Avaliação sistemática de técnicas de bi-agrupamento de dados

Victor Alexandre Padilha

Victor Alexandre Padilha

Avaliação sistemática de técnicas de bi-agrupamento de dados

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Ricardo José Gabrielli Barreto Campello

USP – São Carlos
Novembro de 2016

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

P123a

Padilha, Victor Alexandre
Avaliação sistemática de técnicas de bi-agrupamento de dados / Victor Alexandre Padilha; orientador Ricardo José Gabrielli Barreto Campello. - São Carlos - SP, 2016.
107 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2016.

1. agrupamento de dados. 2. bi-agrupamento de dados. 3. expressão gênica. I. Campello, Ricardo José Gabrielli Barreto, orient. II. Título.

Victor Alexandre Padilha

A systematic comparative evaluation of biclustering techniques

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Ricardo José Gabrielli Barreto Campello

USP – São Carlos
November 2016

Aos meus pais, pelo apoio incondicional durante a realização deste trabalho.

AGRADECIMENTOS

Aos meus pais, Osmário e Walkiria, e demais familiares por sempre acreditarem em meu potencial e serem compreensivos, principalmente nos momentos em que não pude estar presente devido à rotina de pós-graduando.

Ao meu orientador, Prof. Dr. Ricardo J. G. B. Campello, por toda a sua seriedade, competência e paciência durante o andamento deste trabalho. Suas contribuições foram importantes para meu amadurecimento tanto na vida acadêmica quanto na vida pessoal.

A todos os professores do Departamento de Ciência da Computação da Universidade Estadual do Centro-Oeste (UNICENTRO), instituição na qual cursei minha graduação, que participaram de maneira direta ou indireta da minha formação. Em especial ao Prof. Dr. Sandro Rautenberg e ao Prof. Dr. Fábio Hernandes, por terem me motivado a seguir na área acadêmica e terem me auxiliado com tudo o que puderam para que eu conseguisse ingressar no ICMC/USP.

Aos meus amigos da minha cidade natal, Irati-PR, pelos momentos de tranquilidade e distração que pudemos passar juntos sempre que voltei para casa. Embora nossos encontros venham tornando-se cada vez mais raros, ainda assim são inesquecíveis. Em especial agradeço a: Enzo Crisigiovanni, André Vosnika, Giovany Pauluk, Ciro Bittencourt, Mateus Zanlorenzi, Guilherme Kosinski, Tiago Batista e Cristiano Ribas.

A todos os amigos e colegas de graduação que conheci nos meus anos em Guarapuava-PR que, embora distantes, ainda mantenho contato e posso contar em todos os momentos. Em especial agradeço a: Luís Simões, Guilherme Leão, Leandro Loma, Gabriel Cecchin e Willian Yassue.

A todos os amigos que conheci durante esses dois anos e alguns meses em São Carlos-SP. Por toda nossa convivência nos churrascos, palquinhos, festas e reuniões. Sem vocês minha caminhada seria muito mais difícil e estressante. Em especial agradeço aos meus colegas do Laboratório de Computação Bioinspirada (Biocom) e aos amigos do Laboratório de Engenharia de *Software* (LabES) e do Laboratório de Otimização (LOt).

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo suporte financeiro (Processo FAPESP nº 2014/08840-0) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo auxílio financeiro inicial deste trabalho.

"Foi o tempo que dedicastes à tua rosa que a fez tão importante."
(Antoine de Saint-Exupéry)

RESUMO

PADILHA, V. A.. **Avaliação sistemática de técnicas de bi-agrupamento de dados**. 2016. 107 f. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

Análise de agrupamento é um problema fundamental de aprendizado de máquina não supervisionado em que se objetiva determinar categorias que descrevam um conjunto de objetos de acordo com suas similaridades ou inter-relacionamentos. Na formulação tradicional do problema, busca-se por partições ou hierarquias de partições contendo grupos cujos objetos são de alguma forma similares entre si e dissimilares aos objetos dos demais grupos, segundo alguma medida direta ou indireta de (dis)similaridade que leva em conta o conjunto completo de atributos que descrevem os objetos na base de dados sob análise. Entretanto, apesar de décadas de aplicações bem sucedidas, existem situações em que a natureza dos agrupamentos contidos nos dados não pode ser representada segundo este tipo de formulação. Em particular, existem situações em que grupos de objetos se caracterizam como tais apenas segundo um subconjunto dos atributos que os descrevem, sendo que tal subconjunto pode ser distinto para cada grupo. Ao contrário de algoritmos de agrupamento tradicionais, algoritmos de bi-agrupamento são capazes de agrupar simultaneamente linhas e colunas de uma matriz de dados. Tais algoritmos produzem bi-grupos formados por subconjuntos de objetos e subconjuntos de atributos de alguma forma fortemente co-relacionados. Esses algoritmos passaram a atrair a atenção da comunidade científica quando se evidenciou a relevância da tarefa de bi-agrupamento em problemas de análise de dados de expressão gênica em bioinformática. Embora em menor grau, as abordagens de bi-agrupamento também têm ganho atenção em outros domínios de aplicação, tais como mineração de textos (*text mining*) e filtragem colaborativa em sistemas de recomendação. O problema é que uma variedade de algoritmos de bi-agrupamento têm sido propostos na literatura baseados em diferentes princípios e suposições sobre os dados, podendo chegar a resultados completamente distintos em uma mesma aplicação. Nesse cenário, torna-se importante a realização de estudos comparativos que possam contrastar o comportamento e desempenho dos diversos algoritmos. Neste trabalho é apresentado um estudo comparativo envolvendo 17 algoritmos de bi-agrupamento (representativos das principais categorias de algoritmos existentes) em coleções de bases de dados tanto de natureza real como simulada, com particular ênfase em problemas de análise de dados de expressão gênica. Diversos aspectos metodológicos e procedimentos para a avaliação experimental foram considerados, a fim de superar as limitações de estudos comparativos anteriores da literatura. Além da comparação em si, todo o arcabouço comparativo pode ser reutilizado para a comparação de outros algoritmos no futuro.

Palavras-chave: agrupamento de dados, bi-agrupamento de dados, expressão gênica.

ABSTRACT

PADILHA, V. A.. **Avaliação sistemática de técnicas de bi-agrupamento de dados**. 2016. 107 f. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

Data clustering is a fundamental problem in the unsupervised machine learning field, whose objective is to find categories that describe a dataset according to similarities between its objects. In its traditional formulation, we search for partitions or hierarchies of partitions containing clusters such that the objects contained in the same cluster are similar to each other and dissimilar to objects from other clusters according to a similarity or dissimilarity measure that uses all the data attributes in its calculation. So, it is supposed that all clusters are characterized in the same feature space. However, there are several applications where the clusters are characterized only in a subset of the attributes, which could be different from one cluster to another. Different than traditional data clustering algorithms, biclustering algorithms are able to cluster the rows and columns of a data matrix simultaneously, producing biclusters formed with strongly related subsets of objects and subsets of attributes. These algorithms started to draw the scientific community's attention only after some studies that show their importance for gene expression data analysis. To a lesser degree, biclustering techniques have also been used in other application domains, such as text mining and collaborative filtering in recommendation systems. The problem is that several biclustering algorithms have been proposed in the past recent years with different principles and assumptions, which could result in different outcomes in the same dataset. So, it becomes important to perform comparative studies that could illustrate the behavior and performance of some algorithms. In this thesis, it is presented a comparative study with 17 biclustering algorithms (which are representative of the main categories of algorithms in the literature) which were tested on synthetic and real data collections, with particular emphasis on gene expression data analysis. Several methodologies and experimental evaluation procedures were taken into account during the research, in order to overcome the limitations of previous comparative studies from the literature. Beyond the presented comparison, the comparative methodology developed could be reused to compare other algorithms in the future.

Key-words: clustering, biclustering, gene expression.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

13AGRI .. Índice Grand Ajustado
13GRI .... Índice Grand
ARI ..... Índice de Rand Ajustado
BBC ..... *Bayesian BiClustering*
BiBit ..... *Bit-Pattern Biclustering Algorithm*
BicAT .... *Biclustering Analysis Toolbox*
Bimax .... *Binary Inclusion-Maximal Biclustering Algorithm*
bp ..... *base pairs*
CCA ..... *Cheng and Church's Algorithm*
CE ..... *Clustering Error*
COALESCE *Combinatorial Algorithm for Expression and Sequence-based Cluster Extraction*
CPB ..... *Correlated Pattern Biclusters*
CSI ..... *Campello Soft Index*
DeBi ..... *Differentially Expressed Biclusters*
DNA ..... ácido desoxirribonucleico
FABIA ... *Factor Analysis for Bicluster Acquisition*
FARI ..... Índice de Rand Ajustado *Fuzzy*
ISA ..... *Iterative Signature Algorithm*
LAS ..... *Large Average Submatrices*
MAFIA .. *MAximal Frequent Itemset Algorithm*
MFI ..... *Maximal Frequent Itemset*
mRNA ... RNA mensageiro
MSR ..... *Mean Squared Residue*
MSSRCC . *Minimum Sum-Squared Residue Coclustering*
NGS ..... *Next Generation Sequencing*
OPSM .... *Order-Preserving Submatrix*
QUBIC ... *QUalitative BIClustering*
RMSE .... *Root Mean Squared Error*
RNA ..... ácido ribonucleico
rRNA .... RNA ribossômico
SAMBA .. *Statistical-Algorithmic Method for Bicluster Analysis*

tRNA .... RNA transportador

xMOTIFs . *Conserved Gene Expression Motifs*

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

1.1 Motivação

Análise de agrupamento de dados é um problema conceitual fundamental em aprendizado de máquina (BISHOP, 2006) e mineração de dados (TAN; STEINBACH; KUMAR, 2006), o qual tem por objetivo encontrar grupos que caracterizam um conjunto de dados, de modo que objetos contidos em um mesmo grupo sejam altamente similares e objetos contidos em grupos diferentes sejam consideravelmente distintos (GAN; MA; WU, 2007). A solução desse problema tipicamente constitui a meta final do procedimento de análise, mas solucionar um problema de agrupamento de dados pode também auxiliar na resolução de outros problemas relacionados, como classificação de padrões e extração de regras a partir de dados (WANG; FU, 2005). Trata-se de um campo interdisciplinar com elementos de disciplinas variadas, tais como estatística, algoritmos e reconhecimento de padrões, possuindo uma variada gama de aplicações, dentre as quais, pode-se citar: recuperação de informação, segmentação de imagens, categorização de textos e identificação de genes e proteínas (XU; WUNSCH II, 2008).

Ao longo das últimas seis décadas vários algoritmos de agrupamento de dados foram propostos, cada qual com suas características e finalidades particulares (vide *surveys* em (XU; WUNSCH II, 2005; BERKHIN, 2006; JAIN, 2010)). De maneira geral, pode-se separar esses algoritmos em três categorias de acordo com o tipo de resultado que são capazes de produzir: hierárquicos, particionais rígidos e particionais com sobreposição. As duas primeiras estão relacionadas no sentido de que um agrupamento hierárquico consiste em uma sequência aninhada de agrupamentos particionais, cada um representando uma partição rígida do conjunto de dados com uma quantidade distinta de subconjuntos mutuamente disjuntos, denominados grupos (do inglês, *clusters*). Dado um conjunto X composto por n objetos x_j, cada qual descrito por um vetor d-dimensional, uma partição rígida consiste em uma coleção $C = \{C_1, C_2, \cdots, C_k\}$ de k subconjuntos, de tal modo que $C_1 \cup C_2 \cup \cdots \cup C_k = X$, $C_i \neq \emptyset$ e $C_i \cap C_j = \emptyset$ para $i, j \in \{1, \cdots, k\}$ e $i \neq j$ (JAIN; DUBES, 1988). Técnicas particionais com sobreposição buscam por partições

probabilísticas ou *fuzzy*, relaxando, de algum modo, a restrição de exclusividade mútua (GAN; MA; WU, 2007).

Algoritmos tradicionais de agrupamento tipicamente buscam por grupos segundo uma medida de (dis)similaridade, a qual leva em conta todos os d atributos dos objetos de dados. Assim, objetos de um mesmo grupo são similares entre si e dissimilares daqueles presentes em outros grupos. Todavia, embora várias aplicações bem sucedidas foram reportadas na literatura nas últimas décadas, existem situações nas quais a natureza dos agrupamentos contidos no conjunto de dados não pode ser representada por meio dessa formulação tradicional. Particularmente, em tais situações, os grupos de objetos se caracterizam apenas segundo um subconjunto dos atributos que os descrevem, podendo tal subconjunto variar para cada grupo (KRIEGEL; KRÖGER; ZIMEK, 2009). Considerando as várias aplicações em que problemas dessa natureza podem ocorrer e suas respectivas características e finalidades, diferentes abordagens de agrupamento foram propostas. De especial interesse deste trabalho são os algoritmos de bi-agrupamento.

Hartigan (1972) propôs o primeiro algoritmo capaz de agrupar linhas e colunas de uma matriz de dados de maneira simultânea, o qual é comumente referenciado como *block clustering*. Posteriormente, Mirkin (1996) denominou esse tipo de agrupamento como *biclustering* (atualmente, também conhecido como *coclustering* ou *two-mode/two-way clustering*). Contudo, apenas após a publicação de Cheng e Church (2000), a qual destaca a importância dessa abordagem em problemas de análise de dados de expressão gênica, que o paradigma de bi-agrupamento de dados tornou-se amplamente conhecido e utilizado.

Dados de expressão gênica resultam de experimentos com tecnologias de *high throughput*, tais como *microarrays* (ZHANG, 2006) e RNA-Seq (WANG; GERSTEIN; SNYDER, 2009). Tipicamente investiga-se o comportamento de milhares de genes, mensurando seus níveis de expressão quando sujeitos a várias condições experimentais (TURNER *et al.*, 2005), tais como amostras de tecidos saudáveis e tecidos cancerígenos ou fases de um processo celular (JIANG; TANG; ZHANG, 2004). Esses dados são normalmente organizados em uma matriz, na qual cada linha representa um gene, cada coluna uma condição experimental, e cada elemento a abundância de RNA mensageiro de um gene sob uma condição experimental específica (MADEIRA; OLIVEIRA, 2004). Técnicas de agrupamento de dados tornaram-se ferramentas muito populares na análise de dados desse tipo, tendo sido utilizadas para anotação funcional de genes, classificação de tecidos, identificação de *motifs*, dentre outros (TANAY; SHARAN; SHAMIR, 2002). Porém, nesses e em outros problemas relacionados, sabe-se que vários padrões de ativação são comuns apenas a um grupo de genes quando sujeitos a subconjuntos de condições experimentais, podendo tais genes comportarem-se de maneira independente nas condições restantes (MADEIRA; OLIVEIRA, 2004; LIU; WANG, 2007; BHATTACHARYA; DE, 2009). Além disso, um gene (ou condição experimental) pode participar em múltiplas vias biológicas em um organismo (GU; LIU, 2008), o que permite a existência de grupos com sobreposição. Consequentemente, esses motivos popularizaram o paradigma de bi-agrupamento de dados no

campo de análise de dados de expressão gênica.

Embora em menor grau, técnicas de bi-agrupamento de dados também são utilizadas em alguns outros domínios de aplicação, tais como: (i) mineração de textos e recuperação de informação, (ii) filtragem colaborativa em sistemas de recomendação e (iii) consultas em grandes bases de dados (MADEIRA; OLIVEIRA, 2004; BUSYGIN; PROKOPYEV; PARDALOS, 2008). Em (i), pode-se buscar por grupos de documentos com propriedades similares quando sujeitos a um mesmo subconjunto de atributos (por exemplo, palavras ou imagens), permitindo a extração de informações relevantes em domínios que compreendem consulta e indexação em motores de busca. Em (ii), é possível identificar subgrupos de consumidores com perfis similares ao comprar ou buscar por determinadas categorias de produtos. Por fim, em (iii), o paradigma de bi-agrupamento pode ser aplicado com a finalidade de reduzir a dimensionalidade de bases de dados compostas por tabelas com milhares de linhas e centenas de colunas.

1.2 Definição do problema e objetivos gerais

Após o renomado trabalho publicado por Cheng e Church (2000), uma variedade de algoritmos de bi-agrupamento vêm sendo propostos, em sua maior parte voltados à análise de dados de expressão gênica (vide *surveys* em (MADEIRA; OLIVEIRA, 2004; TANAY; SHARAN; SHAMIR, 2005; BUSYGIN; PROKOPYEV; PARDALOS, 2008; PONTES; GIRÁLDEZ; AGUILAR-RUIZ, 2015)). É importante ressaltar que tais algoritmos são baseados em diferentes princípios matemáticos e/ou algorítmicos, bem como em diferentes suposições acerca dos dados e dos padrões buscados, o que normalmente acarreta em resultados diferentes em um mesmo cenário. Essa diversidade de métodos torna difícil a escolha de um algoritmo para determinado tipo de aplicação. Portanto, destaca-se a importância da realização de estudos comparativos, que investiguem o comportamento e desempenho de diversos algoritmos, com a finalidade de revelar, para cada classe de problema, aqueles mais apropriados.

Para comparar algoritmos, tipicamente leva-se em consideração a sua eficiência computacional e a sua acurácia. A primeira pode ser avaliada por meio de análises teóricas de complexidade assintótica ou por meio de experimentos controlados medindo o tempo de processamento e utilização de memória em diferentes bases de dados e parametrizações. Por sua vez, a tarefa de avaliar a efetividade de um algoritmo é um problema mais complexo. Ao contrário de problemas de aprendizado supervisionado, nos quais se dispõe de rótulos de classe para os objetos considerados, (bi-)agrupamento é uma abordagem não supervisionada e, devido a isso, em cenários envolvendo dados reais, normalmente não se dispõe de uma solução conhecida a priori.

Segundo Horta e Campello (2014), existem basicamente quatro abordagens para avaliar a efetividade de algoritmos de bi-agrupamento:

- A primeira delas baseia-se na interpretação dos resultados feita por um especialista de domínio, o qual pode utilizar ferramentas auxiliares de visualização de dados (por exemplo, gráficos de coordenadas paralelas ou *heatmaps*) e conhecimento prévio acerca dos genes e condições experimentais considerados. Todavia, essa abordagem baseia-se em critérios subjetivos, diretamente dependentes da base de dados em particular, tornando-a impraticável em cenários nos quais se deseja avaliar sistematicamente vários algoritmos em uma grande quantidade de bases de dados e sujeitos a diversas parametrizações.

- Outro método baseia-se na comparação dos resultados encontrados com conhecimento prévio documentado. Em análise de dados de expressão gênica, existem genes anotados usando ontologias (por exemplo, *Gene Ontology*) (ASHBURNER *et al.*, 2000). Dessa maneira, ao aplicar um algoritmo em uma base de dados real, torna-se possível uma análise de enriquecimento dos bi-grupos encontrados, por meio da qual obtém-se *p-values*, os quais indicam o grau de aleatoriedade de cada bi-grupo. É importante destacar que este método considera apenas os genes envolvidos em cada bi-grupo durante a avaliação, ou seja, não considera as condições experimentais.

- A terceira abordagem tem como base a utilização de índices de avaliação internos (YANG *et al.*, 2005; CANO *et al.*, 2007; GREMALSCHI; ALTUN, 2008; LEE; LEE; JUN, 2011; AYADI; ELLOUMI; HAO, 2012), os quais dependem única e exclusivamente dos dados e da solução que se deseja avaliar. Entretanto, tais critérios podem ser utilizados como função objetivo a ser otimizada por algoritmos de bi-agrupamento, o que desqualifica tais funções para avaliação dos próprios algoritmos que as utilizam. Ademais, ao se aplicar diferentes algoritmos em uma mesma base de dados real, com o uso de índices internos de avaliação pode-se apenas afirmar qual dentre as soluções é a melhor em termos relativos, mas nada pode ser dito sobre a qualidade da melhor solução em termos absolutos, ou seja, em relação aos padrões (em princípio desconhecidos) que de fato supõe-se que existam nos dados.

- Por fim, em situações nas quais uma solução de referência é conhecida, pode-se utilizar índices para validação externa de bi-agrupamentos. Tais índices são capazes de medir a correspondência entre as soluções encontradas com aquelas previamente conhecidas, sem a necessidade de qualquer tipo de suposição. Consequentemente, essa abordagem é particularmente interessante para estudos comparativos. Todavia, a principal dificuldade em aplicá-la advém da escassez de bases de dados reais rotuladas, sendo tipicamente utilizada apenas em cenários compostos por bases de dados sintéticas.

Independentemente da abordagem utilizada para avaliação de resultados, grande parte dos trabalhos na literatura de bi-agrupamento de dados baseia-se na introdução de um novo algoritmo (PRELIĆ *et al.*, 2006), geralmente acompanhada de uma avaliação do mesmo frente a alguns já existentes (HORTA; CAMPELLO, 2014). Paralelamente a esses trabalhos, alguns estudos

comparativos foram desenvolvidos e reportados em (PRELIĆ *et al.*, 2006; BOZDAĞ; KUMAR; ÇATALYÜREK, 2010; EREN *et al.*, 2013). No entanto, todos eles apresentam severas limitações, dentre elas: (i) conjuntos pouco representativos de algoritmos foram comparados, (ii) coleções limitadas de bases de dados foram consideradas e (iii) avaliações de resultados foram realizadas por meio de índices externos comprovadamente problemáticos (HORTA; CAMPELLO, 2014). Desse modo, considerando tais fatos, os objetivos do presente trabalho foram:

- Investigar uma coleção mais representativa de diferentes algoritmos de bi-agrupamento existentes na literatura.
- Compilar coleções de dados reais e sintéticos que estivessem disponíveis publicamente e permitissem análise de bi-agrupamento de dados em cenários biológicos.
- Avaliar comparativamente o desempenho dos algoritmos investigados em diferentes cenários, utilizando tanto dados sintéticos como dados reais, no primeiro caso avaliando as soluções obtidas por meio de dois índices de validação externa comprovadamente robustos, e no segundo caso avaliando por meio de metodologias bem estabelecidas adotadas na área de bioinformática.

1.3 Contribuições

Em resumo, as principais contribuições deste trabalho foram:

1. O desenvolvimento de uma metodologia inteiramente nova que pode ser utilizada futuramente para a comparação de novos algoritmos de bi-agrupamento. Tal metodologia envolve uma coleção variada de cenários relevantes para aplicações reais de bi-agrupamento, bem como a utilização de abordagens e medidas para avaliação que são justificadamente mais adequadas para cada cenário.
2. A proposta de uma nova coleção de bases de dados sintéticas, contendo bi-agrupamentos que seguem uma estrutura particular, conhecida como *checkerboard*, assumida por certos algoritmos, a qual não foi levada em conta em estudos comparativos anteriores disponíveis na literatura.
3. A comparação dos algoritmos em bases de dados reais em relação à sua acurácia para o agrupamento de amostras de câncer, o qual é um problema que não foi abordado em estudos comparativos anteriores.
4. O desenvolvimento da biblioteca biclustlib (https://bitbucket.org/padilha/biclustlib), a qual contém a implementação de alguns dos algoritmos investigados neste trabalho na linguagem Python e continuará sendo expandida, por meio da inclusão de outras técnicas de bi-agrupamento e também de medidas de validação.

5. A escrita do seguinte artigo científico, o qual foi submetido para um periódico internacional de destaque e está em processo de revisão:

    - PADILHA, V. A.; CAMPELLO, R. J. G. B. *A systematic comparative evaluation of biclustering techniques*.

1.4 Organização do trabalho

Além deste capítulo introdutório, este trabalho está organizado da seguinte forma. No Capítulo 2 são explicados os fundamentos biológicos relacionados às bases de dados de expressão gênica. No Capítulo 3 são introduzidos os conceitos fundamentais sobre o paradigma de biagrupamento de dados para o entendimento desta dissertação. No Capítulo 4 são apresentados os experimentos realizados com bases de dados sintéticas. No Capítulo 5 são relatados os experimentos executados em coleções de bases de dados reais. Por fim, no Capítulo 6, são apresentadas as conclusões.

CAPÍTULO

2

EXPRESSÃO GÊNICA

Diversas informações importantes a respeito de um organismo podem ser extraídas a partir de experimentos com dados de expressão gênica. Tais informações podem ser relevantes em várias tarefas, por exemplo: descobrir funções desconhecidas de genes, ampliar o conhecimento referente a diferentes processos biológicos, identificar potenciais alvos de uma droga, dentre outros (HARRINGTON; ROSENOW; RETIEF, 2000).

Neste capítulo é apresentada uma visão geral e simplificada dos fundamentos biológicos necessários para familiarizar o leitor com a natureza dos dados de expressão gênica, os quais constituem a principal motivação para o desenvolvimento de algoritmos de bi-agrupamento. Na Seção 2.1 são introduzidos os ácidos nucleicos e o processo de expressão gênica. Na Seção 2.2 é explicado como dados de expressão gênica são coletados por meio de *microarrays*, os quais consistem em tecnologias de *high throughput* capazes de monitorar o comportamento de milhares de genes em amostras biológicas. As bases de dados utilizadas nos experimentos do Capítulo 5 são provenientes de estudos realizados com tais tecnologias. Por fim, na Seção 2.3 são feitas as considerações finais.

2.1 Fundamentos da biologia molecular

Células consistem em unidades básicas das quais os organismos vivos são constituídos, e que são originadas a partir de outras células predecessoras (ALBERTS *et al.*, 2014). Toda a informação genética de uma célula é armazenada em moléculas chamadas ácidos nucleicos (CLARK, 2005). Os organismos vivos contêm dois tipos de ácidos nucleicos, a saber, ácido desoxirribonucleico (DNA) e ácido ribonucleico (RNA) (SETUBAL; MEIDANIS, 1997; ZHANG, 2006).

O DNA é composto por duas fitas ligadas em uma estrutura amplamente referenciada como dupla hélice, onde cada fita consiste em um arranjo linear de unidades denominadas

nucleotídeos, compostos por um açúcar, um fosfato e uma base nitrogenada (HUMAN GENOME PROGRAM, 1992). No total, existem quatro diferentes bases nitrogenadas no DNA: adenina (A), guanina (G), citosina (C) e timina (T). As duas fitas de uma molécula de DNA são ligadas por suas bases, satisfazendo duas regras: adenina liga-se apenas com a timina e a citosina liga-se apenas com a guanina (JASKOWIAK, 2016). O número de pares de bases ligadas em uma molécula de DNA pode ser utilizado como uma medida de comprimento, nomeada *base pairs* (bp) (HUMAN GENOME PROGRAM, 1992; SETUBAL; MEIDANIS, 1997; ZHANG, 2006; BLACKBURN *et al.*, 2006; JASKOWIAK, 2016). Um exemplo da estrutura do DNA, contendo oito bp, é apresentado na Figura 1[1].

Figura 1 – Exemplo da estrutura de dupla hélice do DNA com oito bp. As bases nitrogenadas estão representadas por suas respectivas iniciais.

Fonte: Adaptado de (BALL, 2007)[2].

De modo geral, a estrutura de um RNA possui diversas similaridades com aquela de um DNA. Entretanto, dentre as principais características que as diferenciam, merecem destaque (SETUBAL; MEIDANIS, 1997; ZHANG, 2006; JASKOWIAK, 2016): (i) no RNA a base timina é substituída pela base uracila (U), que também liga-se com a adenina; (ii) o RNA geralmente não apresenta a estrutura de dupla fita presente no DNA, sendo normalmente encontrado como uma única fita; e (iii) diferente do DNA, que possui apenas a função de codificar as informações de um organismo, o RNA pode ser de três diferentes tipos: RNA mensageiro, RNA transportador ou RNA ribossômico. Eles serão explicados em conjunto com o processo de expressão gênica, descrito a seguir.

Genes consistem em regiões de uma molécula de DNA onde são encontradas as informações genéticas necessárias para produzir as proteínas, as quais possuem diversas funções em um organismo, desde formar os componentes estruturais de tecidos e células até fornecer enzimas para processos bioquímicos (HUMAN GENOME PROGRAM, 1992). O processo de expressão gênica consiste na leitura e utilização da informação genética contida em um gene para produzir

[1] É importante destacar que este é um exemplo meramente ilustrativo. O genoma humano, por exemplo, contém cerca de três bilhões de bp (HUMAN GENOME PROGRAM, 1992; LODISH *et al.*, 2003).

[2] A imagem original está disponibilizada sob a licença CC BY-SA 3.0 (https://creativecommons.org/licenses/by-sa/3.0/).

proteínas (LODISH *et al.*, 2003). Em suma, tal processo pode ser descrito por três procedimentos distintos (SETUBAL; MEIDANIS, 1997; JASKOWIAK, 2016):

1. No procedimento de **replicação**, uma molécula de DNA é duplicada, gerando duas moléculas de DNA contendo exatamente a mesma informação genética. Esse procedimento permite que um organismo produza, a partir de uma única célula, bilhões de outras células contendo a mesma informação genética da célula original.

2. No procedimento de **transcrição**, é identificada a região de um gene no DNA e realizada uma cópia do mesmo em uma molécula de RNA, denominada RNA mensageiro (mRNA). Tal molécula contém a sequência presente em uma das fitas do gene, porém substituindo a base timina pela base uracila. Após isso, o mRNA é enviado ao ribossomo, o qual consiste em uma estrutura interna da célula formada por proteínas e moléculas de RNA ribossômico (rRNA), responsável pela produção de novas proteínas em um processo denominado tradução.

3. No procedimento de **tradução**, o mRNA é lido no interior do ribossomo em códons[3], os quais são responsáveis por codificar diferentes aminoácidos. Durante o processamento o RNA transportador (tRNA) realiza a ligação do aminoácido codificado por cada códon à proteína em formação. Todo esse procedimento é repetido até que ocorra a leitura de um codón que indique o término da síntese proteica.

Todos os passos descritos acima fazem parte do dogma central da biologia molecular, o qual estabelece como ocorre o fluxo de informações genéticas do DNA, do RNA e das proteínas dentro de uma célula (SETUBAL; MEIDANIS, 1997; ZHANG, 2006).

2.2 *Microarrays*

De modo geral, um *microarray* consiste tipicamente em uma pastilha de plástico, vidro ou *nylon*, organizada em um formato de grade onde, em cada célula da grade, segmentos de uma fita de DNA, relacionados a um gene, são depositados ou impressos na pastilha (ZHANG, 2006). Em suma, um experimento de *microarray* consiste nos seguintes passos (JASKOWIAK, 2016):

1. **Preparação de amostras:** moléculas de mRNA coletadas a partir de amostras biológicas são rotuladas, tipicamente com algum material fluorescente, para permitir posterior detecção.

