Método para Auxiliar a Interpretação de *Clusters* de Expressão Gênica considerando Sumarização Automática

Daniane S. de Paula[1], **Alessandra A. Macedo**[1]

[1]Programa Interunidades em Bioinformática
Grupo de Informática Biomédica
Departamento de Computação e Matemática (DCM)
Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP-USP)
Av. dos Bandeirantes, 3900 – Campus da USP – Ribeirão Preto-SP – Brasil

daniane@usp.br, ale.alaniz@usp.br

Resumo. *Este artigo apresenta o método SARI para auxiliar a análise de interações entre genes ou produtos gênicos. Esta análise é realizada por meio de consultas a literatura, que é sumarizada devido a sua grande extensão, visando o benefício do leitor. Apresenta-se também cenários de aplicação do SARI, descrevendo desde a escolha de um conjunto de dados de expressão gênica até a sumarização dos textos que tentam caracterizar os clusters gênicos formados e uma forma genérica de utilizar a sumarização no contexto biomédico, não considerando o método utilizado para a obtenção do grupo de genes.*

1. Introdução

Experimentos clássicos da genética revelaram que todas as células de um organismo possuem o mesmo conteúdo de DNA [Strachan and Read 1999]. Apesar de possuírem exatamento o mesmo DNA, as células de um organismo complexo se diferenciam e executam funções diferentes. As células executam as funções necessárias para a manutenção da vida do organismo ao expressar genes diferentes, os quais são apropriados para cada situação, tecido, etc. Genes são segmentos de DNA, que contêm informações para codificar as proteínas e RNAs necessários para o funcionamento da célula. Os padrões de expressão gênica se alteram de acordo com o estado fisiológico da célula, assim genes são ativados ou inativados nos processos de crescimento, divisão, respostas ao ambiente. Pode-se monitorar a expressão gênica utilizando técnicas de *microarrays* de DNA. Dados provenientes de *microarrays* representam o nível de atividade de milhares de genes simultaneamente em um ambiente bioquímico. A possibilidade de medir como os genes se comportam em um momento contribuiu para o entendimento de processos celulares, tratamento e diagnóstico de doenças e desenvolvimento de drogas [Kankar et al. 2002]. Cada experimento de *microarray* possui uma quantidade enorme de dados. Um dos principais objetivos da análise de *microarrays* é agrupar genes com perfil de expressão gênica similares. O desenvolvimento de técnicas de análise de proteínas, DNA e RNA tem gerado o crescimento exponencial de dados biomoleculares. Para a promoção de avanços científicos, é fundamental a transformação dos dados gerados em informação e conhecimento.

Clustering é um tipo de aprendizado de máquina não-supervisionado utilizado na análise de *microarrays* de DNA. Algoritmos de *clustering* agrupam dados de acordo com similaridades, contudo métodos não-supervisionados exigem análises posteriores dos

grupos gerados [Monard and Baranauskas 2003]. Técnicas como *clustering* envolvem grande quantidade de dados, os quais necessitam ser analisados dentro de um contexto, implicando em eventuais consultas a literatura *online*. Na Internet, o enorme volume de dados e de literatura disponível dificulta a pesquisa de informações. Por exemplo, o desuso da nomenclatura gênica oficial é um dos problemas mais comuns para a busca de informações de genes e seus produtos em trabalhos científicos. Há casos de (i) artigos com nomenclatura obsoleta e (ii) autores que não especificam se fazem referência ao gene ou a proteína resultante, etc [Splendore 2005]. Nesse cenário, a utilização de buscas avançadas, ferramentas de relacionamento automático de informações e sumários automáticos podem se tornar interessantes. A sumarização automática extrai conteúdo de uma fonte de informação e apresenta somente o assunto mais importante. Considerando o grande volume de publicações científicas, a tarefa de identificar, selecionar e analisar textos de interesse tornou-se uma tarefa difícil. Assim, utilizar sumários é um recurso interessante, uma vez que possibilita obter o conteúdo relevante de um texto, de forma condensada [Mani 2001] [Pardo et al. 2002].

O sistema LAKE é um trabalho relacionado que utiliza TF/IDF, a posição dos termos no documento e um classificador *Naive Bayes* para selecionar as candidatas as frases-chave do documento a ser sumarizado [D'Avanzo et al. 2004]. Nenkova(05) discute o impacto da frequência de termos na sumarização e o papel da frequência em um sistema de sumarização. O trabalho conclui que a frequência no texto fonte é um forte indicativo de uma palavra estar em um sumário feito por um humano. No entanto, a frequência não explica completamente as escolhas humanas. O SUMMARIST identifica e interpreta tópicos centrais do texto original e gera sumários [Nenkova and Vanderwende 2005].

