Daniele Yumi Sunaga

Aplicação de técnicas de validação estatística e biológica em agrupamento de dados de expressão gênica

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática Aplicada da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática Aplicada.

Curitiba
2006

Daniele Yumi Sunaga

Aplicação de técnicas de validação estatística e biológica em agrupamento de dados de expressão gênica

Dissertação de Mestrado apresentado ao Programa de Pós-Graduação em Informática Aplicada da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática Aplicada.

Área de Concentração: Metodologia e Técnicas de Computação

Orientador: Prof. Dr. Júlio Cesar Nievola

Curitiba
2006

Sunaga, Daniele Yumi Sunaga
Aplicação de técnicas de validação estatística e biológica em agrupamento de dados de expressão gênica. Curitiba, 2006.

Dissertação de Mestrado - Pontifícia Universidade Católica do Paraná Programa de Pós-Graduação em Informática Aplicada.

1. Agrupamento de dados 2. Expressão Gênica 3. Validação de agrupamento I.Pontifícia Universidade Católica do Paraná. Centro de Ciências Exatas e Tecnologia. Programa de Pós-Graduação em Informática Aplicada II - t

Ao meu ditian (*in memorian*).

Agradecimentos

Ao TECPAR e IBMP por todo o apoio sem o qual não seria possível a realização deste trabalho.

Ao Dr. Eng. Júlio César Nievola, pela sua orientação e incentivo.

Ao Dr. Eng. Milton Pires Ramos, a quem dedico agradecimento especial pela sua orientação, incentivo e seus conselhos.

Ao Eng. Júlio Cezar Zanoni pela grande ajuda na correção do trabalho.

Agradeço ao Leonardo por toda a sua ajuda, colaboração, dedicação, paciência e companhia.

A minha família e a todos que de certa forma contribuíram para a realização deste trabalho.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

A	Adenina
C	Citosina
CTWC	*Coupled Two-Way Clustering*
CVE	*Clustering and Validation Environment*
DNA	Ácido desoxirribonucléico
ESS	*Error Sum of Squares*
G	Guanina
GEMS	Gene Expression Mining Server
GO	*Gene Ontology*
IBMP	Instituto de Biologia Molecular do Paraná
MPSS	*Massively Parallel Signature Sequence technology*
ORF	*Open Read Frames*
PRIMA	*Promoter Integration in Microarray Analysis*
PUCPR	Pontifícia Universidade Católica do Paraná
RT-PCR	*Reverse-Transcription Polymerase Chain Reaction*
SAGE	*Serial Analisys of Gene Expression*
SAMBA	*Statistical-Algorithmic Method for Biclustering Analysis*
SOM	*Self Organizing Maps*

T	Timina
TANGO	*Tool for Analysis of GO Enrichment*
TECPAR	Instituto de Tecnologia do Paraná
TFs	Fatores de transcrição
U	Uracila
mRNA	RNA mensageiro
rRNA	RNA ribossômico
tRNA	RNA de transferência ou transportador

Resumo

O crescimento exponencial dos dados de expressão gênica provenientes da tecnologia de microarranjo de DNA é acompanhado pelo aumento da necessidade de ferramentas computacionais eficientes que auxiliem o processo de análise e interpretação desses dados.
Técnicas de agrupamento têm se revelado ferramentas valiosas na análise desses dados porque possibilitam a identificação de padrões entre os milhares de genes viabilizando a elucidação de questões biológicas, como funções que estes genes desempenham no organismo e processos biológicos que estão envolvidos.
Neste trabalho foram aplicados os algoritmos de agrupamento unidimensional k-médio, SOM e o algoritmo SAMBA da abordagem de agrupamento bidimensional. Estes algoritmos foram aplicados com diferentes parâmetros e em duas diferentes bases de dados de expressão gênica.
De um modo geral, a análise dos resultados de agrupamento não é uma tarefa trivial, é altamente delicada, podendo ser até mesmo subjetiva pois geralmente envolve uma grande quantidade de experimentos. Por esta razão, a escolha da melhor solução de agrupamento foi feita com o auxílio de diferentes técnicas de validação estatística e biológica.
A combinação da aplicação de diferentes algoritmos de agrupamento, variações de parâmetros, bases de dados e diferentes técnicas de validação estatística e biológica, permitiu concluir vantagens e desvantagens dos algoritmos de agrupamento e verificar quais os índices estatísticos são corroborados pelo processo de validação biológica. Além disso, a formação de grupos de genes de alta homogeneidade e a associação de funções biológicas e fatores de transcrição a esses grupos sugerem o envolvimento desses genes em uma mesma função ou processo biológico.

