior número de ESTs concordantes entre todos (ver Tabela 11). No entanto, verifica-se que o grupo *CL26* possui 405 ESTs a mais que o seu grupo correspondente no agrupamento de referência, configurando um erro crasso por parte da ferramenta. Assim, nesta análise, buscou-se compreender por que a ferramenta agrupou 405 ESTs a mais no grupo *CL26*, sendo que este deveria ter somente 138 ESTs. Para isso, as ESTs incorretas foram comparadas ao grupo *00248* com o BLASTN. Como os grupos discrepantes são praticamente iguais, escolheu-se somente o grupo *CL26* do XSACT para esta análise. Apesar do grupo discrepante do CAP3 possuir vários grupos a menos, ele é um subconjunto dos demais, e, portanto, pode se valer dos resultados assim como os outros.

Inicialmente, foi gerada a seqüência consenso do grupo *00248* utilizando o CAP3. Em seguida foi gerado um arquivo no formato FASTA com as 405 ESTs incorretas do grupo *CL26*. O arquivo das ESTs incorretas foi então comparado à seqüência consenso do grupo *00248* utilizando o BLASTN. Das 405 ESTs comparadas, 9,38% (38 ESTs) não produziram alinhamentos, 89,63% (363 ESTs) produziram alinhamentos com percentual de identidade maior ou igual a 85%, e 75,8% (307 ESTs) produziram alinhamentos com percentual de identidade maior ou igual a 90%. Esses resultados demonstraram que as ESTs incorretas possuem um nível de similaridade elevado com o grupo *00248*.

Porém, como o BLASTN produz alinhamento local, ainda foi investigado se a cobertura do alinhamento local era representativa em relação ao tamanho da EST. Portanto, a cobertura foi

calculada para cada EST dividindo-se o número de identidades do alinhamento local pelo tamanho da EST. Das ESTs que produziram alinhamentos, 76,54% (310 ESTs) apresentaram uma cobertura maior ou igual a 50%, e 63,95% (259 ESTs) apresentaram uma cobertura maior ou igual a 75%. Assim, os alinhamentos locais gerados pelo BLASTN demonstraram que existe um grau de similaridade relativamente alto entre as ESTs e a seqüência consenso do grupo *00248*, e que este grau de similaridade é representativo em relação ao tamanho das ESTs. Apesar do agrupamento de referência nos apontar que estas ESTs foram agrupadas indevidamente, vê-se que as ferramentas não apresentaram nenhum comportamento aberrante considerando que a única informação da qual dispõem para realizar os agrupamentos é a seqüência da EST.

4.8.2 Anotações dos grupos *00248*, *00910* e *00064*

A análise anterior mostrou que as ESTs agrupadas incorretamente no grupo *CL26* compartilham um nível de similaridade relativamente alto com o grupo *00248*. Adicionalmente, a Tabela 11 nos mostra que as ESTs incorretas do grupo *CL26* estão em grupos do agrupamento de referência que estão distribuídos em vários cromossomos. Com o objetivo de compreender esta configuração de grupos com alto grau de similaridade de seqüência espalhados em cromossomos diferentes, os grupos *00248*, *00910* e *00064* (os três maiores) foram pesquisados no genoma, conforme descrito no tópico 3.10 (página 35), a fim de verificar suas anotações.

A Tabela 12 apresenta os dois primeiros resultados do BLASTN para cada um dos contigs dos grupos. No grupo *00248*, o alinhamento de maior escore (1939) do Contig1 ocorreu no cromossomo 5 na região do gene ACTBP2, que codifica a proteína beta actina. No grupo *00910*, o CAP3 gerou três contigs. Os alinhamentos com maior escore dos contigs Contig1 e Contig2 ocorreram no cromossomo 17 na região do gene ACTG1, que codifica a proteína actina gama. O Contig3 alinhou em uma região do cromossomo 3 que possui as anotações do locus LOC643897

e do gene SYN2, o primeiro um locus similar à proteína beta actina e o segundo o gene da sinapsina II. No grupo *00064*, o alinhamento com maior escore (618) do Contig1 ocorreu no cromossomo 1 em uma região que não apresentou gene anotado no Entrez Gene, mas que apresentou a anotação de uma predição (hmm208267) relacionada à proteína gama actina 1 (AAH00292.1) no mapa de modelos *AB initio*. Com exceção do segundo alinhamento do Contig1 do grupo *00248*, que alinhou na região do gene PLEKHB2, todos os outros alinhamentos ocorreram em regiões com anotações de genes ou loci relacionados à proteína actina.

As actinas são proteínas que estão envolvidas em vários processos como a mobilidade da célula e a manutenção do citoesqueleto. Entre os eucariotos as actinas são bem conservadas e são encontradas em todas as células, sendo que a maioria dos organismos tem múltiplos genes que a codificam. Seqüências de aminoácidos de actinas de espécies diferentes normalmente compartilham 90% de identidade. Nos vertebrados há três isoformas da actina denominadas alfa, beta e gama. A alfa actina é expressa somente em tecidos musculares. As actinas beta e gama coexistem na maioria dos tipos celulares como componentes do citoesqueleto, e como mediadoras da mobilidade interna da célula.

Conforme demonstrado através das anotações, apesar dos grupos estarem situados em cromossomos diferentes, todos estão localizados em genes codificadores da proteína actina. Assim, ao juntar os grupos 00248, 00910 e 00064 as ferramentas agruparam ESTs originadas de genes duplicados no genoma, denominados parálogos. Genes parálogos representam um problema para as ferramentas de agrupamento de ESTs, pois normalmente possuem similaridade elevada quando originados de um evento de duplicação recente do ponto de vista evolutivo, e isso torna difícil o trabalho da ferramenta, que não consegue ajustar seus parâmetros de forma a diferenciar os erros inerentes dos dados de ESTs das mutações sofridas pelos genes.