2. **Hibridização:** as moléculas de mRNA rotuladas no passo anterior são depositadas no *microarray*. A hibridização consiste no processo de tais moléculas ligarem-se aos seus

[3] Um códon consiste em uma sequência de três nucleotídeos, sendo que cada nucleotídeo possui uma base nitrogenada em sua composição.

complementos de acordo com a regra de ligação entre pares de bases, que foi explicada na Seção 2.1 deste capítulo. Após isso, o *microarray* é lavado, com a finalidade de remover moléculas que não foram hibridizadas.

3. **Escaneamento:** por fim, o *microarray* é escaneado e uma imagem do mesmo é gerada contendo a intensidade dos sinais emitidos pelos materiais fluorescentes de uma amostra. Tais intensidades são assumidas como os níveis de expressão dos genes contidos no *microarray*, e são posteriormente quantificadas por meio de técnicas de processamento de imagens.

Existem diversos métodos para a manufatura de *microarrays* (HARRINGTON; ROSENOW; RETIEF, 2000; JASKOWIAK, 2016). Neste trabalho, as bases de dados reais utilizadas são provenientes de estudos realizados com dois tipos de *microarray* mais conhecidos: cDNA (SCHENA *et al.*, 1995; DUGGAN *et al.*, 1999) e Affymetrix (LOCKHART *et al.*, 1996; DALMA-WEISZHAUSZ *et al.*, 2006). Dentre as principais diferenças entre eles, destacam-se (JASKOWIAK, 2016): (i) *microarrays* Affymetrix permitem a análise de apenas uma amostra por vez, enquanto que em *microarrays* de cDNA existe a necessidade de duas amostras: interesse e controle; e (ii) o nível de expressão quantificado para cada gene a partir de um *microarray* Affymetrix é absoluto, enquanto que o nível de expressão obtido de um *microarray* cDNA para um gene j é relativo e tipicamente calculado como:

$$e_j = \log\left(\frac{\text{Cy5}}{\text{Cy3}}\right) \tag{2.1}$$

onde Cy5 e Cy3 indicam, respectivamente, a intensidade dos corantes fluorescentes previamente utilizados para rotular as amostras de interesse e de controle. Portanto, em um *microarray* cDNA, um valor de expressão positivo indica que um gene se expressou mais na amostra de interesse do que na de controle, um valor de expressão negativo aponta que um gene teve maior expressão na amostra de controle do que na de interesse, enquanto que um valor de expressão igual a zero indica que não houve diferença na expressão de um gene entre as amostras de interesse e de controle.

Devido às diferenças supracitadas, os *microarrays* cDNA e Affymetrix são normalmente referenciados na literatura como *double-channel microarray* e *single-channel microarray*, respectivamente.

Por fim, é importante mencionar que, além dos *microarrays*, existem outros métodos capazes de mensurar a expressão de uma grande quantidade de genes. Dentre eles, um que vem ganhando bastante popularidade é o RNA-Seq (WANG; GERSTEIN; SNYDER, 2009), que consiste na utilização de tecnologias *Next Generation Sequencing* (NGS) para amostrar com um menor viés as moléculas de mRNA de uma célula e medir de maneira mais precisa o nível de expressão dos genes investigados (TRAPNELL; PACHTER; SALZBERG, 2009). Entretanto, neste trabalho não foram utilizadas bases de dados de RNA-Seq, uma vez que ainda são poucos

os trabalhos da literatura que aplicam algoritmos de (bi-)agrupamento em tal cenário comparado com a grande quantidade de estudos com dados gerados a partir de *microarrays*.

2.3 Considerações

Neste capítulo foram apresentados os conceitos básicos da biologia molecular e de como dados de expressão gênica são coletados, com a finalidade de introduzir ao leitor a principal aplicação de algoritmos de bi-agrupamento de dados. Primeiramente, foram brevemente discutidos os ácidos nucleicos (DNA e RNA) e as regras que permitem o processo de hibridização de suas moléculas. Em seguida, foi descrito o processo de expressão gênica, o qual faz parte do dogma central da biologia molecular. Ademais, a tecnologia de *microarray* foi explicada, bem como os passos envolvidos em seus experimentos. Finalmente, foram mencionadas as principais diferenças entre os dois tipos de *microarray* mais utilizados (cDNA e Affymetrix).

CAPÍTULO

3

BI-AGRUPAMENTO DE DADOS

Neste capítulo são apresentados os conceitos fundamentais sobre o paradigma de bi-agrupamento de dados para o entendimento deste trabalho. Uma breve visão acerca da complexidade do problema é fornecida na Seção 3.1. Na Seção 3.2 é introduzida a notação utilizada. Os tipos de padrões normalmente buscados em dados de expressão gênica são apresentados na Seção 3.3. Na Seção 3.4 é feita uma revisão de vários dos algoritmos propostos na literatura de bi-agrupamento de dados, os quais foram comparados neste trabalho, categorizados segundo os tipos de heurísticas que utilizam. Eles consistem dos mesmos algoritmos compreendidos em estudos comparativos anteriores (PRELIĆ *et al.*, 2006; BOZDAĞ; KUMAR; ÇATALYÜREK, 2010; EREN *et al.*, 2013), além de outros com destaque na literatura e/ou com implementações publicamente disponíveis. Na Seção 3.5, é explicado como os resultados obtidos por meio desses algoritmos podem ser avaliados. Na Seção 3.6 são discutidos os principais trabalhos comparativos existentes na literatura de bi-agrupamento de dados e suas limitações. Considerações finais são apresentadas na Seção 3.7.

3.1 Complexidade do problema

A complexidade do problema de bi-agrupamento de dados depende diretamente da sua formulação e da função objetivo a qual se deseja otimizar (KRIEGEL; KRÖGER; ZIMEK, 2009). Entretanto, quase todas as suas variantes são problemas NP-Completos (MADEIRA; OLIVEIRA, 2004), uma vez que, para uma matriz de dados de expressão gênica contendo n genes e m condições experimentais, se tem complexidade computacional da ordem de 2^{n+m} para uma combinação completa das linhas e colunas da mesma (JIANG; TANG; ZHANG, 2004). Devido a isso, boa parte dos algoritmos existentes na literatura são baseados em abordagens heurísticas para a identificação de bi-grupos nos dados. Outros utilizam enumeração exaustiva para essa tarefa porém tipicamente apresentam uma alta complexidade computacional no pior caso, sendo necessária a imposição de restrições para tornar a enumeração factível (MADEIRA;

OLIVEIRA, 2004).

3.2 Notação

Primeiramente, uma matriz de dados é denotada como $A = (R, C)$, sendo $R = \{1, 2, \cdots, n\}$ o conjunto de linhas, $C = \{1, 2, \cdots, m\}$ o conjunto de colunas e a_{ij} o nível de expressão do gene i quando sujeito à condição experimental j. Dessa maneira, um bi-grupo consiste em uma submatriz (I, J), sendo $I \subseteq R$ um conjunto de linhas e $J \subseteq C$ um conjunto de colunas, na qual todos os genes em I apresentam comportamentos similares sob todas as condições experimentais em J e vice-versa. Ademais, a_{iJ}, a_{Ij} e a_{IJ} representam as médias da linha i, da coluna j e de todos os elementos de um bi-grupo, respectivamente.

3.3 Tipos de bi-grupos

Madeira e Oliveira (2004), em um *survey* sobre bi-agrupamento de dados amplamente difundido na literatura, identificaram quatro tipos de bi-grupos normalmente buscados ao se trabalhar com análise de dados de expressão gênica: (i) bi-grupos com valores constantes, (ii) bi-grupos com linhas ou colunas constantes, (iii) bi-grupos com valores coerentes, e (iv) bi-grupos com evoluções coerentes, os quais serão detalhados, respectivamente, nas subseções 3.3.1 a 3.3.4. Posteriormente, Aguilar-Ruiz (2005) sumarizou formalmente os três primeiros tipos de padrões em outros dois tipos: bi-grupos aditivos e bi-grupos multiplicativos. Estes serão explicados na subseção 3.3.5.

Ademais, é importante ressaltar que os tipos (i), (ii) e (iii) se referem diretamente aos valores numéricos contidos nas células de uma submatriz representando um bi-grupo (Figura 2a–e). Por sua vez, o tipo (iv) considera tais células como símbolos, os quais podem ser nominais (Figura 2f–2h), definir uma ordem linear entre as colunas do bi-grupo (Figura 2i), ou representar regulações positivas ou negativas dos genes considerados (Figura 2j) (MADEIRA; OLIVEIRA, 2004).

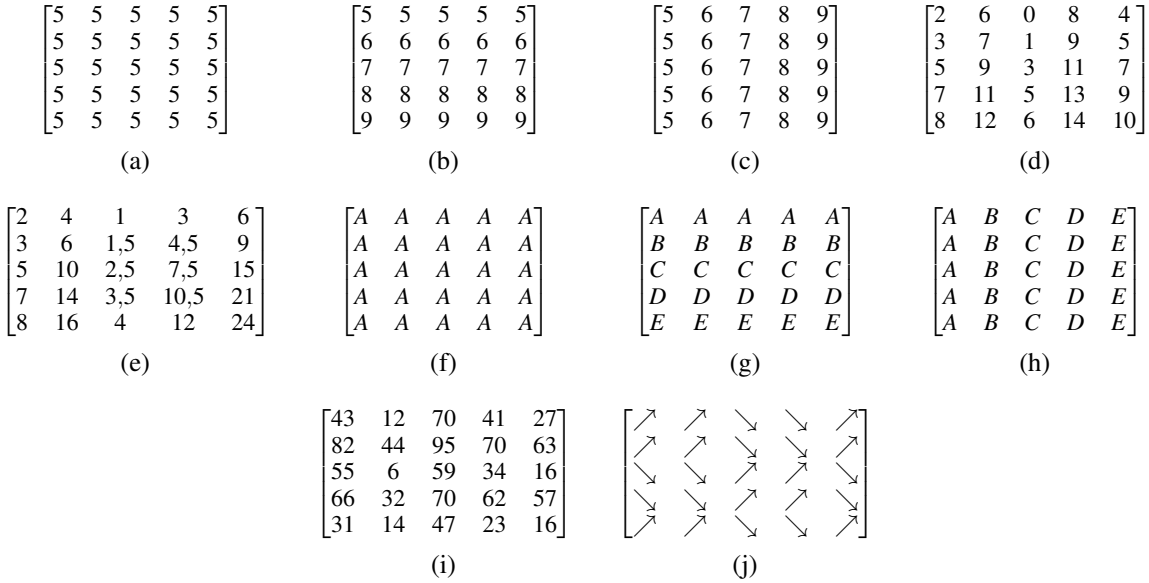
3.3.1 Bi-grupos com valores constantes

Em um bi-grupo constante, para todo gene $i \in I$ e toda condição experimental $j \in J$, se tem como valor um mesmo número (Equação 3.1).

$$a_{ij} = \mu \tag{3.1}$$

É importante mencionar que esse tipo de bi-grupo é de ocorrência muito rara em dados de expressão gênica reais, uma vez que em matrizes dessa natureza tipicamente ocorre a existência de ruídos, valores ausentes e variações sistemáticas geradas a partir do procedimento experimental

Figura 2 – Exemplos dos tipos de bi-grupos considerados ao se trabalhar com análise de dados de expressão gênica.



Fonte: Baseado na Figura 1 de Madeira e Oliveira (2004).

(JIANG; TANG; ZHANG, 2004). A função objetivo comumente utilizada para identificar bi-grupos constantes é a variância (MADEIRA; OLIVEIRA, 2004), pois tais bi-grupos idealmente se caracterizam por submatrizes com variância nula. Um exemplo é apresentado na Figura 2a.

3.3.2 Bi-grupos com linhas ou colunas constantes

Um bi-grupo com linhas constantes é definido como uma submatriz onde qualquer elemento pode ser obtido por meio de uma dentre as seguintes expressões:

$$a_{ij} = \mu + \alpha_i, \tag{3.2}$$

$$a_{ij} = \mu \cdot \alpha_i, \tag{3.3}$$

sendo μ o valor característico do bi-grupo e α_i um valor de ajuste para a linha $i \in I$. Um exemplo desse tipo é demonstrado na Figura 2b.

De maneira semelhante, em um bi-grupo com colunas constantes qualquer elemento pode ser obtido por uma das duas equações abaixo:

$$a_{ij} = \mu + \beta_j, \tag{3.4}$$

$$a_{ij} = \mu \cdot \beta_j, \tag{3.5}$$

onde μ possui a mesma definição dada acima e β_j denota um valor de ajuste para a coluna $j \in J$. Um exemplo é apresentado na Figura 2c.

Segundo Madeira e Oliveira (2004), diferente dos bi-grupos com valores inteiramente constantes, bi-grupos com linhas ou colunas constantes não podem ser identificados aplicando-se diretamente o cálculo da variância. Para que isso seja possível, os autores sugerem a normalização das linhas ou das colunas da matriz de dados por meio da média de cada linha e da média de cada coluna, respectivamente. Como consequência, bi-grupos tais como aqueles apresentados nas Figuras 2b e 2c seriam transformados naquele apresentado na Figura 2a.

3.3.3 Bi-grupos com valores coerentes

Este tipo de bi-grupo pode ser visto como uma generalização dos bi-grupos inteiramente constantes ou com linhas ou colunas constantes. Especificamente, o valor de cada elemento de um bi-grupo com valores coerentes pode ser obtido por um modelo aditivo (Equação 3.6) ou por um modelo multiplicativo (Equação 3.7). A principal diferença entre tais modelos e os anteriores decorre da existência de dois valores de ajuste, um para a linha e outro para a coluna do elemento em questão.

$$a_{ij} = \mu + \alpha_i + \beta_j \tag{3.6}$$

$$a_{ij} = \mu \cdot \alpha_i \cdot \beta_j \tag{3.7}$$

As Equações (3.6) e (3.7), μ, α_i e β_j denotam, respectivamente, o valor típico dentro do bi-grupo e os fatores de ajuste para a linha i e para a coluna j. Exemplos do primeiro e do segundo são apresentados nas Figuras 2d e 2e. As Equações (3.2) e (3.4) são casos especiais da Equação (3.6) quando $\beta_j = 0$ e $\alpha_i = 0$, nessa ordem. Por sua vez, as Equações (3.3) e (3.5) são casos especiais da Equação (3.7) quando $\beta_j = 1$ e $\alpha_i = 1$, respectivamente.

Na ausência de ruídos, o valor de cada elemento de uma submatriz representando um bi-grupo com valores coerentes aditivos pode ser obtido pela média da sua linha, da sua coluna e do bi-grupo (MADEIRA; OLIVEIRA, 2004). Assim, na Equação (3.6), se obtém $\mu = a_{IJ}$, $\alpha_i = a_{iJ} - a_{IJ}$ e $\beta_j = a_{Ij} - a_{IJ}$, resultando em:

$$a_{ij} = a_{iJ} + a_{Ij} - a_{IJ}. \tag{3.8}$$

Como exemplo, considere o bi-grupo apresentado na Figura 2d. Nessa submatriz, $a_{IJ} = 7$ e, ao indexar as linhas de cima para baixo e as colunas da esquerda para a direita, se obtém $\{a_{1J}, a_{2J}, a_{3J}, a_{4J}, a_{5J}\} = \{4,5,7,9,10\}$ e $\{a_{I1}, a_{I2}, a_{I3}, a_{I4}, a_{I5}\} = \{5,9,3,11,7\}$. Devido à inexistência de ruídos, qualquer um de seus elementos pode ser obtido pela Equação (3.8).

Como em cenários de aplicação reais a presença de ruídos é muito frequente, Cheng e Church (2000) introduziram o conceito de resíduo, o qual quantifica a diferença entre o valor

real do elemento a_{ij} na matriz de dados e o seu valor predito a partir da média da sua linha, da sua coluna e do bi-grupo:

$$\begin{aligned} \text{resíduo}(a_{ij}) &= a_{ij} - (a_{iJ} + a_{Ij} - a_{IJ}) \\ &= a_{ij} - a_{iJ} - a_{Ij} + a_{IJ}. \end{aligned} \tag{3.9}$$

Para avaliar a qualidade de um bi-grupo (I,J) com valores coerentes, os autores propuseram a medida *Mean Squared Residue* (MSR):

$$\text{MSR}(I,J) = \frac{1}{|I| \cdot |J|} \cdot \sum_{i \in I, j \in J} \text{resíduo}(a_{ij})^2. \tag{3.10}$$

Segundo Madeira e Oliveira (2004), o MSR quantifica o nível de coerência entre as linhas e colunas de um bi-grupo. Posteriormente, Aguilar-Ruiz (2005) conduziu um estudo acerca dessa medida, provando que a mesma não é adequada para a identificação de bi-grupos que apresentem padrões multiplicativos, uma vez que ela depende diretamente da variância dos fatores de escala. Assim, o autor afirmou que métodos de bi-agrupamento baseados na MSR podem não encontrar bi-grupos que sigam esses padrões quando a variância dos valores dos genes for alta.

3.3.4 Bi-grupos com evoluções coerentes

Bi-grupos com evoluções coerentes possuem como principal característica a representação de seus elementos como símbolos. Desse modo, tipicamente se busca por: (i) bi-grupos com genes conservados em um único estado sob as mesmas condições experimentais, podendo tal estado ser o mesmo para todos os genes (Figura 2f) ou variar de gene para gene (Figura 2g); (ii) bi-grupos cujos genes comportem-se de maneira similar em um mesmo conjunto de estados sob as mesmas condições experimentais (Figura 2h); (iii) bi-grupos que induzam uma ordem linear entre suas colunas, tal como aquele na Figura 2i; ou (iv) bi-grupos que possuam genes positivamente/negativamente regulados de forma coerente em algum subconjunto de colunas (Figura 2j).

3.3.5 Bi-grupos aditivos e bi-grupos multiplicativos

Os tipos de bi-grupos numéricos previamente apresentados podem ser descritos a partir de outros dois tipos de padrões que os sumarizam: aditivos e multiplicativos (AGUILAR-RUIZ, 2005). Um bi-grupo aditivo é determinado como uma submatriz onde o valor de cada elemento da mesma pode ser obtido por meio de uma constante ξ_j, referente a uma condição experimental j, somada a outra constante π_i, relacionada ao comportamento de um gene i. Assim, os valores contidos em um bi-grupo aditivo seguem a Equação (3.11).

$$a_{ij} = \pi_i + \xi_j \tag{3.11}$$

Considerando que na Equação (3.6) o valor de μ é constante para todos os elementos do bi-grupo, um bi-grupo aditivo pode ser visto como uma generalização de um bi-grupo com valores coerentes aditivos ao se assumir $\pi_i = \mu + \alpha_i$ e $\xi_j = \beta_j$, onde μ, α_i e β_j seguem as definições da Equação (3.6). Logo, um exemplo de bi-grupo aditivo pode ser visto na Figura 2d.

Por sua vez, um bi-grupo multiplicativo pode ser definido como uma submatriz onde cada elemento da mesma é determinado por um valor ψ_j, constante para cada coluna j, multiplicado por uma constante π_i, que representa o comportamento de um gene i. Assim, os valores de um bi-grupo multiplicativo seguem a Equação (3.12).

$$a_{ij} = \pi_i \cdot \psi_j \tag{3.12}$$

De maneira similar aos bi-grupos aditivos, um bi-grupo multiplicativo pode ser visto como uma generalização de um bi-grupo com valores coerentes multiplicativos ao se considerar $\pi_i = \mu \cdot \alpha_i$ e $\psi_j = \beta_j$, com μ, α_i e β_j seguindo as definições utilizadas na Equação (3.7). Portanto, um exemplo de bi-grupo multiplicativo pode ser conferido na Figura 2e.

Por fim, Aguilar-Ruiz (2005) e Pontes, Giráldez e Aguilar-Ruiz (2015) afirmam que uma submatriz pode ainda apresentar comportamento aditivo e multiplicativo simultaneamente. Assim, é dito que um bi-grupo apresenta um comportamento aditivo-multiplicativo quando seus elementos podem ser calculados como:

$$a_{ij} = \pi_i \cdot \psi_j + \xi_j \tag{3.13}$$

sendo π_i a constante que representa o comportamento do gene i, e as constantes ψ_j e ξ_j os valores de ajuste de uma condição experimental j. Os bi-grupos aditivos e bi-grupos multiplicativos podem ser vistos como casos especiais deste quando $\psi_j = 1$ e $\xi_j = 0$, respectivamente. Um exemplo de bi-grupo aditivo-multiplicativo é apresentado na Figura 3, onde $\{\pi_1, \pi_2, \pi_3, \pi_4, \pi_5\} = \{4,5,1,9,10\}$, $\{\psi_1, \psi_2, \psi_3, \psi_4, \psi_5\} = \{1;1{,}5;2{,}0;2{,}5;3{,}0\}$ e $\{\xi_1, \xi_2, \xi_3, \xi_4, \xi_5\} = \{1,2,3,4,5\}$.

Figura 3 – Exemplo de um bi-grupo aditivo-multiplicativo.

$$\begin{bmatrix} 5 & 8 & 11 & 14 & 17 \\ 6 & 9{,}5 & 13 & 16{,}5 & 20 \\ 2 & 3{,}5 & 5 & 6{,}5 & 8 \\ 10 & 15{,}5 & 21 & 26{,}5 & 32 \\ 11 & 17 & 23 & 29 & 35 \end{bmatrix}$$

3.4 Algoritmos

Considerando a alta complexidade do problema de bi-agrupamento de dados (conforme discutido na Seção 3.1), os algoritmos tipicamente baseiam-se em diversas abordagens heurísticas para sua solução (MADEIRA; OLIVEIRA, 2004). Devido a isso, nesta seção são apresentados

os algoritmos estudados, organizados conforme as heurísticas em que são fundamentados. Em 3.4.1 são apresentados algoritmos baseados em heurísticas gulosas. Em 3.4.2 é explicada uma abordagem baseada no paradigma de divisão e conquista. Em 3.4.3 são discutidos alguns algoritmos baseados na adaptação de modelos matemáticos ou estatísticos aos dados. Por fim, em 3.4.4, são apresentadas certas técnicas de bi-agrupamento baseadas em enumeração exaustiva.

3.4.1 Algoritmos gulosos

Algoritmos gulosos são aqueles que realizam, a cada iteração, a melhor escolha local na esperança de que a mesma leve a uma solução ótima global (CORMEN *et al.*, 2009). Em bi-agrupamento de dados, tais algoritmos são tipicamente baseados na inserção/remoção de linhas/colunas de uma matriz de dados, buscando maximizar alguma função de ganho localmente (MADEIRA; OLIVEIRA, 2004). Embora soluções globalmente ótimas não sejam garantidas, algoritmos gulosos de bi-agrupamento normalmente possuem como vantagem uma menor complexidade computacional.

Cheng and Church's Algorithm (CCA)

Cheng e Church (2000) propuseram o primeiro algoritmo guloso para o problema de bi-agrupamento de dados. Na formulação utilizada por eles, um bi-grupo é definido como uma submatriz com resíduo quadrático médio (Equação 3.10) menor ou igual a um limiar $\delta \geq 0$ pré-estabelecido. Baseado nisso, o CCA inicia com toda a matriz de dados e, a cada iteração, remove as linhas e colunas com os maiores valores de MSR, até que a restrição mencionada seja satisfeita. Em seguida, linhas e colunas previamente removidas são adicionadas de volta à submatriz obtida, desde que o MSR da mesma não seja aumentado.

Na literatura, o CCA se mostrou eficiente para encontrar bi-grupos inteiramente constantes, com linhas ou colunas constantes, e com valores coerentes. Entretanto, em aplicações nas quais se deseja investigar a existência de múltiplos bi-grupos, aqueles já encontrados em execuções anteriores precisam ser *mascarados* com valores aleatórios para que não sejam re-encontrados ao se executar novamente o algoritmo à procura de novos bi-grupos. É importante salientar que a inserção desses valores pode degradar consideravelmente o resultado final do algoritmo em cenários que apresentem alto grau de sobreposição entre bi-grupos.

Order-Preserving Submatrix (OPSM)

Ben-Dor *et al.* (2003) definiram um bi-grupo como uma OPSM. Nessa formulação, uma submatriz (I,J) é considerada uma OPSM caso satisfaça um modelo completo (J,π), onde $\pi = (j_1, j_2, \cdots, j_s)$ define uma permutação das colunas em J. Dessa maneira, é dito que uma linha da matriz de dados satisfaz (J,π) se os valores contidos nos seus s índices análogos, quando arranjados conforme a ordem imposta por π, crescem estritamente.

Conforme o valor de s cresce, uma enumeração exaustiva de todos os modelos completos existentes nos dados se torna inviável. Devido a isso, os autores propuseram o conceito de modelos parciais, os quais são expandidos iterativamente, na esperança de que convirjam para os melhores modelos completos. Um modelo parcial de ordem (a,b) especifica os índices dos a menores elementos $< j_1,\cdots,j_a >$ e dos b maiores elementos $< j_{s-b+1},\cdots,j_s >$ de um modelo completo (J,π).

O algoritmo proposto inicia gerando todos os modelos parciais de ordem $(1,1)$. Dentre esses, os l melhores são selecionados e todas as possíveis expansões dos mesmos para modelos parciais de ordem $(2,1)$ são geradas. O procedimento é repetido por meio da criação de modelos de ordem $(2,2)$, $(3,2)$, $(3,3)$, $(4,3)$ e assim sucessivamente, até que l modelos de ordem $(\lceil s/2 \rceil, \lfloor s/2 \rfloor)$ sejam obtidos, dentre os quais o melhor é retornado. Esse procedimento é capaz de encontrar bi-grupos com evoluções coerentes, assim como aquele apresentado na Figura 2i, onde $\pi = (2,5,4,1,3)$.

Conserved Gene Expression Motifs (xMOTIFs)

Murali e Kasif (2003) introduziram o conceito de xMOTIF. Considerando uma matriz de dados de expressão gênica, um xMOTIF corresponde a um subconjunto de genes simultaneamente conservados em um mesmo subconjunto de condições experimentais. É dito que um gene se mantém conservado em um subconjunto de condições experimentais caso ele esteja em um mesmo estado em todas elas, sendo que um estado corresponde a um intervalo de valores estatisticamente significativo[1]. Um exemplo de xMOTIF é apresentado na Figura 2g.

Complementarmente, os autores apresentaram um algoritmo guloso estocástico capaz de encontrar xMOTIFs. Partindo de um subconjunto de colunas (denominadas sementes e selecionadas de maneira uniforme) de tamanho n_s, tal algoritmo sorteia, para cada coluna, um conjunto D (dito conjunto discriminante), com tamanho s_d, de outras colunas. Um bi-grupo contém uma linha caso ela esteja no mesmo estado em uma semente e em todas as colunas do conjunto discriminante correspondente. Em seguida, é realizada uma busca por todas as colunas da matriz de dados que concordem com as linhas selecionadas e seus respectivos estados. Um xMOTIF é descartado se a fração de colunas que casam com ele, em relação ao total existente na base de dados, é menor do que um parâmetro pré-determinado α. Para cada semente, são selecionados n_d conjuntos discriminantes. Dentre todos os xMOTIFs encontrados, aquele que possui a maior quantidade de linhas é retornado.

Iterative Signature Algorithm (ISA)

Em (BERGMANN; IHMELS; BARKAI, 2003) um bi-grupo é definido como um subconjunto de genes co-regulados em um subconjunto de condições experimentais. Com isso, os

[1] Intervalo que contém uma quantidade maior de valores de expressão do que o esperado caso os mesmos fossem gerados segundo uma distribuição uniforme.

autores propuseram o ISA, o qual, a partir de um subconjunto inicial de genes (denominado semente), avalia todas as condições experimentais segundo uma função (denominada função de limiar), mantendo aquelas que extrapolem um determinado parâmetro t_c. De maneira similar, os genes são avaliados segundo a mesma função, sendo mantidos aqueles que ultrapassem outro parâmetro, denotado por t_g. Esse procedimento ocorre de maneira iterativa até que as linhas e colunas do bi-grupo convirjam (ou seja, não mudem). De acordo com a formulação do método, embora a quantidade de sementes iniciais possíveis para uma base de dados seja imensa, elas tipicamente convergem para uma quantidade fixa e relativamente pequena de bi-grupos para um dado par (t_g, t_c). Dessa maneira, os autores sugerem a execução do algoritmo para uma quantidade suficientemente grande de sementes, a fim de encontrar todos os bi-grupos existentes. Essa técnica foi comprovada eficaz na identificação de bi-grupos com evoluções coerentes, compostos por genes positivamente ou negativamente regulados (Figura 2j).

Minimum Sum-Squared Residue Coclustering (MSSRCC)

Em (CHO *et al.*, 2004; CHO; DHILLON, 2008) foi proposto o algoritmo MSSRCC, o qual gera bi-agrupamentos nos quais as linhas e colunas da base de dados original são particionadas em k_r e k_c grupos, respectivamente, os quais são combinados de modo a originar $k_r \cdot k_c$ bi-grupos sem sobreposição. Tais bi-agrupamentos são amplamente referidos como *checkerboard* (MADEIRA; OLIVEIRA, 2004), uma vez que lembram uma estrutura similar à de um tabuleiro.

Dada uma matriz de dados A, a formulação do MSSRCC busca minimizar a seguinte função objetivo:

$$\sum_{(I,J)} \sum_{i \in I, j \in J} \text{resíduo}(a_{ij})^2, \tag{3.14}$$

sendo o resíduo definido em (3.9). Para a resolução do problema acima os autores propuseram um procedimento capaz de convergir para um mínimo local, por meio do decréscimo monotônico da função objetivo.

Primeiramente, o MSSRCC atribui rótulos às linhas e colunas de maneira aleatória. Em seguida, executa dois passos iterativamente até sua convergência. No primeiro deles, o algoritmo busca para cada linha o grupo com o qual a mesma possua maior similaridade, segundo a medida de resíduo. Após isso, os rótulos das linhas são atualizados. No passo seguinte, procede de modo similar para as colunas. Como o MSSRCC utiliza a medida de resíduo no cálculo de sua função objetido, os bi-grupos reportados seguem o padrão de valores coerentes aditivos (Figura 2d).