Apoiado na mesma problemática proposta do presente trabalho, o MedMeSH Summarizer auxilia pesquisadores a criar referências cruzadas experimentais e analíticas de microarrays. O resultado apresentado pelo MedMeSH é uma lista ordenada por frequência de termos MeSH para a lista de genes que o usuário inseriu na entrada [Kankar et al. 2002]. Seguindo as motivações deste artigo e do MedMeSH, imagina-se que pesquisadores estão continuamente interessando em fazer comparação de novos resultados com fatos biológicos previamente conhecidos, teorias bem estabelecidas e resultados anteriores. Bases de dados com literatura biológica e médica fornecem um grande depósito de conhecimento para estas comparações. No entanto, o tamanho dessas bases torna a tarefa de fazer referências cruzadas lenta, tediosa e desencorajadora.

Este trabalho apresenta o método SARI (**S**umarização Automática de **A**rtigos Científicos para **R**epresentar o significado de **I**nterações Gênicas), o qual foi desenvolvido com objetivo de auxiliar na definição de significado a grupos de genes que interagem na maquinaria celular. Para alcançar esse objetivo o SARI foi proposto pela composição dos seguintes processos: (i) análise de dados de expressão gênica; (ii) consulta a literatura científica, na busca de informações que expliquem os resultados do processo (i); e (iii) apresentação sumarizada dos resultados. Na tentativa de validar o SARI, o método foi instanciado pelas seguintes etapas no cenário de auxílio a análise de *clusters* de expressão gênica: conversão de formato; *clustering*; consultas de interações gênicas na literatura; busca e recuperação do conteúdo dos artigos no PubMed e sumarização de artigos para facilitar a visualização dos resultados. Um outro cenário experimentado foi a pesquisa de um conjunto de genes para buscar interações entre eles. Neste cenário, não foi consi-

derada a maneira como os dados de expressão gênica foram obtidos ou agrupados, pois essas etapas ficaram sob responsabilidade do usuário.

Em relação a materiais e métodos, diferentes abordagens de sumarização automática foram investigadas para verificar suas adaptações a artigos científicos que contêm nomenclatura de genes. Nesta proposta, a sumarização foi guiada pela presença dos nomes dos genes nas sentenças. O artigo está organizado do seguinte modo: a Seção 2 descreve o método SARI e aplicações; a Seção 3 apresenta os experimentos e resultados; e a Seção 4 apresenta conclusões.

2. Método SARI e Aplicações

O método SARI auxilia a interpretação da interação entre genes ou produtos gênicos. O SARI, ilustrado na Figura 1, possui quatro passos principais: (1a) obtenção de dados gênicos, que são submetidos a processos de análise de dados ou (1b) inserção de um conjunto de genes; (2) consultas a literatura científica, que reforcem ou contradigam os resultados da análise de dados; (3) sumarização automática dos textos consultados e (4) apresentação ao usuário a literatura de interesse de forma condensada.

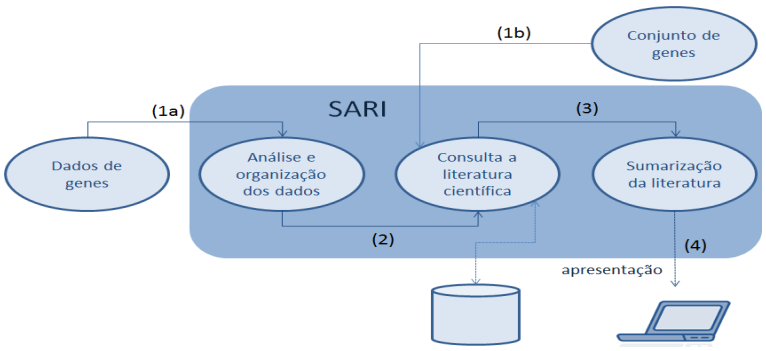

Figura 1. Método SARI.