Palavras-chave: Agrupamento, expressão gênica, validação estatística e validação biológica.

Abstract

The exponential growth of gene expression data resulting from DNA microarray technology is accompanied by an increase of demand of efficient computational tools that hel p the analysis process and the interpretation of these data.
Clustering techniques are useful tools for analysis of gene expression data because they can identify patterns among gene expression profiles that can potentially hold m eaning biological information, such as gene function and the biological processes involved.
In this work it was applied the unidimensional k-means and SOM clustering algorithms and bidimensional SAMBA algorithm, with different parameters and on two different da tabases of gene expression.
Clustering results analysis is very sensible and even subjective because they involve a huge set of heterogeneous factors. For this reason, the choice of best clustering solution was made using different statistical and biologic validations techniques.
The combination of different clustering algorithms, parameters, databases and statistical and biological validation techniques confirmed some advantages and disadvantage s of clustering algorithms and also showed which statistical indexes were corroborated by the biological validation process. Furthermore, high homogeneity gene expression clusters were formed. With the biological function and transcription factor binding site determination we were able to evidence the relation of these genes in the same function or biologi cal process.

Keywords: Clustering, gene expression, statistical validation, biological validation.

Capítulo 1

Introdução

Este trabalho apresenta a análise de técnicas de validação estatística e biológica aplicadas nos resultados dos agrupamentos de dados de expressão gênica nas diferentes abordagens de agrupamento unidimensional e bidimensional.

1.1 Motivação

A biologia molecular tem avançado consideravelmente nos últimos anos marcada pelo desenvolvimento de métodos eficientes de seqüenciamento de DNA e posterior genômica funcional.

A genômica funcional corresponde à caracterização funcional dos genes. Embora os métodos tradicionais da genética ainda trabalhem com experimentos "um gene, um produto", a complexidade dos organismos sugere a mudança de paradigma para uma abordagem holística, onde as funções biológicas são desempenhadas em conjunto, em que um gene está associado a várias proteínas e participa de diferentes processos biológicos.

A tecnologia de microarranjo de DNA permite a análise simultânea da atividade de milhares de genes e até mesmo a atividade de genomas completos. Ela tem sido adotada como ferramenta padrão da genômica funcional, sendo amplamente utilizada para estudar genes envolvidos em doenças como o câncer, Alzheimer, Parkinson e diabetes; ela também é usada para identificar a variabilidade de organismos mutantes ou transgênicos e ainda para estudar a resposta de células quando estão sob ação de uma droga específica.

A motivação inicial deste trabalho surgiu da necessidade da utilização de métodos computacionais eficientes condizentes com a abordagem holística da técnica de microarranjo e com potencial para auxiliar na interpretação dos resultados. Estes resultados geralmente correspondem à resposta de milhares de genes submetidos às mesmas condições experimentais, por exemplo, estresse nutricional, variações de temperatura, etc. Tais res-

postas ou expressão dos genes são representadas em valores numéricos que aumentam ou diminuem dependendo de quanto o gene se manifesta em determinada condição.

A interpretação desses resultados através de técnicas computacionais pode siginificar o adiantamento de meses ou até anos de trabalho nos laboratórios de biologia, não somente pela performance e desempenho dos programas computacionais, mas também por possibilitar a análise simultânea de todos os genes e suas relações, o que é impossível fazer manualmente. O auxílio computacional ainda garante a organização das informações dos milhares de genes em bancos de dados, podendo relacioná-los com todas as informações pertinentes ao experimento realizado, com informações de outros experimentos, com a caracterização funcional desses genes na literatura, etc.

Análises computacionais mais complexas podem revelar associações entre os genes de um experimento: se um gene só trabalha mediante a expressão de outro, se vários genes trabalham juntos para executar a mesma função, se alguns genes se mantém inalterados quando outros se expressam, se a expressão de alguns genes é responsável pela inibição de outros, ou ainda, inferir que genes com função desconhecida exercem a mesma função que aqueles que apresentaram comportamento de expressão semelhante.