QUalitative BIClustering (QUBIC)

Em (LI *et al.*, 2009) a matriz de entrada é representada de uma maneira qualitativa, onde cada elemento da matriz é apresentado como um valor inteiro. Para isso, os autores supõem que os valores de expressão de cada gene são gerados segundo uma distribuição normal, tal que elementos positivamente ou negativamente regulados devem estar distantes da média. Com

isso, assumem que os $q \cdot m$ menores ou maiores valores de expressão de um gene representam tais elementos, os quais podem ser classificados em uma quantidade *ranks* pré-determinada, denotada por r. Elementos positivamente (ou negativamente) regulados são ordenados decrescentemente (ou crescentemente), sendo seus respectivos *ranks* atribuídos segundo essa ordem. Como tipicamente deseja-se que $r << m$, tem-se elementos com valores similares atribuídos a um mesmo *rank*.

Dados os conceitos supracitados, Li *et al.* (2009) introduziram o algoritmo QUBIC, o qual representa a matriz discretizada por meio de um grafo G, onde cada gene é um vértice e uma aresta entre dois genes possui ponderação igual à quantidade de condições experimentais para as quais eles apresentam os mesmos valores inteiros. Com isso, o QUBIC busca encontrar todos os bi-grupos (I,J) que maximizem $\min\{|I|,|J|\}$ e que satisfaçam um limiar de consistência $0 < c \leq 1$, o qual diz respeito à fração mínima de linhas com o mesmo valor inteiro, em relação a $|I|$, para cada coluna em J.

O algoritmo QUBIC forma uma quantidade desejada de bi-grupos a partir de um conjunto de sementes S, o qual consiste inicialmente no conjunto de todas as arestas ordenadas pelos seus respectivos pesos. Uma aresta e é considerada uma semente se, e somente se: (i) pelo menos um de seus genes não ocorre em nenhum bi-grupo já encontrado, ou (ii) ambos os genes foram inseridos em diferentes bi-grupos, (I_1,J_1) e (I_2,J_2), com $I_1 \cap I_2 = \emptyset$ e $w(e) \geq \max\{|I_1|,|I_2|\}$, onde $w(e)$ denota a ponderação de e. Assim, dada uma semente inicial, um bi-grupo (I,J) é formado em dois passos. No primeiro deles, é identificado o conjunto J de colunas idênticas entre os genes pertencentes à semente e, em seguida, são adicionadas em I, dentre todas as linhas da matriz de dados, aquelas cujos elementos concordem com J, mantendo a consistência do bi-grupo igual a um. No segundo passo, a restrição de consistência é relaxada, de modo que são inseridas em (I,J) colunas e linhas que não acarretem violação do limiar c. Ao final, todos os bi-grupos encontrados são retornados. O QUBIC é capaz de encontrar padrões semelhantes àquele apresentado na Figura 2h.

Correlated Pattern Biclusters (CPB)

Bozdağ, Parvin e Çatalyürek (2009) propuseram o algoritmo CPB, o qual busca por bi-grupos que possuam linhas com uma alta correlação de Pearson em relação a uma linha de referência considerando um conjunto de colunas aleatoriamente selecionado. Um bi-grupo (I,J) é formado de maneira iterativa, alternando entre a inserção de linhas e colunas no mesmo. Durante a atualização do conjunto I, são adicionadas as linhas que possuam correlação de Pearson acima de um limiar previamente definido, denotado por ρ, com o vetor formado pela média das colunas em J ao considerar os elementos em I. Para atualização do conjunto J, primeiramente a linha i que possui correlação mínima em I é encontrada, todas as colunas que alcancem um valor de *Root Mean Squared Error* (RMSE) inferior ao RMSE de i são incluídas no bi-grupo. O processo é repetido até que os conjuntos I e J convirjam (ou seja, não mudem). Por meio da inicialização

aleatória da linha de referência e do conjunto de colunas, vários bi-grupos com valores coerentes, segundo um padrão aditivo ou multiplicativo, podem ser encontrados (EREN *et al.*, 2013).

Large Average Submatrices (LAS)

Shabalin *et al.* (2009) apresentaram a técnica LAS, a qual busca por submatrizes cujas médias de seus elementos sejam significativamente positivas. Inicialmente, o algoritmo assume que cada elemento a_{ij} de uma matriz de expressão gênica A pode ser representado pela seguinte função linear:

$$a_{ij} = \sum_{k=1}^{K} \gamma_k \cdot f(i \in I_k \wedge j \in J_k) + \varepsilon_{ij}, \tag{3.15}$$

onde K indica o número de bi-grupos existentes, γ_k representa a contribuição do bi-grupo k para o valor de a_{ij}, f é uma função indicadora que retorna um quando a condição entre parênteses é satisfeita e zero caso contrário, e ε_{ij} é uma variável aleatória $N(0,1)$. Sob tal formulação, caso $K = 0$, obtém-se um modelo nulo, no qual a matriz A consiste em $n \cdot m$ variáveis aleatórias independentes $N(0,1)$.

Considerando o modelo nulo supracitado, os autores propuseram uma função para avaliação de um bi-grupo no qual a média de seus elementos assume um valor real $\tau > 0$. Tal função é definida como:

$$S(I,J) = -\log\left[\binom{n}{|I|} \cdot \binom{m}{|J|} \cdot \Phi(-\tau \cdot \sqrt{|I| \cdot |J|})\right], \tag{3.16}$$

onde Φ é a função de distribuição acumulada de uma normal e, nessa equação, representa a probabilidade de que a média de $|I| \cdot |J|$ normais independentes exceda τ. Por sua vez, o termo $\binom{n}{|I|} \cdot \binom{m}{|J|}$ representa a quantidade de submatrizes de dimensão $|I| \times |J|$ existentes em uma matriz de dimensão $n \times m$.

Em seu primeiro passo, o algoritmo inicia selecionando r linhas e c colunas da matriz de dados aleatoriamente para definir os conjuntos I e J iniciais. Os valores de r e c são sorteados dentre $\{1,2,\cdots,\lceil n/2 \rceil\}$ e $\{1,2,\cdots,\lceil m/2 \rceil\}$, respectivamente. A cada iteração, o LAS seleciona r linhas cuja soma seja maximizada nas colunas em J e atualiza o conjunto I. Em seguida, procede de maneira similar para atualizar o conjunto J. Após a convergência da submatriz buscada, o LAS aplica nela um passo de aprimoramento, no qual o número de linhas e de colunas da mesma é adaptado, a fim de maximizar S localmente. Para encontrar um bi-grupo, todo o procedimento descrito é repetido T vezes, sendo retornado aquele que possuir o maior valor para S. Em seguida, a média de seus elementos é subtraída da matriz original. O LAS é executado repetidamente até que um número desejado de bi-grupos seja encontrado ou até que o último bi-grupo reportado não satisfaça um limiar mínimo, previamente definido, para S.

Combinatorial Algorithm for Expression and Sequence-based Cluster Extraction (COALESCE)

Huttenhower *et al.* (2009) introduziram o COALESCE, que é capaz de encontrar bi-grupos com genes positivamente ou negativamente regulados. O algoritmo inicia a construção de um bi-grupo por meio de um par de genes, os quais possuem correlação máxima em todas as condições experimentais da base de dados. Em seguida, dois passos são executados até a convergência do novo bi-grupo. No primeiro deles, cada condição experimental é submetida a um teste z, o qual compara a diferença entre a média de todos os seus valores e a média dos seus valores para os genes contidos no bi-grupo. Assim, são escolhidas todas as condições cujos *p-values* satisfaçam um nível de significância p_e. No passo seguinte, baseando-se na distribuição de valores observados para as condições experimentais selecionadas, a probabilidade a posteriori de cada gene pertencer ao bi-grupo em formação é calculada, e adiciona-se ao mesmo apenas os genes cujas probabilidades estejam acima de um limiar p_g. Após a convergência, subtrai-se de cada condição experimental a média dos valores de expressão dos genes do bi-grupo reportado e inicia-se a busca por outro.

3.4.2 Algoritmos de divisão e conquista

Algoritmos baseados em divisão e conquista possuem como característica a resolução recursiva de um problema, através da aplicação de três passos distintos: divisão do problema considerado em instâncias menores da mesma natureza, resolução de cada uma delas recursivamente e, por fim, combinação dos resultados obtidos, constituindo assim, uma solução para o problema original (CORMEN *et al.*, 2009). Esse tipo de abordagem possui o potencial de ser computacionalmente eficiente em vários cenários. Todavia, no contexto de bi-agrupamento, existe a possibilidade de que bi-grupos de alta qualidade sejam divididos antes que possam ser identificados (MADEIRA; OLIVEIRA, 2004).

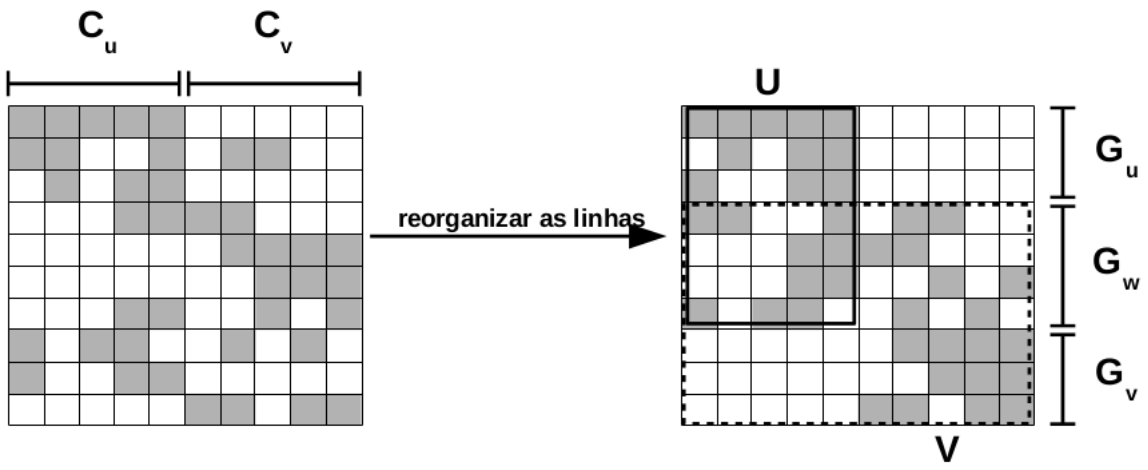
Binary Inclusion-Maximal Biclustering Algorithm (Bimax)

Prelić *et al.* (2006), em seu trabalho comparativo, introduziram o Bimax, o qual recebe como entrada uma matriz binária, com valores iguais a um quando um gene expressa-se em uma condição experimental e zero caso contrário, e busca por submatrizes cujos elementos sejam todos iguais a um. Iniciando com toda a matriz de dados, o Bimax divide a mesma em três submatrizes, sendo uma delas composta apenas por valores iguais a zero. Desse modo, o método é aplicado recursivamente nas duas restantes, até que aquelas compostas apenas por elementos iguais a um sejam encontradas.

Na Figura 4 é apresentado um exemplo dos passos do Bimax. É importante observar que inicialmente o conjunto de colunas é dividido em dois subconjuntos, C_U e C_V, ao se tomar como modelo alguma linha da matriz de dados (nesse exemplo, a primeira). Em seguida, as linhas restantes são reorganizadas, de modo que aquelas que possuem valor igual a um apenas para C_U ocorram primeiro, seguidas daquelas com células iguais a um para C_U e C_V, e por fim

as com valores um somente para C_V. Caso as submatrizes U e V possuam linhas em comum (representadas pelo conjunto G_W), o algoritmo toma um cuidado especial, para que em V sejam gerados apenas bi-grupos que possuam pelo menos uma coluna em comum com C_V. Caso contrário, U e V podem ser processadas de maneira independente.

Figura 4 – Exemplo de um passo do algoritmo Bimax. As células escurecidas representam o valor um e as em branco o valor zero.



Fonte: Baseado na Figura 1 de Prelić *et al.* (2006).

Considerando a restrição do Bimax a dados binários, é necessário aplicar algum tipo de discretização no conjunto de dados de entrada. Para isso, Prelić *et al.* (2006) estabeleceram um limiar no formato:

$$\frac{(\min(A) + \max(A))}{2}, \tag{3.17}$$

onde $\min(A)$ e $\max(A)$ denotam os valores mínimo e máximo da matriz de dados, nessa ordem. Com isso, elementos acima do limiar recebem o valor um e os restantes zero. Por conseguinte, o Bimax é capaz de encontrar bi-grupos com evoluções coerentes tal como aquele na Figura 2f.

3.4.3 Algoritmos de identificação de parâmetros de distribuição

Segundo Madeira e Oliveira (2004), algoritmos deste tipo assumem um determinado modelo estatístico sobre os dados e buscam adaptar os parâmetros do mesmo iterativamente, de modo que um determinado critério seja minimizado.

Plaid

Lazzeroni e Owen (2002), em seu estudo, apresentaram um modelo capaz de representar as interações entre bi-grupos para os quais um mesmo elemento a_{ij} da matriz de dados pertence. Dessa maneira, cada célula a_{ij} é representada como um somatório de funções lineares (denominadas camadas), correspondentes aos bi-grupos aos quais ela pertence. Tal formulação foi

nomeada *plaid model*, sendo definida formalmente como:

$$a_{ij} = \sum_{k=0}^{K} \theta_{ijk} \cdot \rho_{ik} \cdot \eta_{jk}, \tag{3.18}$$

onde K denota a quantidade de camadas (bi-grupos), θ_{ijk} especifica a contribuição de cada bi-grupo k para o valor de a_{ij}, e θ_{ij0} consiste em uma camada que cobre toda a matriz de dados, sendo responsável por representar possíveis variações que não sejam específicas de bi-grupo algum (por exemplo, ruído). Por sua vez, as variáveis ρ_{ik} e η_{jk} assumem valores binários, os quais indicam, respectivamente, se uma determinada linha i ou coluna j pertence à camada k.

A utilização da variável θ_{ijk} tornou o *plaid model* flexível o suficiente para identificar diferentes tipos de bi-grupos. Quando $\theta_{ijk} = \mu_k$, torna-se possível encontrar bi-grupos constantes. Se $\theta_{ijk} = \mu_k + \alpha_{ik}$, pode-se obter bi-grupos com linhas constantes. Com $\theta_{ijk} = \mu_k + \beta_{jk}$, o modelo é capaz de identificar bi-grupos com colunas constantes. Por fim, caso $\theta_{ijk} = \mu_k + \alpha_{ik} + \beta_{jk}$, bi-grupos com valores coerentes são reconhecidos (MADEIRA; OLIVEIRA, 2004).

Em (TURNER; BAILEY; KRZANOWSKI, 2005), foi apresentado um algoritmo capaz de ajustar o *plaid model*, descrito acima, a uma matriz de dados. Assumindo que $K-1$ camadas foram adaptadas, tal algoritmo busca por uma K-ésima camada que minimize a soma dos erros quadráticos, dada por:

$$Q = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{m} (Z_{ij} - \theta_{ijK} \cdot \rho_{iK} \cdot \eta_{jK})^2, \tag{3.19}$$

onde

$$Z_{ij} = a_{ij} - \theta_{ij0} - \sum_{k=1}^{K-1} \theta_{ijk} \cdot \rho_{ik} \cdot \eta_{jk}, \tag{3.20}$$

ou seja, o resíduo de uma célula a_{ij}.

Primeiramente, os valores iniciais para as variáveis ρ_i e η_j são obtidos a partir de resultados gerados pelo algoritmo *k-means* (TAN; STEINBACH; KUMAR, 2006), com $k = 2$, aplicado de maneira independente para as linhas e colunas da matriz de dados. O bi-grupo inicial é formado a partir dos objetos pertencentes aos menores grupos em cada resultado. Posteriormente, os valores de ρ_i e η_j são otimizados por meio de um procedimento baseado no método *Binary Least Squares* (CHATURVEDI; CARROLL, 1994).

Mais tarde, Turner *et al.* (2005) propuseram duas extensões para o algoritmo descrito acima. A primeira delas buscou tornar o modelo parcialmente supervisionado, por meio da utilização de conhecimento prévio acerca dos genes e/ou das condições experimentais consideradas no processo de bi-agrupamento. A segunda teve como objetivo adaptar a formulação do *plaid model*, tornando-a aplicável em cenários cujos dados possuam, além das dimensões dos genes e das condições experimentais, uma dimensão temporal.

Spectral

Kluger *et al.* (2003) desenvolveram um algoritmo capaz de encontrar bi-agrupamentos *checkerboard* compostos por bi-grupos inteiramente constantes. Considerando uma matriz de expressão gênica A normalizada, o algoritmo encontra a decomposição em valores singulares da mesma no formato $A = U\Sigma V^T$ e, dentre os autovetores em V associados aos l maiores autovalores em Σ, seleciona os l_{best} capazes de gerar um padrão *checkerboard*. Após isso, as linhas de A são projetadas nesses l_{best} autovetores e o algoritmo *k-means*, com $k = k_r$, é aplicado, e assim, um agrupamento para as linhas é obtido. Para a matriz U, um procedimento similar é realizado para produzir um agrupamento com k_c grupos para as colunas de A.

No passo seguinte do algoritmo, os k_r grupos obtidos a partir da projeção das linhas nos melhores autovetores de V e os k_c grupos obtidos a partir da projeção das colunas nos melhores autovetores de U são combinados tal como um produto cartesiano, produzindo uma quantidade $k_r \cdot k_c$ de bi-grupos sem sobreposição.

Bayesian BiClustering (BBC)

Gu e Liu (2008) introduziram uma abordagem capaz de adaptar um modelo Bayesiano hierárquico, nomeado BBC, a uma matriz de dados. Para apenas um bi-grupo, o procedimento assume o *plaid model*. Para múltiplos bi-grupos, a formulação do método restringe a existência de sobreposição, de modo que ela ocorra apenas nas linhas ou nas colunas da matriz de dados, uma vez que, segundo os autores, bi-grupos encontrados pelo *plaid model* original tendem a se sobrepor em uma grande extensão. Para realizar a inferência de bi-grupos, Gu e Liu (2008) utilizaram a técnica de amostragem de Gibbs. Assim, os valores amostrados representam o grau de pertinência de uma linha/coluna a um bi-grupo. Quando se deseja encontrar partições rígidas, limiares para os mesmos podem ser estabelecidos (EREN *et al.*, 2013).

Factor Analysis for Bicluster Acquisition (FABIA)

Em (HOCHREITER *et al.*, 2010), cada bi-grupo de uma matriz de dados A é representado como a multiplicação de um vetor λ_i, o qual possui valores diferentes de zero para genes que pertencem ao bi-grupo i e zero caso contrário, com um vetor z_i^T, definido de maneira similar para as condições experimentais. Uma matriz de dados A é modelada como a soma de k camadas (bi-grupos) e uma camada de ruído Υ:

$$A = \sum_{i=1}^{k} \lambda_i z_i^T + \Upsilon = \Lambda Z + \Upsilon, \tag{3.21}$$

onde $\Lambda \in \mathbb{R}^{n \times k}$ possui como colunas todos os vetores λ_i e $Z \in \mathbb{R}^{k \times m}$ tem como linhas os vetores z_i^T. Para ajustar a Equação (3.21) aos dados, os autores utilizaram uma abordagem baseada em um modelo de análise de fatores, em conjunto com um algoritmo *Expectation Maximization* variacional (GIROLAMI, 2001; PALMER *et al.*, 2005) a fim de selecionar os parâmetros mais

adequados para tal modelo. Ao fim, as associações dos genes e das condições experimentais aos bi-grupos são *fuzzy*, representadas pelos valores absolutos presentes nos vetores λ_i e z_i, nessa ordem. Para a obtenção de partições rígidas, o FABIA aplica limiares nos elementos de tais vetores.

3.4.4 Algoritmos de enumeração exaustiva

Algoritmos baseados nesta abordagem partem do princípio que os melhores bi-grupos em uma matriz de dados podem ser identificados apenas pela geração de todas as possíveis combinações de linhas e colunas da mesma. Por conseguinte, eles tipicamente encontram soluções ótimas para a formulação considerada. Entretanto, devido à complexidade do problema, para que a enumeração seja factível, normalmente são impostas restrições no tamanho das submatrizes buscadas (MADEIRA; OLIVEIRA, 2004).

Statistical-Algorithmic Method for Bicluster Analysis (SAMBA)

Tanay, Sharan e Shamir (2002) propuseram o procedimento SAMBA, o qual representa os dados de entrada por meio de um grafo bipartido ponderado $G = (U, V, E)$, sendo U o conjunto de condições experimentais, V o conjunto de genes, e $(u, v) \in E$ se, e somente se, o nível de expressão do gene v altera-se significativamente sob a condição u. Desse modo, os bi-grupos existentes em uma base de dados correspondem a subgrafos bipartidos completos (bicliques) no formato $H = (U', V', E')$.

Dentre todos os bicliques, o SAMBA busca identificar aqueles de ponderação máxima, assumindo que tal ponderação corresponde à significância estatística de cada um. Adicionalmente, o algoritmo impõe uma restrição no grau dos vértices representando genes, considerando apenas aqueles que não excedam uma constante d. Essa restrição existe por dois motivos: (i) para evitar um pior caso exponencial, e (ii) geralmente, genes que apresentam expressão elevada com alta frequência não são interessantes, pois participam de vários processos, não manifestando um foco específico.

O SAMBA pode ser executado com base em dois modelos estatísticos referentes ao resultado esperado, definidos em (TANAY; SHARAN; SHAMIR, 2002). O modelo mais simples assume que as arestas de G podem ocorrer de maneira independente e equiprovável. Com isso, os autores demonstraram como calcular um limite superior da probabilidade de ocorrência de um bi-grupo. O modelo mais sofisticado assume que cada aresta (u, v) é uma variável de Bernoulli, estimando a probabilidade de ocorrência de cada uma por meio de um processo de Monte Carlo, considerando grafos com sequências de graus idênticas à de G que incluem (u, v). Dessa forma, os autores introduziram um método para ponderar as arestas, levando em conta a direção em que a expressão de cada gene foi alterada (positivamente ou negativamente), de modo que um bi-grupo corresponda a um subgrafo de máxima verossimilhança.

A busca por bi-grupos através do SAMBA ocorre em três fases. Na primeira delas, o grafo bipartido é formado, e os pesos de suas arestas calculados segundo algum dos modelos comentados anteriormente. É considerado que um gene se expressou positivamente ou negativamente sob uma condição experimental caso o seu nível de expressão, com média zero e desvio-padrão unitário, está acima de um ou abaixo de menos um, respectivamente. No segundo passo, para cada vértice $v \in U \cup V$, é aplicado um algoritmo capaz de examinar todos os seus subconjuntos de vértices adjacentes, com a finalidade de identificar os k melhores bicliques que o intersectam. Para diminuir o tempo de processamento, apenas subconjuntos que possuam tamanho dentro de um intervalo (N_1, N_2) são levados em conta. Por fim, um procedimento guloso é aplicado, com a finalidade de melhorar o resultado final. Tal procedimento consiste em executar a melhor ação em cada bi-grupo (adição/remoção de um único vértice) repetidamente, até que não seja possível aperfeiçoá-lo. Complementarmente, para evitar a existência de bi-grupos muito semelhantes, é efetuada uma filtragem na saída obtida, removendo aqueles que possuam mais do que $L\%$ dos seus vértices em comum com outro previamente gerado. Na Figura 2j é ilustrado o tipo de bi-grupo encontrado por esse algoritmo.

Bit-Pattern Biclustering Algorithm (BiBit)

Rodriguez-Baena, Perez-Pulido e Aguilar-Ruiz (2011) apresentaram o BiBit, o qual busca por bi-grupos de tamanho máximo em matrizes binárias, onde cada elemento com valor igual a um indica que um gene expressou-se em uma condição experimental e zero indica o contrário, formados a partir de padrões obtidos ao se aplicar o operador lógico $\wedge$ (AND) a todos os $n \cdot (n-1)/2$ pares possíveis de linhas.

Preliminarmente, o conjunto de dados de entrada é transformado em uma matriz de valores inteiros. Para isso, cada linha é dividida em palavras de *bits* de um certo tamanho, sendo cada palavra transformada em sua representação decimal inteira. Em seguida, cada par de linhas é utilizado para produzir um padrão ρ, por meio do operador $\wedge$. Caso a quantidade de dígitos iguais a um em ρ seja maior ou igual a um parâmetro referente ao número mínimo de colunas de um bi-grupo, o algoritmo aplica o mesmo operador para ρ e toda linha i restante da matriz de dados, sendo inseridas na submatriz considerada aquelas que satisfaçam a restrição $i \wedge \rho = \rho$. Complementarmente, para ser incluída no bi-agrupamento final, uma submatriz deve satisfazer um limiar concernente à quantidade mínima de linhas que um bi-grupo deve conter. Esse algoritmo encontra bi-grupos semelhantes aos que resultam do Bimax (Figura 2f).

Considerando que o BiBit exige como entrada uma matriz de dados cujos elementos sejam valores binários, os autores desse algoritmo propuseram também um procedimento de discretização. Primeiramente, a base de dados é normalizada para média zero e desvio-padrão unitário. Em seguida, valores dentro do intervalo $[-3, 3]$ são divididos em doze níveis distintos, igualmente espaçados, com valores abaixo de -3 pertencendo ao primeiro nível e valores maiores que 3 pertencendo ao último. A base de dados original é discretizada por meio da conversão

de cada um de seus valores de expressão em um inteiro $z \in \{0, 1, \cdots, 11\}$, correspondendo ao nível que pertence. Após isso, para cada nível z uma nova matriz binária é gerada, a qual contém valor igual a um para elementos cujos valores inteiros sejam maiores ou iguais a z, e zero caso contrário. O BiBit é executado em todas as bases de dados binárias geradas, e os resultados são combinados, gerando assim uma solução final.

Differentially Expressed Biclusters (DeBi)

Serin e Vingron (2011), em seu trabalho, desenvolveram o algoritmo DeBi, o qual toma por base uma abordagem para descoberta de regras de associação amplamente utilizada em mineração de dados, conhecida como *Maximal Frequent Itemset* (MFI) (BURDICK; CALIMLIM; GEHRKE, 2001). De modo geral, o DeBi ocorre em três passos, a serem descritos a seguir.

Primeiramente, deriva-se a partir da matriz de entrada uma matriz com valores binários, denotada por A^+ e que representa as expressões positivas dos respectivos genes sobre as condições experimentais consideradas. Cada elemento a^+_{ij} de A^+ é definido como:

$$a^+_{ij} = \begin{cases} 1 & \text{se o gene } i \text{ expressa-se positivamente na condição experimental } j \\ 0 & \text{caso contrário.} \end{cases} \tag{3.22}$$

Caso se deseje buscar por bi-grupos negativamente regulados, uma matriz A^- pode ser definida de maneira similar.

No segundo passo, o algoritmo busca identificar grupos de genes com suporte acima de um limiar c_1 e cardinalidade acima de um limiar c_2. O suporte de um gene i é definido como a fração de condições experimentais para as quais i assume valor um. Para um conjunto de genes, o suporte é calculado a partir do vetor binário resultante da aplicação da operação lógica $\wedge$, elemento a elemento, entre todos os seus membros. Assim, diz-se que um conjunto de genes é (c_1, c_2)-frequente, se o mesmo satisfaz as restrições mencionadas no início deste parágrafo.

Para a identificação de MFIs cujos suportes estejam acima do limiar c_1, o DeBi utiliza o *MAximal Frequent Itemset Algorithm* (MAFIA) (BURDICK; CALIMLIM; GEHRKE, 2001), o qual realiza uma busca em profundidade em uma árvore constituída pelos conjuntos possíveis de genes, aplicando técnicas eficientes de poda, com a finalidade de evitar o pior caso de uma enumeração exaustiva. Na primeira aplicação, é utilizado um valor de c_1 igual ao valor máximo de suporte encontrado para um gene na matriz, o qual é reduzido em execuções subsequentes. Os genes que compõem os MFIs encontrados são removidos da matriz de dados. Desse modo, cada execução do MAFIA utiliza as matrizes modificadas na iteração anterior. Todo esse procedimento é repetido enquanto o valor de c_1 considerado for maior do que um parâmetro c_{min} pré-determinado.

Ao final, são realizadas inserções/remoções de genes nos bi-grupos encontrados. Para isso, o DeBi utiliza o teste exato de Fisher para avaliar a significância da associação de um gene

a um bi-grupo. Se tal avaliação resultar em um *p-value* abaixo de um parâmetro α, o gene é incluído no bi-grupo. Caso contrário, é excluído do mesmo. Adicionalmente, são removidos do bi-agrupamento final todos os bi-grupos que possuam mais do que $L\%$ dos seus elementos em comum com outros de maior tamanho.

3.5 Avaliação de resultados

A forma de avaliação dos resultados obtidos em análises experimentais depende diretamente do tipo de dados considerado e da própria tarefa de análise em questão. Devido a isso, o restante desta seção está organizada conforme segue. Em 3.5.1 é comentado sobre a avaliação de resultados em bases de dados que compreendem cenários reais onde não se dispõe de uma solução de referência na forma de bi-grupos pré-conhecidos e rotulados. Em 3.5.2 são apresentados índices de validação externa para avaliação de bi-agrupamentos em bases de dados com bi-grupos pré-conhecidos.

3.5.1 Validação sem uma solução de referência

Validação para agrupamento de genes

Ao trabalhar com bases de dados reais, existem diversas formas adotadas na literatura para avaliação de resultados, que podem variar para cada tipo de problema tratado. Por exemplo, para o problema de agrupamento de genes, existem bases de dados reais compostas por genes anotados em ontologias (ASHBURNER *et al.*, 2000), uma forma de organização hierárquica dos genes de um determinado organismo em categorias conhecidas. Com isso, torna-se possível mensurar a compatibilidade entre tais categorias e os grupos identificados por algum algoritmo de agrupamento. Uma forma de avaliar tal compatibilidade consiste em analisar a significância biológica dos grupos por meio de enriquecimento de genes, o qual é capaz de fornecer *p-values* que indicam o grau de aleatoriedade dos grupos encontrados (EREN *et al.*, 2013). Em particular, quanto menor o *p-value* de um bi-grupo for, maior a evidência de que os genes pertencentes a ele participam de um mesmo processo biológico (LI *et al.*, 2009). No presente trabalho, tais *p-values* foram calculados por meio de uma distribuição hipergeométrica (Equação 3.23), assim como feito em (TANAY; SHARAN; SHAMIR, 2002; PRELIĆ *et al.*, 2006; LI *et al.*, 2009; SERIN; VINGRON, 2011; EREN *et al.*, 2013).