No método SARI, o primeiro passo para analisar possíveis interações gênicas e proteicas é obter os dados biológicos(ver Figura 1(1a)). Pesquisadores podem obter dados biológicos em experimentos feitos em laboratórios próprios ou parceiros. Nesses experimentos, um grande volume de dados biológicos é gerado, os quais podem ser armazenados em grandes bases de dados online. Por exemplo, existem dados de expressão gênica disponíveis em bases de dados como o GEO (*Gene Expression Omnibus*) ou o *ArrayExpress*. Após obter os dados biológicos, o SARI encontra-se "rico em dados, mas pobre em informação". Assim, a análise de dados deve ser executada, a partir de métodos que permitam descrever fatos, detectar padrões e desenvolver explicações [Levine 1996]. Com a análise de dados, espera-se obter informação útil (ver Figura 1(2)).

A consulta a literatura objetiva verificar os resultados obtidos na análise de dados (ver Figura 1(2)). Essa etapa pode ser também abastecida pela entrada de informações classificadas por processos externos ao SARI (ver Figura 1(1b)). Em ambas as situações, é importante relacionar os dados de expressão gênica com informações biológicas presentes na literatura para verificar a fundamentação científica dos genes agrupados. Ao estabelecer uma relação entre os dados de expressão com informações externas, consegue-se agregar conhecimento ou fazer novas descobertas sobre os processos biológicos [Babu 2004].

O conhecimento prévio publicado na literatura pode reforçar ou contradizer os resultados obtidos na análise de dados. Desta maneira, pode-se também gerar novos focos de estudo para resultados obtidos, mas não presentes no estado da arte. Algumas questões que podem ser abordadas após a análise de dados são predizer: sítios de ligação, interações e funções gênicas e proteicas, módulos conservados e redes regulatórias.

Ao consultar um assunto de interesse na literatura, a quantidade de informação retornada pode ser bastante extensa. Consequentemente, é difícil para o usuário assimilar tanta informação sem se sobrecarregar ou até mesmo ficar perdido. Portanto, propôs-se uma aplicação de sumarização automática de textos como uma das etapas do método SARI (ver Figura 1(3)). A sumarização automática permite reduzir a quantidade de conteúdo textual, sem que a informação principal do texto seja perdida. O método de sumarização utilizada é uma adaptação do método de palavras-chave, o qual está baseado na premissa de que autor do texto usa algumas palavras-chave para expressar suas idéias e essas palavras se repetem ao longo do texto. No entanto, o interesse neste trabalho é identificar as interações nos artigos, e não o tópico central da publicação. Portanto, palavras-chave são os nomes dos genes ou quaisquer palavras, que identifiquem um gene, como símbolo oficial, símbolos e nomes alias, símbolos e nomes prévios ou não oficiais. Logo, a sumarização consistiu em: (a) identificar sentenças relevantes que possuam os nomes ou símbolos ou palavras que identificam um gene; (b) extrair do texto original as sentenças de interesse identificadas; e (c) justapor as sentenças para compor o sumário.

Finalmente, a última etapa do SARI é apresentar os sumários ao usuário (ver Figura 1(4)). Além dos sumários, elaborou-se uma rede das interações identificadas. Assim, além do texto exibe-se uma visualização gráfica dos resultados. Os nós da rede são os genes e as arestas que ligam dois genes são os sumários dos artigos cujos as interações estão descritas, como mostrado na Figura 2. Assim, obtém-se uma integração visual e textual que auxiliará na atribuição de significado aos clusters. A seguir, cenários de instanciação do método SARI nas aplicações SARI-DU e SARI-MD são apresentados.

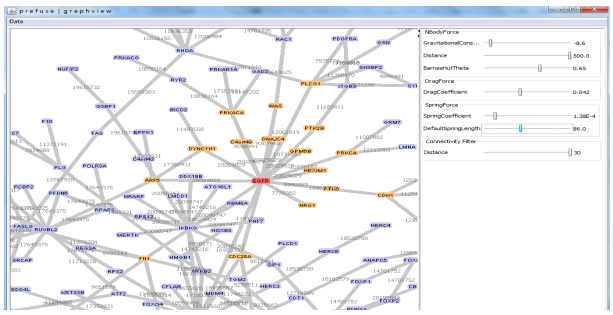

Figura 2. Exemplo de Apresentação de uma Rede de Interações.