Técnicas de agrupamento têm sido adotadas como ferramenta padrão de análise de dados de expressão gênica justamente por responder, de maneira eficiente, à maioria dessas questões biológicas. Basicamente, são utilizadas para identificar subconjuntos de genes com comportamento similar sob diferentes condições experimentais. Porém, os resultados dos agrupamentos podem auxiliar na elucidação de questões biológicas, por exemplo, a identificação de genes participantes de um mesmo processo biológico [Dra03].

Outros fatores motivadores deste trabalho foram a necessidade da análise dos dados do IBMP (Instituto de Biologia Molecular do Paraná) e a utilização adequada das técnicas de agrupamento, já que cada algoritmo possui características específicas que, quando exploradas e combinadas com as caracteristicas dos dados de entrada, podem oferecer melhores resultados. A abordagem do agrupamento bidimensional, por exemplo, ainda é pouco explorada na área de biologia molecular embora revele grande aplicabilidade para as especificidades dos dados de microarranjo de DNA [GGD, ATS02, WK05].

Este trabalho é fruto da sinergia surgida no início de 2003 entre pesquisadores do TECPAR (Instituto de Tecnologia do Paraná), do IBMP como patrocinadores deste trabalho e da PUCPR (Pontifícia Universidade Católica do Paraná), grupo com grande potencial para o desenvolvimento e valorização da pesquisa em Bioinformática para o Estado do Paraná e para o Brasil.

O grupo de pesquisa do IBMP, pioneiro na aplicação da técnica de microarranjo de DNA no país, participa de vários projetos de caracterização funcional de genes. Na

grande maioria são projetos que estudam organismos causadores de doenças com grande impacto na saúde pública brasileira, como a Doença de Chagas, por exemplo.

As conclusões e o conhecimento obtido a partir deste trabalho servirão para contribuir com as pesquisas de genômica funcional realizadas no IBMP.

1.2 Objetivo Principal

O objetivo principal deste trabalho é comparar o comportamento de técnicas de validação estatística e biológica (*in silico*) aplicadas em diferentes algoritmos de agrupamento de dados. Desta comparação, espera-se identificar se a resposta de ambas as técnicas coincidem, se elas se complementam ou não revelam nenhuma associação.

1.3 Objetivos Específicos

- Analisar o comportamento dos k-médias e SOM (*Self Organizing Maps*) de agrupamento unidimensional de dados;
- Analisar o comportamento do algoritmo SAMBA (*Statistical-Algorithmic Method for Biclustering Analysis*) de agrupamento bidimensional de dados;
- Avaliar os resultados obtidos com o auxílio de técnicas de validação estatística;
- Avaliar os resultados obtidos com o auxílio de técnicas de validação biológica;
- Verificar se existe relação entre os resultados das técnicas de validação estatística e biológica.

1.4 Organização do trabalho

Este trabalho está organizado em seis capítulos. O Capítulo 2 apresenta a descrição do problema e os principais conceitos de biologia. O Capítulo 3 descreve os fundamentos da abordagem de agrupamento unidimensional e bidimensional de dados e os algoritmos correspondentes. O Capítulo 4 apresenta a metodologia do trabalho. O Capítulo 5 descreve os resultados obtidos e as discussões e o Capítulo 6 apresenta a conclusão do trabalho, com os benefícios alcançados e os trabalhos futuros relacionados à proposta. No final do trabalho é disponibilizado um material mais detalhado dos principais conceitos de biologia molecular e celular.

Capítulo 2

Descrição do Problema

Neste capítulo é apresentado o problema de análise de dados de expressão gênica através de técnicas de agrupamento de dados e alguns dos principais conceitos de biologia molecular, com o intuito de fornecer ao leitor uma contextualização desta área de estudo envolvida no trabalho.

Em anexo, é disponibilizado um material mais detalhado dos principais conceitos de biologia molecular e biologia celular úteis para o entendimento da etapa de validação biológica e das bases de dados utilizadas.