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{c}{i} \cdot \binom{n-c}{y-i}}{\binom{n}{y}} \tag{3.23}$$

Na Equação (3.23), n indica o número total de genes na base de dados, c indica a quantidade de genes da base que pertencem a uma determinada categoria existente na ontologia,

y é o tamanho do grupo cujo *p-value* está sendo calculado e k indica o número de genes do grupo que pertencem à categoria analisada. Deve-se observar que um mesmo grupo pode conter genes de diferentes categorias, o que pode acarretar em vários *p-values* distintos associados a ele.

Validação para agrupamento de condições experimentais

Problemas que possuem como objetivo o agrupamento de condições experimentais são, em princípio, menos complexos, uma vez que existem várias bases de dados reais disponíveis cujas condições experimentais foram rotuladas (JASKOWIAK; CAMPELLO; COSTA, 2013). Desse modo, pode-se avaliar a compatibilidade entre os grupos de condições experimentais rotulados e os bi-grupos obtidos por algoritmos de bi-agrupamento utilizando-se medidas convencionais de comparação de partições.

Seja X um conjunto composto por n objetos x_j. Uma partição para X composta por k grupos pode ser representada através de uma matriz P, de dimensão $k \times n$, onde cada elemento P_{ij} representa o grau de pertinência do objeto x_j ao i-ésimo grupo. Tal matriz deve satisfazer três propriedades (HORTA; CAMPELLO, 2015):

$$P_{ij} \in [0,1] \qquad \forall i \in \{1,\cdots,k\}, \forall j \in \{1,\cdots,n\}, \tag{3.24a}$$

$$\sum_{j=1}^{n} P_{ij} > 0 \qquad \forall i \in \{1,\cdots k\}, \tag{3.24b}$$

$$\sum_{i=1}^{k} P_{ij} > 0 \qquad \forall j \in \{1,\cdots n\}. \tag{3.24c}$$

Com isso, quatro tipos de agrupamento podem ser definidos: (i) agrupamento possibilístico, determinado por $N_p = \{P \in \mathbb{R}^{k \cdot n} \mid P \text{ satisfaz } (3.24)\}$; (ii) agrupamento *fuzzy*/probabilístico, descrito por $N_f = \{P \in N_p \mid \sum_{i=1}^{k} P_{ij} = 1 \quad \forall j\}$; (iii) agrupamento particional não exclusivo, estabelecido por $N_{pne} = \{P \in N_p \mid P_{ij} \in \{0,1\} \quad \forall i,j\}$; e (iv) agrupamento particional exclusivo, determinado por $N_{pe} = N_f \cap N_{pne} = \{P \in N_p \mid P_{ij} \in \{0,1\} \text{ e } \sum_{i=1}^{k} P_{ij} = 1 \quad \forall i,j\}$ (ANDERSON *et al.*, 2010; CAMPELLO, 2010; HORTA; CAMPELLO, 2015).

Horta e Campello (2015) revisaram 28 medidas para comparação de partições nos diferentes domínios de agrupamento de dados descritos acima, discutiram as principais características e limitações de cada uma, e apresentaram um estudo experimental considerando importantes aspectos para avaliação de resultados de agrupamento de dados a partir de uma perspectiva prática. Ademais, os autores propuseram o Índice Grand (13GRI) e sua versão ajustada para aleatoriedade (13AGRI), capazes de mensurar a semelhança entre partições possibilísticas. Ao fim do trabalho, os resultados obtidos indicaram que o 13AGRI e outro índice, nomeado originalmente como Índice de Rand Ajustado *Fuzzy* (FARI) (BROUWER, 2009), foram superiores aos demais para os domínios de agrupamento descritos acima. Portanto, neste trabalho, ao avaliar as soluções encontradas por algoritmos de bi-agrupamento em bases de dados reais cujas condições experimentais foram rotuladas, as medidas FARI e 13AGRI foram utilizadas, uma vez que tais

soluções estão inseridas no domínio de agrupamentos particionais não exclusivos, pois podem conter grupos sobrepostos (ou seja, grupos com condições experimentais compartilhadas) ou, até mesmo, condições experimentais não agrupadas, as quais foram consideradas como *singletons*[2] durante a avaliação.

No restante desta subseção são apresentadas as medidas FARI e 13AGRI. Para isso, define-se uma matriz de co-associação $J^P = P^T P$ de dimensão $n \times n$, onde, no domínio de agrupamentos particionais não exclusivos, cada elemento J^P_{ij} indica a quantidade de vezes que dois objetos x_i e x_j pertencem a um mesmo grupo em P.

Índice de Rand Ajustado *Fuzzy* (FARI)

Boa parte das medidas mais conhecidas para a comparação de agrupamentos particionais exclusivos baseiam-se na contagem de pares de objetos (HORTA; CAMPELLO, 2015). Tipicamente, nesse tipo de abordagem, ao comparar duas partições $P \in \mathbb{R}^{k\times n}$ e $Q \in \mathbb{R}^{l\times n}$, quatro variáveis são definidas: a, que aponta a quantidade de pares de objetos em um mesmo grupo em P e Q (ou seja, $a = \sum_{i<j} J^P_{ij} \cdot J^Q_{ij}$)[3]; b, a qual indica o número de pares de objetos em um mesmo grupo em P mas em diferentes grupos em Q (ou seja, $b = \sum_{i<j} J^P_{ij} \cdot (1 - J^Q_{ij})$); c, que representa o número de pares de objetos em diferentes grupos em P mas em um mesmo grupo em Q (ou seja, $c = \sum_{i<j} (1 - J^P_{ij}) \cdot J^Q_{ij}$); e d, que determina a quantidade de pares de objetos em diferentes grupos tanto em P quanto em Q (ou seja, $d = \sum_{i<j} (1 - J^P_{ij}) \cdot (1 - J^Q_{ij})$). A partir delas, diversas medidas podem ser definidas (ALBATINEH; NIEWIADOMSKA-BUGAJ; MIHALKO, 2006). Dentre elas, uma das mais conhecidas e utilizadas é o Índice de Rand Ajustado (ARI) (HUBERT; ARABIE, 1985):

$$ARI(P,Q) = \frac{a - \frac{(a+c)\cdot(a+b)}{a+b+c+d}}{\frac{(a+c)+(a+b)}{2} - \frac{(a+c)\cdot(a+b)}{a+b+c+d}}. \tag{3.25}$$

Em (BROUWER, 2009), o autor propôs uma extensão para o ARI, tornando o mesmo apto a comparar agrupamentos possibilísticos através da redefinição das variáveis a, b, c e d em:

$$\hat{a} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \hat{J}^P_{ij} \cdot \hat{J}^Q_{ij} - n}{2}, \tag{3.26a}$$

$$\hat{b} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \hat{J}^P_{ij} \cdot (1 - \hat{J}^Q_{ij})}{2}, \tag{3.26b}$$

$$\hat{c} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} (1 - \hat{J}^P_{ij}) \cdot \hat{J}^Q_{ij}}{2}, \tag{3.26c}$$

$$\hat{d} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} (1 - \hat{J}^P_{ij}) \cdot (1 - \hat{J}^Q_{ij})}{2}, \tag{3.26d}$$

[2] Um *singleton* consiste em um grupo contendo apenas um objeto.

[3] Neste caso é importante notar que, no domínio de agrupamentos particionais exclusivos, cada elemento J^P_{ij} (ou J^Q_{ij}) assume valor um se dois objetos pertencem a um mesmo grupo em P (ou Q) e zero caso contrário.

onde as matrizes $\hat{J}^P$ e $\hat{J}^Q$ consistem, respectivamente, em J^P e J^Q normalizadas de modo que possuam linhas com norma unitária. Ao substituir as variáveis originais da fórmula do ARI pelas suas respectivas redefinições, dá-se origem ao FARI.

Índice Grand Ajustado (13AGRI)

Para as medidas 13GRI e 13AGRI, Horta e Campello (2015) utilizaram, além da matriz de co-associação J^P de uma partição $P \in \mathbb{R}^{k\times n}$, uma segunda matriz no formato $S^P = P^T(\mathbb{1}_k - \mathbb{I}_k)P$, onde $\mathbb{1}_k$ é uma matriz $k \times k$ com todos os seus elementos iguais a um e $\mathbb{I}_k$ é a matriz identidade $k \times k$. Intuitivamente, no domínio de agrupamentos particionais não exclusivos, cada elemento S^P_{ij} indica a quantidade de vezes que dois objetos x_i e x_j pertencem a grupos diferentes em P.

Dadas duas partições $P \in \mathbb{R}^{k\times n}$ e $Q \in \mathbb{R}^{l\times n}$ para uma base de dados X, Horta e Campello (2015) definiram o 13GRI conforme a Equação (3.27), onde $T^P = J^P + S^P$ e $T^Q = J^Q + S^Q$.

$$\text{13GRI}(P,Q) = \frac{\sum_{i<j}\min\{J^P_{ij}, J^Q_{ij}\} + \sum_{i<j}\min\{S^P_{ij}, S^Q_{ij}\}}{\max\{\sum_{i<j}T^P_{ij}, \sum_{i<j}T^Q_{ij}\}} \tag{3.27}$$

No mesmo trabalho, Horta e Campello (2015) ajustaram o 13GRI para aleatoriedade, com a finalidade tornar esse índice capaz de apresentar uma avaliação com valor esperado nulo quando soluções de agrupamento geradas aleatoriamente fossem comparadas, dando origem ao 13AGRI:

$$\text{13AGRI}(P,Q) = \frac{13GRI(P,Q) - E[13GRI]_{P,Q}}{1 - E[13GRI]_{P,Q}}, \tag{3.28}$$

onde $E[13GRI]_{P,Q}$ indica o valor esperado para o 13GRI ao comparar as partições P e Q. Para detalhes referentes ao modelo nulo assumido para o índice, bem como o cálculo de $E[13GRI]_{P,Q}$, recomenda-se a leitura do artigo original.

Considerações

É importante notar que as formas de validação mencionadas nesta subseção possuem uma grande desvantagem, pois levam em conta apenas uma das dimensões do problema, independente da dimensão restante. Ou seja, nessas abordagens tipicamente avalia-se a qualidade de grupos de genes ou grupos de condições experimentais frente a algum conhecimento prévio de domínio, documentado por meio de algum tipo de rotulação externa. Com o intuito de complementar esse tipo de análise durante a avaliação de um algoritmo, pode-se recorrer a cenários nos quais se conheça previamente a estrutura dos bi-grupos contidos nos dados, tal como será descrito a seguir.

3.5.2 Validação externa com uma solução de referência

Em agrupamento de dados, índices de validação externa são utilizados para medir a eficácia de um algoritmo por meio da comparação da estrutura de uma solução encontrada pelo mesmo com outra conhecida de antemão (JAIN; DUBES, 1988). Diferente dos métodos apresentados na subseção anterior, esse tipo de abordagem tem como característica não necessitar de suposições prévias acerca dos padrões de expressão dos genes (HORTA; CAMPELLO, 2014).

Até onde se sabe, em bi-agrupamento de dados são raríssimas as bases de dados reais com bi-grupos conhecidos e rotulados. Para avaliar algoritmos dessa maneira, bases de dados sintéticas são tipicamente usadas na literatura. Para isso, se torna necessária a aplicação de índices que sejam capazes de aferir adequadamente a compatibilidade entre um bi-agrupamento gerado por um algoritmo com aquele implantado nos dados sintéticos.

Em um estudo recente, Horta e Campello (2014) investigaram oito propriedades desejáveis que devem ser satisfeitas por medidas de avaliação externa de bi-agrupamentos. Além disso, revisaram 14 medidas existentes na literatura, provando quais propriedades cada uma apresenta. As comparações, de caráter tanto teórico como empírico, demonstraram que boa parte dessas medidas (dentre as quais aquelas utilizadas nos estudos comparativos anteriores de algoritmos de bi-agrupamento, a saber, (PRELIĆ *et al.*, 2006; BOZDAĞ; KUMAR; ÇATALYÜREK, 2010; EREN *et al.*, 2013)) possuem sérias limitações que podem acarretar em resultados contra-intuitivos e até mesmo enganosos em várias situações. Ao fim do trabalho, os autores recomendaram a utilização de duas medidas dentre as 14 avaliadas, as quais satisfazem sete das oito propriedades formalizadas e apresentaram os melhores resultados na fase experimental, sendo, portanto, superiores às demais. Tendo por base tais fatos, no presente trabalho, quando foram realizadas análises experimentais que envolveram coleções de bases de dados cujas soluções eram conhecidas, foram utilizadas as duas medidas sugeridas por Horta e Campello (2014), a saber, *Clustering Error* e *Campello Soft Index*, que serão descritas à frente nesta seção.

Outra medida de validação externa para agrupamentos em subespaços recentemente publicada, a qual poderia ser aplicada no contexto de bi-agrupamento de dados, chama-se E4SC (GÜNNEMANN *et al.*, 2011). Como parte preliminar do presente trabalho, foram desenvolvidas as devidas provas para as propriedades investigadas por Horta e Campello (2014), o que tornou possível verificar que apenas três delas eram satisfeitas. Devido a esse fato, a E4SC não foi utilizada. As provas e as propriedades são apresentadas no Apêndice A.

No restante da presente subseção são apresentadas as duas medidas recomendadas por Horta e Campello (2014). Para isso, considera-se a existência de dois bi-agrupamentos, $B = \{B_i\}_{i=1}^{k} = \{(I_i, J_i)\}_{i=1}^{k}$ e $\hat{B} = \{\hat{B}_i\}_{i=1}^{q} = \{(\hat{I}_i, \hat{J}_i)\}_{i=1}^{q}$, os quais representam, respectivamente, o bi-agrupamento encontrado por um algoritmo (contendo k bi-grupos) e a solução de referência conhecida para a base de dados utilizada (contendo q bi-grupos). As duas medidas variam no intervalo $[0,1]$, sendo que valores mais altos indicam uma maior equivalência da solução

encontrada por um algoritmo com a solução de referência.

Clustering Error (CE)

A medida CE foi originalmente proposta no estudo de Patrikainen e Meila (2006) em conjunto com outras três medidas para avaliação de agrupamentos em subespaços. Denotando por $N_{i,j}$ e $\hat{N}_{i,j}$ a quantidade de bi-grupos que um elemento a_{ij} pertence nos bi-agrupamentos B e $\hat{B}$, o tamanho de um conjunto formado pela união de grupos sobrepostos pode ser definido como:

$$|U| = \sum_{i,j} \max\{N_{i,j}, \hat{N}_{i,j}\}. \tag{3.29}$$

Considerando uma matriz de confusão M na qual cada célula representa a quantidade de elementos a_{ij} compartilhados entre bi-grupos de soluções distintas, uma permutação de suas linhas e colunas é encontrada, de modo que a soma dos elementos na diagonal de M seja maximizada[4], cujo valor é denotado por $d_{\max}$. Em (PATRIKAINEN; MEILA, 2006) a medida CE foi introduzida como um índice de dissimilaridade. A forma utilizada nesta dissertação condiz com aquela apresentada por Horta e Campello (2014) que, para fins comparativos, transformaram-na em um índice de similaridade:

$$\mathrm{CE}(B, \hat{B}) = \frac{d_{\max}}{|U|}. \tag{3.30}$$

Campello Soft Index (CSI)

Horta e Campello (2014) utilizaram uma abordagem capaz de representar um bi-agrupamento por meio de um agrupamento particional não exclusivo para permitir a aplicação de medidas de validação externa desenvolvidas para o segundo caso em cenários envolvendo o primeiro. Considerando uma matriz de dados de dimensão $n \times m$, cada elemento a_{ij} da mesma é visto como um objeto dentro de um conjunto $O = \{o_1, o_2, \cdots, o_{n \cdot m}\}$, onde o_1 atua como a_{11}, o_2 como a_{21} e assim por diante. Seguindo essa ideia, os autores definiram um mapeamento no formato:

$$\pi(i, j) = i + n \cdot (j - 1) \qquad \forall i \in \{1, \cdots, n\}, \forall j \in \{1, \cdots, m\}. \tag{3.31}$$

Assim, um bi-grupo $B_i = (I_i, J_i)$ pode ser transformado em um grupo P_i conforme demonstrado pela Equação (3.32).

$$P_i = \bigcup_{x \in I_i, y \in J_i} \{o_{\pi(x,y)}\} \tag{3.32}$$

Ao considerar um bi-agrupamento B constituído por k bi-grupos, torna-se possível obter uma partição não exclusiva P composta por k grupos. É interessante observar que, em vários

[4] Caso necessário, M é transformada em uma matriz quadrada por meio da inserção de linhas ou colunas preenchidas inteiramente com zeros.

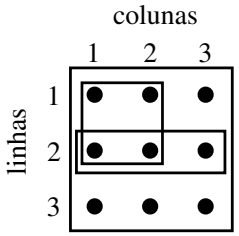
cenários de aplicação, alguns elementos da matriz de dados original podem não fazer parte de nenhum bi-grupo encontrado. Desse modo, Horta e Campello (2014) sugeriram a inserção de *singletons* em P representando tais elementos, resultando em:

$$P = \{P_1, P_2, \cdots, P_k, P_{k+1}, \cdots, P_{k+z}\}, \tag{3.33}$$

onde grupos com índice maior do que k representam tais *singletons*.

Como exemplo, considere o bi-agrupamento apresentado na Figura 5, composto pelos bi-grupos $B_1 = (I_1, J_1) = (\{1,2\}, \{1,2\})$ e $B_2 = (I_2, J_2) = (\{2\}, \{1,2,3\})$. Aplicando o mapeamento descrito na Equação (3.31) em todas as entradas da matriz, obtém-se como resultado um novo conjunto $O = \{o_1, o_2, \cdots, o_9\}$. Em seguida, por meio das Equações (3.32) e (3.33), a partição $P = \{P_1, P_2, \cdots, P_6\}$ é gerada, onde $P_1 = \{o_1, o_2, o_4, o_5\}$, $P_2 = \{o_2, o_5, o_8\}$, $P_3 = \{o_3\}$, $P_4 = \{o_6\}$, $P_5 = \{o_7\}$ e $P_6 = \{o_9\}$.

Figura 5 – Matriz de dados 3×3 contendo dois bi-grupos.



Fonte: Baseado na Figura 1 de Horta e Campello (2014).

Após a conversão dos bi-agrupamentos B (sob avaliação) e $\hat{B}$ (referência) em P e $\hat{P}$, alguma medida de validação externa para agrupamentos particionais não exclusivos pode ser aplicada. Horta e Campello (2014), baseados em seu estudo teórico e empírico, sugeriram o CSI (CAMPELLO, 2010).

Sejam $\alpha_P(o_i, o_j)$ e $\alpha_{\hat{P}}(o_i, o_j)$ a quantidade de grupos nos quais o_i e o_j participam conjuntamente em P e $\hat{P}$, nessa ordem. Define-se também $\beta_P(o_i)$ e $\beta_{\hat{P}}(o_i)$ como a quantidade de grupos que o objeto o_i pertence em P e $\hat{P}$, respectivamente, menos um. Com isso, os acordos ($a_{P,\hat{P}}$) e desacordos ($d_{P,\hat{P}}$) agregados aos dois objetos i e j para as partições P e $\hat{P}$, podem ser definidos como:

$$a_{P,\hat{P}}(o_i, o_j) = \min\{\alpha_P(o_i, o_j), \alpha_{\hat{P}}(o_i, o_j)\} + \sum_{l \in \{i,j\}} \min\{\beta_P(o_l), \beta_{\hat{P}}(o_l)\}, \tag{3.34}$$

$$d_{P,\hat{P}}(o_i, o_j) = |\alpha_P(o_i, o_j) - \alpha_{\hat{P}}(o_i, o_j)\}| + \sum_{l \in \{i,j\}} |\beta_P(o_l) - \beta_{\hat{P}}(o_l)|. \tag{3.35}$$

Ao somar os valores de $a_{P,\hat{P}}(o_i, o_j)$ e $d_{P,\hat{P}}(o_i, o_j)$ para todos os possíveis pares ordenados (o_i, o_j) $(i < j)$, obtém-se suas respectivas medidas globais, $a_{P,\hat{P}}$ e $d_{P,\hat{P}}$. Desse modo, o CSI é formulado como:

$$\text{CSI}(P, \hat{P}) = \frac{a_{P,\hat{P}}}{a_{P,\hat{P}} + d_{P,\hat{P}}}. \tag{3.36}$$

3.6 Estudos comparativos em bi-agrupamento de dados

Prelić *et al.* (2006) apresentaram o primeiro trabalho comparativo em bi-agrupamento de dados da literatura. Ao todo, foram comparados seis algoritmos (Bimax, CCA, ISA, OPSM, SAMBA e xMOTIFs) em uma coleção de bases de dados sintéticas contendo bi-grupos constantes positivamente regulados ou bi-grupos com colunas constantes positivamente regulados e em duas bases de dados reais. Em dados sintéticos, os resultados foram avaliados segundo duas medidas de validação externa propostas pelos próprios autores, as quais satisfazem apenas uma dentre as oito propriedades desejáveis estudadas em (HORTA; CAMPELLO, 2014). Em dados reais, os bi-grupos encontrados foram avaliados através de enriquecimento de grupos de genes através da *Gene Ontology* (ASHBURNER *et al.*, 2000) e de redes de interação proteína-proteína.

Bozdağ, Kumar e Çatalyürek (2010) revisaram seis algoritmos de bi-agrupamento da literatura, comparando cinco deles (CCA, CPB, MSSRCC, OPSM e SAMBA) em três bases de dados reais e em bases de dados sintéticas contendo bi-grupos aditivos e multiplicativos. Para validação dos resultados, foram realizados enriquecimentos de bi-grupos em bases de dados reais e, para as bases sintéticas, os bi-agrupamentos encontrados pelos algoritmos foram avaliados a partir de duas medidas de validação externa propostas pelos autores, sendo que uma delas satisfaz apenas duas dentre as oito propriedades estudadas por Horta e Campello (2014), e ainda assim apenas em cenários que a sobreposição entre bi-grupos não existe, enquanto que a outra viola todas as oito.

Finalmente, Eren *et al.* (2013) desenvolveram um estudo comparativo incluindo doze algoritmos (BBC, Bimax, CCA, COALESCE, CPB, FABIA, ISA, OPSM, Plaid, QUBIC, Spectral e xMOTIFs), oito bases de dados reais, e uma coleção de bases de dados sintéticas compostas por seis diferentes modelos de bi-grupos (constantes, positivamente regulados, aditivos, multiplicativos, aditivos-multiplicativos e *plaid*). Os autores investigaram qual modelo era melhor identificado por cada algoritmo, resultando em uma comparação mais justa em relação aos trabalhos anteriores. Para avaliação dos resultados em bases de dados sintéticas, duas medidas baseadas no coeficiente de Jaccard (JACCARD, 1908) foram propostas, as quais obedecem apenas uma das oito propriedades desejáveis em (HORTA; CAMPELLO, 2014). Em bases de dados reais, os bi-grupos encontrados foram enriquecidos, assim como nos trabalhos anteriores.

Observa-se portanto que, à parte da utilização de coleções limitadas de bases de dados e/ou de algoritmos, os trabalhos comparativos anteriores em (PRELIĆ *et al.*, 2006; BOZDAĞ; KUMAR; ÇATALYÜREK, 2010; EREN *et al.*, 2013) utilizaram medidas de validação externa de bi-agrupamento de dados que podem levar a resultados questionáveis, o que é objeto de estudo do presente trabalho.

3.7 Considerações

Neste capítulo foram apresentados os conceitos de bi-agrupamento de dados fundamentais para o entendimento do presente trabalho. Primeiramente a complexidade do problema foi brevemente introduzida. Após isso, foram comentados os principais tipos de padrões de bi-grupos que podem ser buscados em bases de dados de expressão gênica. Em seguida, foram apresentados 17 algoritmos do estado da arte na área de bi-agrupamento de dados, os quais foram utilizados durante a fase experimental deste estudo, bem como as medidas que foram escolhidas e utilizadas para avaliá-los. Por fim, os três trabalhos comparativos mais importantes da literatura foram sumarizados e suas principais limitações, que são alvo principal do presente estudo, foram destacadas.

CAPÍTULO

4

EXPERIMENTOS COM BASES DE DADOS SINTÉTICAS

No presente capítulo são apresentados os experimentos realizados com bases de dados sintéticas. Na Seção 4.1 são introduzidas as duas coleções de bases de dados utilizadas. Na Seção 4.2 é apresentada a metodologia adotada para as execuções dos algoritmos e para análise de resultados. Na Seção 4.3 é introduzido o Experimento 1, no qual os algoritmos foram executados em uma coleção de bases de dados baseada naquela proposta em (EREN *et al.*, 2013). Na Seção 4.4 é exposto o Experimento 2, proposto nesta dissertação, o qual engloba um conjunto de bases de dados contendo bi-agrupamentos seguindo padrão *checkerboard*, projetadas especificamente para a comparação de algoritmos que assumem tal estrutura. Considerações referentes aos resultados obtidos são feitas na Seção 4.5.

Outra coleção de base de dados, amplamente referenciada e utilizada na literatura de bi-agrupamento de dados, é aquela criada e disponibilizada por (PRELIĆ *et al.*, 2006). A mesma não foi utilizada neste trabalho, pois possui uma série de limitações. Primeiramente, ela é baseada no modelo artificial proposto por Ihmels *et al.* (2002), o qual foi utilizado para avaliar um algoritmo predecessor ao ISA, o que claramente impõe um viés, uma vez que os padrões nela contidos correspondem particularmente com aqueles buscados apenas por parte dos algoritmos investigados. Além disso, Hochreiter *et al.* (2010) argumentam que a distribuição dos dados nas bases é bimodal, diferente do observado em diversas bases de dados de expressão gênica reais. Por fim, algumas das bases de dados, na ausência de ruídos, são originalmente binárias, o que está aquém das bases de dados reais obtidas a partir de experimentos com *microarrays* utilizadas em boa parte dos estudos da literatura de bi-agrupamento.

4.1 Bases de dados

Coleção 1

Para o primeiro experimento com bases de dados sintéticas, foi gerada uma coleção com 1680 bases tal como aquela proposta em (EREN *et al.*, 2013). Para uma primeira etapa tal coleção apresenta 20 bases de dados, constituídas de 500 linhas e 200 colunas, geradas para cada um dentre seis tipos diferentes de padrões que podem ocorrer em bi-grupos (constantes, positivamente regulados, aditivos, multiplicativos, aditivos-multiplicativos e *plaid*) com a finalidade de identificar aqueles detectados por cada algoritmo. Cada base de dados contém um bi-grupo de dimensão 50×50, cujos valores base de cada linha e de ajuste para cada coluna foram amostrados a partir da distribuição $N(0,1)$ e combinados conforme a respectiva definição de cada padrão apresentado no Capítulo 3. As únicas exceções foram as bases que continham bi-grupos constantes ou positivamente regulados, os quais foram gerados segundo a Equação (3.1) com $\mu = 0$ e $\mu = 5$, respectivamente. As células restantes de cada base de dados, que não pertenciam aos bi-grupos implantados (*background*), tiveram seus valores gerados independentemente e identicamente distribuídos a partir de $N(0,1)$.

Para uma segunda etapa, cada algoritmo foi estudado em três cenários distintos em bases de dados contendo bi-grupos seguindo os tipos de padrões que ele se mostrou capaz de detectar nos experimentos da primeira etapa. Tais cenários consistem em:

- Influência de ruído: as bases de dados criadas inicialmente e mencionadas acima são perturbadas ao somar valores gerados segundo uma distribuição $N(0,\sigma^2)$ em suas células, onde $\sigma \in \{0{,}0; 0{,}25; 0{,}50; 0{,}75; 1{,}0\}$ representa o nível de ruído. Para cada valor de σ existem 20 bases de dados.

- Influência do número de bi-grupos: as bases de dados neste cenário são formadas por 500 linhas, 250 colunas e k bi-grupos de dimensão 50×50 sem sobreposição e sem a presença de ruídos, com $k \in \{1,2,3,4,5\}$. Para cada valor de k foram geradas 20 bases de dados.

- Influência de sobreposição entre bi-grupos: as bases de dados neste cenário contêm 500 linhas, 200 colunas e dois bi-grupos de dimensão 50×50 os quais compartilham elementos em uma região de sobreposição $d \times d$, com $d \in \{0,10,20,30\}$. Para cada valor de d foram geradas 20 bases de dados.

Coleção 2

Para o segundo experimento com bases de dados sintéticas foi proposta uma nova coleção com 200 bases de dados contendo bi-agrupamentos *checkerboard*, onde as linhas e as colunas de uma matriz são particionadas em uma quantidade k_r e k_c de grupos, nessa ordem, dando origem a $k_r \cdot k_c$ bi-grupos sem sobreposição. Esta coleção foi utilizada para a

avaliação dos algoritmos MSSRCC e Spectral, pois ambos obrigam todas as linhas e colunas de uma matriz de dados a serem bi-agrupadas em submatrizes disjuntas. Uma vez que esses algoritmos buscam por bi-agrupamentos *checkerboard*, eles seriam penalizados pelas medidas de validação externa quando avaliados nas bases de dados da Coleção 1, a qual não contém bases de dados com bi-agrupamentos que seguem essa estrutura. Assim, uma comparação com os demais algoritmos seria injusta, uma vez que o MSSRCC e o Spectral não conseguiriam atingir resultados expressivos para as medidas CE e CSI.

Nesta coleção, as bases de dados foram geradas com 500 linhas, 250 colunas, $k_r = 10$ e $k_c = 5$. Os tipos de padrões considerados foram de bi-grupos constantes e bi-grupos aditivos, os quais condizem, respectivamente, com as formulações dos algoritmos Spectral e MSSRCC. Assim como na Coleção 1, foram geradas 20 bases de dados para cada tipo de bi-grupo. Para o modelo constante, foi sorteado um valor para cada bi-grupo segundo a distribuição $N(0,1)$. Para o modelo aditivo, assim como na Coleção 1, os valores base de cada linha e de ajuste para cada coluna em cada bi-grupo foram amostrados de $N(0,1)$ e combinados conforme a definição desse modelo no Capítulo 3.

Para uma segunda etapa, apenas o cenário de ruído é simulado, neste caso seguindo a mesma metodologia e os mesmos níveis de ruído descritos para a Coleção 1. Os cenários restantes não foram investigados para esta coleção pois, conforme comentado acima, os algoritmos MSSRCC e Spectral não são capazes de identificar bi-grupos sobrepostos e, ao aumentar a quantidade de grupos nas linhas ou nas colunas, cada bi-grupo contido nas bases iria conter menos elementos, não refletindo um cenário favorável a esses algoritmos, podendo surgir bi-grupos compostos por quantidades muito pequenas de linhas ou colunas.