SARI-DU. A primeira aplicação do SARI, *SARI-DU*, gerou resultados com sumarização apresentada em um único documento. A Figura 3 apresenta os mecanismos dessa instanciação. Seguindo o método SARI, o primeiro passo foi obter um conjunto de dados de expressão gênica. Utilizou-se dados da base de dados de expressão gênica GEO, conforme apresentado na Figura 3**(a)**. Houve um pré-processamento dos dados para utilizá-los com as ferramentas da Weka [Hall et al. 2009]. A Weka manipula arquivos no formato ARFF (*Attribute-Relation File Format*) e o GEO disponibiliza seus dados no formato SOFT. Após adquirir o conjunto de dados do GEO, foi necessário aplicar um

algoritmo de conversão do formato SOFT para formato ARFF (Figura 3**(b)**). O arquivo ARFF possui duas seções: cabeçalho (nome da relação) e dados (uma lista de atributos).

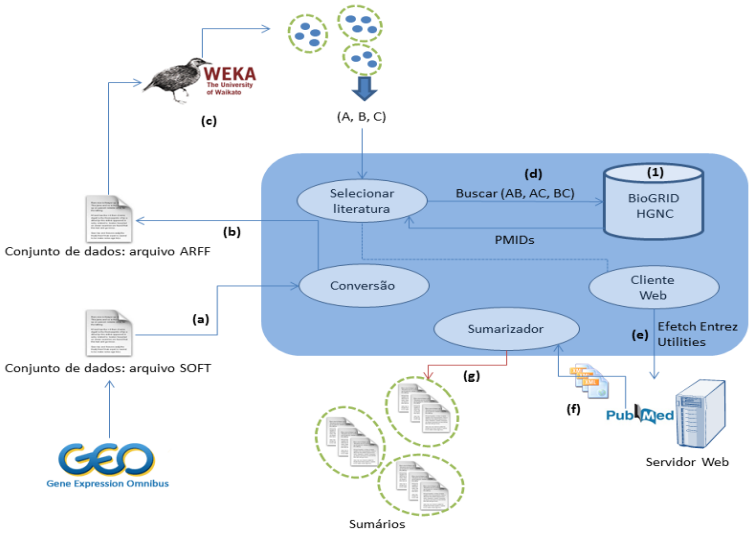

Figura 3. Instanciação do método SARI com sumarização em documento único - SARI-DU.

O próximo passo do SARI (Figura 3**(c)**), foi aplicar o algoritmo de *clustering* k-means da Weka[1]. Nessa etapa, a entrada foi o conjunto de dados de expressão gênica anteriormente pré-processada e a saída os genes agrupados. Para a aplicação SARI-DU, criou-se um banco de dados seguindo o modelo de entidade-relacionamento para armazenar as seguintes informações da base de dados de interações gênicas e proteicas BioGRID [Stark et al. 2005]: (i) nome, símbolo oficial e *aliases*; (ii) quais genes interagem entre si e em qual publicação científica essas interações estão demonstradas; e (iii) as informações de nomenclatura gênica do HGNC (comitê que regulamenta nomes e símbolos para genes humanos).

Após agrupar os genes com o algoritmo de *clustering* da Weka, consultou-se o banco de dados para identificar artigos que contivessem interações entre os genes (Figura 3**(d)**). Dado um *cluster* que contenha os genes [*A, B, C*], procurou-se no banco de dados as possíveis combinações de interações sem repetições dos genes do *cluster*, ou seja, artigos que apresentassem textualmente as interações para as combinações: *AB*, *AC*, *BC*. O resultado da consulta foi uma lista de PMIDs (*PubMed Identifier*) de artigos. Com os PMIDS, a etapa **(e)** na Figura 3 representa a consulta ao PubMed utilizando a ferramenta EFetch do PubMed.

O uso do EFetch forneceu os arquivos XML dos artigos, os quais passaram pelo processo de sumarização. Foi necessário também um processamento dos arquivos XML, para que o texto de interesse fosse extraído. Aplicou-se uma abordagem superficial de sumarização nos artigos recuperados (Figura 3**(g)**). Uma interação entre dois genes pode estar relatada em diversos artigos, mas nessa abordagem utilizou-se apenas um artigo para cada interação, escolhido arbitrariamente. Todas as sentenças classificadas como

[1]A Weka possui vários algoritmos para aprendizado de máquina, incluindo algoritmos de *clustering* [Hall et al. 2009]

de interesse fizeram parte do sumário apresentado em um único documento sem taxa de compressão.