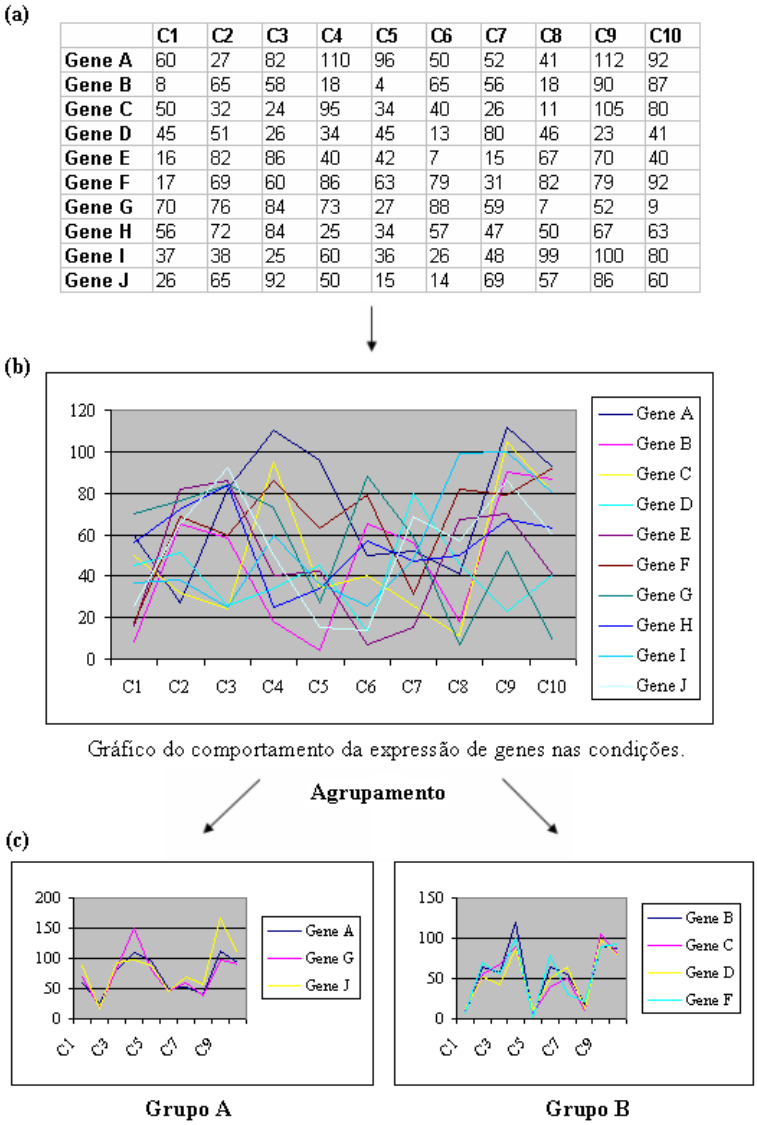
2.1 Análise de dados de expressão gênica

A Figura 2.1 ilustra a necessidade da busca por padrões na grande quantidade de dados resultantes da técnica de microarranjo de DNA.

Os dados do microarranjo de DNA são geralmente representados por uma matriz, onde os registros correspondem aos genes e as colunas correspondem às condições. O conteúdo da matriz é o nível de expressão de cada gene sobre cada condição em que ele foi submetido, conforme ilustrado na Figura 2.1 (a). Esses níveis podem ser absoluto, relativo ou ainda, normalizado. Cada vetor corresponde ao padrão de expressão de um gene sobre todas as condições [SS01a].

O gráfico (b) demonstra o comportamento da expressão de todos os genes da matriz, no entanto, a grande quantidade de dados impossibilita a identificação de padrões sem o auxílio computacional. Por esta razão são aplicadas técnicas de mineração de dados, como o agrupamento, por exemplo.

O agrupamento de dados é um passo chave neste processo de análise de dados resultantes da técnica de microarranjo de DNA. A idéia é formar grupos de genes com padrão de expressão similar (co-regulados), conforme Figura 2.1 (c). Desta forma, in-



	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Gene A	60	27	82	110	96	50	52	41	112	92
Gene B	8	65	58	18	4	65	56	18	90	87
Gene C	50	32	24	95	34	40	26	11	105	80
Gene D	45	51	26	34	45	13	80	46	23	41
Gene E	16	82	86	40	42	7	15	67	70	40
Gene F	17	69	60	86	63	79	31	82	79	92
Gene G	70	76	84	73	27	88	59	7	52	9
Gene H	56	72	84	25	34	57	47	50	67	63
Gene I	37	38	25	60	36	26	48	99	100	80
Gene J	26	65	92	50	15	14	69	57	86	60

Figura 2.1: Agrupamento de dados de expressão gênica

ferir que genes pertencentes ao mesmo grupo estejam envolvidos num mesmo processo biológico.