Tabela 12 - Anotações dos grupos 00248, 00910 e 00064.

Grupos do Agrup. Ref.			Resultados do BLASTN				Anotações			
Id.	**Contig**	**Tam.**	**Seqüência**	**Score (Bits)**	**E-Value**	**% Ident.**	**Cr.**	**GeneID**	**Símbolo Oficial**	**Nome do Gene**
00248 (Cr5)	Contig1	1445	ref|NT_006713.14|Hs5_6870	1939	0	92%	5	62	ACTBP2	actin, beta pseudogene 2
			ref|NT_022135.15|Hs2_22291	1814	0	92%	2	55041	PLEKHB2	pleckstrin homology domain containing, family B (evectins) member 2
00910 (Cr17)	Contig1	907	ref|NT_024871.11|Hs17_25027	718	0	100%	17	71	ACTG1	actin, gamma 1
			ref|NT_004610.18|Hs1_4767	708	0	91%	1	644961	LOC644961	similar to cytoplasmic beta-actin
	Contig2	578	ref|NT_024871.11|Hs17_25027	1076	0	99%	17	71	ACTG1	actin, gamma 1
			ref|NT_011875.11|HsY_12032	143	1e-31	87%	Y	74	ACTGP2	actin, gamma pseudogene 2
	Contig3	1910	ref|NT_022517.17|Hs3_22673	1744	0	89%	3	643897	LOC643897	similar to cytoplasmic beta-actin
								6854	SYN2	synapsin II
			ref|NT_024871.11|Hs17_25027	1715	0	100%	17	71	ACTG1	actin, gamma 1
00064 (Cr1)	Contig1	656	ref|NT_004836.17|Hs1_4993	618	8e-175	91%	1			
			ref|NT_011875.11|HsY_12032	523	4e-146	89%	Y	74	ACTGP2	actin, gamma pseudogene 2
								414754	LOC414754	actin, gamma pseudogene

5 DISCUSSÃO

5.1 Agrupamento de Referência

A abordagem de avaliação dos agrupamentos de ESTs baseada no agrupamento de referência forneceu uma base de comparação única que permitiu avaliar objetivamente as ferramentas de agrupamento. O genoma humano foi escolhido por já estar totalmente seqüenciado e bem anotado. Como a construção do agrupamento de referência utiliza um genoma completo, as ferramentas de agrupamento só podem ser avaliadas com essa abordagem em organismos com o genoma seqüenciado. Porém, existe a alternativa de utilizar genomas de organismos modelos relacionados ao organismo de interesse para supor como seria o desempenho das ferramentas.

Na construção do agrupamento de referência, é necessária a sobreposição sem ambigüidade das ESTs no genoma. Dois requisitos básicos são fundamentais para as ferramentas utilizadas neste procedimento: velocidade e capacidade de alinhar seqüências com junções de éxons. De maneira geral, como os genomas possuem muita informação, se a ferramenta não for rápida o suficiente a pesquisa das ESTs se torna inviável. Por outro lado, os alinhamentos devem levar em consideração a presença de íntrons no DNA genômico, sendo capazes de inferir quais éxons foram unidos na transcrição de determinada EST. No presente trabalho as ferramentas BLAST e est2genome foram utilizadas de maneira combinada para mapear as ESTs no genoma. A tentativa de combinar as ferramentas ssahaEST e est2genome não foi viável pelo fato do ssahaEST, mesmo sendo mais rápido que o BLAST, não detectar uma grande quantidade de alinhamentos, em função de ser muito estringente, penalizando bases não concordantes (“mismatches”) entre as ESTs e o genoma. No entanto, existem outras ferramentas descritas na

literatura que combinam essas duas características e podem ser testadas na construção do agrupamento de referência (Wheelan et al., 2001; Ogasawara e Morishita, 2002; Wu et al., 2005).

5.2 Métricas de comparação de agrupamentos

Após a seleção das três bibliotecas de EST humanas, diferindo quanto ao número total de leituras e tecidos, procedeu-se à execução das cinco ferramentas de agrupamento de ESTs. Uma das principais tarefas foi a seleção de medidas de comparação entre os agrupamentos das ferramentas e o agrupamento de referência.

Entre as métricas de comparação de agrupamentos utilizadas neste trabalho, o percentual de concordância perfeita foi aquela que se mostrou mais rigorosa. Esta métrica é bastante sensível a pequenas diferenças entre os agrupamentos. Na biblioteca CTRONCO_38K, enquanto a média do percentual de ESTs concordantes do CAP3 foi de 91,16%, o percentual de concordância perfeita foi de 83,28%, uma diferença de quase 8%. Já na biblioteca CEREBRO_15K, onde o CAP3 obteve um resultado inferior ao das outras ferramentas, enquanto que a média do percentual de ESTs concordantes foi de 85,29%, o percentual de concordância perfeita foi de 70,13%, uma diferença de 15,16%. Assim, esta métrica, por ser mais sensível que as demais, se mostrou mais adequada para a comparação de agrupamentos com alta similaridade.