SARI-MD. Pesquisadores podem trabalhar com um conjunto de dados expressão gênica já agrupado. Também há diversos algoritmos de *clustering* que podem ser aplicados, além do k-means. Portanto, diferentemente do cenário da aplicação do SARI-DU, desejava-se que a análise de interações gênicas por meio de sumarização fosse aplicada não importando os métodos pelos quais foram obtidos os *clusters* de genes. Para a segunda instanciação do método SARI, *SARI-MD*, deseja-se obter sumários de artigos que contenham interações entre genes, sendo invisível o método utilizado para chegar nesse conjunto de genes.

Outra ampliação de abrangência do método SARI, foi considerar que uma interação entre dois genes pode estar descrita em vários artigos. Assim, a sumarização na aplicação SARI-MD foi projetada para suportar múltiplos documentos. Utilizando-se múltiplos documentos, há uma cobertura maior da literatura sobre determinada interação. Na sumarização codificada para esta aplicação, adotou-se o conceito de taxa de compressão. Como foram utilizados múltiplos documentos, sem taxa de compressão poderia ocorrer a produção de sumários muito extensos, o que contradiria o objetivo do método. Para selecionar dentre todas as sentenças de interesse (que contém palavras-chave), quais pertencerão ao sumário utilizou-se três critérios: quantidade de palavras-chave, data da publicação do artigo e tamanho da sentença. As sentenças em que aparecem mais vezes nomes, símbolos ou alias são consideradas mais importantes. As sentenças que pertencem a artigos mais recentes contém uma informação mais atualizada. As sentenças menores foram privelegiadas, pois elas transmitem informação de uma forma mais ágil e concisa para o usuário. Portanto, a aplicação SARI-MD é mais abrangente, já que recupera, sumariza e apresenta conteúdo variado da literatura científica. Além disso, o segundo cenário permite mais liberdade ao usuário, que pode agrupar os genes seguindo métodos que lhe pareça mais convenientes.

3. Experimentação e Resultados

As principais etapas do método SARI são análise de dados com *clustering* e processamento da literatura para apresentação sumarizada. Nesta seção são apresentados os experimentos realizados e os resultados obtidos para essas principais etapas do método SARI nas duas aplicações apresentadas anteriormente, SARI-DU e SARI-MD. Não há experimentos de *clustering* na aplicação SARI-MD, pois o usuário tem a liberdade de obter os *clusters* gênicos da maneira que lhe for mais conveniente.

3.1. Análise de Dados com Clusterização

Na experimentação da aplicação SARI-DU, utilizou-se o seguinte conjunto de dados de expressão gênica proveniente do GEO: *NOTCH antagonist SAHM1 effect on T-ALL cell lines*. Este conjunto de dados é originário de uma pesquisa que relata o processo de desenho de peptídeos, os quais são alvos críticos na interface proteína-proteína do complexo NOTCH. Proteínas NOTCH participam de vias conservadas que regulam a diferenciação, proliferação e morte celular. Normalmente, a duração e a força da sinalização do NOTCH é rigidamente controlada. Quando ocorrem mutações de perda de função, são observadas diversas doenças. Já mutações de ganho de função na via NOTCH são relacionadas

ao desenvolvimento de câncer. A ativação inapropriada do receptor NOTCH está diretamente ligada a várias patogêneses, inclusive a leucemia linfoblástica aguda. O tratamento de células leucêmicas com o peptídeo SAHM1 resultou na supressão dos genes NOTCH ativados. É demonstrado que o peptídeo SAHM1 previne a montagem do complexo de transcrição ativo [Moellering et al. 2009].

O conjunto de dados de expressão gênica foi convertido do formato SOFT para ARFF. No conjunto de expressão gênica, aplicou-se o algoritmo de *clustering* k-means da Weka. Os parâmetros utilizados foram similares aos da ferramenta de *clustering* e visualização do GEO, o valor escolhido para *k* foi 15, logo os genes foram agrupados em 15 *clusters* distintos e utilizou-se a distância euclidiana para calcular a distância entre os genes. A quantidade de genes agrupados para cada um dos *cluster* foi: 2952, 989, 1043, 1209, 757, 1106, 258, 1054, 73, 1025, 1909, 938, 1119, 1068 e 1340 respectivamente. O *cluster* 1 possui mais genes e o *cluster* 9 agrupou o menor número de genes. A atribuição de significado a esses clusters está na Seção 3.2.