Técnicas de classificação também são muito utilizadas para a análise de dados de expressão gênica, mas com potencial maior para diagnósticos, tendo revelado bons resultados para a determinação de tipos de cânceres [ABDY99].

Pelo fato da tarefa de agrupamento ser de origem exploratória, isto é, devido ao seu objetivo ser o de conhecer as características de um conjunto de dados, torna-se difícil avaliar se determinado resultado é correto ou não. Isto porque, para avaliar se um determinado resultado está correto ou não, deve-se compará-lo com um resultado conhecido. Por esta razão a análise de agrupamento é altamente delicada podendo ser

subjetiva.

Neste trabalho foram aplicados cinco índices de validação estatística, além dos índices de homogeneidade e separação entre os grupos. Muitos desses índices têm sido aplicados para a validação dos resultados de agrupamentos de dados de expressão gênica. As vantagens desses índices de validação é a habilidade para comparar soluções e medir o progresso do desempenho de um algoritmo. O problema é que essas técnicas podem não refletir exatamente a intuição que um biólogo pode ter. Por esta razão, técnicas de validação biológicas também foram utilizadas. A idéia é maximizar a validação dos agrupamentos e comparar se as técnicas de validação estatísticas confirmam os resultados das técnicas de validação biológicas e vice-versa.

2.2 Conceitos Básicos de Biologia Molecular

O DNA (Ácido desoxirribonucléico) é o material genético da maioria dos seres vivos, está organizado na célula na forma de filamentos espiralados chamados cromossomo. Alguns trechos do DNA correspondem a genes, que são as unidades fundamentais da hereditariedade. Embora grande e aparentemente complexa, a molécula de DNA é simples, composta por uma combinação de apenas quatro unidades químicas chamadas nucleotídeos ou bases, são eles: A (Adenina), G (Guanina), C (Citosina) e T (Timina) [MS94]. A disposição desses nucleotídeos na seqüência de DNA é que determina as instruções que dão origem a um organismo. No entanto, estas instruções não são diretamente utilizadas para a síntese protéica, devem ser previamente transcritas em outra estrutura, denominada mRNA (RNA mensageiro), que posteriormente são processadas por mecanismos moleculares chamados de replicação, transcrição e tradução, que compõem os fundamentos do "dogma central da biologia molecular".

O dogma central da biologia molecular estabelece que o DNA atua como molde para se replicar, ele também é transcrito em RNA, e o RNA é traduzido em proteína. O RNA é semelhante ao DNA, também é formado por uma cadeia de nucleotídeos: A (Adenina), C (Citosina), G (Guanina) e U (Uracila), ao invés de Timina. Existem três tipos de RNAs de acordo com sua função e/ou estrutura: o mRNA, o tRNA (RNA de transferência ou transportador) e o rRNA (RNA ribossômico).

Conforme mostrado na Figura 2.2, as proteínas originam-se da tradução da informação contida no DNA. Elas são as macromoléculas mais abundantes nas células. De maneira geral, as proteínas desempenham as seguintes funções no organismo: estrutural, enzimática, hormonal, de defesa, nutritivo, coagulação sangüínea e transporte. Alguns exemplos de proteínas: colágeno, lipases , insulina, hemoglobina, etc.

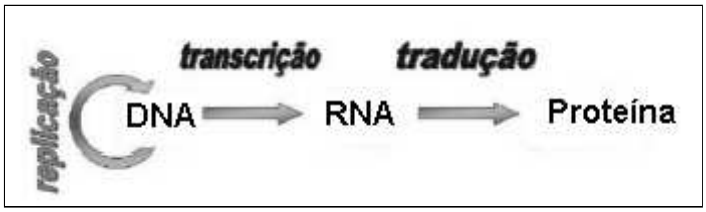


Figura 2.2: Dogma central da biologia molecular.

As unidades constituintes das proteínas são os aminoácidos. Embora existam apenas vinte tipos de aminoácidos, longas repetições de seqüências múltiplas permitem dezenas de milhares de combinações de aminoácidos para formar uma grande variedade de proteínas. No entanto, o conhecimento da seqüência de aminoácidos que compõem uma proteína é apenas o primeiro passo no complexo caminho da descrição de sistemas biológicos. O segredo está em desvendar o mistério de como os genes produzem as proteínas e como funciona a interação entre elas. As proteínas são a chave para descobrir como uma doença se desenvolve no organismo. Entendendo o funcionamento delas e como interagem, será possível desenvolver medicamentos que atuem em alvos específicos sem causar sequer efeitos colaterais [Zah03].