A média do coeficiente de Jaccard por grupo e do percentual de ESTs concordantes por grupo se comportaram de maneira semelhante, inclusive com valores próximos entre si. Na verdade, o percentual de ESTs concordantes é um caso especial do coeficiente de Jaccard. Este é dado por **p/(p + q + r)**, onde **p** é o número de variáveis que aparece nos dois objetos, **q** é o número de variáveis que aparece somente no objeto ***i***, e **r** é o número de variáveis que aparece somente no objeto ***j***. Já o percentual de ESTs concordantes, utilizando os mesmos termos do coeficiente de Jaccard, seria dado por **p/(p + q) * 100**. Se o percentual de ESTs concordantes

fosse expresso em forma de coeficiente, este seria dado por **p/(p + q)**. Como pode ser visto, a única diferença seria o termo **r**. Traduzindo para a situação real, dado um grupo do agrupamento de referência e um grupo da ferramenta, enquanto o coeficiente de Jaccard considera o número de ESTs comuns aos dois grupos (**p**) em relação ao número de ESTs comuns aos dois grupos (**p**), mais as ESTs que só aparecem no grupo do agrupamento de referência (**q**), mais as ESTs que só aparecem no grupo da ferramenta (**r**), o percentual de ESTs concordantes desconsidera o número de ESTs que só aparecem no grupo da ferramenta (**r**). Portanto, o coeficiente de Jaccard é mais rigoroso que o percentual de ESTs concordantes, mesmo os dois tendo apresentado valores bastante próximos para os agrupamentos deste trabalho.

De maneira geral, as métricas utilizadas na avaliação dos agrupamentos se mostraram bastante consistentes entre si, ou seja, na mesma biblioteca, os desempenhos das ferramentas, umas em relação às outras, foram compatíveis e consistentes.

5.3 Avaliação das ferramentas

Na biblioteca CTRONCO_38K, os agrupamentos gerados pelas ferramentas apresentaram boa qualidade quando comparados ao agrupamento de referência. O agrupamento gerado pelo ESTate, que obteve o pior resultado, apresentou um percentual de concordância perfeita de 79,85%. O agrupamento gerado pelo XSACT, que obteve o melhor resultado, apresentou um percentual de concordância perfeita de 83,29%. Calculando-se a diferença entre o melhor e o pior agrupamento, verifica-se que o XSACT superou o ESTate em apenas 3,44%. Na biblioteca FIGADO_10K, as ferramentas obtiveram desempenhos ainda melhores. O agrupamento gerado pelo ESTate, que obteve o pior resultado, apresentou um percentual de concordância perfeita de 93,66%. O agrupamento gerado pelo d2_cluster, que obteve o melhor resultado, apresentou um percentual de concordância perfeita de 95,24%. Calculando-se a diferença entre os dois

percentuais, verifica-se que o d2_cluster superou o ESTate em apenas 1,58%. As diferenças entre os melhores e os piores resultados nas bibliotecas CTRONCO_38K e FIGADO_10K não permitiram afirmar que houve uma ferramenta que se destacou por apresentar uma vantagem qualitativa significante.

Se fossemos considerar os resultados das métricas somente para as bibliotecas CTRONCO_38K e FIGADO_10K, a avaliação seria que as ferramentas CAP3, d2_cluster, TGICL e XSACT estão em um patamar de qualidade levemente superior ao ESTate. Essa avaliação se baseia no fato de que as quatro ferramentas citadas anteriormente apresentaram resultados bastante próximos entre si nas três métricas e o ESTate apresentou sempre o pior resultado, com uma diferença um pouco maior em relação ao restante das ferramentas. Na biblioteca CTRONCO_38K, por exemplo, a diferença entre o XSACT e o TGICL, que apresentaram respectivamente o melhor e o pior percentual de concordância perfeita do subgrupo das quatro ferramentas, ficou em 0,46%. Já a diferença entre o TGICL e o ESTate ficou em 2,98%. Aplicando a mesma lógica para a mesma métrica na biblioteca FIGADO_10K, a maior diferença entre o subgrupo ficou em 0,2%, enquanto que a diferença entre a pior ferramenta do subgrupo e o ESTate ficou em 1,55%. Esta leve inferioridade do ESTate é facilmente verificada nos gráficos das métricas.

Apesar de a ferramenta ESTate ter demonstrado desempenho inferior ao restante das ferramentas nas duas bibliotecas, esta diferença não é suficiente para se julgar, em termos absolutos, se uma ferramenta é melhor que as demais. Dentro desta perspectiva, o critério de seleção da ferramenta de agrupamento deve levar em consideração outros fatores que não a potencial acuidade, como por exemplo, o tempo total de execução.

No caso da biblioteca CEREBRO_15K, os agrupamentos gerados pelas ferramentas CAP3 e TGICL foram severamente penalizados pelo elevado número de singletons. Os agrupamentos

gerados pelas ferramentas CAP3 e TGICL produziram, respectivamente, percentuais de concordância perfeita de 70,13% e 66,62%, enquanto que o agrupamento do XSACT, que obteve o melhor resultado, produziu um percentual de concordância perfeita de 79,46%. O XSACT foi 9,33% melhor que o CAP3, e 12,84% melhor que o TGICL. Nesta biblioteca, a diferença de desempenho entre a melhor ferramenta e as duas piores foi expressivo, configurando uma inferioridade das ferramentas CAP3 e TGICL. Também ficou evidente que a biblioteca pode afetar drasticamente o resultado de uma ferramenta.

Considerando os resultados das métricas para as três bibliotecas, a avaliação é de que as ferramentas estão divididas em quatro patamares de qualidade, com o XSACT e o d2_cluster no primeiro patamar, apresentando os melhores desempenhos, o ESTate no segundo, apresentando desempenho levemente inferior ao primeiro patamar, o CAP3 no terceiro, apresentando desempenho levemente inferior ao segundo patamar, e o TGICL no quarto e último, apresentando desempenho levemente inferior ao patamar anterior. As médias dos percentuais de concordância perfeita para cada ferramenta, que foram apresentados na Tabela 9, tornam bastante clara essa divisão.