3.2. Experimentação de Consulta a Literatura com Sumarização Automática

Para a experimentação da sumarização na aplicação SARI-DU, precisa-se identificar os genes de um mesmo *cluster* que possuem interação. As interações são buscadas na base de dados pelo número identificador PubMed de um artigo, o qual contém a descrição da interação entre esses genes. O número de interações confirmadas por artigos científicos do PubMed em cada *cluster*: i. 2194 de 2952, ii. 220 de 989, iii. 203 de 1043, iv. 317 de 1209, v. 98 de 757, vi. 255 de 1106, vii. 25 de 258, viii. 290 de 1054, ix. 3 de 73, x. 218 de 1025, xi. 996 de 1909, xii. 173 de 938, xiii. 377 de 1119, xiv. 202 de 1068 e xv. 482 de 1340. Cada interação está contida em um artigo, em alguns casos, um mesmo artigo pode conter mais de um par de genes interagindo. Nesse caso, o artigo é contabilizado duas vezes, uma vez que as sentenças de interesse irão mudar de acordo com a mudança dos genes. Com esses números de interações por *cluster*, pode-se observar a grande quantidade de informação na literatura envolvida na análise de um conjunto de dados de expressão gênica clusterizado. Por exemplo, observou-se que o algoritmo de *clustering* k-means, agrupou os genes *HCN4* e *RYR2* no mesmo *cluster*. No entanto, segundo as informações do BioGRID não há artigo do PubMed que demonstre interação entre esses genes ou entre seus produtos. Já a interação entre *HCN4* e *HCN2*, definida pelo SARI, aparece em um artigo científico com PMID 12034718.

As informações de interesse para sumarização dos artigos foram o título e o resumo. Considerando esses metadados, o sumarizador compõe o vetor de palavras-chave para cada artigo. O vetor de palavras-chave é formado pelo símbolo oficial, nome oficial, símbolos *aliases*, nomes *aliases*, símbolos prévios e nomes prévios. Um exemplo de um vetor de palavras-chave do artigo é: *[stathmin 1, STMN1, oncoprotein 18, MGC138870, MGC138869, C1orf215, Lag, LAP18, FLJ32206, OP18, PR22, PP19, PP17, SMN, chromosome 1 open reading frame 215, stathmin 1 oncoprotein 18, heat shock 70kDa protein 8, HSPA8, HSC70, HSP73, HSPA10, HSC54, HSC71, HSP71, MGC29929, MGC131511, NIP71, LAP1, heat shock 70kD protein 8]*. Esse vetor contém o nome, o símbolo oficial dos genes STMN1 e HSPA8, os seus nomes e os símbolos *aliases* e prévios. O processo de sumarização identifica sentenças que possuem uma ou mais palavras-chave do vetor. As sentenças identificadas são concatenadas e as sentenças que não contém palavras-chave são descartadas.

Dois experimentos foram realizados para medir a redução dos textos utilizando os *clusters* e as interações. No primeiro experimento, o vetor de palavras-chave possuía somente nomes e símbolos oficiais dos genes ((Sumarização sem *alias*)). No segundo, o vetor de palavras possuía nomes e símbolos oficiais e não oficiais ((Sumarização com *alias*)). Os dois experimentos foram feitos com o mesmo conjunto de dados *NOTCH antagonist SAHM1 effect on T-ALL cell lines*, do GEO. Nos experimentos de sumarização sem *alias* e com *alias*, o texto original é a concatenção do título e do resumo do artigo.

Sumarização sem *alias*. No experimento sem considerar os *aliases*, o vetor de palavras-chave usado para identificar as sentenças de interesse continha apenas nomes e símbolos oficiais dos genes que estão interagindo. Assim, o sumário foi formado somente com sentenças que continham a nomenclatura oficial do par de genes de interesse. A Figura 4(a) apresenta a quantidade média de sentenças nos textos originais e nos sumários. O tamanho médio dos textos fontes é 9,4 sentenças. Com a sumarização, obtém-se textos com tamanho médio de 4,7 sentenças. O *cluster* 9 possui 73 genes e apenas três pares de genes interagindo. Ao utilizar somente nomes e símbolos oficiais, nenhum dos três artigos do *cluster* 9 formou sumário, pois título e resumo não citam o nome ou símbolo oficial dos genes que possuem a interação caracterizada no artigo. Logo, infere-se que os autores usaram nomes e símbolos não oficiais para identificar os genes.

Sumarização com *alias*. No experimento de sumarização do SARI considerando os *alias* dos genes, o vetor de palavras-chave continha nomes e símbolos oficiais dos genes, assim como outros nomes e símbolos não oficiais para identificar os genes. Na Figura 4(b), observa-se a quantidade média de sentenças nos textos originais e nos sumários, a quantidade média de sentenças no texto original é 9,4, no sumário foi 6,2.