A obtenção de um panorama geral dessa interação gene-proteína e proteína-proteína, além de outros mecanismos celulares, tem sido possível graças às técnicas da genômica funcional.

Genômica funcional é a busca pelo entendimento da funcionalidade dos genes através da análise de seus níveis de expressão. Expressão, porque a maneira como o gene se "expressa"" através dos seus produtos, do RNA e, indiretamente, através das proteínas.

Apesar de todos os genes estarem linearmente arranjados ao longo dos cromossomos e presentes em todas as células, nem todos se expressam da mesma forma em células diferentes, em outras palavras, os genes são "ligados" e "desligados" em tipos celulares diferentes.

Portanto, a expressão gênica permite que células geneticamente idênticas tornem-se morfológica, química e funcionalmente diferentes [Slo02].

Sabe-se que os genes são expressos em padrões espaciais e temporais, mas ainda não foram revelados os códigos que regulam essas expressões. Estes padrões permitem que o nível de expressão de genes possa ser estudado sob várias condições experimentais: antes e após a aplicação de uma droga, em diferentes instantes de tempo, diferentes tratamentos e diferentes tecidos (normal e tumoral), etc. Decifrar estes códigos é um problema científico importante na era pós-genômica.

Diversas técnicas têm sido propostas para obtenção da expressão dos genes, MPSS (*Massively Parallel Signature Sequence technology*) [BJB+00], SAGE (*Serial Analisys of Gene Expression*) [VEVK95] , Real-time RT-PCR (*Reverse-Transcription Polymerase Chain Reaction*) [WMFV99] e Microarranjo de DNA [MSB95].

Os dados de expressão gênica analisados neste trabalho são provenientes da técnica de microarranjo de DNA. Optou-se por esta técnica por sua grande vantagem de permitir a detecção simultânea de milhares de genes transcritos em um único experimento e porque é uma técnica amplamente utilizada no IBMP, onde pretende-se aplicar as conclusões deste trabalho

2.3 Tecnologia de microarranjo de DNA

A tecnologia de microarranjo de DNA [MSB95], também denominada de Chip de DNA ou Biochip, é especialmente apropriada para estudos de comparação da expressão gênica em diferentes tecidos ou diferentes condições que uma população de células possa estar submetida.

A preparação do microarranjo é feita a partir de uma lâmina de vidro de dimensões mínimas. Esta lâmina apresenta uma coleção de pequenos poços, ou *spots*, onde uma amostra específica de DNA representando um gene distinto é depositada. Como resultado, conforme ilustrado na etapa de Preparação da Sonda na Figura 2.3, o microarranjo conterá em cada *spot* uma amostra representativa de um gene, geralmente denominada sonda.

O passo seguinte é a preparação do alvo, que consiste em uma condição diferenciada da sonda. A Figura 2.3 exemplifica um experimento com tecido sadio como sonda ou condição controle, sendo comparado com um tecido tumoral, que corresponde ao alvo ou condição variante.

Os dois tecidos são marcados com grupamentos químicos distintos de natureza fluorescente. Os marcadores fluorescentes absorvem a radiação e emitem energia em outro comprimento de onda, que é prontamente captado por um leitor óptico. Como cada marcador emite radiação em um comprimento de onda diferente, pode ser avaliado o quanto a sonda respondeu ao alvo em qualquer *spot* do arranjo. Esta correspondência entre sonda e alvo é baseada na complementaridade entre as duas fitas da molécula de DNA (consultar o material com os conceitos básicos de biologia no Apêndice B).

Computacionalmente, cores são atribuídas às faixas de emissão de cada um dos marcadores, a amostra controle geralmente recebe a cor verde e a situação variante recebe a cor vermelha. A cor de cada ponto indica a situação na qual o gene foi expresso e a intensidade do seu brilho é proporcional à sua intensidade de expressão. Sendo assim,

um *spot* que tenha cor verde representa um predomínio relativo da expressão gênica na amostra controle, o que implica em uma sub-expressão na condição variante. Ao contrário, um *spot* vermelho indica maior expressão gênica na condição variante, dizendo-se que o gene está relativamente superexpresso nesta condição. O amarelo denota uma situação intermediária de igual expressão gênica nas duas condições. E a cor cinza indica ausência de sinal ou sinal de baixa qualidade. A imagem de um microarranjo com as cores de cada *spot* é ilustrada na Figura 2.3.