4.2 Metodologia

Os parâmetros considerados para cada experimento relatado, neste capítulo, estão de acordo, sempre que possível, com aqueles recomendados originalmente pelos autores de cada algoritmo e com aqueles utilizados no estudo comparativo de Eren *et al.* (2013). Comentários sobre os parâmetros e valores eventualmente adaptados para os experimentos relatados nesta dissertação estão presentes no Apêndice B.

Para os algoritmos que requerem a quantidade desejada k de bi-grupos, o número correto existente nas bases foi informado. Em outros casos, os quais não exigem como parâmetro o número k de bi-grupos, uma comparação direta entre os resultados destes com os resultados daqueles que requerem k seria injusta, uma vez que eles seriam penalizados pelas medidas CE e CSI sempre que apresentassem uma quantidade de bi-grupos maior do que o esperado. Devido a isso, adotou-se o mesmo procedimento iterativo para filtragem de resultados descrito em (PRELIĆ *et al.*, 2006; EREN *et al.*, 2013). A cada iteração desse procedimento, o maior dentre os bi-grupos ainda não selecionados, que possui uma proporção menor do que o de elementos

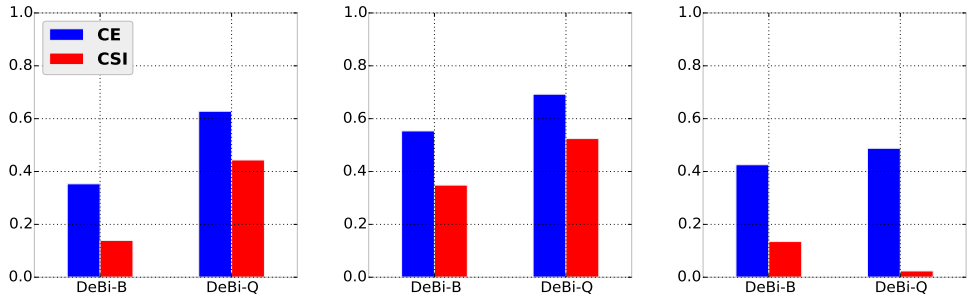
em comum com qualquer outro bi-grupo previamente selecionado, é escolhido. Esse processo iterativo é executado até que a quantidade desejada de bi-grupos seja alcançada ou até que não existam bi-grupos que satisfaçam o limiar máximo de sobreposição. Assim como em (PRELIĆ *et al.*, 2006; EREN *et al.*, 2013), foi assumido $o = 0{,}25$. A única exceção para o valor de tal limiar foi para os experimentos com níveis variáveis de sobreposição quando $d = 30$, onde os bi-grupos implantados nos dados se sobrepõem em 30 linhas e 30 colunas de maneira simultânea, atingindo uma proporção de elementos em comum de 0,36. Portanto, neste caso, foi assumido $o = 0{,}36$.

No que diz respeito aos algoritmos determinísticos, cada um foi executado uma vez sobre cada base de dados. Como os cenários estudados compreendiam quantidades fixas de bases de dados para diferentes níveis de ruído, sobreposição ou quantidade de bi-grupos, os resultados foram reportados como uma média das medidas CE e CSI para cada nível.

Para algoritmos que possuem qualquer tipo de aleatoriedade envolvida, foram realizadas 30 execuções para cada base de dados considerada e, ao final, foi tirada uma média dos valores obtidos para as medidas de validação. Como as bases de dados utilizadas em cada nível de cada cenário foram geradas a partir de um mesmo mecanismo, as médias obtidas para cada algoritmo foram tomadas, e intervalos de confiança para a média foram estimados com 95% de precisão entre as diferentes bases de dados.

Dada a quantidade de parâmetros testados para alguns algoritmos, uma análise exaustiva dos resultados seria impraticável. Para contornar tal situação, em cada experimento reportado foram realizados testes preliminares considerando as bases de dados nos níveis mais extremos de cada cenário estudado (ruído, sobreposição e número de bi-grupos), a fim de possibilitar a escolha de uma combinação de parâmetros mais adequada para as execuções restantes. Assim, para cada algoritmo, foi escolhida aquela combinação que apresentava bons resultados com intervalos de confiança mais informativos nesses testes preliminares. Como exemplo, considere os resultados apresentados na Figura 6 para o algoritmo DeBi. Como esse algoritmo exige uma matriz binária, o mesmo foi testado com os procedimentos de discretização dos algoritmos Bimax e QUBIC (os quais são explicados no Capítulo 3), conforme discutido no Apêndice B. Nessa figura, DeBi-B indica os resultados do DeBi quando utilizado o procedimento de discretização do Bimax e DeBi-Q aponta os resultados do DeBi com a discretização do QUBIC. Considerando os desempenhos apresentados é possível observar que, em sua maioria, os melhores resultados para os níveis extremos dos cenários investigados foram obtidos com a configuração DeBi-Q. Portanto, essa foi a escolhida para esse algoritmo para os experimentos nos níveis restantes de cada cenário.

Figura 6 – Resultados do algoritmo DeBi nos níveis mais extremos de cada cenário estudado com bases de dados sintéticas.



(a) Nível máximo de ruído (b) Quantidade máxima de bi-grupos (c) Nível máximo de sobreposição

4.3 Experimento 1

Seleção dos modelos de bi-grupos

Conforme comentado na Seção 4.1, neste experimento foi utilizada a Coleção 1. Portanto, a primeira etapa consistiu em aplicar 15 dentre os 17 algoritmos[1] estudados em todas as bases de dados geradas para cada um dos modelos de bi-grupos considerados, a fim de encontrar aqueles detectados por cada algoritmo. Esta primeira etapa é importante, pois comparar diferentes técnicas de bi-agrupamento em respeito a apenas um modelo de bi-grupo provavelmente acarretaria em resultados enganosos (EREN *et al.*, 2013). Os resultados estão dispostos na Figura 7.

Nos experimentos subsequentes, os quais investigaram bases de dados com diferentes níveis de ruído, sobreposição e número de submatrizes implantadas, cada algoritmo foi executado apenas para os modelos de bi-grupos para os quais alcançou um valor de, pelo menos, 0,8 para qualquer uma dentre as duas medidas de validação externa nesta etapa inicial. Quando tal critério não foi satisfeito para nenhum modelo de bi-grupo, o algoritmo em questão foi aplicado para aquele modelo que apresentou os melhores resultados nesta etapa inicial.

Dentre todos os métodos de bi-agrupamento, o CPB foi aquele capaz de identificar o maior número de modelos, atingindo valores acima de 0,8 para as duas medidas de validação externa em quatro dentre os seis tipos de bi-grupos: constantes, multiplicativos, aditivos e aditivos-multiplicativos. Este resultado era esperado, uma vez que a heurística desse algoritmo utiliza a correlação de Pearson para a formação de bi-grupos e todas as submatrizes desses quatro tipos de padrões apresentam linhas perfeitamente correlacionadas quando comparadas aos elementos de *background*.

O CCA atingiu sua melhor performance nas bases de dados contendo bi-grupos constan-

[1] Os algoritmos MSSRCC e Spectral não foram incluídos, pois ambos forçam o bi-agrupamento de todas as linhas e colunas de uma matriz de dados em submatrizes disjuntas, sendo então penalizados pelas medidas de validação externa ao reportar bi-grupos compostos apenas por elementos que não pertencem à solução desejada (*background*).

tes. Tal comportamento pode ser explicado, devido ao fato de tal modelo apresentar um resíduo quadrático médio (que é o critério utilizado pelo CCA para a formação dos seus bi-grupos) perfeito, ou seja, igual a zero.

O BBC e o OPSM tiveram um melhor desempenho para bi-grupos aditivos. O BBC foi projetado inicialmente para a detecção do modelo *plaid* (Equação 3.18), o qual pode ser visto como uma generalização do modelo aditivo (MADEIRA; OLIVEIRA, 2004). O OPSM obteve bons resultados neste tipo de modelo pois bi-grupos aditivos preservam uma ordem linear entre suas colunas.

Os algoritmos restantes foram capazes de detectar o modelo positivamente regulado. Dentre eles, dois foram também capazes de detectar outros padrões: Plaid, que identificou bi-grupos aditivos (sendo válidos os mesmos argumentos utilizados para o BBC) e o xMOTIFs, que encontrou bi-grupos constantes (que, após a discretização exigida por esse algoritmo, condizem com o modelo de evoluções coerentes buscado por ele).

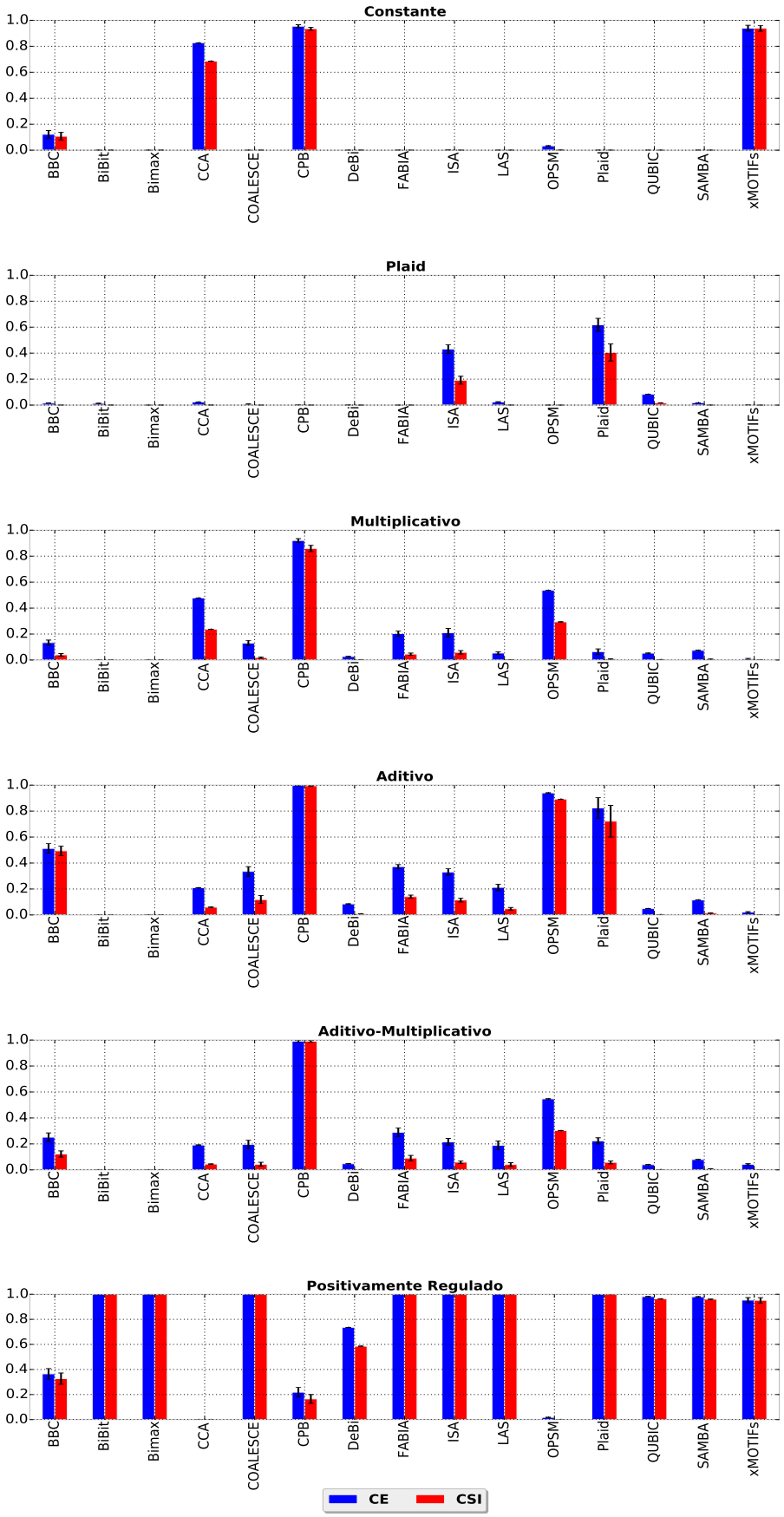
Influência de ruído

Os resultados deste cenário estão dispostos na Figura 8. Alguns algoritmos apresentaram um comportamento robusto, atingindo altos valores para as medidas CSI e CE em todos os níveis de ruído. São os casos de: COALESCE, FABIA, ISA, LAS, Plaid (este em bi-grupos positivamente regulados) e SAMBA. Outros algoritmos, tais como BiBit, Bimax e QUBIC foram também relativamente robustos, mas se mostraram sensíveis aos níveis mais altos de ruído, conforme refletido por uma queda em suas performances. O BiBit e o Bimax não foram capazes de encontrar bi-grupos que respeitassem as restrições de quantidade mínima de linhas e de colunas impostas por eles. O QUBIC foi afetado pelo seu procedimento de discretização, utilizado como pré-processamento. Para cada gene, tal procedimento leva em conta um quantil superior e um quantil inferior das suas condições experimentais para ser capaz de diferenciar valores de expressão extremos dos demais. Entretanto, nas bases de dados com maior influência de ruídos, tal procedimento pode ser muito restritivo, e assim, algumas linhas ou colunas dos bi-grupos existentes podem não ser encontradas por esse algoritmo.

O CCA e o CPB obtiveram bons resultados apenas nos níveis mais baixos de ruído. O primeiro algoritmo possivelmente não recuperou algumas linhas e colunas dos bi-grupos reais devido ao acréscimo nos seus valores de MSR (Equação 3.10) após as bases de dados serem perturbadas. O segundo algoritmo utiliza a correlação de Pearson para definir seus bi-grupos. Conforme a quantidade de ruído cresce em uma base de dados, as tendências dos níveis de expressão dos genes correspondem cada vez menos, tornando as correlações difíceis de serem distinguidas.

Com relação ao BBC, em todos os níveis de ruído, percebeu-se que em algumas execuções um bi-grupo equivalente ou muito próximo daquele implantado era retornado. Entretanto,

Figura 7 – Resultados dos algoritmos para diferentes modelos de bi-grupos.



nas restantes, essa técnica apresentou bi-grupos muito maiores ou discrepantes ao original. Com o algoritmo DeBi ocorreu um problema similar; vários dos bi-grupos encontrados por ele possuíam mais colunas do que os existentes.

Por fim, o OPSM e o xMOTIFs identificaram bi-grupos apenas na ausência de ruídos. O OPSM é muito sensível a ruídos, pois qualquer perturbação nos dados pode facilmente alterar a ordem linear das colunas do bi-grupo implantado, violando as suposições feitas por esse algoritmo. Por sua vez, o xMOTIFs exige que os dados sejam discretizados. Após o passo de discretização foi observado que, devido ao ruído, os bi-grupos implantados geralmente contêm mais de um valor discreto em suas linhas, desobedecendo assim o modelo de evoluções coerentes assumido por esse algoritmo.

Influência do número de bi-grupos

Os resultados deste cenário estão dispostos na Figura 9. A maioria dos algoritmos foram pouco influenciados pelo crescente número de bi-grupos nas bases de dados. São os casos de: BiBit, Bimax, COALESCE, CPB, FABIA, ISA, LAS, SAMBA e xMOTIFs, os quais atingiram, em geral, valores acima de 0,8 para CE e CSI. Outros algoritmos, como CCA, OPSM e Plaid apresentaram soluções gradativamente piores conforme k aumentou, o que mostra que mesmo quando o número de bi-grupos em uma base de dados é conhecido, recuperá-los precisamente pode ser uma tarefa desafiadora para alguns algoritmos.

Os algoritmos BBC e DeBi apresentaram comportamentos similares a aqueles observados no experimento anterior. Em geral, o primeiro reportou vários bi-grupos consideravelmente maiores do que os originais, enquanto que o segundo incluiu colunas extras em suas submatrizes, as quais não faziam parte da solução desejada.

O QUBIC conseguiu seus melhores desempenhos nas bases de dados com apenas um bi-grupo. Conforme o número de bi-grupos foi aumentado, alguns dos bi-grupos implantados nos dados foram recuperados mais de uma vez. Como as medidas CE e CSI punem soluções de bi-agrupamento em tais situações (HORTA; CAMPELLO, 2014), não se podia esperar que esse algoritmo alcançasse avaliações próximas do valor máximo (unitário).

Influência de sobreposição

Os resultados deste cenário estão dispostos na Figura 10. Em geral, os algoritmos investigados foram altamente influenciados pelo nível crescente de sobreposição entre os bi-grupos implantados nas bases de dados. As técnicas BiBit, Bimax e CPB (este último, em bi-grupos constantes, aditivos e aditivos-multiplicativos) foram as únicas exceções as quais apresentaram, na maioria dos casos, resultados acima de 0,8 para as duas medidas de validação entre os diferentes níveis de sobreposição. O SAMBA também foi capaz de obter bons resultados, exceto para o nível mais alto de sobreposição.

Figura 8 – Resultados dos algoritmos para bases de dados ruidosas.

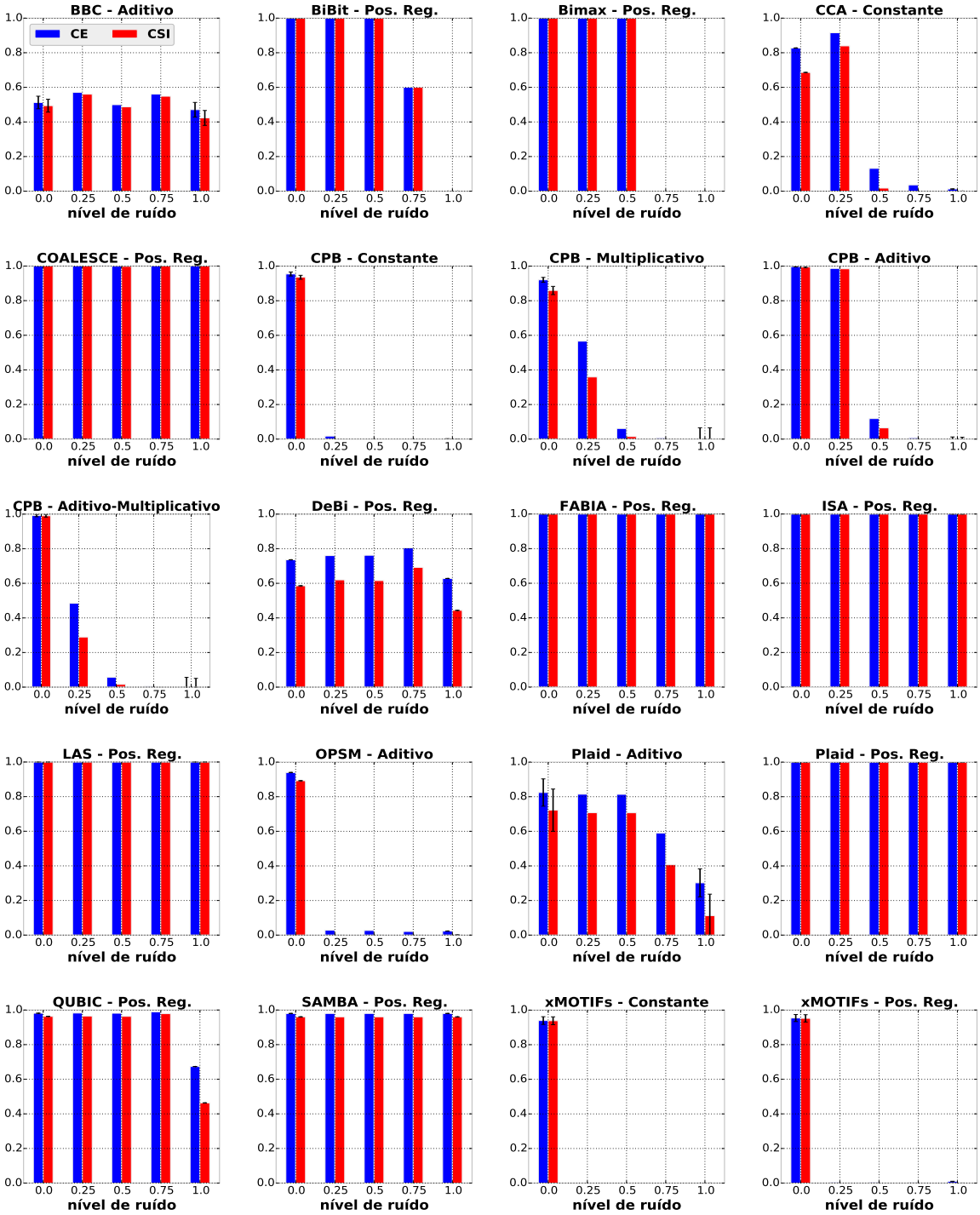
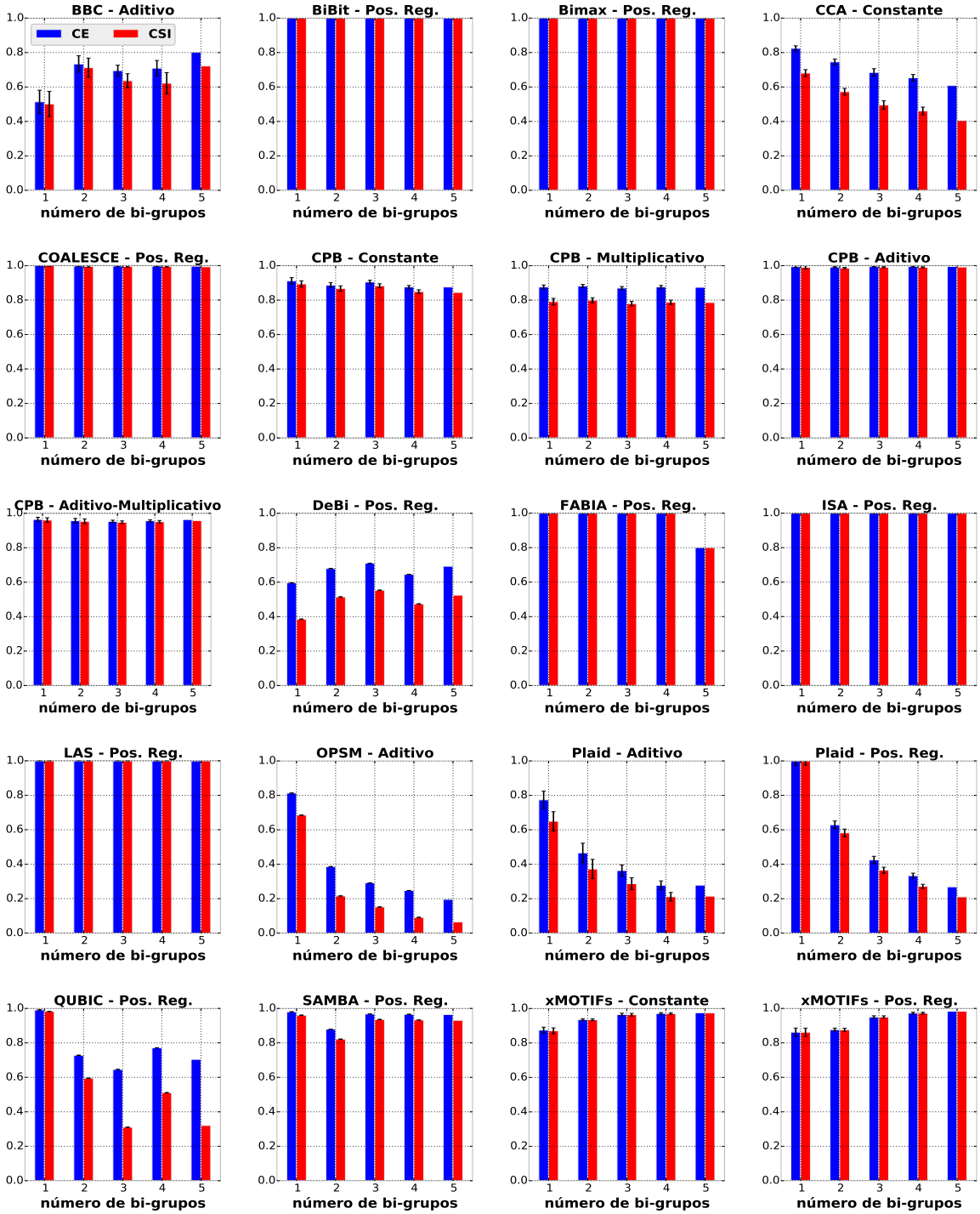


Figura 9 – Resultados dos algoritmos para bases de dados com diferentes números de bi-grupos.



Neste cenário, não era esperado que o BBC, o xMOTIFs e o CCA obtivessem bons desempenhos. A razão é que o BBC e o xMOTIFs não permitem a existência de bi-grupos sobrepostos em suas formulações. O CCA, por sua vez, mascara cada bi-grupo encontrado com valores aleatórios, com a finalidade de garantir que sua heurística determinística possa encontrar diferentes submatrizes cada vez que ele é executado. Embora tais valores possuam uma pequena probabilidade de formar qualquer padrão detectável, eles podem interferir diretamente no processo de bi-agrupamento, especialmente quando existe sobreposição entre bi-grupos (YANG *et al.*, 2002; YANG *et al.*, 2003; YANG *et al.*, 2005).

O Plaid tende a reportar bi-grupos que se sobrepõem em uma grande extensão (GU; LIU, 2008). Isso é refletido pela melhora das suas soluções com o aumento da sobreposição em bi-grupos positivamente regulados. O QUBIC também reportou suas melhores performances nos níveis mais altos. Ao analisar os bi-agrupamentos reportados por este algoritmo, foi observado que em algumas bases de dados dos níveis mais baixos de sobreposição esse algoritmo reportou o mesmo bi-grupo real duas vezes, sendo então penalizado pelas medidas CE e CSI. Com relação ao algoritmo DeBi, assim como nos cenários anteriores, a maioria dos bi-grupos reportados possuíam mais colunas do que aqueles inseridos nos dados.

O COALESCE foi capaz de identificar os dois bi-grupos implantados apenas na ausência de sobreposição. Conforme a sobreposição aumentou, a maior das submatrizes reportadas por esse algoritmo geralmente consistia em uma união das duas submatrizes (bi-grupos) implantadas nos dados, o que acarretou na inclusão de elementos da base que não pertenciam ao bi-agrupamento real.

4.4 Experimento 2

Seleção dos modelos de bi-grupos

Conforme comentado na Seção 4.1, neste experimento foi utilizada a Coleção 2 para a avaliação dos algoritmos MSSRCC e Spectral, os quais não foram incluídos na análise anterior. Seguindo a metodologia do Experimento 1, a primeira fase deste segundo experimento consistiu em aplicar os dois algoritmos nas bases de dados geradas para os bi-grupos constantes e bi-grupos aditivos, com o intuito de selecionar o melhor para cada técnica para as execuções posteriores. Os resultados estão dispostos na Figura 11.

Pelos resultados apresentados pode-se perceber que, conforme esperado, nesta etapa cada algoritmo atingiu seu melhor desempenho no tipo de padrão para o qual foi projetado. Ou seja, o modelo selecionado para o Spectral foi o de bi-grupos inteiramente constantes enquanto que o escolhido para o MSSRCC foi o de bi-grupos aditivos.

Figura 10 – Resultados dos algoritmos para bases de dados com sobreposição entre bi-grupos.

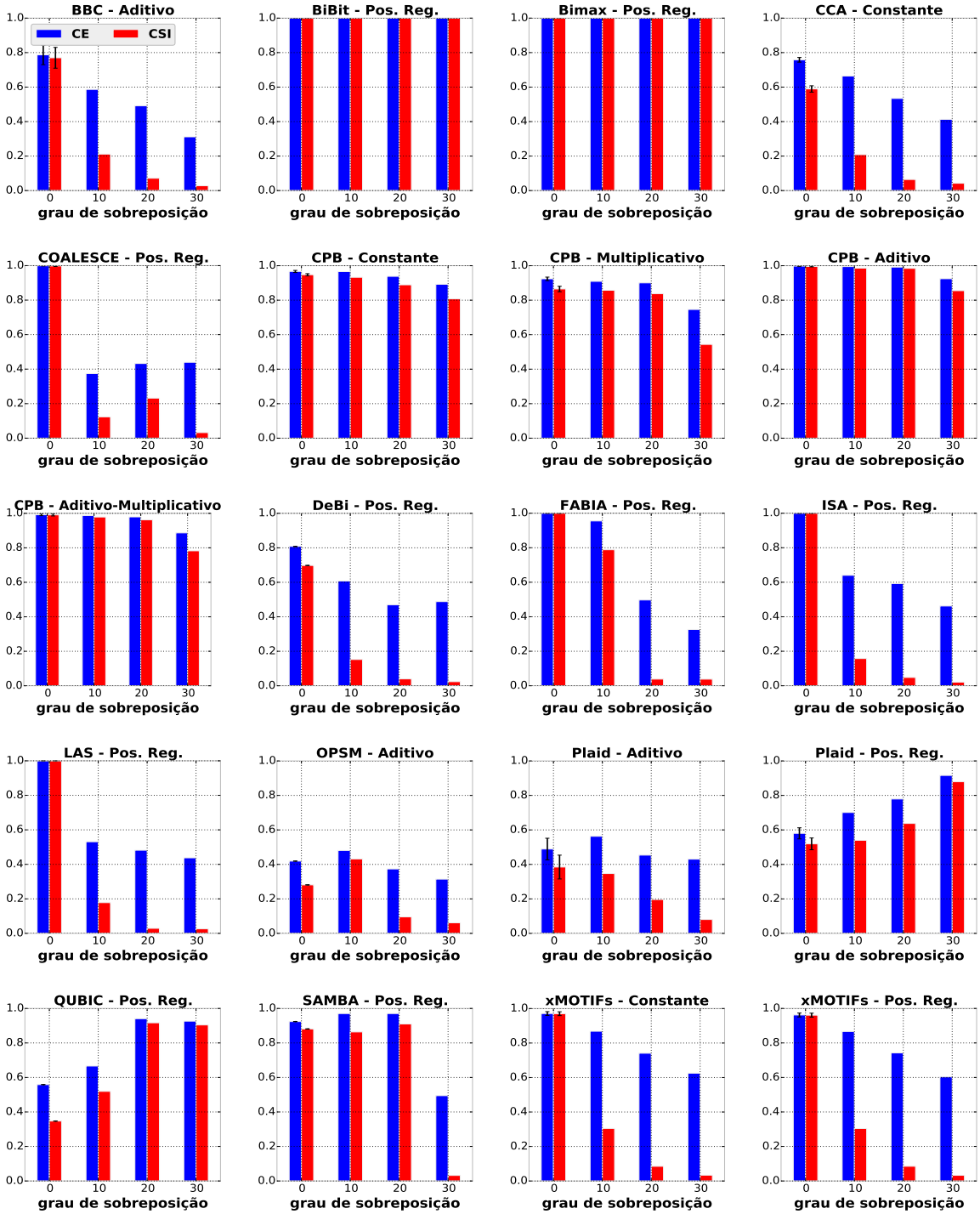
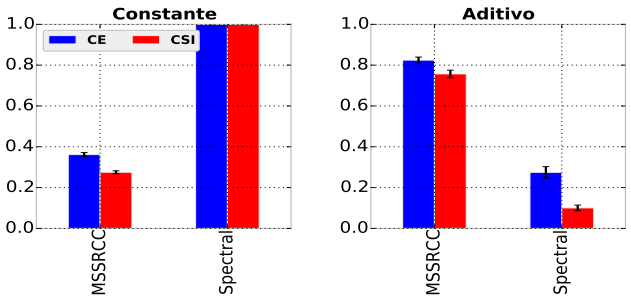


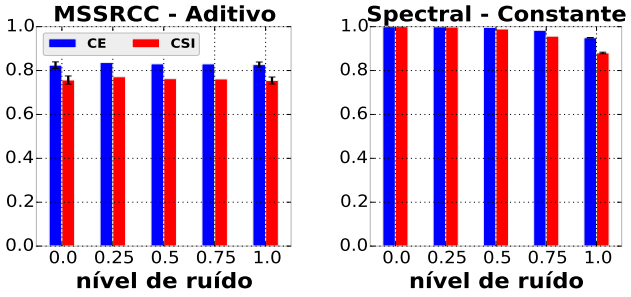
Figura 11 – Resultados dos algoritmos MSSRCC e Spectral para diferentes modelos de bi-grupos.