As ferramentas também foram avaliadas quanto ao desempenho computacional (ver Tabela 6). O XSACT processou um arquivo com 36.219 ESTs em aproximadamente 16 minutos, seguido pelo ESTate que gastou aproximadamente 33 minutos. O CAP3 gastou 1 hora e 6 minutos, seguido pelo TGICL que gastou 1 hora e 22 minutos. Por último, o d2_cluster gastou 2 horas e 43 minutos. Os dados mostraram que o XSACT foi bastante superior ao restante das ferramentas. O ESTate, apesar de ter levado aproximadamente o dobro do tempo do XSACT, também obteve um desempenho bastante superior ao restante das ferramentas. CAP3 e TGICL ficaram em um nível intermediário, e o d2_cluster ficou bastante atrás.

Considerando a avaliação de qualidade juntamente com a avaliação do desempenho computacional das ferramentas, o XSACT foi a ferramenta que se saiu melhor, pois apesar de ter ficado no mesmo patamar de qualidade do d2_cluster, apresentou desempenho computacional muito superior. A segunda melhor ferramenta foi o d2_cluster, que apesar de ter apresentado o pior desempenho computacional, ficou classificada no primeiro patamar de qualidade. É importante notar que o critério qualidade foi considerado mais importante que o critério desempenho computacional nesta avaliação, mas nada impede que as ferramentas sejam classificadas com ponderações diferentes para estes critérios dependendo da necessidade. Em seguida, na ordem, tivemos o ESTate, o CAP3 e o TGICL. O ESTate superou o CAP3 e o TGICL nos dois critérios, e o CAP3 superou o TGICL nos dois critérios.

Além dos dois critérios utilizados anteriormente, existem outros, não considerados neste trabalho, que poderiam agregar valor à avaliação, tornando-a mais completa. Do ponto de vista computacional, por exemplo, a utilização de memória pelas ferramentas seria interessante. A solução de muitos problemas computacionais passa pela negociação entre eficiência e utilização de memória. No caso de grandes volumes de ESTs, saber a quantidade de memória que cada ferramenta consome é um dado importante para o dimensionamento do parque computacional.

A classificação das ferramentas apresentada neste trabalho de forma alguma pode ser encarada de maneira absoluta e deve ser utilizada com prudência. Vários aspectos devem ser considerados, o primeiro deles é o organismo. A Genômica tem se encarregado de caracterizar os genomas sob os mais variados aspectos, e por isso sabemos que os genomas das várias espécies guardam características próprias. Genomas de bactérias, por exemplo, possuem menos seqüências repetitivas em comparação ao genoma humano. Mesmo entre genomas eucarióticos existe uma variabilidade considerável. Em plantas, particularmente, observa-se uma grande incidência de retrotransposons e a ocorrência de duplicações genômicas (poliploidia) que dificultam ainda mais

o agrupamento de ESTs. Assim, não podemos garantir que as ferramentas obtenham sempre os mesmos desempenhos independentemente do genoma.

Outro aspecto diz respeito à própria biblioteca de ESTs, que é específica de acordo com o tecido e as condições da célula. Resultados deste trabalho mostraram como características intrínsecas das bibliotecas podem afetar radicalmente os desempenhos das ferramentas. Dessa forma, não há como garantir que, assim como o CAP3 e o TGICL foram afetados pela biblioteca CEREBRO_15K, as outras ferramentas também não seriam afetadas por uma biblioteca de outro tecido.

5.4 Super estimativa de singletons pelo CAP3 e TGICL

Na biblioteca CEREBRO_15K, as ferramentas CAP3 e TGICL super estimaram a quantidade de singletons em detrimento dos grupos não-singletons. O gráfico da Figura 15 mostrou que os percentuais de concordância perfeita de grupos não-singletons das ferramentas CAP3 e TGICL ficaram bem abaixo dos percentuais das outras ferramentas, prejudicando os seus resultados globais.

Procurou-se investigar se a alteração de alguns parâmetros que influenciam no agrupamento das ESTs poderiam melhorar o resultado das ferramentas. No CAP3 as opções *limite do comprimento da sobreposição* e *limite do percentual de identidade da sobreposição* foram testadas com parâmetros diferentes. Conforme demonstrado nos resultados, a alteração do valor da opção *limite do comprimento da sobreposição* não modificou os resultados da ferramenta. A alteração do valor da opção *limite do percentual de identidade da sobreposição* chegou a diminuir o número de singletons, porém de maneira insignificante (0,45%).

No TGICL as opções *comprimento mínimo de sobreposição* e *percentual de identidade mínimo para sobreposições* foram testadas com parâmetros diferentes. A alteração do valor da

opção *comprimento mínimo de sobreposição* surtiu pouco efeito sobre o número de singletons, reduzindo-o em apenas 0,5%. Já a alteração da opção *percentual de identidade mínimo para sobreposições* modificou de maneira mais significativa o número de singletons, reduzindo-o em 16,14%. Vale ressaltar que mesmo com essa redução o número de singletons do TGICL continuou elevado em relação ao número de singletons do agrupamento de referência, permanecendo 94,61% maior. Ainda assim, o impacto dessa redução na qualidade do agrupamento foi avaliada através do percentual de concordância perfeita. Verificou-se uma melhora pequena, com o percentual passando de 66,62% para 67,84%.