As aplicações desta técnica são amplas, variando desde a análise de patologias como o câncer, estudo do efeito de fármacos no organismo humano, análise de tecidos submetidos a uma determinada condição de estresse, chegando até a genotipagem e abordagens proteômicas.

O material de biologia, em anexo, explica o funcionamento da técnica de microarranjo com mais detalhes e exemplos. Um suplemento completo da técnica de microarranjo de DNA está disponível no site da revista *Nature Genetics* [Nat06].

A última etapa do experimento de microarranjo consiste na análise dos resultados, que exige o auxílio de técnicas computacionais por envolver a complexidade de uma grande quantidade de dados.

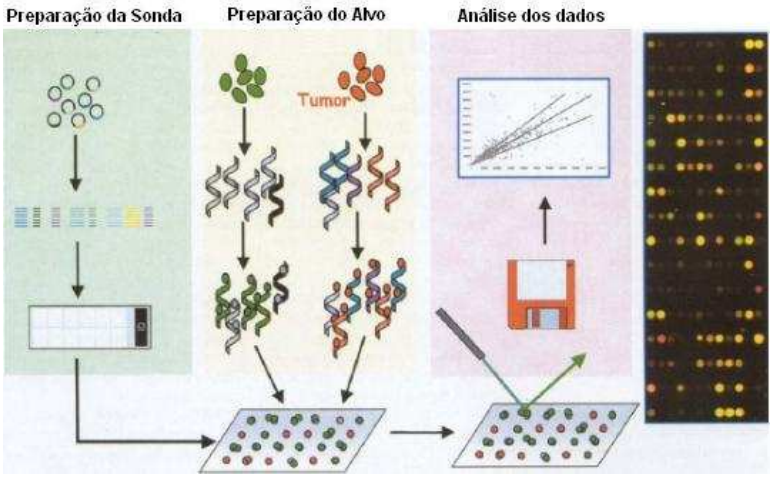


Figura 2.3: Esquema de um experimento de análise de expressão gênica utilizando microarranjo de DNA [KBR06].

Normalização dos dados de microarranjo de DNA

Uma etapa bastante importante de pré-processamento dos dados de microarranjo é a chamada normalização. Muitas vezes, os diferentes atributos que representam os padrões dos dados se apresentam em escalas diferentes. Quando os intervalos de valores

dos atributos diferem muito, pode ser que um atributo domine o resultado do agrupamento. Para solucionar este problema, é comum a padronização dos dados de forma que os atributos estejam na mesma escala.

No caso de dados de expressão gênica, usando como referência os *spots* de genes controles (sabidamente expressos ou reprimidos nos tecidos ou células estudados), o que se busca é, basicamente, retirar a influência de manchas espúrias (*background*) e de variações do processo de hibridação dos valores de cada *spot*. Desta forma, após a normalização, torna-se possível a comparação de *spots* de uma mesma lâmina ou de experimentos diferentes.

Um exemplo típico de normalização é a transformação logarítmica dos valores brutos do microarranjo. Há várias razões para a transformação logarítmica.

Primeiro, oferece valores que são mais facilmente interpretáveis do ponto de vista biológico. Considerando dois genes com valores de intensidade de *background* igual a 1.000 na amostra controle. Uma próxima medida dos dois mesmos genes na condição interesse registra valores de intensidade de *background* de 100 e 10.000, respectivamente. Se for considerado o valor absoluto da diferença entre os valores controle e os valores dos dois experimentos, o resultado é um gene muito mais expresso:

$$10.000 - 1.000 = 9.000 >> 1.000 - 100 = 900$$

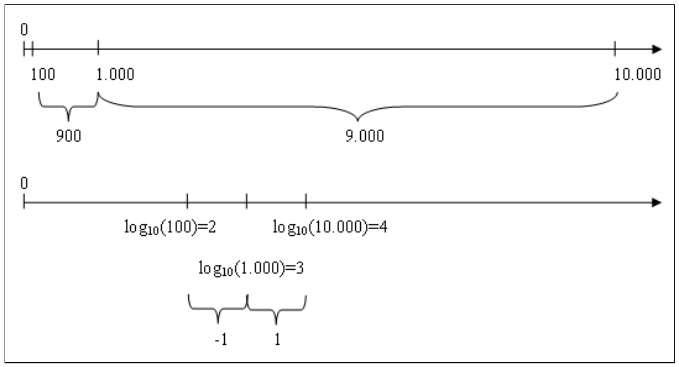


Figura 2.4: Transformação logarítmica dos dados brutos do microarranjo.