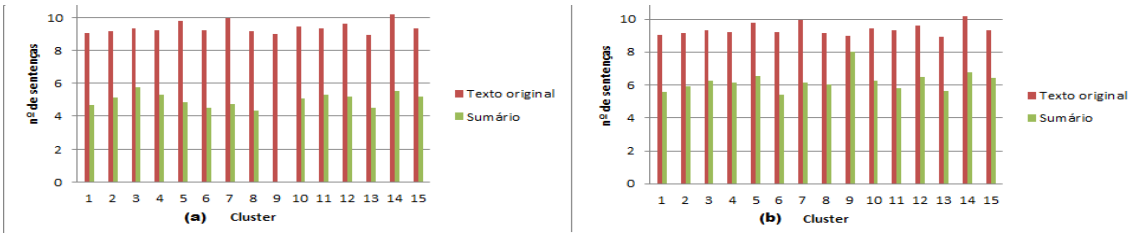

Figura 4. Quantidade média de sentenças nos textos originais e nos sumários: (a) sem alias e (b) com alias.

A Figura 5 ilustra a quantidade de artigos, que segundo o BioGRID, possuem a interação entre dois genes, porém nenhum sumário foi formado. A coluna *oficiais* indica o vetor formado pelo nome e símbolo oficial do par de genes que interagem; a coluna *oficiais e alias* representa o vetor de palavras-chave e contém nomenclatura oficial e *alias*. Pode-se observar que ao utilizar exclusivamente a nomenclatura oficial, há uma grande quantidade de sumários não formados. Porém, quando utiliza-se os *aliases*, o número de sumários não formados diminui. Essa constatação indica que a nomenclatura oficial é pouco utilizada e os pseudônimos dos genes estão presentes na literatura. Por exemplo, observa-se na Figura 4(a), que existem não sumários no *cluster* 9. Essa situação não ocorreu no experimento com *alias*. Isso pode demonstrar que muitos artigos citam os genes sem utilizar a nomenclatura oficial.

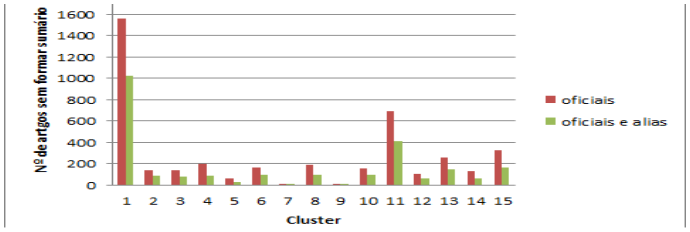

Figura 5. Quantidade de textos que não formaram sumários.

Um resumo dos resultados indica que, quando foi utilizado vetor de palavras-chave somente com nomenclatura oficial, a quantidade média de sentenças foi 4,7. Utilizando alias, o valor foi 6,2. Como o intuito é reduzir a quantidade de texto apresentado, pode-se afirmar que a sumarização com palavras-chave oficiais foi mais eficiente na redução dos textos. No entanto, ao comparar a quantidade de sumários não formados ao utilizar somente a nomenclatura oficial, obteve-se 275,1, e ao utilizar os *alias*, 163,3. Observa-se que ignorar a nomenclatura não oficial resulta em grande perda de informação. Assim, pode-se concluir que quanto mais específicas são as palavras-chave, melhores serão os resultados na redução da quantidade de texto. Observa-se também que não se pode ignorar termos de nomenclatura não oficiais, pois são amplamente usados.

Na instanciação do método SARI-MD, o método pelo qual os genes foram agrupados não é considerado. Portanto, foram feitos somente experimentos de sumarização automática. Buscou-se interações entre os seguintes genes: *MLL, MEN1, CREBBP, PPIE, XAB2, XPA, BRCA1*. Nesse grupo de sete genes, foram encontradas sete interações gênicas. Aplicou-se o método de sumarização com taxa de compressão de 20% e obteve-se a quantidade de sentenças apresentada na Figura 6(b). Pode-se observar a quantidade de sentenças obtidas em todos os artigos fonte, que apresentam a interação entre dois genes de interesse. Há também a quantidade de sentenças que possuem palavras-chave e a quantidade de sentenças resultante após aplicar uma taxa de compressão de 20%. A média da quantidade de sentenças nos sumário é 4,25. Na Figura 6(a), apresenta-se a quantidade de sentenças resultante quando a taxa de compressão aplicada é de 10%. Neste caso, os sumário ficaram pequenos, a média de sentenças foi 1,75.