Assim, ficou claro que o relaxamento dos critérios de agrupamento do CAP3 para a biblioteca CEREBRO_15K produziu alterações insignificantes no número de singletons e não foi capaz de melhorar o desempenho da ferramenta. No caso do TGICL, o relaxamento dos critérios de agrupamento surtiu algum efeito, porém a redução no número de singletons ficou aquém do necessário para melhorar significativamente o resultado da ferramenta.

No TGICL as ESTs são agrupadas com base em comparações par-a-par, utilizando o software mgblast, que é uma versão modificada do megablast, e depois são montadas utilizando o CAP3. Assim, o resultado inferior apresentado pelo TGICL na mesma biblioteca que o CAP3 não foi mera coincidência. Provavelmente, boa parte do resultado do TGICL se deve ao desempenho inferior do CAP3. É possível que a diminuição no número singletons do TGICL verificada nas análises se deva tão somente à primeira fase (comparações com o mgblast) do agrupamento das ESTs, que fornece os grupos para montagem pelo CAP3.

Ao que parece, alguma característica das ESTs da biblioteca CEREBRO_15K invalidou o mecanismo de detecção de sobreposições pelo CAP3, afetando também os resultados do TGICL. As razões para tanto podem estar relacionadas intrinsecamente com o padrão de expressão do tecido, ou alternativamente, poderia estar relacionada ao processo experimental de construção da

biblioteca de cDNA ou na própria geração das ESTs. No entanto, essas hipóteses envolvem variáveis que fogem ao nosso controle e por isso não são facilmente verificáveis. Assim, o nosso espectro de investigação está restrito principalmente aos dados de seqüências dos quais dispomos.

Essa análise teve um papel importante ao demonstrar que as ferramentas são suscetíveis a características intrínsecas à biblioteca, as quais são alheias aos parâmetros ajustáveis das ferramentas. Assim, uma investigação futura com o objetivo de descobrir quais aspectos afetam essas ferramentas é de extrema importância, pois podem auxiliar o desenvolvimento de novas estratégias nos algoritmos para minimizar estes problemas.

5.5 Perfil dos singletons incorretos

A análise do perfil dos singletons incorretos permitiu avaliar como estes estão distribuídos nos grupos do agrupamento de referência. Não foi detectada nenhuma concentração preferencial de singletons incorretos em grupos específicos de determinado tamanho. Estes foram distribuídos pelos grupos do agrupamento de referência de acordo com o esperado, ou seja, o número de singletons por grupo diminuiu à medida que o tamanho do grupo do agrupamento de referência foi aumentando.

Na biblioteca CEREBRO_15K, a super estimativa de singletons pelo CAP3 e TGICL alterou os perfis de distribuição dos singletons incorretos destas em relação aos perfis das outras ferramentas. Houve um balanço maior entre grupos pequenos e grupos com tamanho maior ou igual a 10 nos perfis do CAP3 e TGICL. Porém, não se pode afirmar que a diferença nos perfis do CAP3 e TGICL caracterizou uma distribuição preferencial dos singletons incorretos, para isso seria necessário um tratamento estatístico mais refinado dos dados. O mais provável é que as alterações verificadas nos dois perfis estejam em consonância com a super estimativa do número de singletons, sem desviar do comportamento esperado.

5.6 Análise de grupos discrepantes

A comparação entre o número de ESTs por grupo encontradas no agrupamento de referência e pelas ferramentas de agrupamento mostrou que, de maneira geral, os grupos não divergiram significativamente em comparação ao agrupamento de referência. Adicionalmente, este tipo de abordagem forneceu uma maneira rápida de se identificar quais os grupos de ESTs formados pelas ferramentas não se comportaram da maneira esperada, os quais foram denominados grupos discrepantes.

A análise do grupo discrepante da biblioteca CTRONCO_38K permitiu a verificação de eventos biológicos que afetam sobremaneira os resultados das ferramentas, as duplicações gênicas. Conforme foi visto na Tabela 11, onde se procedeu à análise detalhada do grupo mais discrepante para esta biblioteca, a ferramenta CAP3 uniu em um único grupo o que seriam nove grupos no agrupamento de referência. Já o restante das ferramentas uniu quinze grupos do agrupamento de referência. É importante ressaltar que para os dois casos estes grupos estão localizados em vários cromossomos, mas as ferramentas os uniram em um único grupo. Do ponto de vista computacional, constatou-se que de fato as ESTs apresentaram um nível de similaridade relativamente alto entre si, e por isso foram colocadas no mesmo grupo pelas ferramentas.

Para este grupo discrepante, uma análise mais detalhada das regiões no genoma humano que foram indevidamente agrupadas pelas ferramentas revelou que estas correspondem a uma família de genes que codificam a proteína actina (beta e gama), distribuídos nos cromossomos 1, 5 e 17. Do ponto de vista biológico, verificou-se que as ferramentas uniram ESTs oriundas de genes parálogos distribuídos por vários cromossomos no mesmo grupo, o que seria incorreto. No entanto, como as ferramentas só dispõem da informação da seqüência, torna-se extremamente difícil separar ESTs de genes parálogos, pois os erros de seqüenciamento das ESTs se confundem com as mutações nos genes.

Uma forma de evitar a união de ESTs de genes parálogos seria elevar o limiar de similaridade exigido para agrupá-las. Entretanto, essa medida, ao passo que separaria ESTs de genes parálogos, também poderiam causar o efeito adverso de separar ESTs que deveriam estar no mesmo grupo. Assim, essa alteração não permite atuar somente nas ESTs de genes parálogos. Conseqüentemente, deve-se buscar um balanço entre separação e união que melhore o resultado global.