Influência de ruído

Os resultados deste cenário estão dispostos na Figura 12. O MSSRCC pouco sofreu com o crescente ruído. Esse algoritmo obteve resultados muito similares nos diferentes níveis. Contudo, em várias das execuções, linhas ou colunas foram erroneamente inseridas em alguns grupos, o que não permitiu que as medidas externas atingissem seus valores máximos. O Spectral, apesar de apresentar uma leve sensibilidade em níveis mais altos de ruído, em geral alcançou bons resultados nas bases de dados, uma vez que a maioria de seus bi-agrupamentos foram avaliados acima de 0,8 para CE e CSI em todos os níveis de ruído. Essa performance pode ser explicada pelo passo de pré-processamento realizado pelo algoritmo, o qual busca normalizar a matriz de entrada, de modo a evidenciar os padrões nela existentes e amenizar possíveis interferências causadas por variações que podem ocorrer durante o procedimento de geração dos dados.

Figura 12 – Resultados dos algoritmos MSSRCC e Spectral para bases de dados ruidosas.



4.5 Considerações

No presente capítulo uma quantidade maior de experimentos com bases de dados sintéticas foi apresentada do que nos estudos comparativos anteriores, utilizando medidas de validação externa comprovadamente mais eficazes (HORTA; CAMPELLO, 2014). Na análise experimental realizada, algumas das principais limitações de cada algoritmo foram bastante aparentes.

Em (PRELIĆ *et al.*, 2006) os algoritmos CCA e xMOTIFs mostraram-se bastante sensíveis à presença de ruídos nos dados, assim como ocorreu nos experimentos reportados

neste capítulo. Entretanto, é importante lembrar que naquele estudo as bases de dados sintéticas continham apenas bi-grupos positivamente regulados, os quais claramente não condizem com o tipo de padrão identificado pelo CCA, conforme comprovado empiricamente neste trabalho. Além disso, Prelić *et al.* (2006) concluem que o Bimax apresentou resultados tão bons quanto os melhores reportados pelos outros algoritmos por eles comparados. Em contraste, no presente estudo, demonstrou-se que tal algoritmo pode não ser eficiente em cenários de alto ruído.

Em (BOZDAĞ; KUMAR; ÇATALYÜREK, 2010) os autores concluem que, dentre os algoritmos investigados, o CPB foi aquele que apresentou a melhor performance. Entretanto, ressalta-se que as bases de dados sintéticas geradas naquele trabalho continham apenas bi-grupos seguindo padrões aditivos ou multiplicativos, o que claramente impõe um viés na avaliação, uma vez que tais modelos condizem com a formulação do CPB, conforme confirmado pelos experimentos deste capítulo. Ademais, os resultados aqui reportados demonstram que os algoritmos BBC e Plaid, os quais não foram comparados em (BOZDAĞ; KUMAR; ÇATALYÜREK, 2010), possuem uma melhor performance em identificar corretamente bi-grupos aditivos do que o CPB em casos mais extremos de ruído.

Em (EREN *et al.*, 2013), os autores concluem que a maioria das técnicas investigadas por eles tiveram suas performances degradadas conforme a quantidade de bi-grupos aumentou nas bases de dados. Em contraste, no presente trabalho, nove algoritmos (BiBit, Bimax, COALESCE, CPB, FABIA, ISA, LAS, SAMBA e xMOTIFs) foram pouco ou nada influenciados pelo número de bi-grupos implantados, dentre os quais, apenas três (BiBit, LAS e SAMBA) não foram investigados naquele estudo. Além disso, Eren *et al.* (2013) mencionam que nenhum método de bi-agrupamento conseguiu separar bi-grupos altamente sobrepostos. Nos experimentos aqui descritos, três algoritmos (BiBit, Bimax e CPB) conseguiram bons resultados em todos os níveis de sobreposição considerados, sendo que dois deles (Bimax e CPB) foram utilizados naquele trabalho comparativo. Esses resultados mostram a importância da utilização de medidas mais apropriadas para validação externa ao testar algoritmos de bi-agrupamento em bases de dados cujos bi-grupos são conhecidos de antemão. Conforme mostrado em (HORTA; CAMPELLO, 2014), as medidas de validação externa empregadas em estudos comparativos anteriores satisfazem no máximo duas dentre as oito propriedades desejáveis, enquanto que cada uma das medidas consideradas neste trabalho satisfazem sete das oito propriedades e, quando utilizadas conjuntamente, obedecem todas as oito. Ademais, para o cenário de sobreposição foi observada a necessidade de relaxar o limiar máximo de sobreposição quando $d = 30$ para algoritmos que necessitam do processo de filtragem de bi-grupos descrito na Seção 4.2. Caso contrário, BiBit e CPB não seriam capazes de atingir resultados expressivos (pois um dos bi-grupos reais seria filtrado das suas soluções, uma vez que tais bi-grupos compartilham uma proporção de 0,36 dos seus elementos um com o outro, valor consideravelmente maior do que o limiar $o = 0{,}25$ utilizado em estudos anteriores).

Ademais, este trabalho introduziu uma nova coleção de bases de dados sintéticas, as

quais são compostas por bi-agrupamentos *checkerboard*. Dentre os estudos comparativos anteriores, Bozdağ, Kumar e Çatalyürek (2010) e Eren *et al.* (2013) levaram em conta os algoritmos MSSRCC e Spectral, respectivamente. Contudo, esses algoritmos foram aplicados em coleções de bases de dados sintéticas cujos bi-agrupamentos de referência não seguiam a estrutura *checkerboard*, não refletindo assim um cenário favorável a eles. Nos experimentos aqui reportados, MSSRCC e Spectral foram testados na nova coleção proposta, a qual permite uma análise mais justa e realista dos seus resultados.

Por fim, os resultados obtidos estão sumarizados nas Tabelas 1 e 2, as quais indicam, para cada algoritmo, quais foram os tipos de bi-grupos identificados e em quais cenários (ruído, número de bi-grupos e sobreposição) o mesmo obteve bons resultados para todos os níveis considerados.

Tabela 1 – Sumário da seleção dos modelos de bi-grupos para cada algoritmo.

Algoritmo	Constante	Multiplicativo	Aditivo	Aditivo-multiplicativo	*Plaid*	Positivamente regulado
BBC			✓			
BiBit						✓
Bimax						✓
CCA	✓					
COALESCE						✓
CPB	✓	✓	✓	✓		
DeBi						✓
FABIA						✓
ISA						✓
LAS						✓
MSSRCC		–	✓	–	–	–
OPSM			✓			
Plaid			✓			✓
QUBIC						✓
SAMBA						✓
Spectral	✓	–		–	–	–
xMOTIFs	✓					✓

✓: indica que o algoritmo da linha é capaz de identificar o tipo de bi-grupo da coluna.
– : indica que o algoritmo da linha não foi testado para o tipo de bi-grupo da coluna.

Tabela 2 – Sumário dos resultados para os diferentes cenários investigados com bases de dados sintéticas.

Algoritmo	Ruído	Número de bi-grupos	Sobreposição
BBC			
BiBit		✓	✓
Bimax		✓	✓
CCA			
COALESCE	✓	✓	
CPB		✓	✓
DeBi			
FABIA	✓	✓	
ISA	✓	✓	
LAS	✓	✓	
MSSRCC	✓	–	–
OPSM			
Plaid	✓		
QUBIC	✓		
SAMBA	✓	✓	
Spectral	✓	–	–
xMOTIFs		✓	

✓: indica que o algoritmo da linha atingiu bons resultados em todos os níveis do cenário da coluna para pelo menos um tipo de bi-grupo.
– : indica que o algoritmo da linha não foi testado para o cenário da coluna.

CAPÍTULO

5

EXPERIMENTOS COM BASES DE DADOS REAIS

Embora os experimentos com bases de dados sintéticas possam fornecer boas intuições sobre os pontos fortes e fracos de cada algoritmo de bi-agrupamento, eles são capazes de simular apenas alguns aspectos de cenários biológicos reais. Portanto, é importante também testar os algoritmos estudados em tais cenários. No presente capítulo são apresentados os experimentos realizados em coleções compostas por dados biológicos reais. Na Seção 5.1 é apresentado o experimento realizado para avaliação da qualidade dos grupos de genes encontrados pelos algoritmos. Na Seção 5.2 é exposto o experimento executado para aferir a capacidade das técnicas de bi-agrupamento na tarefa de agrupamento de condições experimentais, que consistem em amostras de tecidos cancerígenos. Na Seção 5.3, são feitas as considerações finais.

5.1 Experimento em agrupamento de genes

Bases de dados

Para o primeiro experimento em cenários reais foi utilizada uma coleção composta por 27 bases de dados, sumarizadas na Tabela 3, dentre as quais:

- Duas são as utilizadas em (PRELIĆ *et al.*, 2006), concernentes aos organismos *Saccharomyces cerevisiae* e *Arabidopsis thaliana*, disponíveis em http://www.tik.ethz.ch/sop/bimax/.

- Oito são aquelas exploradas em (EREN *et al.*, 2013), obtidas a partir do repositório *Gene Expression Omnibus*[1] (EDGAR; DOMRACHEV; LASH, 2002), e relacionadas aos organismos *Homo sapiens*, *Rattus norvegicus*, *Caenorhabditis elegans* e *Mus musculus*.

[1] http://www.ncbi.nlm.nih.gov/geo/

Boa parte delas possuem valores ausentes e, assim como naquele estudo, os mesmos foram estimados através do pacote de *software* pcaMethods (STACKLIES *et al.*, 2007).

- Por fim, 17 bases que foram compiladas e pré-processadas por Jaskowiak, Campello e Costa (2013), referentes ao organismo *Saccharomyces cerevisiae*. No pré-processamento, os autores removeram genes com mais de 10% de seus valores ausentes e, através do procedimento de *fold-change* (FACELI; CARVALHO; SILVA JR, 2004), selecionaram quantidades próximas de 1000 genes para cada base de dados, por meio da remoção daqueles pouco informativos. Tais bases de dados estão disponíveis em http://lapad-web.icmc.usp.br/repositories/ieee-tcbb-2013/index.html.

Tabela 3 – Descrição das bases de dados utilizadas para agrupamento de genes.

Nome	Genes	Condições Experimentais	Referência
arabidopsis	734	69	Prelić *et al.* (2006)
saccharomyces	2993	173	
GDS181	12424	84	Eren *et al.* (2013)
GDS589	8752	122	
GDS1027	15872	154	
GDS1319	22584	123	
GDS1406	12432	87	
GDS1490	12439	150	
GDS3715	12581	110	
GDS3716	22225	42	
alpha factor	1099	18	Jaskowiak, Campello e Costa (2013)
cdc 15	1086	24	
cdc 28	1044	17	
elutriation	935	14	
1mM menadione	1050	9	
1M sorbitol	1030	7	
1.5mM diamide	1038	8	
2.5mM DTT	991	8	
constant 32nM H2O2	976	10	
diauxic shift	1016	7	
complete DTT	962	7	
heat shock 1	988	8	
heat shock 2	999	7	
nitrogen depletion	1011	10	
YPD 1	1011	12	
YPD 2	1022	10	
yeast sporulation	1171	7	

Metodologia

Os parâmetros considerados nos experimentos em agrupamento de genes, sempre que possível, estão de acordo com os sugeridos pelos autores originais de cada algoritmo de bi-agrupamento e nos estudos de Prelić *et al.* (2006) e Eren *et al.* (2013). Comentários adicionais

sobre os parâmetros e valores adaptados para os experimentos relatados nesta dissertação estão presentes no Apêndice B.

Em bases de dados de expressão gênica reais, Eren *et al.* (2013) utilizaram 30 e 500 como entrada para algoritmos que exigem como parâmetro o número de bi-grupos ou número de sementes, respectivamente. Entretanto, acredita-se que tal abordagem resulta em uma comparação injusta entre essas duas categorias de algoritmos, uma vez que aqueles que recebem como parâmetro uma quantidade de sementes tiveram maiores chances de encontrar soluções compostas por mais bi-grupos. Portanto, nos experimentos aqui relatados, ambos os valores foram definidos como 500, com algumas exceções (as quais serão explicadas abaixo). Após as execuções dos algoritmos, cada um dos bi-agrupamentos encontrados foi filtrado pelo mesmo procedimento adotado em (PRELIĆ *et al.*, 2006; AYADI; ELLOUMI; HAO, 2012), removendo bi-grupos altamente sobrepostos e selecionando apenas os 100 maiores restantes.

Como exceções ao procedimento acima, os algoritmos xMOTIFs, MSSRCC e Spectral não permitem que exista sobreposição entre submatrizes em suas formulações, então eles não necessitam de qualquer filtragem. xMOTIFs foi então executado para encontrar 100 bi-grupos, que é o mesmo número esperado após a filtragem por sobreposição aplicada nas soluções dos outros algoritmos (mencionado acima). MSSRCC e Spectral foram executados para buscar 100 grupos nas linhas e dois grupos nas colunas, seguindo os experimentos em (CHO *et al.*, 2004; CHO; DHILLON, 2008). Foram buscados apenas dois grupos nas colunas devido ao número de colunas das bases de dados ser muito menor do que o de linhas na maioria dos casos. O BBC também não permite a existência de sobreposição. Entretanto, ele não conseguiu encontrar bi-agrupamentos em boa parte das bases de dados quando o número desejado era de 100 bi-grupos. Portanto, foi empregada uma metodologia similar àquela utilizada em (GU; LIU, 2008; EREN *et al.*, 2013). Assim, esse algoritmo foi executado para buscar quantidades de bi-grupos no intervalo $[30, 60]$ com passos de 5. Finalmente, o pacote de *software* do FABIA restringe o número máximo de bi-grupos buscados como o valor mínimo entre a quantidade de linhas e a quantidade de colunas em uma base de dados, caso contrário a complexidade assintótica dele é severamente aumentada.

Os bi-grupos de cada solução de bi-agrupamento resultante foram submetidos a um processo de enriquecimento de genes através da *Gene Ontology*. Para isso foi utilizado o pacote de *software* clusterProfiler (YU *et al.*, 2012), o qual realiza o teste hipergeométrico apresentado na Equação (3.23). As categorias de genes foram selecionadas a partir da *Biological Process Ontology* e um bi-grupo foi considerado enriquecido se o seu respectivo *p-value* ajustado por meio da correção de Benjamini e Hochberg (HOCHBERG; BENJAMINI, 1990) estava abaixo de 0,05 (EREN *et al.*, 2013). Para cada bi-grupo, foi considerada a categoria com o menor *p-value* ajustado (LI *et al.*, 2009; BOZDAĞ; KUMAR; ÇATALYÜREK, 2010).

Para a análise dos enriquecimentos encontrados, foram consideradas duas abordagens. A primeira delas é a mesma utilizada em (PRELIĆ *et al.*, 2006; LIU; WANG, 2007; SERIN;

VINGRON, 2011; LI *et al.*, 2009; AYADI; ELLOUMI; HAO, 2012; EREN *et al.*, 2013), a qual baseia-se na análise da porcentagem de bi-grupos enriquecidos em relação ao total de bi-grupos encontrados por cada algoritmo em cada base de dados. Entretanto, uma possível limitação desta abordagem decorre de a mesma não passar uma noção clara de quais algoritmos produziram melhores bi-grupos (ou seja, com menores *p-values*), uma vez que as porcentagens analisadas foram produzidas a partir de bi-agrupamentos compostos por diferentes quantidades de submatrizes e os resultados são analisados a partir de uma quantidade pequena de níveis de significância (geralmente quatro ou cinco). Portanto, empregou-se também uma segunda abordagem para avaliação dos resultados, a qual consiste naquela utilizada em (JASKOWIAK; CAMPELLO; COSTA, 2014). Para dois agrupamentos de genes, denotados por r_1 e r_2, tal abordagem baseia-se em contar a quantidade de vezes que r_1 apresentou grupos de genes com melhores *p-values* que r_2 e vice-versa, combinando tais quantias por meio da Equação (5.1), onde a função # retorna o número de vezes que a condição passada como argumento é verdadeira.

$$\text{Comparação}(r_1, r_2) = \log\left(\frac{\#(r_1 < r_2)}{\#(r_2 < r_1)}\right) \tag{5.1}$$

Deve-se notar que a ordem de comparação das listas r_1 e r_2 altera apenas o sinal do resultado mas não sua magnitude. Assim, conforme apresentado acima, $\text{Comparação}(r_1, r_2)$ retornará valores positivos se r_1 for melhor que r_2 e negativos caso contrário.

Algoritmos determinísticos foram executados apenas uma vez para cada configuração experimental em cada base de dados. Por sua vez, algoritmos não determinísticos foram executados um total de 30 vezes. Posteriormente, nas duas análises de enriquecimentos utilizadas, foi considerado para cada algoritmo em cada base de dados a solução de bi-agrupamento com a melhor proporção de bi-grupos enriquecidos.

Resultados

Na Tabela 4 são apresentados os resultados acumulados[2] de cada algoritmo para as 27 bases de dados consideradas, contendo a quantidade de bi-grupos após a filtragem daqueles altamente sobrepostos na segunda coluna e o número de bi-grupos que apresentaram enriquecimento em um nível de significância de 0,05 na terceira coluna. Complementarmente, o gráfico de barras apresentado na Figura 13 reflete, para cada algoritmo investigado, a porcentagem de bi-grupos enriquecidos para todas as bases de dados em cinco diferentes níveis de significância.

Todos os algoritmos foram capazes de encontrar bi-grupos enriquecidos. Os algoritmos BBC, Bimax, COALESCE, FABIA, ISA e Plaid apresentaram os melhores resultados nesta análise, uma vez que mais de 80% dos bi-grupos reportados por eles apresentaram enriquecimento em um nível de significância de 0,05. Todavia, é importante informar que apenas os algoritmos BBC, CCA, CPB, FABIA e SAMBA conseguiram encontrar bi-grupos enriquecidos em todas as

[2] Isto é, cada valor apresentado na segunda e na terceira coluna da Tabela 4 refere-se à quantidade de bi-grupos somada a partir de todas as soluções de bi-agrupamento de cada algoritmo para as 27 bases de dados.

bases de dados. O Plaid, embora tenha apresentado 100% de enriquecimento para seus bi-grupos filtrados em um nível de significância de 0,05, não encontrou bi-grupo algum em todas as suas execuções para duas dentre as 27 bases de dados: *arabidopsis* e *elutriation*.

O CPB reportou várias submatrizes com um número pequeno de linhas altamente correlacionadas em pequenos subconjuntos de colunas. A maioria desses padrões não são relevantes e normalmente ocorrem ao acaso em uma base de dados (EREN, 2012). O OPSM encontrou alguns bi-grupos altamente significativos (com *p-values* abaixo de 10^{-50}). No entanto, muitos deles eram formados por quantidades pequenas de colunas, geralmente entre duas e quatro. Por fim, o CCA e o xMOTIFs obtiveram alguns dos piores resultados nesta análise. Isso deve-se ao fato de ambos terem encontrado vários bi-grupos constituídos por grupos pequenos de genes, os quais não foram enriquecidos com nenhum termo da *Gene Ontology*.

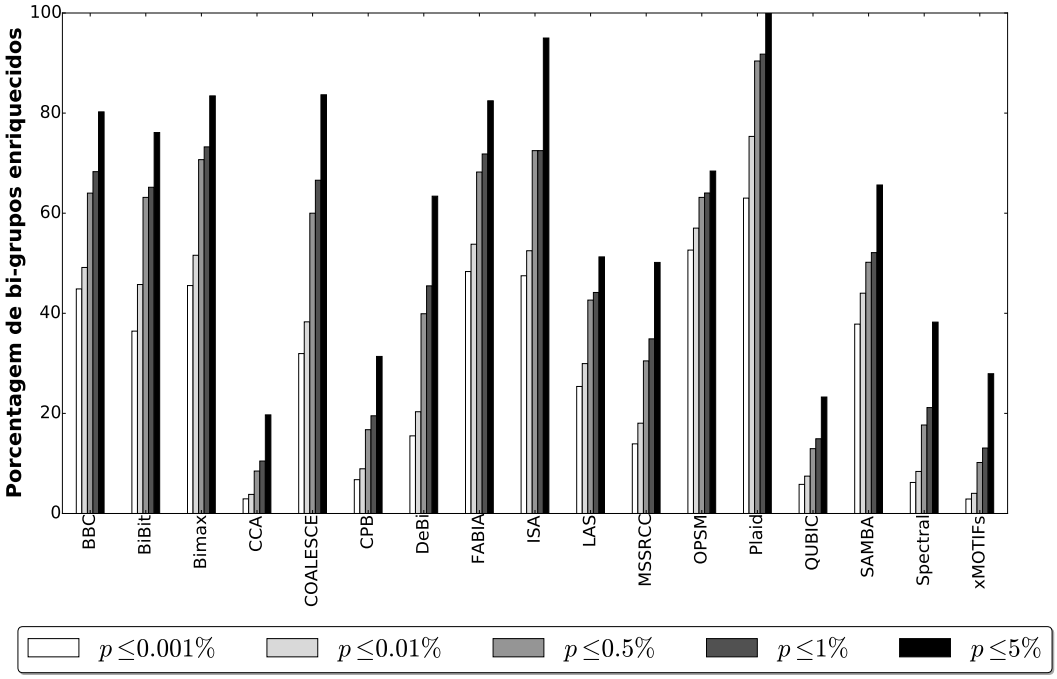
Tabela 4 – Resultados acumulados para os algoritmos nas 27 bases de dados.

Algoritmo	Bi-grupos remanescentes do procedimento de filtragem	Bi-grupos enriquecidos
BBC	653	524
BiBit	247	188
Bimax	314	262
CCA	2700	532
COALESCE	410	343
CPB	1332	418
DeBi	664	421
FABIA	1026	846
ISA	40	38
LAS	197	101
MSSRCC	3680	1846
OPSM	114	78
Plaid	146	146
QUBIC	911	212
SAMBA	259	170
Spectral	5000	1912
xMOTIFs	795	222

A fim de complementar a análise discutida acima, os grupos de genes enriquecidos foram comparados com a medida introduzida na Equação (5.1). Os resultados de tal comparação estão dispostos na Figura 14 na forma de *heatmap*, onde cada célula indica a quantidade de bases de dados para as quais o algoritmo da linha apresentou melhores bi-grupos do que o algoritmo da coluna. Quanto mais avermelhada/azulada a cor de uma célula, melhor/pior o algoritmo da linha foi quando comparado ao da coluna.

Conforme pode ser observado, os algoritmos BBC, Bimax, COALESCE, FABIA, ISA e Plaid, que obtiveram alguns dos melhores resultados na análise anterior, mais uma vez apresentaram melhores enriquecimentos em relação à maioria dos outros algoritmos, exceto quando comparados às técnicas OPSM, SAMBA e LAS, as quais são agora as melhores de acordo com a medida na Equação (5.1). Tais resultados mostram a importância de se comparar soluções de bi-agrupamento através de mais de uma perspectiva, uma vez que muitos daqueles algoritmos

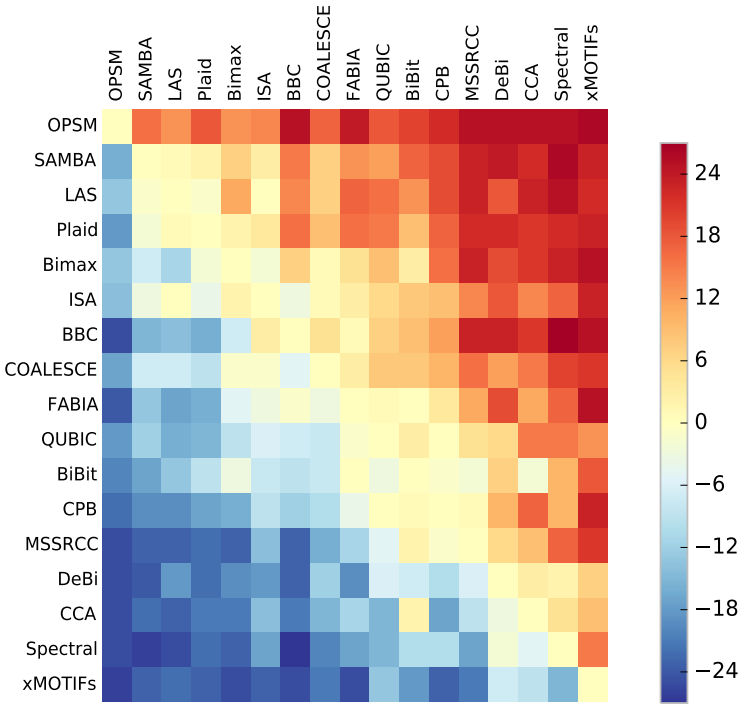
Figura 13 – Porcentagem de bi-grupos enriquecidos para cada algoritmo em cinco diferentes níveis de significância.



que apresentaram as melhores porcentagens de enriquecimento (o qual foi o principal critério utilizado em estudos comparativos anteriores) nem sempre apresentaram os melhores *p-values*.

Por fim, algoritmos como o CCA, CPB e xMOTIFs, apresentaram novamente alguns dos piores resultados.

Figura 14 – Comparação entre *p-values* dos bi-grupos encontrados entre todos os pares de algoritmos.



5.2 Experimento em agrupamento de amostras

Bases de dados

Para o segundo experimento em cenários reais, foi utilizada uma coleção de 35 bases de dados, compiladas e pré-processadas por Souto *et al.* (2008), cujas condições experimentais consistem em amostras rotuladas de tecidos cancerígenos como, por exemplo: próstata, cérebro, pele, pulmão e medula óssea. Durante o pré-processamento, os autores removeram genes contendo mais de 10% de valores ausentes e aplicaram o procedimento de *fold-change* em cada base de dados a fim de selecionar cerca de 10% dos seus genes. Por fim, genes que ainda possuíam valores ausentes tiveram os mesmos substituídos por suas respectivas médias. Esta coleção está sumarizada na Tabela 5 e pode ser obtida em http://algorithmics.molgen.mpg.de/Supplements/CompCancer/.

Tabela 5 – Descrição das bases de dados utilizadas para agrupamento de amostras de câncer.

Nome	Genes	Amostras	Classes
alizadeh-v1	1095	42	2
alizadeh-v2	2093	62	3
alizadeh-v3	2093	62	4
armstrong-v1	1081	72	2
armstrong-v2	2194	72	3
bhattacharjee	1543	203	5
bittner	2201	38	2
bredel	1739	50	3
chen	85	179	2
chowdary	182	104	2
dyrskjot	1203	40	3
garber	4553	66	4
golub-v1	1868	72	2
golub-v2	1868	72	3
gordon	1626	181	2
khan	1069	83	4
laiho	2202	37	2
lapointe-v1	1625	69	3
lapointe-v2	2496	110	4
liang	1411	37	3
nutt-v1	1377	50	4
nutt-v2	1070	28	2
nutt-v3	1152	22	2
pomeroy-v1	857	34	2
pomeroy-v2	1379	42	5
ramaswamy	1363	190	14
risinger	1771	42	4
shipp	798	77	2
singh	339	102	2
su	1571	174	10
tomlins-v1	2315	104	5
tomlins-v2	1288	92	4
west	1198	49	2
yeoh-v1	2526	248	2
yeoh-v2	2526	248	6

Metodologia

Para algoritmos que recebem qualquer informação referente à quantidade de bi-grupos buscados, foi informada a quantidade real de classes existentes em cada base de dados. Para aqueles que buscam por bi-agrupamentos *checkerboard*, informou-se a mesma quantidade de grupos de genes e de amostras, sendo esta quantidade igual ao número real de classes de cada base de dados.

Para cada algoritmo em cada base de dados foi reportado o melhor resultado avaliado pelas medidas FARI e 13AGRI. Conforme já discutido no Capítulo 3, tais medidas foram empregadas neste experimento pois as bases de dados utilizadas possuem rótulos para suas condições experimentais, os quais podem ser vistos como soluções de referência para agrupamento de dados. Nesta avaliação, considerando os algoritmos que são determinísticos, apenas uma execução foi realizada. Em relação aos não determinísticos, cada um foi executado 30 vezes em cada base de dados e a média das avaliações para cada medida foi reportada.

É importante notar que boa parte das técnicas de bi-agrupamento de dados investigadas nesta dissertação não exigem que todas as linhas ou colunas de uma matriz de dados sejam agrupadas. Devido a isso, nos experimentos relatados nesta seção, houveram algumas soluções de bi-agrupamento que não incluíram todas as amostras de câncer (colunas) de uma base de dados. Em tais casos, as amostras não agrupadas foram consideradas como *singletons* durante a fase de avaliação.