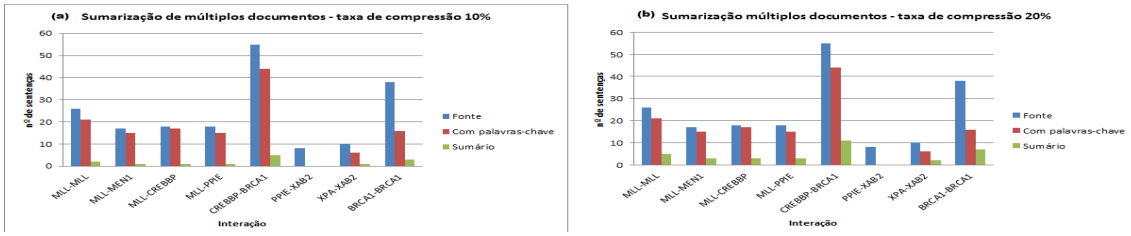

Figura 6. Quantidade de sentenças nos textos e nos sumários: taxas de compressão (a) 10% e (b) 20%.

4. Conclusão

Algoritmos de *clustering* são um clássico na detecção de interações e co-regulação gênica. Neste artigo, apresentou-se uma maneira de analisá-los e outra de relacioná-los por meio da literatura com sumarização automática de artigos científicos. O método SARI foi desenvolvido para agrupar essas duas tarefas, agrupar e sumarizar. Desse modo, a principal

contribuição do trabalho é o auxílio a validação e a atribuição de significado aos *clusters* gerados a partir de dados de expressão gênica. Quando a literatura científica indica relacionamento entre genes de um *cluster*, pode-se inferir que o *cluster* não foi formado por aleatoriedade. Quando um *cluster* aponta relacionamento entre genes que nunca foram citados na literatura, pode ser um novo foco de estudo ou um indicativo de problemas no algoritmo de *clustering*. Como trabalhos futuros pretende-se fortalecer a atribuição de significado a cluster através da análise de características (por exemplo: genes centralizadores dos clusters): dos genes agrupados em *cluster* (intra-cluster) e dos genes separados (inter-cluster).

Referências

Babu, M. M. (2004). *Computational Genomics*, chapter 11 An Introduction to Microarray Data Analysis. Horizon press.

D'Avanzo, E., Magnini, B., and Vallin, A. (2004). Keyphrase extraction for summarization purposes: The LAKE system at DUC-2004. In *DUC2004*.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11:10–18.

Kankar, P., Adak, S., Sarkar, A., Murari, K., Sharma, G., and Expression, G. (2002). Medmesh summarizer: text mining for gene clusters. In *Proceedings of the Second SIAM International Conference on Data Mining*.

Levine, J. H. (1996). *Introduction to data analysis: rules of evidence volume I: well-behaved variables*. Dartmouth College, Dartmouth.

Mani, I. (2001). *Automatic Summarization*. Natural Language Processing. John Benjamins Publishing Company.

Moellering, R. E., Cornejo, M., Davis, T. N., Del Bianco, C., Aster, J. C., Blacklow, S. C., Kung, A. L., Gilliland, D. G., Verdine, G. L., and Bradner, J. E. (2009). Direct inhibition of the notch transcription factor complex. *Nature*, 462(7270):182–8.

Monard, M. C. and Baranauskas, J. A. (2003). *Sistemas Inteligentes: Fundamentos e Aplicações*, chapter Conceitos sobre aprendizado de máquina. Editora Manole Ltda.

Nenkova, A. and Vanderwende, L. (2005). The impact of frequency on summarization. TechReport MSR-TR-2005-101, MSR-TR-2005-101.

Pardo, T., Rino, L., and Nunes, M. (2002). Extractive summarization: how to identify the gist of a text. In *the Proceedings of the 1st International Information Technology Symposium–I2TS*, pages 1–6. Citeseer.

Splendore, A. (2005). Para que existem as regras de nomenclatura genética? *Revista Brasileira de Hematologia e Hemoterapia*, 27:148–152.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2005). BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34(suppl 1):D535–D539.

Strachan, T. and Read, A. P. (1999). *Human Molecular Genetics*. Garland Science, 2 edition.