Relembrando a terminologia introduzida por Burke (Burke et al., 1999) em relação aos possíveis erros de classificação das ESTs, temos que o erro tipo I ocorre quando ESTs do mesmo gene são colocadas em grupos diferentes, enquanto que o erro tipo II ocorre quando ESTs de genes distintos são colocadas no mesmo grupo. No caso descrito acima, temos que a incidência de parálogos induz a ocorrência do erro do tipo II, e qualquer tentativa para diminuir a sua incidência acarreta no aumento do erro do tipo I. De maneira geral, tentativas para mitigar um tipo de erro irão incrementar o outro tipo, o que é característico de qualquer processo de inferência estatística.

6 CONCLUSÕES E DIRECIONAMENTOS FUTUROS

Em face dos resultados obtidos, concluímos que a abordagem de comparação de ferramentas de agrupamento baseada no agrupamento de referência se mostrou bastante útil ao fornecer uma base de comparação única entre as ferramentas. Além disso, esta abordagem também permitiu que os comportamentos das métricas fossem comparados, fornecendo informações sobre as situações mais adequadas a cada uma delas.

A análise dos resultados das métricas de comparação de agrupamentos permitiu verificar que:

- as ferramentas apresentaram resultados satisfatórios na produção de grupos em relação ao agrupamento de referência;
- as ferramentas produziram resultados próximos entre si;
- a biblioteca de ESTs pode afetar drasticamente o resultado das ferramentas de agrupamento, como foi o caso do CAP3 e TGICL na biblioteca 0626;
- à luz dos critérios de avaliação adotados, a ferramenta XSACT foi aquela que consistentemente teve o melhor desempenho, entretanto, a diferença não foi significativa o bastante para se adotar uma ferramenta em particular.

A análise dos grupos discrepantes permitiu verificar que os eventos de duplicação gênica induzem as ferramentas de agrupamento de ESTs a erro, unindo transcritos diferentes no mesmo grupo. Vale ressaltar, no entanto, que as ferramentas não têm como tratar esse problema somente com base na informação de seqüência.

A própria natureza das ESTs indica que as ferramentas de agrupamento estão sujeitas a uma taxa inerente de erros quando somente a informação de seqüência é fornecida. Quais seriam alternativas para mitigar estes problemas? Primeiramente, seria interessante a possibilidade de se

verificar, sem qualquer informação adicional sobre o organismo estudado, quando um grupo possuiria um número maior de ESTs do que o esperado. Existem distribuições estatísticas esperadas para o nível de expressão gênica de uma célula (Kuznetsov et al., 2002) através das quais poderia se fazer um ajuste das distribuições produzidas pelas ferramentas para se identificar grupos discrepantes. Outra alternativa seria a criação de uma nova geração de ferramentas de agrupamento que não levassem em consideração somente a informação das seqüências, mas também informações adicionais sobre qual a provável função biológica das ESTs, incorporando, por exemplo, resultados de anotação genômica produzidos por outras ferramentas de Bioinformática, como o BLAST para a busca de similaridade contra bancos não redundantes, ou como o programa HMMER para predição de domínios no banco Pfam. Esta opção seria um contraponto à própria necessidade de se agrupar as ESTs, qual seja, a diminuição do número de seqüências para facilitar o processo de anotação genômica. Com o crescente aumento do poder computacional, esta opção torna-se viável para a construção de agrupamentos mais condizentes com a realidade. Vale ressalvar, no entanto, que em alguns casos somente as informações sobre função não são suficientes para reconstruir a realidade biológica. Como foi visto neste trabalho, alguns grupos estavam em cromossomos diferentes, apesar de todos estarem relacionados com a actina.

Outros aspectos relevantes para a análise de agrupamentos de ESTs seriam importantes para a criação de um modelo geral para melhor entender o processo. A avaliação das ferramentas em outros organismos permitiria verificar se o desempenho destas seria afetado. Talvez fosse possível fazer uma separação entre aspectos gerais que afetam o agrupamento, e aspectos ligados ao organismo. Adicionalmente, a inclusão de mais métricas tornaria a comparação das ferramentas mais consistente e confiável, além de permitir que fossem comparadas às métricas já existentes.

7 REFERÊNCIAS BIBLIOGRÁFICAS

Adams, M. D., J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, R. F. Moreno, and . "**Complementary DNA sequencing: expressed sequence tags and human genome project**." Science 252.5013 (1991): 1651-56.

Alberts, B, A Johnson, J Lewis, M Raff, K Roberts, and P Walter. "**Manipulating Proteins, DNA, and RNA**." Molecular Biology of THE CELL. 4th ed. New York: Garland Science, 2002. 469-546.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. "**Basic local alignment search tool**." J.Mol.Biol. 215.3 (1990): 403-10.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. "**Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**." Nucleic Acids Res. 25.17 (1997): 3389-402.

Anderson, S. "**Shotgun DNA sequencing using cloned DNase I-generated fragments**." Nucleic Acids Res. 9.13 (1981): 3015-27.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp, and D. L. Wheeler. "**GenBank**." Nucleic Acids Res. 30.1 (2002): 17-20.

Boguski, M. S., T. M. Lowe, and C. M. Tolstoshev. "**dbEST--database for "expressed sequence tags"**." Nat.Genet. 4.4 (1993): 332-33.

Boguski, M. S., C. M. Tolstoshev, and D. E. Bassett, Jr. "**Gene discovery in dbEST**." Science 265.5181 (1994): 1993-94.

Brandenberger, R., H. Wei, S. Zhang, S. Lei, J. Murage, G. J. Fisk, Y. Li, C. Xu, R. Fang, K. Guegler, M. S. Rao, R. Mandalam, J. Lebkowski, and L. W. Stanton. "**Transcriptome characterization elucidates signaling networks that control human ES cell growth and differentiation**." Nat.Biotechnol. 22.6 (2004): 707-16.