Resultados

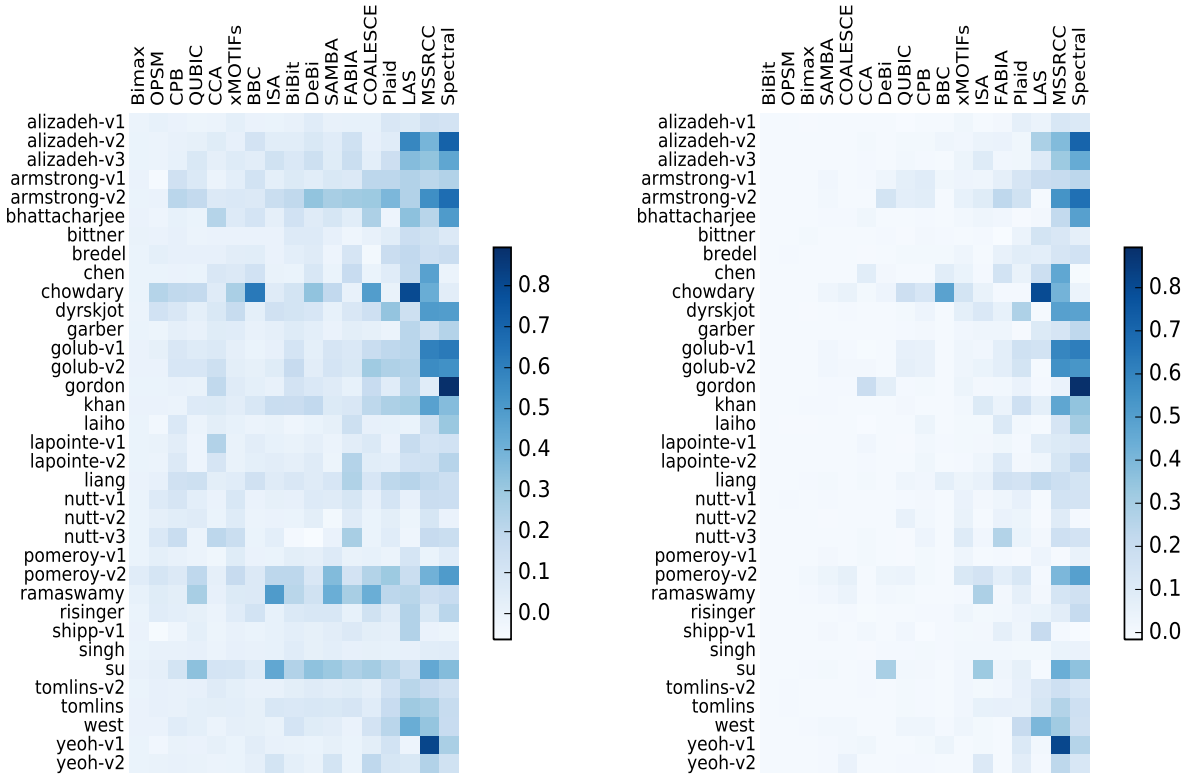
Os resultados obtidos nesta comparação estão dispostos na Figura 15 na forma de *heatmaps*, onde cada linha representa uma base de dados, cada coluna representa um algoritmo e, quanto mais escura/clara a cor de uma célula, melhor/pior um determinado algoritmo comportou-se em uma base de dados. É importante observar que os dois *heatmaps* possuem ordens diferentes em suas colunas, uma vez que elas foram ordenadas de acordo com a performance de cada algoritmo para cada medida em todas as bases de dados.

Os algoritmos de bi-agrupamento que atingiram os melhores resultados neste experimento foram o Spectral e o MSSRCC, os quais obrigam que todas as linhas e todas as colunas de uma matriz de dados sejam bi-agrupadas. É importante mencionar que cada um desses dois algoritmos atingiu os mesmos valores para as suas soluções com ambas as medidas, uma vez que os resultados retornados por eles nesta avaliação consistem em agrupamentos particionais exclusivos e, em tal situação, as medidas FARI e 13AGRI são equivalentes ao Índice de Rand Ajustado (Equação 3.25) (HORTA; CAMPELLO, 2015).

Boa parte dos algoritmos restantes não atingiram resultados satisfatórios para o agrupamento de amostras de câncer, possivelmente devido à presença de vários *singletons* em algumas das soluções retornadas por eles, acarretando em penalizações por parte das medidas de validação

externa. Entretanto, em certos casos, alguns deles produziram partições de agrupamento melhores quando comparados às técnicas mencionadas acima. Por exemplo: o LAS nas bases de dados chowdary e west, o BBC em chowdary e o ISA em ramaswamy.

Figura 15 – Resultados das medidas FARI e 13AGRI para cada algoritmo em cada base de dados de câncer.



(a) FARI.

(b) 13AGRI.

5.3 Considerações

Neste capítulo foram reportados experimentos em um número mais representativo de bases de dados reais do que em estudos comparativos anteriores. Foram feitas duas análises distintas de resultados, sendo a primeira focada na validação dos grupos de genes encontrados pelos algoritmos frente ao conhecimento biológico representado por meio de ontologias, enquanto que a segunda buscou mensurar a correspondência entre grupos de amostras de câncer encontrados pelos algoritmos com rotulações conhecidas de antemão.

Com relação à tarefa de agrupamento de genes, todos os algoritmos estudados reportaram bi-grupos enriquecidos. Entretanto, apenas cinco deles foram capazes de reportar enriquecimentos em todas as bases de dados, sendo que dois (BBC e FABIA) obtiveram mais de 80% dos seus bi-grupos enriquecidos. Ao considerar a comparação descrita pela Equação (5.1), três técnicas de bi-agrupamento (LAS, OPSM e SAMBA), as quais obtiveram entre 50% e 70% de enriquecimento na primeira análise, apresentaram melhores *p-values* do que as demais, o que

indica a importância de, além das proporções de bi-grupos enriquecidos, se levar também em conta a qualidade dos seus grupos de genes conforme indicado pelos seus respectivos *p-values*.

Para o agrupamento de amostras de câncer, a maioria das técnicas de bi-agrupamento de dados apresentaram resultados insatisfatórios. As únicas exceções foram os algoritmos MSSRCC e Spectral. De certa forma, tal comportamento era esperado, pois nesse cenário ambos produzem agrupamentos particionais exclusivos, evitando possíveis penalizações pelas medidas de validação decorrentes da existência de objetos não agrupados (*singletons*) ou grupos com amostras compartilhadas. Além disso, em seus estudos originais, ambos foram desenvolvidos com foco principal em agrupamento de dados de expressão gênica provenientes de tecidos cancerígenos.

CAPÍTULO

6

CONCLUSÕES

Nesta dissertação foram estudados e comparados 17 algoritmos do estado da arte na literatura de bi-agrupamento de dados. Os experimentos comparativos foram conduzidos em dois cenários distintos: o primeiro envolvendo bases de dados sintéticas onde se conhecia de antemão as soluções contidas em cada uma delas; e o segundo, em que se investigou a aplicabilidade dos algoritmos em bases de dados de expressão gênica reais, onde não se disponibilizava de bi-grupos pré-conhecidos, sendo necessário avaliar os resultados sob duas perspectivas distintas (agrupamento de genes e agrupamento de amostras).

No Experimento 1 do Capítulo 4 foram apresentados os resultados obtidos de 15 dentre os 17 algoritmos em uma coleção de bases de dados sintéticas gerada segundo a proposta de Eren *et al.* (2013). Nela, foram investigados três cenários distintos: (i) ruído, (ii) número de bi-grupos e (iii) sobreposição entre bi-grupos. Pelas avaliações feitas, nenhum dos algoritmos obteve alta performance para os três casos. Entretanto, alguns conseguiram alcançar bons resultados em dois deles, como: COALESCE, FABIA, ISA, LAS e SAMBA mostraram-se robustos em (i) e (ii), enquanto que BiBit, Bimax e CPB saíram-se bem em (ii) e (iii). Além disso, ao escolher um algoritmo de bi-agrupamento de dados para uma aplicação, deve-se também levar em conta os tipos de padrões que o mesmo é capaz de identificar. Na maioria das situações, os tipos de bi-grupos que se supõe que existam nos dados não são conhecidos. Desse modo, conforme observado por Eren *et al.* (2013), pode ser razoável em um primeiro momento optar por algoritmos que sejam capazes de identificar mais de um tipo de padrão (por exemplo, CPB e Plaid, de acordo com os experimentos reportados nesta dissertação).

No Experimento 2 do Capítulo 4 foram relatados os resultados dos algoritmos MSSRCC e Spectral, não inclusos no Experimento 1, em uma coleção de bases de dados sintéticas proposta neste trabalho contendo bi-agrupamentos *checkerboard*. Ambos conseguiram identificar exatamente os tipos de padrões para os quais eles foram projetados. Além disso, os dois mostraram-se robustos à presença de ruídos nos dados, uma vez que as avaliações do MSSRCC pouco variaram

entre os diferentes níveis de ruído, e o Spectral obteve avaliações muito próximas do ideal, tendo um leve descréscimo apenas nos níveis mais altos.

No Capítulo 5 foram discutidos os experimentos em dados de expressão gênica reais. Primeiramente, os algoritmos foram avaliados com relação aos seus agrupamentos de genes em uma coleção de 27 bases que foram utilizadas em três estudos distintos. Nesta tarefa, os algoritmos LAS, OPSM e SAMBA mostraram-se opções interessantes, uma vez que tiveram entre 50% e 70% dos seus bi-grupos enriquecidos e foram superiores aos demais algoritmos quando comparados em relação aos *p-values* associados às suas soluções. Para a tarefa de agrupamento de amostras de câncer, os algoritmos foram aplicados em uma coleção contendo 35 bases de dados. Durante a análise de resultados, ficou evidente que o MSSRCC e Spectral foram superiores às demais técnicas de bi-agrupamento.

Ademais, houve uma contribuição significativa no estabelecimento de uma metodologia geral de avaliação de algoritmos de bi-agrupamento em seus diferentes cenários e diferentes aspectos. Em suma, foram compiladas várias escolhas de configurações e procedimentos para a avaliação experimental a partir da literatura, criticando cada uma de forma a selecionar aquelas julgadas mais adequadas, descartando as demais e propondo outras em seu lugar. Mais do que a comparação em si, todo o arcabouço comparativo pode ser reusado para a comparação de outros algoritmos de bi-agrupamento no futuro.

Finalmente, em trabalhos futuros, pode-se investigar possíveis combinações de partes dos algoritmos investigados e outros da literatura, de modo a apresentar novas técnicas de bi-agrupamento que possam tirar vantagem dos pontos positivos de alguns deles. Assim, as performances das novas abordagens podem ser estudadas através da metodologia experimental estabelecida neste trabalho. Outra possibilidade seria de testar as técnicas investigadas em bases de dados geradas a partir de RNA-Seq, dado que poucos trabalhos da literatura investigaram algoritmos de bi-agrupamento em tal tarefa. Ademais, pode-se estender todo o procedimento metodológico adotado para a comparação de algoritmos de bi-agrupamento em outros domínios de aplicação, tais como: mineração de textos, recuperação de informação ou filtragem colaborativa em sistemas de recomendação.

REFERÊNCIAS

AGUILAR-RUIZ, J. S. Shifting and scaling patterns from gene expression data. **Bioinformatics**, Oxford Univ Press, v. 21, n. 20, p. 3840–3845, 2005. Citado 3 vezes nas páginas 36, 39 e 40.

ALBATINEH, A. N.; NIEWIADOMSKA-BUGAJ, M.; MIHALKO, D. On similarity indices and correction for chance agreement. **Journal of Classification**, Springer, v. 23, n. 2, p. 301–313, 2006. Citado na página 55.

ALBERTS, B.; JOHNSON, A.; LEWIS, J.; MORGAN, D.; RAFF, M.; ROBERTS, K.; WALTER, P. (Ed.). **Molecular Biology of the Cell**. Sixth. [S.l.]: Garland Science, 2014. Citado na página 29.

ANDERSON, D. T.; BEZDEK, J. C.; POPESCU, M.; KELLER, J. M. Comparing fuzzy, probabilistic, and possibilistic partitions. **IEEE Transactions on Fuzzy Systems**, IEEE, v. 18, n. 5, p. 906–918, 2010. Citado na página 54.

ASHBURNER, M.; BALL, C. A.; BLAKE, J. A.; BOTSTEIN, D.; BUTLER, H.; CHERRY, J. M.; DAVIS, A. P.; DOLINSKI, K.; DWIGHT, S. S.; EPPIG, J. T.; MIDORI, A. H.; HILL, D. P.; ISSEL-TARVER, L.; KASARKIS, A.; LEWIS, S.; MATESE, J. C.; RICHARDSON, J. E.; RINGWALD, M.; RUBIN, G. M.; SHERLOCK, G. Gene ontology: tool for the unification of biology. **Nature genetics**, Nature Publishing Group, v. 25, n. 1, p. 25–29, 2000. Citado 3 vezes nas páginas 26, 53 e 60.

AYADI, W.; ELLOUMI, M.; HAO, J. K. Bicfinder: a biclustering algorithm for microarray data analysis. **Knowledge and Information Systems**, Springer, v. 30, n. 2, p. 341–358, 2012. Citado 3 vezes nas páginas 26, 81 e 82.

BALL, M. P. **DNA replication split**. 2007. Disponível em: <https://commons.wikimedia.org/w/index.php?curid=2497221>. Acesso em: 1 ago. 2016. Citado na página 30.

BARKOW, S.; BLEULER, S.; PRELIĆ, A.; ZIMMERMANN, P.; ZITZLER, E. Bicat: a biclustering analysis toolbox. **Bioinformatics**, Oxford Univ Press, v. 22, n. 10, p. 1282–1283, 2006. Citado na página 105.

BEN-DOR, A.; CHOR, B.; KARP, R.; YAKHINI, Z. Discovering local structure in gene expression data: the order-preserving submatrix problem. **Journal of Computational Biology**, Mary Ann Liebert, Inc., v. 10, n. 3-4, p. 373–384, 2003. Citado na página 41.

BERGMANN, S.; IHMELS, J.; BARKAI, N. Iterative signature algorithm for the analysis of large-scale gene expression data. **Physical Review E**, APS, v. 67, n. 3, p. 031902, 2003. Citado na página 42.

BERKHIN, P. A survey of clustering data mining techniques. In: **Grouping Multidimensional Data**. [S.l.]: Springer, 2006. p. 25–71. Citado na página 23.

BHATTACHARYA, A.; DE, R. K. Bi-correlation clustering algorithm for determining a set of co-regulated genes. **Bioinformatics**, Oxford Univ Press, v. 25, n. 21, p. 2795–2801, 2009. Citado na página 24.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. [S.l.]: Springer, 2006. Hardcover. ISBN 0387310738. Citado na página 23.

BLACKBURN, G. M.; GAIT, M. J.; LOAKES, D.; WILLIAMS, D. M. (Ed.). **Nucleic acids in chemistry and biology**. [S.l.]: Royal Society of Chemistry, 2006. Citado na página 30.

BOZDAĞ, D.; KUMAR, A. S.; ÇATALYÜREK, Ü. V. Comparative analysis of biclustering algorithms. In: ACM. **Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology**. [S.l.], 2010. p. 265–274. Citado 7 vezes nas páginas 27, 35, 57, 60, 76, 77 e 81.

BOZDAĞ, D.; PARVIN, J. D.; ÇATALYÜREK, Ü. V. A biclustering method to discover co-regulated genes using diverse gene expression datasets. In: **Bioinformatics and Computational Biology**. [S.l.]: Springer, 2009. p. 151–163. Citado na página 44.

BROUWER, R. K. Extending the rand, adjusted rand and jaccard indices to fuzzy partitions. **Journal of Intelligent Information Systems**, Springer, v. 32, n. 3, p. 213–235, 2009. Citado 2 vezes nas páginas 54 e 55.

BURDICK, D.; CALIMLIM, M.; GEHRKE, J. Mafia: A maximal frequent itemset algorithm for transactional databases. In: IEEE. **Proceedings of the 17th International Conference on Data Engineering**. [S.l.], 2001. p. 443–452. Citado na página 52.

BUSYGIN, S.; PROKOPYEV, O.; PARDALOS, P. M. Biclustering in data mining. **Computers & Operations Research**, Elsevier, v. 35, n. 9, p. 2964–2987, 2008. Citado na página 25.

CAMPELLO, R. J. G. B. Generalized external indexes for comparing data partitions with overlapping categories. **Pattern Recognition Letters**, Elsevier, v. 31, n. 9, p. 966–975, 2010. Citado 2 vezes nas páginas 54 e 59.

CANO, C.; ADARVE, L.; LÓPEZ, J.; BLANCO, A. Possibilistic approach for biclustering microarray data. **Computers in biology and medicine**, Elsevier, v. 37, n. 10, p. 1426–1436, 2007. Citado na página 26.

CHATURVEDI, A.; CARROLL, J. D. An alternating combinatorial optimization approach to fitting the indclus and generalized indclus models. **Journal of Classification**, Springer, v. 11, n. 2, p. 155–170, 1994. Citado na página 48.

CHENG, Y.; CHURCH, G. M. Biclustering of expression data. In: **Proceedings of the 8th International Conference on Intelligence Systems for Molecular Biology**. [S.l.]: AAAI Press, 2000. v. 8, p. 93–103. Citado 4 vezes nas páginas 24, 25, 38 e 41.

CHO, H.; DHILLON, I. S. Coclustering of human cancer microarrays using minimum sum-squared residue coclustering. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, IEEE Computer Society Press, v. 5, n. 3, p. 385–400, 2008. Citado 3 vezes nas páginas 43, 81 e 107.

CHO, H.; DHILLON, I. S.; GUAN, Y.; SRA, S. Minimum sum-squared residue co-clustering of gene expression data. In: SIAM. **Proceedings of the 2004 SIAM International Conference on Data Mining**. [S.l.], 2004. v. 3, p. 3. Citado 2 vezes nas páginas 43 e 81.

CLARK, D. (Ed.). **Molecular Biology**. [S.l.]: Elsevier, 2005. Citado na página 29.

CORMEN, T. H.; LEISERSON, C. E.; RIVEST, R. L.; STEIN, C. **Introduction to Algorithms**. Third. [S.l.]: MIT press, 2009. Citado 2 vezes nas páginas 41 e 46.

CSÁRDI, G.; KUTALIK, Z.; BERGMANN, S. Modular analysis of gene expression data with r. **Bioinformatics**, Oxford Univ Press, v. 26, n. 10, p. 1376–1377, 2010. Citado 2 vezes nas páginas 105 e 107.

DALMA-WEISZHAUSZ, D. D.; WARRINGTON, J.; TANIMOTO, E. Y.; MIYADA, C. G. The affymetrix genechip® platform: An overview. **Methods in enzymology**, Elsevier, v. 410, p. 3–28, 2006. Citado na página 32.

DUGGAN, D. J.; BITTNER, M.; CHEN, Y.; MELTZER, P.; TRENT, J. M. Expression profiling using cdna microarrays. **Nature genetics**, Nature Publishing Group, v. 21, p. 10–14, 1999. Citado na página 32.

EDGAR, R.; DOMRACHEV, M.; LASH, A. E. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. **Nucleic acids research**, Oxford Univ Press, v. 30, n. 1, p. 207–210, 2002. Citado na página 79.

EREN, K. **Application of Biclustering Algorithms to Biological Data**. Dissertação (Mestrado) — The Ohio State University, 2012. Citado 2 vezes nas páginas 83 e 107.

EREN, K.; DEVECI, M.; KÜÇÜKTUNÇ, O.; ÇATALYÜREK, Ü. V. A comparative analysis of biclustering algorithms for gene expression data. **Briefings in Bioinformatics**, Oxford Univ Press, v. 14, n. 3, p. 279–292, 2013. Citado 21 vezes nas páginas 27, 35, 45, 49, 53, 57, 60, 63, 64, 65, 66, 67, 76, 77, 79, 80, 81, 82, 89, 105 e 107.

FACELI, K.; CARVALHO, A. C. P. L. F.; SILVA JR, W. A. Evaluation of gene selection metrics for tumor cell classification. **Genetics and Molecular Biology**, SciELO Brasil, v. 27, n. 4, p. 651–657, 2004. Citado na página 80.

GAN, G.; MA, C.; WU, J. **Data Clustering: Theory, Algorithms, and Applications**. [S.l.]: Siam, 2007. Citado 2 vezes nas páginas 23 e 24.

GIROLAMI, M. A variational method for learning sparse and overcomplete representations. **Neural computation**, MIT Press, v. 13, n. 11, p. 2517–2532, 2001. Citado na página 49.

GREMALSCHI, S.; ALTUN, G. Mean squared residue based biclustering algorithms. In: SPRINGER. **International Symposium on Bioinformatics Research and Applications**. [S.l.], 2008. p. 232–243. Citado na página 26.

GU, J.; LIU, J. S. Bayesian biclustering of gene expression data. **BMC Genomics**, BioMed Central Ltd, v. 9, n. Suppl 1, p. S4, 2008. Citado 4 vezes nas páginas 24, 49, 73 e 81.

GÜNNEMANN, S.; FÄRBER, I.; MÜLLER, E.; ASSENT, I.; SEIDL, T. External evaluation measures for subspace clustering. In: ACM. **Proceedings of the 20th ACM International Conference on Information and Knowledge Management**. [S.l.], 2011. p. 1363–1372. Citado 3 vezes nas páginas 57, 99 e 102.

HARRINGTON, C. A.; ROSENOW, C.; RETIEF, J. Monitoring gene expression using dna microarrays. **Current opinion in Microbiology**, Elsevier, v. 3, n. 3, p. 285–291, 2000. Citado 2 vezes nas páginas 29 e 32.

HARTIGAN, J. A. Direct clustering of a data matrix. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 67, n. 337, p. 123–129, 1972. Citado na página 24.

HOCHBERG, Y.; BENJAMINI, Y. More powerful procedures for multiple significance testing. **Statistics in medicine**, Wiley Online Library, v. 9, n. 7, p. 811–818, 1990. Citado na página 81.

HOCHREITER, S.; BODENHOFER, U.; HEUSEL, M.; MAYR, A.; MITTERECKER, A.; KASIM, A.; KHAMIAKOVA, T.; SANDEN, S. V.; LIN, D.; TALLOEN, W.; BIJNENS, L.; GÖHLMANN, H. W. H.; SHKEDY, Z.; CLEVER, D.-A. Fabia: factor analysis for bicluster acquisition. **Bioinformatics**, Oxford Univ Press, v. 26, n. 12, p. 1520–1527, 2010. Citado 3 vezes nas páginas 49, 63 e 105.

HORTA, D.; CAMPELLO, R. J. G. B. Similarity measures for comparing biclusterings. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, IEEE, v. 11, n. 5, p. 942–954, 2014. Citado 13 vezes nas páginas 25, 26, 27, 57, 58, 59, 60, 70, 75, 76, 99, 100 e 103.

____. Comparing hard and overlapping clusterings. **Journal of Machine Learning Research**, v. 16, p. 2949–2997, 2015. Citado 4 vezes nas páginas 54, 55, 56 e 86.

HUBERT, L.; ARABIE, P. Comparing partitions. **Journal of classification**, Springer, v. 2, n. 1, p. 193–218, 1985. Citado na página 55.

HUMAN GENOME PROGRAM. **Primer on Molecular Genetics**. [S.l.: s.n.], 1992. Citado na página 30.

HUTTENHOWER, C.; MUTUNGU, K. T.; INDIK, N.; YANG, W.; SCHROEDER, M.; FORMAN, J. J.; TROYANSKAYA, O. G.; COLLER, H. A. Detailing regulatory networks through large scale data integration. **Bioinformatics**, Oxford Univ Press, v. 25, n. 24, p. 3267–3274, 2009. Citado na página 46.

IHMELS, J.; FRIEDLANDER, G.; BERGMANN, S.; SARIG, O.; ZIV, Y.; BARKAI, N. Revealing modular organization in the yeast transcriptional network. **Nature genetics**, Nature Publishing Group, v. 31, n. 4, p. 370–377, 2002. Citado na página 63.

JACCARD, P. Nouvelles recherches sur la distribution florale. **Bulletin de la Socit Vaudoise de Sciences Naturelles**, p. 44, 1908. Citado na página 60.

JAIN, A. K. Data clustering: 50 years beyond k-means. **Pattern Recognition Letters**, Elsevier, v. 31, n. 8, p. 651–666, 2010. Citado na página 23.

JAIN, A. K.; DUBES, R. C. **Algorithms for Clustering Data**. [S.l.]: Prentice-Hall, Inc., 1988. Citado 2 vezes nas páginas 23 e 57.

JASKOWIAK, P. A. **On the evaluation of clustering results: measures, ensembles, and gene expression data analysis**. Tese (Doutorado) — Universidade de São Paulo, 2016. Citado 3 vezes nas páginas 30, 31 e 32.

JASKOWIAK, P. A.; CAMPELLO, R. J. G. B.; COSTA, I. G. Proximity measures for clustering gene expression microarray data: a validation methodology and a comparative analysis. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, IEEE Computer Society Press, v. 10, n. 4, p. 845–857, 2013. Citado 2 vezes nas páginas 54 e 80.

____. On the selection of appropriate distances for gene expression data clustering. **BMC Bioinformatics**, BioMed Central, v. 15, n. 2, p. 1, 2014. Citado na página 82.

JIANG, D.; TANG, C.; ZHANG, A. Cluster analysis for gene expression data: A survey. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 16, n. 11, p. 1370–1386, 2004. Citado 3 vezes nas páginas 24, 35 e 37.

KAISER, S.; LEISCH, F. **A Toolbox for Bicluster Analysis in R**. [S.l.], 2008. Citado na página 105.

KLUGER, Y.; BASRI, R.; CHANG, J. T.; GERSTEIN, M. Spectral biclustering of microarray data: coclustering genes and conditions. **Genome research**, Cold Spring Harbor Lab, v. 13, n. 4, p. 703–716, 2003. Citado na página 49.

KRIEGEL, H.-P.; KRÖGER, P.; ZIMEK, A. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. **ACM Transactions on Knowledge Discovery from Data**, ACM, v. 3, n. 1, p. 1, 2009. Citado 2 vezes nas páginas 24 e 35.

LAZZERONI, L.; OWEN, A. Plaid models for gene expression data. **Statistica Sinica**, Institute of Statistical Science, Academia Sinica, v. 12, n. 1, p. 61–86, 2002. Citado na página 47.

LEE, Y.; LEE, J.; JUN, C.-H. Stability-based validation of bicluster solutions. **Pattern Recognition**, Elsevier, v. 44, n. 2, p. 252–264, 2011. Citado na página 26.

LI, G.; MA, Q.; TANG, H.; PATERSON, A. H.; XU, Y. Qubic: a qualitative biclustering algorithm for analyses of gene expression data. **Nucleic Acids Research**, Oxford Univ Press, p. gkp491, 2009. Citado 5 vezes nas páginas 43, 44, 53, 81 e 82.

LIU, X.; WANG, L. Computing the maximum similarity bi-clusters of gene expression data. **Bioinformatics**, Oxford Univ Press, v. 23, n. 1, p. 50–56, 2007. Citado 3 vezes nas páginas 24, 81 e 82.

LOCKHART, D. J.; DONG, H.; BYRNE, M. C.; FOLLETTIE, M. T.; GALLO, M. V.; CHEE, M. S.; MITTMANN, M.; WANG, C.; KOBAYASHI, M.; HORTON, H.; BROWN, E. L. Expression monitoring by hybridization to high-density oligonucleotide arrays. **Nature biotechnology**, v. 14, n. 13, p. 1675–1680, 1996. Citado na página 32.

LODISH, H.; BERK, A.; MATSUDAIRA, P.; KAISER, C. A.; KRIEGER, M.; SCOTT, M. P.; ZIPRSKY, L.; DARNELL, J. **Molecular cell biology**. Fifth. [S.l.]: W. H. Freeman, 2003. Citado 2 vezes nas páginas 30 e 31.

MADEIRA, S. C.; OLIVEIRA, A. L. Biclustering algorithms for biological data analysis: a survey. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, IEEE, v. 1, n. 1, p. 24–45, 2004. Citado 15 vezes nas páginas 24, 25, 35, 36, 37, 38, 39, 40, 41, 43, 46, 47, 48, 50 e 68.

MIRKIN, B. G. **Mathematical Classification and Clustering**. [S.l.]: Springer, 1996. Citado na página 24.

MURALI, T.; KASIF, S. Extracting conserved gene expression motifs from gene expression data. In: **Pacific Symposium on Biocomputing**. [S.l.: s.n.], 2003. v. 8, p. 77–88. Citado 2 vezes nas páginas 42 e 106.

PALMER, J.; KREUTZ-DELGADO, K.; RAO, B. D.; WIPF, D. P. Variational EM algorithms for non-gaussian latent variable models. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2005. p. 1059–1066. Citado na página 49.

PATRIKAINEN, A.; MEILA, M. Comparing subspace clusterings. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 18, n. 7, p. 902–916, 2006. Citado na página 58.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; MATTHIEU, B.; PERROT, M.; DUCHESNAY, É. Scikit-learn: Machine learning in python. **The Journal of Machine Learning Research**, JMLR. org, v. 12, p. 2825–2830, 2011. Citado na página 105.

PONTES, B.; GIRÁLDEZ, R.; AGUILAR-RUIZ, J. S. Biclustering on expression data: A review. **Journal of biomedical informatics**, Elsevier, v. 57, p. 163–180, 2015. Citado 2 vezes nas páginas 25 e 40.

PRELIĆ, A.; BLEULER, S.; ZIMMERMANN, P.; WILLE, A.; BÜHLMANN, P.; GRUISSEM, W.; HENNIG, L.; THIELE, L.; ZITZLER, E. A systematic comparison and evaluation of biclustering methods for gene expression data. **Bioinformatics**, Oxford Univ Press, v. 22, n. 9, p. 1122–1129, 2006. Citado 19 vezes nas páginas 26, 27, 35, 46, 47, 53, 57, 60, 63, 65, 66, 75, 76, 79, 80, 81, 82, 105 e 106.

RODRIGUEZ-BAENA, D. S.; PEREZ-PULIDO, A. J.; AGUILAR-RUIZ, J. S. A biclustering algorithm for extracting bit-patterns from binary datasets. **Bioinformatics**, Oxford Univ Press, v. 27, n. 19, p. 2738–2745, 2011. Citado 2 vezes nas páginas 51 e 107.

SCHENA, M.; SHALON, D.; DAVIS, R. W.; BROWN, P. O. Quantitative monitoring of gene expression patterns with a complementary dna microarray. **Science**, The American Association for the Advancement of Science, v. 270, n. 5235, p. 467, 1995. Citado na página 32.

SERIN, A.; VINGRON, M. Debi: Discovering differentially expressed biclusters using a frequent itemset approach. **Algorithms for Molecular Biology**, v. 6, n. 1, p. 18, 2011. Citado 4 vezes nas páginas 52, 53, 81 e 82.

SETUBAL, J. C.; MEIDANIS, J. **Introduction to computational molecular biology**. [S.l.]: PWS Pub., 1997. Citado 3 vezes nas páginas 29, 30 e 31.

SHABALIN, A. A.; WEIGMAN, V. J.; PEROU, C. M.; NOBEL, A. B. Finding large average submatrices in high dimensional data. **The Annals of Applied Statistics**, JSTOR, p. 985–1012, 2009. Citado na página 45.

SHARAN, R.; MARON-KATZ, A.; SHAMIR, R. Click and expander: a system for clustering and visualizing gene expression data. **Bioinformatics**, Oxford Univ Press, v. 19, n. 14, p. 1787–1799, 2003. Citado na página 105.

SOUTO, M. C. P. de; COSTA, I. G.; ARAUJO, D. S. A. de; LUDERMIR, T. B.; SCHLIEP, A. Clustering cancer gene expression data: a comparative study. **BMC Bioinformatics**, BioMed Central, v. 9, n. 1, p. 1, 2008. Citado na página 85.

STACKLIES, W.; REDESTIG, H.; SCHOLZ, M.; WALTHER, D.; SELBIG, J. pcamethods—a bioconductor package providing pca methods for incomplete data. **Bioinformatics**, Oxford Univ Press, v. 23, n. 9, p. 1164–1167, 2007. Citado na página 80.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. [S.l.]: Addison Wesley, 2006. Citado 2 vezes nas páginas 23 e 48.

TANAY, A.; SHARAN, R.; SHAMIR, R. Discovering statistically significant biclusters in gene expression data. **Bioinformatics**, Oxford Univ Press, v. 18, n. suppl 1, p. S136–S144, 2002. Citado 3 vezes nas páginas 24, 50 e 53.

____. Biclustering algorithms: A survey. **Handbook of Computational Molecular Biology**, Chapman and Hall/CRC Boca Raton (Florida), v. 9, p. 26–1, 2005. Citado na página 25.

TRAPNELL, C.; PACHTER, L.; SALZBERG, S. L. Tophat: discovering splice junctions with rna-seq. **Bioinformatics**, Oxford Univ Press, v. 25, n. 9, p. 1105–1111, 2009. Citado na página 32.

TURNER, H.; BAILEY, T.; KRZANOWSKI, W. Improved biclustering of microarray data demonstrated through systematic performance tests. **Computational Statistics & Data Analysis**, Elsevier, v. 48, n. 2, p. 235–254, 2005. Citado na página 48.

TURNER, H. L.; BAILEY, T. C.; KRZANOWSKI, W. J.; HEMINGWAY, C. A. Biclustering models for structured microarray data. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, IEEE Computer Society Press, v. 2, n. 4, p. 316–329, 2005. Citado 2 vezes nas páginas 24 e 48.

WANG, L.; FU, X. **Data Mining with Computational Intelligence**. [S.l.]: Springer, 2005. Citado na página 23.

WANG, Z.; GERSTEIN, M.; SNYDER, M. Rna-seq: a revolutionary tool for transcriptomics. **Nature Reviews Genetics**, Nature Publishing Group, v. 10, n. 1, p. 57–63, 2009. Citado 2 vezes nas páginas 24 e 32.

XU, R.; WUNSCH II, D. C. Survey of clustering algorithms. **IEEE Transactions on Neural Networks**, IEEE, v. 16, n. 3, p. 645–678, 2005. Citado na página 23.

____. **Clustering**. [S.l.]: John Wiley & Sons, 2008. Citado na página 23.

YANG, J.; WANG, H.; WANG, W.; YU, P. S. Enhanced biclustering on expression data. In: IEEE. **Proceedings of the Third IEEE Symposium on Bioinformatics and Bioengineering**. [S.l.], 2003. p. 321–327. Citado na página 73.

____. An improved biclustering method for analyzing gene expression profiles. **International Journal on Artificial Intelligence Tools**, World Scientific, v. 14, n. 05, p. 771–789, 2005. Citado 2 vezes nas páginas 26 e 73.

YANG, J.; WANG, W.; WANG, H.; YU, P. S. δ-clusters: Capturing subspace correlation in a large data set. In: IEEE. **Proceedings of the 18th International Conference on Data Engineering**. [S.l.], 2002. p. 517–528. Citado na página 73.

YU, G.; WANG, L.-G.; HAN, Y.; HE, Q.-Y. clusterprofiler: an r package for comparing biological themes among gene clusters. **Omics: a journal of integrative biology**, Mary Ann Liebert Inc., v. 16, n. 5, p. 284–287, 2012. Citado na página 81.

ZHANG, A. **Advanced analysis of gene expression microarray data**. [S.l.]: World Scientific, 2006. Citado 4 vezes nas páginas 24, 29, 30 e 31.

APÊNDICE

A

PROPRIEDADES DA MEDIDA E4SC

O presente apêndice foi escrito com duas finalidades. A primeira delas é de introduzir a medida de validação externa E4SC, proposta por Günnemann *et al.* (2011) (Seção A.1). A segunda é de apresentar as propriedades para medidas de validação externa em bi-agrupamento de dados investigadas em (HORTA; CAMPELLO, 2014) e demonstrar, por meio de provas teóricas, quais delas são satisfeitas pela E4SC (Seção A.2).

A.1 E4SC

Em (GÜNNEMANN *et al.*, 2011), a medida E4SC foi definida através dos conceitos de revocação e precisão. Considerando dois bi-grupos $B_{sol} = (I_{sol}, J_{sol})$ e $B_{ref} = (I_{ref}, J_{ref})$ os quais dizem respeito, respectivamente, a uma submatriz gerada por uma solução de bi-agrupamento e a uma submatriz pertencente a um bi-agrupamento de referência, tais conceitos foram determinados pelos autores como:

$$\text{revocação}(B_{sol}, B_{ref}) = \frac{|t(B_{sol}) \cap t(B_{ref})|}{t(B_{ref})} = \text{precisão}(B_{ref}, B_{sol}), \tag{A.1}$$

sendo a função t definida como:

$$t(B_l) = \{(i, j) | i \in I_l \wedge j \in J_l\}, \tag{A.2}$$

para um bi-grupo qualquer $B_l = (I_l, J_l)$.

Para avaliar a correspondência entre os bi-grupos B_{sol} e B_{ref}, foi utilizada a média harmônica, apresentada na Equação (A.3).

$$\text{F1}_{\text{SC}}(B_{sol}, B_{ref}) = \frac{2 \cdot \text{revocação}(B_{sol}, B_{ref}) \cdot \text{precisão}(B_{sol}, B_{ref})}{\text{revocação}(B_{sol}, B_{ref}) + \text{precisão}(B_{sol}, B_{ref})} \tag{A.3}$$

Como a medida F1_{SC}, apresentada acima, é capaz de comparar apenas dois bi-grupos, os autores originais viram a necessidade de estendê-la, para que fosse possível a comparação de

dois bi-agrupamentos. Portanto, seja $B = \{B_i\}_{i=1}^{k}$ uma solução de bi-agrupamento encontrada por um algoritmo qualquer e composta por k bi-grupos, e $\hat{B} = \{\hat{B}_i\}_{i=1}^{q}$ uma solução de referência constituída por q bi-grupos para a base de dados considerada. Tal extensão foi formulada como:

$$\mathrm{F1}_{\mathrm{SC}}^{\mathrm{Clus}}(B,\hat{B}) = \frac{1}{|B|} \cdot \sum_{B_i \in B} \max_{\hat{B}_j \in \hat{B}} \{\mathrm{F1}_{\mathrm{SC}}(B_i,\hat{B}_j)\}. \tag{A.4}$$

É importante observar que a medida apresentada acima não é simétrica, ou seja, o valor de $\mathrm{F1}_{\mathrm{SC}}^{\mathrm{Clus}}(B,\hat{B})$ é diferente daquele calculado para $\mathrm{F1}_{\mathrm{SC}}^{\mathrm{Clus}}(\hat{B},B)$, o que motivou a combinação desses dois possíveis resultados em uma média harmônica, gerando assim, a medida E4SC (Equação A.5).

$$\mathrm{E4SC}(B,\hat{B}) = \frac{2 \cdot \mathrm{F1}_{\mathrm{SC}}^{\mathrm{Clus}}(B,\hat{B}) \cdot \mathrm{F1}_{\mathrm{SC}}^{\mathrm{Clus}}(\hat{B},B)}{\mathrm{F1}_{\mathrm{SC}}^{\mathrm{Clus}}(B,\hat{B}) + \mathrm{F1}_{\mathrm{SC}}^{\mathrm{Clus}}(\hat{B},B)} \tag{A.5}$$

A.2 Provas de propriedades da E4SC

Na presente seção, as oito propriedades investigadas em (HORTA; CAMPELLO, 2014) são enunciadas, e as devidas provas teóricas para a medida E4SC são apresentadas.

Para as provas apresentadas a seguir, algumas definições devem ser levadas em conta. Horta e Campello (2014) determinam que dois bi-grupos B_i e B_j são equivalentes (isto é, $B_i \equiv B_j$) se, e somente se, forem formados pelas mesmas linhas e colunas. Além disso, os autores definem que dois bi-agrupamentos B e $\hat{B}$ são equivalentes (isto é, $B \equiv \hat{B}$) se, e somente se, possuem a mesma quantidade de bi-grupos e uma bijeção entre B e $\hat{B}$ existe.

Propriedade 1. Seja $\hat{B} = \{\hat{B}_i\}_{i=1}^{q}$ um bi-agrupamento de referência. Seja $B = \{\hat{B}_i\}_{i=1}^{q} \cup \{B_i\}_{i=q+1}^{k}$ uma solução de bi-agrupamento encontrada por um algoritmo qualquer. $\{B_i\}_{i=q+1}^{k}$ é um conjunto de bi-grupos espúreos em B (isto é, bi-grupos que contêm apenas elementos que não foram bi-agrupados em $\hat{B}$). Seja $\overline{B}$ um bi-agrupamento equivalente a B, com exceção de que um ou mais bi-grupos espúreos foram incrementados em tamanho e continuam espúreos. É dito que uma medida S é sensível ao tamanho de bi-grupos espúreos se, e somente se, $S(B,\hat{B}) > S(\overline{B},\hat{B})$.

Proposição 1. A medida E4SC nem sempre é sensível ao tamanho de bi-grupos espúreos.

Demonstração. Considerando $\hat{B} = \{\hat{B}_1\}$, $B = \{\hat{B}_1, B_2\}$, $\overline{B} = \{\hat{B}_1, \overline{B}_2\}$ e $B_2 \subset \overline{B}_2$. Tem-se:

$$\mathrm{F1}_{\mathrm{SC}}^{\mathrm{Clus}}(\hat{B},B) = \mathrm{F1}_{\mathrm{SC}}^{\mathrm{Clus}}(\hat{B},\overline{B}) = \mathrm{F1}_{\mathrm{SC}}(\hat{B}_1,\hat{B}_1) = 1,$$

$$\mathrm{F1}_{\mathrm{SC}}^{\mathrm{Clus}}(B,\hat{B}) = \mathrm{F1}_{\mathrm{SC}}^{\mathrm{Clus}}(\overline{B},\hat{B}) = \frac{1}{2} \cdot \mathrm{F1}_{\mathrm{SC}}(\hat{B}_1,\hat{B}_1) = \frac{1}{2}.$$

Assim, obtém-se $\mathrm{E4SC}(B,\hat{B}) = \mathrm{E4SC}(\overline{B},\hat{B})$, violando a condição imposta pela Propriedade 1.

□

Propriedade 2. Sejam B e $\hat{B}$ dois bi-agrupamentos. Assume-se que B contém uma quantidade menor de bi-grupos do que $\hat{B}$ e que cada bi-grupo de B é equivalente a algum bi-grupo de $\hat{B}$. Em outras palavras, assume-se que B é um subconjunto próprio de $\hat{B}$. É dito que uma medida penaliza soluções que não cobrem todos os bi-grupos de referência se, e somente se, seu valor ao comparar B e $\hat{B}$ for menor do que 1.

Proposição 2. A medida E4SC penaliza soluções de bi-agrupamento que não cobrem todos os bi-grupos de referência.

Demonstração. Considerando B e $\hat{B}$ conforme descritos na Propriedade 2. Como $B \subset \hat{B}$ então, para todo $B_i \in B$, há um $\hat{B}_j \in \hat{B}$ tal que $B_i \equiv \hat{B}_j$, resultando em $\mathrm{F1}_{\mathrm{SC}}(B_i, \hat{B}_j) = 1$ e $\mathrm{F1}_{\mathrm{SC}}^{\mathrm{Clus}}(B, \hat{B}) = 1$. Por outro lado, existe pelo menos um elemento $\hat{B}_i \in \hat{B}$ tal que $\hat{B}_i \not\equiv B_j$ e $\mathrm{F1}_{\mathrm{SC}}(B_i, \hat{B}_j) < 1$, para todo $B_j \in B$. Consequentemente $\mathrm{F1}_{\mathrm{SC}}^{\mathrm{Clus}}(\hat{B}, B) < 1$ e $\mathrm{E4SC}(B, \hat{B}) < 1$. □

Propriedade 3. Sejam B e $\hat{B}$ dois bi-agrupamentos e seja Z os elementos da base de dados que não foram bi-agrupados por $\hat{B}$. Seja $\overline{B}$ um bi-agrupamento que difere de B apenas pela adição de elementos de Z nos bi-grupos de B e/ou pela criação de outros bi-grupos com elementos apenas de Z. É dito que uma medida S penaliza soluções por área sem intersecção se, e somente se, $S(B, \hat{B}) > S(\overline{B}, \hat{B})$.

Proposição 3. A medida E4SC não penaliza soluções por área sem intersecção.

Demonstração. A medida E4SC não satisfaz a Propriedade 1, a qual é um caso especial desta. □

Propriedade 4. Sejam $B = \{B_1\}$ e $\hat{B} = \{\hat{B}_i\}_{i=1}^{q}$ dois bi-agrupamentos, tal que $q > 1$, $t(B_1) = \bigcup_{i=1}^{q} t(\hat{B}_i)$ e $\hat{B}$ não possui bi-grupos sobrepostos. É dito que uma medida S penaliza soluções pela cobertura de múltiplos bi-grupos se, e somente se, $S(B, \hat{B}) < 1$.

Proposição 4. A medida E4SC penaliza soluções por cobertura de múltiplos bi-grupos.

Demonstração. Sejam B e $\hat{B}$ dois bi-agrupamentos, tais como descritos na Propriedade 4. Como $t(B_1) = \bigcup_{i=1}^{q} t(\hat{B}_i)$, observa-se que:

$$\text{revocação}(B_1, \hat{B}_i) = \text{precisão}(\hat{B}_i, B_1) = \frac{|t(B_1) \cap t(\hat{B}_i)|}{|t(\hat{B}_i)|} = \frac{|t(\hat{B}_i)|}{|t(\hat{B}_i)|} = 1,$$

$$\text{precisão}(B_1, \hat{B}_i) = \text{revocação}(\hat{B}_i, B_1) = \frac{|t(\hat{B}_i) \cap t(B_1)|}{|t(B_1)|} = \frac{|t(\hat{B}_i)|}{|t(B_1)|} < 1,$$

para todo $\hat{B}_i \in \hat{B}$. Assim, $\mathrm{F1}_{\mathrm{SC}}(B_1, \hat{B}_i) < 1$ e $\mathrm{F1}_{\mathrm{SC}}(\hat{B}_i, B_1) < 1$, para todo $\hat{B}_i \in \hat{B}$. Consequentemente, obtém-se $\mathrm{F1}_{\mathrm{SC}}^{\mathrm{Clus}}(B, \hat{B}) < 1$ e $\mathrm{F1}_{\mathrm{SC}}^{\mathrm{Clus}}(\hat{B}, B) < 1$, resultando em $\mathrm{E4SC}(B, \hat{B}) < 1$. □

Propriedade 5. Seja $\hat{B}$ um bi-agrupamento de referência com bi-grupos não sobrepostos. Seja B um bi-agrupamento que possui um ou mais bi-grupos que casam perfeitamente com bi-grupos em $\hat{B}$. Seja $\overline{B}$ um bi-agrupamento equivalente a B, com a exceção de que possui um ou mais bi-grupos de B replicados. Assim, uma medida S penaliza soluções com bi-grupos repetidos se, e somente se, $S(B,\hat{B}) > S(\overline{B},\hat{B})$.

Proposição 5. A medida E4SC nem sempre penaliza soluções por bi-grupos repetidos.

Demonstração. Sejam os bi-agrupamentos $B = \{B_1\}$, $\overline{B} = \{\overline{B}_1, \overline{B}_2\}$ e $\hat{B} = \{\hat{B}_1, \hat{B}_2\}$, tal que $B_1 \equiv \overline{B}_1 \equiv \overline{B}_2 \equiv \hat{B}_1 \not\equiv \hat{B}_2$. Tem-se:

$$\mathrm{F1}^{\mathrm{Clus}}_{\mathrm{SC}}(\hat{B},B) = \mathrm{F1}^{\mathrm{Clus}}_{\mathrm{SC}}(\hat{B},\overline{B}) = \frac{1}{2},$$

$$\mathrm{F1}^{\mathrm{Clus}}_{\mathrm{SC}}(B,\hat{B}) = \mathrm{F1}_{\mathrm{SC}}(B_1,\hat{B}_1) = 1,$$

$$\mathrm{F1}^{\mathrm{Clus}}_{\mathrm{SC}}(\overline{B},\hat{B}) = \frac{1}{2}\cdot(\mathrm{F1}_{\mathrm{SC}}(\overline{B}_1,\hat{B}_1) + \mathrm{F1}_{\mathrm{SC}}(\overline{B}_2,\hat{B}_1)) = \frac{1}{2}\cdot(1+1) = 1.$$

Assim, obtém-se $\mathrm{E4SC}(B,\hat{B}) = \mathrm{E4SC}(\overline{B},\hat{B})$, violando a condição imposta pela Propriedade 5. □

Propriedade 6. Uma medida S é dita simétrica se, e somente se, $S(B,\hat{B}) = S(\hat{B},B)$.

Proposição 6. A medida E4SC é simétrica.

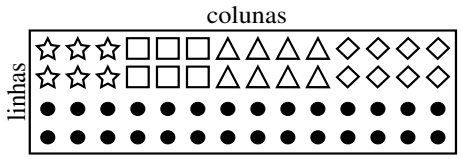
Segundo Günnemann *et al.* (2011) a propriedade de simetria é válida para a E4SC justamente pelo modo como essa medida foi projetada. A prova é trivial e foi, portanto, omitida neste texto.

Propriedade 7. Sejam B e $\hat{B}$ bi-agrupamentos sem sobreposição. Seja $B_i \in B$ um bi-grupo contendo apenas elementos dos bi-grupos $\hat{B}_{ma(i)} \in \hat{B}$ e $\hat{B}_{mi(i)} \in \hat{B}$ tal que $|t(B_i) \cap t(B_{ma(i)})| > |t(B_i) \cap t(B_{mi(i)})|$. Em outras palavras, $\hat{B}_{ma(i)}$ é a categoria majoritária em B_i e os elementos restantes são de $\hat{B}_{mi(i)}$. Sejam B_j, $\hat{B}_{ma(j)}$ e $\hat{B}_{mi(j)}$ analogamente definidos, tal que $mi(i) \neq mi(j)$, $mi(i) \neq ma(j)$ e $mi(j) \neq ma(i)$. Seja $\overline{B}$ um bi-agrupamento equivalente a B, exceto que $x > 0$ elementos pertencentes à categoria minoritária em B_i foram trocados por x elementos da categoria minoritária em B_j. É dito que S é uma medida que penaliza soluções menos homogêneas se, e somente se, $S(B,\hat{B}) \geq S(\overline{B},\hat{B})$, tal que $S(B,\hat{B}) = S(\overline{B},\hat{B})$ se, e somente se, $|t(B_i) \cap t(\hat{B}_{mi(i)})| = |t(B_j) \cap t(\hat{B}_{mi(j)})|$.

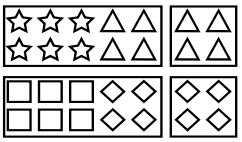
Proposição 7. A medida E4SC nem sempre penaliza soluções menos homogêneas.

Demonstração. Sejam B e $\overline{B}$ os bi-agrupamentos representados nas Figuras 16b e 16c e a solução de referência $\hat{B}$ cujos bi-grupos são representados pelos diferentes conjuntos de formas geométricas na Figura 16a exceto pelos círculos escuros, os quais representam elementos que não foram bi-agrupados. Tem-se $\mathrm{E4SC}(B,\hat{B}) = \mathrm{E4SC}(\overline{B},\hat{B}) = 0{,}71$. □

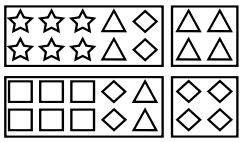
Figura 16 – Exemplo da diferença de homogeneidade entre soluções.



(a) Base de dados 4×14 com 4 bi-grupos.



(b) Bi-agrupamento B.



(c) Bi-agrupamento $\overline{B}$.

Fonte: Baseado na Figura 13 de Horta e Campello (2014).

Propriedade 8. Uma medida S obedece as condições necessária e suficiente para o valor máximo se S é tal que: $S(B,\hat{B}) = 1$ se, e somente se, B e $\hat{B}$ são bi-agrupamentos equivalentes.

Proposição 8. Existem dois bi-agrupamentos não equivalentes B e $\hat{B}$, para os quais $\text{E4SC}(B,\hat{B}) = 1$.

Demonstração. Sejam os bi-agrupamentos $B = \{B_1, B_2\}$ e $\hat{B} = \{\hat{B}_1\}$, tal que $B_1 \equiv B_2 \equiv \hat{B}_1$. Tomando um caminho semelhante ao da demonstração apresentada para a Propriedade 5, encontra-se $\text{F1}_{\text{SC}}^{\text{Clus}}(B,\hat{B}) = \text{F1}_{\text{SC}}^{\text{Clus}}(\hat{B},B) = 1$, resultando em $\text{E4SC}(B,\hat{B}) = 1$. □

Ao todo, a medida E4SC satisfaz apenas três das oito propriedades e, devido a isso, não foi utilizada neste trabalho. A Tabela 6 sumariza quais são apresentadas pelas medidas CE, CSI e E4SC. Para as provas referentes às duas primeiras em conjunto com uma detalhada análise empírica, recomenda-se a leitura do trabalho original de Horta e Campello (2014).

Tabela 6 – Propriedades satisfeitas pelas medidas CE, CSI e E4SC.

Propriedade	CE	CSI	E4SC
1	✓	✓	
2	✓	✓	✓
3	✓	✓	
4	✓	✓	✓
5	✓	✓	
6	✓	✓	✓
7		✓	
8	✓		

APÊNDICE

B

PARÂMETROS E IMPLEMENTAÇÕES DOS ALGORITMOS

No presente apêndice são apresentados comentários acerca dos parâmetros e implementações utilizados para os algoritmos revisados no Capítulo 3. Devido a isso, a organização foi feita conforme segue. Na Seção B.1 é comentado brevemente sobre as implementações e pacotes de *software* usados, os quais estão publicamente disponíveis. Na Seção B.2 são discutidas as configurações de parâmetros consideradas.

B.1 Implementações

A maioria dos códigos usados durante os experimentos estavam disponíveis na linguagem R, por meio dos pacotes biclust (KAISER; LEISCH, 2008), fabia (HOCHREITER *et al.*, 2010) e isa2 (CSÁRDI; KUTALIK; BERGMANN, 2010). Além desses, foram também utilizadas as bibliotecas scikit-learn (PEDREGOSA *et al.*, 2011) e biclustlib (desenvolvida pelo autor desta dissertação e disponível em https://bitbucket.org/padilha/biclustlib), escritas em Python. Outras implementações foram obtidas através da *Biclustering Analysis Toolbox* (BicAT) (BARKOW *et al.*, 2006), do *software* Expander (SHARAN; MARON-KATZ; SHAMIR, 2003) e de rotinas especializadas disponibilizadas pelos autores originais de cada algoritmo. Na Tabela 7 é apresentado um sumário das implementações utilizadas.

B.2 Parâmetros

Boa parte dos parâmetros utilizados para os algoritmos durante os experimentos relatados nesta dissertação estão de acordo com os valores originalmente sugeridos pelos seus respectivos autores. Foram também consideradas as combinações apresentadas em (PRELIĆ *et al.*, 2006; EREN *et al.*, 2013) ao utilizar as bases de dados provenientes de tais estudos. Ademais, alguns

Tabela 7 – Implementações utilizadas nos experimentos.

Algoritmo	Implementação	Disponibilidade
BBC	C	http://www.people.fas.harvard.edu/ junliu/BBC/
BiBit	Java	http://eps.upo.es/bigs/BiBit_algorithm.html
Bimax	R	https://cran.r-project.org/web/packages/biclust/index.html
CCA	R	https://cran.r-project.org/web/packages/biclust/index.html
COALESCE	C++	http://libsleipnir.bitbucket.org/
CPB	C	http://www.bmi.osu.edu/hpc/software/cpb/index.html
DeBi	C++	http://www.molgen.mpg.de/ serin/debi/main.html
FABIA	R	https://www.bioconductor.org/packages/release/bioc/html/fabia.html
ISA	R	https://cran.r-project.org/web/packages/isa2/
LAS	Python	https://bitbucket.org/padilha/biclustlib
MSSRCC	C++	http://www.cs.utexas.edu/users/dml/Software/cocluster.html
OPSM	Java	http://www.tik.ethz.ch/sop/bicat/
Plaid	R	https://cran.r-project.org/web/packages/biclust/index.html
QUBIC	C	http://csbl.bmb.uga.edu/ maqin/bicluster/
SAMBA	Java	http://acgt.cs.tau.ac.il/expander/
Spectral	Python	http://scikit-learn.org/stable/
xMOTIFs	R	https://cran.r-project.org/web/packages/biclust/index.html

valores de parâmetros foram ajustados, conforme descrito a seguir, com a finalidade de atingir resultados mais significativos.

O algoritmo xMOTIFs exige como entrada uma matriz de dados discretizada. Os autores originais desse algoritmo propuseram um método que é baseado em intervalos estatisticamente significativos assumindo uma distribuição uniforme como a hipótese nula. Neste trabalho foi necessário relaxar os *p-values* sugeridos em (MURALI; KASIF, 2003), pois em várias bases de dados não foi possível encontrar intervalos que pudessem satisfazer tais valores. Portanto, para cada base de dados, buscou-se pelo menor *p-value* entre 10^{-10} e 10^{-1} que pudesse ser satisfeito.

Ainda com relação ao xMOTIFs, foram alterados também o tamanho dos conjuntos discriminantes gerados pelo algoritmo e o parâmetro que define a fração mínima de colunas da base de dados que um bi-grupo deve conter. Sobre o primeiro, cada conjunto discriminante é utilizado para determinar as linhas de um bi-grupo por meio da seleção daquelas que estão no mesmo nível para todas as colunas contidas no conjunto. Os autores originais sugeriram valores no intervalo entre sete e dez para tal parâmetro. Entretanto, na coleção de bases de dados sintéticas em que foi aplicado, esse algoritmo não conseguiu reportar bi-grupos efetivamente. Portanto, o tamanho de tais conjuntos foi variado em um intervalo maior, variando entre três e dez em todos os experimentos relatados nesta dissertação. Com relação ao parâmetro de fração mínima, o valor utilizado no artigo original do xMOTIFs não foi informado. Nas bases de dados de Prelić *et al.* (2006) foi utilizado o valor 0,1 tal como sugerido naquele trabalho. Para as bases de dados restantes, esse parâmetro foi relaxado para 0,05 pois caso contrário o algoritmo não conseguiria retornar bi-grupos para a maioria dos experimentos.

Para o algoritmo CPB existe um parâmetro, denotado por κ no artigo original, o qual controla a razão entre a quantidade de linhas e colunas de um bi-grupo. Neste trabalho foi

utilizado $\kappa = 1$ em todos os experimentos, uma vez que em (EREN, 2012) é mostrado que qualquer valor diferente para essa configuração faz com que o algoritmo não seja capaz de encontrar bi-grupos de diferentes formatos, interferindo assim na qualidade dos seus resultados.

Algoritmos que exigem como parâmetro de entrada a quantidade mínima de linhas e de colunas de um bi-grupo receberam como entrada o número correto de linhas e de colunas dos bi-grupos implantados em bases de dados sintéticas. Para os cenários em que se estudou a aplicabilidade de tais procedimentos em bases reais, foi utilizado o valor padrão dois.

O MSSRCC recebe dois limiares como parâmetro, referidos pelos autores originais como limiar de atualização em lote e limiar de busca local, os quais são responsáveis por guiar a busca e o refinamento dos bi-grupos. Além dos valores sugeridos em (CHO; DHILLON, 2008), foi também considerado o valor 10^{-3} para a atualização em lote e -10^{-6} para a busca local, os quais são os valores padrão na implementação disponibilizada pelos autores do algoritmo.

O DeBi, conforme explicado no Capítulo 3, exige como entrada uma matriz de dados discretizada. Neste trabalho, tal algoritmo foi testado com os passos de pré-processamento do Bimax e do QUBIC (os quais estão descritos nas respectivas subseções de cada algoritmo no Capítulo 3).

Em bases de dados reais, o COALESCE foi testado com e sem a normalização de dados disponível em sua implementação, o LAS foi executado com e sem a transformação sugerida por seus autores como etapa de pré-processamento, e o BiBit foi executado apenas para a matriz binária pertencente ao nível máximo de discretização (conforme o procedimento descrito na respectiva seção deste algoritmo no Capítulo 3), uma vez que quanto mais alto tal nível mais especializados os bi-grupos encontrados serão (RODRIGUEZ-BAENA; PEREZ-PULIDO; AGUILAR-RUIZ, 2011).

Para a coleção de bases de dados gerada segundo a metodologia proposta por Eren *et al.* (2013), os parâmetros t_g e t_c sugeridos no trabalho que propõe o ISA foram muito restritivos e não permitiram com que esse algoritmo detectasse qualquer modelo de bi-grupo. Portanto, neste trabalho, foram também testados os parâmetros padrão do pacote isa2 (CSÁRDI; KUTALIK; BERGMANN, 2010), os quais consideram todas as possíveis combinações de t_g e t_c no intervalo $[1{,}0; 3{,}0]$ com passos de 0,5.