Burke, J., D. Davison, and W. Hide. "**d2_cluster: a validated method for clustering EST and full-length cDNAsequences**." Genome Res. 9.11 (1999): 1135-42.

Chevreux, B., T. Pfisterer, B. Drescher, A. J. Driesel, W. E. Muller, T. Wetter, and S. Suhai. "**Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs**." Genome Res. 14.6 (2004): 1147-59.

Christoffels, A., Gelder A. van, G. Greyling, R. Miller, T. Hide, and W. Hide. "**STACK: Sequence Tag Alignment and Consensus Knowledgebase**." Nucleic Acids Res. 29.1 (2001): 234-38.

Ewing, B. and P. Green. "**Base-calling of automated sequencer traces using phred. II. Error probabilities**." Genome Res. 8.3 (1998): 186-94.

Ewing, B., L. Hillier, M. C. Wendl, and P. Green. "**Base-calling of automated sequencer traces using phred. I. Accuracy assessment**." Genome Res. 8.3 (1998): 175-85.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, and . "**Whole-genome random sequencing and assembly of Haemophilus influenzae Rd**." Science 269.5223 (1995): 496-512.

Gardner, R. C., A. J. Howarth, P. Hahn, M. Brown-Luedi, R. J. Shepherd, and J. Messing. "**The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing**." Nucleic Acids Res. 9.12 (1981): 2871-88.

Giergerih, R, S Jurtz, and J Stoye. "**Efficient implementation of lazy suffix trees**." Softw.Pract.Exper. 33 (2003): 1035-49.

Gotoh, O. "**An improved algorithm for matching biological sequences**." J.Mol.Biol. 162.3 (1982): 705-08.

Green, E. D. "**Strategies for the systematic sequencing of complex genomes**." Nat.Rev.Genet. 2.8 (2001): 573-83.

Guy St.C.Slater. "**Algorithms for the Analysis of Expressed Sequence Tags**." Diss. University of Cambridge, 2000.

Heber, S., M. Alekseyev, S. H. Sze, H. Tang, and P. A. Pevzner. "**Splicing graphs and EST assembly problem**." Bioinformatics. 18 Suppl 1 (2002): S181-S188.

Hide, W, J Burke, and R Miller. "**A Novel Approach Towards a Comprehensive Consensus Representation of the Expressed Human Genome**". Proceedings of the Eighth Workshop on Genome Informatics Tokyo, Japan: Universal Academy Press Inc., 1997.V 187-96.

Hide, W., J. Burke, and D. B. Davison. "**Biological evaluation of d2, an algorithm for high-performance sequence comparison**." J.Comput.Biol. 1.3 (1994): 199-215.

Hu, G., B. Modrek, H. M. Riise Stensland, J. Saarela, P. Pajukanta, V. Kustanovich, L. Peltonen, S. F. Nelson, and C. Lee. "**Efficient discovery of single-nucleotide polymorphisms in coding regions of human genes**." Pharmacogenomics.J. 2.4 (2002): 236-42.

Huang, X. "**A contig assembly program based on sensitive detection of fragment overlaps**." Genomics 14.1 (1992): 18-25.

Huang, X. and A. Madan. "**CAP3: A DNA sequence assembly program**." Genome Res. 9.9 (1999): 868-77.

Jaccard, S. "**Nouvelles recherches sur la distribution florale.**" Bull.Soc.Vaud.Sci.Nat.44 (1908): 223-70.

Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. "**Repbase Update, a database of eukaryotic repetitive elements**." Cytogenet.Genome Res. 110.1-4 (2005): 462-67.

Kalyanaraman, A., S. Aluru, S. Kothari, and V. Brendel. "**Efficient clustering of large EST data sets on parallel computers**." Nucleic Acids Res. 31.11 (2003): 2963-74.

Khan, A. S., A. S. Wilcox, M. H. Polymeropoulos, J. A. Hopkins, T. J. Stevens, M. Robinson, A. K. Orpana, and J. M. Sikela. "**Single pass sequencing and physical and genetic mapping of human brain cDNAs**." Nat.Genet. 2.3 (1992): 180-85.

Kuznetsov, V. A., G. D. Knott, and R. F. Bonner. "**General statistics of stochastic process of gene expression in eukaryotic cells**." Genetics 161.3 (2002): 1321-32.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, Bastide M. de la, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A.

Wetterstrand, A. Patrinos, M. J. Morgan, Jong P. de, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, and Y. J. Chen. "**Initial sequencing and analysis of the human genome**." Nature 409.6822 (2001): 860-921.

Lee, C. "**Generating consensus sequences from partial order multiple sequence alignment graphs**." Bioinformatics. 19.8 (2003): 999-1008.

Liolios, K., N. Tavernarakis, P. Hugenholtz, and N. C. Kyrpides. "**The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide**." Nucleic Acids Res. 34.Database issue (2006): D332-D334.

Maglott, D., J. Ostell, K. D. Pruitt, and T. Tatusova. "**Entrez Gene: gene-centered information at NCBI**." Nucleic Acids Res. 33.Database issue (2005): D54-D58.

Malde, K., E. Coward, and I. Jonassen. "**Fast sequence clustering using a suffix array algorithm**." Bioinformatics. 19.10 (2003): 1221-26.

Manber, U. and G. Myers. "**Suffix arrays: a new method for on-line string searches**." SIAM J.Comput.22 (1993): 935-48.

Miller, C., J. Gurd, and A. Brass. "**A RAPID algorithm for sequence database comparisons: application to the identification of vector contamination in the EMBL databases**." Bioinformatics. 15.2 (1999): 111-21.

Modrek, B. and C. Lee. "**A genomic view of alternative splicing**." Nat.Genet. 30.1 (2002): 13-19.

Modrek, B., A. Resch, C. Grasso, and C. Lee. "**Genome-wide detection of alternative splicing in expressed sequences of human genes**." Nucleic Acids Res. 29.13 (2001): 2850-59.

Mott, R. "**EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA**." Comput.Appl.Biosci. 13.4 (1997): 477-78.

Needleman, S. B. and C. D. Wunsch. "**A general method applicable to the search for similarities in the amino acid sequence of two proteins**." J.Mol.Biol. 48.3 (1970): 443-53.

Ning, Z., A. J. Cox, and J. C. Mullikin. "**SSAHA: a fast search method for large DNA databases**." Genome Res. 11.10 (2001): 1725-29.

Ogasawara, J. and S. Morishita. "**Fast and sensitive algorithm for aligning ESTs to human genome**." Proc.IEEE Comput.Soc.Bioinform.Conf. 1 (2002): 43-53.

Pearson, W. R. and D. J. Lipman. "**Improved tools for biological sequence comparison**." Proc.Natl.Acad.Sci.U.S.A 85.8 (1988): 2444-48.

Pertea, G., X. Huang, F. Liang, V. Antonescu, R. Sultana, S. Karamycheva, Y. Lee, J. White, F. Cheung, B. Parvizi, J. Tsai, and J. Quackenbush. "**TIGR Gene Indices clustering tools**

(TGICL): a software system for fast clustering of large EST datasets." Bioinformatics. 19.5 (2003): 651-52.

Picoult-Newberg, L., T. E. Ideker, M. G. Pohl, S. L. Taylor, M. A. Donaldson, D. A. Nickerson, and M. Boyce-Jacino. "**Mining SNPs from EST databases**." Genome Res. 9.2 (1999): 167-74.

Pontius, JU, L Wagner, and GD Schuler. "**UniGene: a Unified View of the Transcriptome**." The NCBI Handbook. Ed. J McEntyre and J. Ostell. National Library of Medicine(US), NCBI, 2003. 1-12.

Pruitt, K. D., T. Tatusova, and D. R. Maglott. "**NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**." Nucleic Acids Res. 33.Database issue (2005): D501-D504.

Ptitsyn, A. "**New Algorithms for EST Clustering**." Diss. University of the Western Cape, 2000.

Quackenbush, J., F. Liang, I. Holt, G. Pertea, and J. Upton. "**The TIGR gene indices: reconstruction and representation of expressed gene sequences**." Nucleic Acids Res. 28.1 (2000): 141-45.

Sanger, F., A. R. Coulson, G. F. Hong, D. F. Hill, and G. B. Petersen. "**Nucleotide sequence of bacteriophage lambda DNA**." J.Mol.Biol. 162.4 (1982): 729-73.

Sikela, J. M. and C. Auffray. "**Finding new genes faster than ever**." Nat.Genet. 3.3 (1993): 189-91.

Smit, AFA, Hubley, R, and Green, P. RepeatMasker Open-3.0. 1996.
Ref Type: Unpublished Work

Smith, T. F. and M. S. Waterman. "**Identification of common molecular subsequences**." J.Mol.Biol. 147.1 (1981): 195-97.

Sutton, G, O White, MD Adams, and AR Kerlavage. "**TIGR Assembler: a new tool for assembling large shotgun sequencing projects.**" Genome Sci.Technol.1 (1995): 9-18.

Torney, DC, C Burkes, D Davidson, and KM Sirkin. "**Computation of D2: a Measure of Sequence Dissimilarity**". New York: Addison-Wesley, 1990.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. bu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, Francesco Di, V, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y.

Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, and M. Nodell. "**The sequence of the human genome**." Science 291.5507 (2001): 1304-51.

Vinga, S. and J. Almeida. "**Alignment-free sequence comparison-a review**." Bioinformatics. 19.4 (2003): 513-23.

Wang, J. P., B. G. Lindsay, J. Leebens-Mack, L. Cui, K. Wall, W. C. Miller, and C. W. dePamphilis. "**EST clustering error evaluation and correction**." Bioinformatics. 20.17 (2004): 2973-84.

Waterman, M. S. "**Databases and Rapid Sequence Analysis**." INTRODUCTION TO COMPUTATIONAL BIOLOGY: Maps, sequences and genomes. First ed. Chapman & Hall/CRC, 1995. 161-81.

Wheelan, S. J., D. M. Church, and J. M. Ostell. "**Spidey: a tool for mRNA-to-genomic alignments**." Genome Res. 11.11 (2001): 1952-57.

Wolfsberg, TG and D Landsman. "**EXPRESSED SEQUENCE TAGS (ESTs)**." Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. Ed. AD Baxevanis and BFF Ouellette. Second Edition ed. John Wiley & Sons, Inc., 2001. 283-301.

Wu, X, W Lee, and C Tseng. "**ESTmapper: Efficiently Aligning DNA Sequences to Genomes**". Fourth IEEE International Workshop on High Performance Computational Biology 2005.

Xu, Q., B. Modrek, and C. Lee. "**Genome-wide detection of tissue-specific alternative splicing in the human transcriptome**." Nucleic Acids Res. 30.17 (2002): 3754-66.

Zhang, Z., S. Schwartz, L. Wagner, and W. Miller. "**A greedy algorithm for aligning DNA sequences**." J.Comput.Biol. 7.1-2 (2000): 203-14.