Contudo, do ponto de vista biológico o fenômeno é o mesmo, ambos os genes registram uma variação de 10 vezes do valor de *background*. É muito conveniente transformar os números para eliminar a desproporção enganosa entre essas duas mudanças relativas e transformação logarítmica atende esse objetivo. Por exemplo, usando a transformação logarítmica na base 10, os valores são transformados em:

$$\begin{aligned} \log_{10}(100) &= 2 \\ \log_{10}(1.000) &= 3 \\ \log_{10}(10.000) &= 4, \end{aligned}$$

tais valores refletem o fato de que o fenômeno que afeta dois genes é o mesmo e somente ocorre em direções opostas. Desta forma, os genes apresentam variação de:

$$2 - 3 = -1,$$

para um gene e:

$$4 - 3 = 1,$$

para o outro gene.

Desta forma os valores refletem o fato que dois genes mudam pela mesma magnitude, porém em direções opostas. Por este motivo é dito que a transformação logarítmica separa a variância da intensidade média.

O segundo argumento para a utilização de transformação logarítmica é para a distribuição de valores em curva. A transformação logarítmica deixa a distribuição simétrica e quase normal.

O terceiro argumento é a conveniência. Se o log é transformado na base 2, a análise e interpretação dos dados é facilitada. Por exemplo, selecionando genes com uma variação de 4 vezes pode ser feito cortando o histrograma no valor da razão $log_2(\text{razão}) = 2$. A partir daí, a base do logarítmo é assumida como sendo = 2 [Dra03].

2.4 Conclusão do capítulo

Neste capítulo foi apresentada a importância da utilização de técnicas de agrupamento de dados para a análise de dados de expressão gênica. Também foram descritos os conceitos básicos de biologia molecular. Conhecer estes conceitos é importante para o entendimento dos próximos capítulos.

O próximo capítulo descreve as etapas do processo de agrupamento de dados, incluindo o funcionamento dos algoritmos k-médias, SOM e SAMBA, de acordo com suas respectivas abordagens undimensional e bidimensional de agrupamento de dados. Também descreve as técnicas de validação estatística e biológica utilizadas.

Capítulo 3

Agrupamento de dados

Agrupamento de dados, do inglês *clustering*, é uma técnica de descoberta de conhecimento que identifica associações ou correlações entre objetos. Neste trabalho a técnica de agrupamento de dados convencional foi denominada de agrupamento unidimensional para distinguir do conceito de agrupamento em duas dimensões, ou bidimensional.

No processo de agrupamento não há informação *a priori* dos grupos ou classes que caracterizam os dados. É um tipo de análise não-supervisionada, pois, ao contrário da classificação, não há como comparar os resultados com modelos conhecidos para saber se o processo ocorreu de forma adequada ou não.

O objetivo do agrupamento é formar grupos de objetos com alta similaridade e baixa similaridade em relação a objetos de outros grupos. O conceito de similaridade está normalmente associado à distância entre os objetos. É identificada através de uma análise de todas as suas características e os objetos são comparados baseados na semelhança dessas características.

Problemas de agrupamento podem ser encontrados em áreas dos mais variados contextos: reconhecimento de padrões, análise espacial de dados, processamento de imagens, classificação de documentos, inferência filogenética, análise de dados de expressão gênica, entre outras.

Para a análise de dados de expressão gênica a tarefa de agrupamento é considerada um passo chave porque viabiliza a detecção de grupos de genes que exibem padrões de expressão similares (co-regulados), ou ainda, que demonstrem expressão diferencial.

Alguns trabalhos pioneiros de mineração de dados de expressão incluem a aplicação do algoritmo hierárquico [MBEB98], k-médias [STC99] e SOM [PTG99].

Em dezembro de 1998 foi publicado o primeiro artigo sobre o uso de técnicas de agrupamento para organização e facilitação da visualização dos dados de microarranjo . Foi utilizada como métrica o coeficiente de correlação de Pearson para comparar os

padrões de expressão gênica durante diferentes tipos de processos biológicos, principalmente exposições a situações extremas e processos de desenvolvimento, sendo que os genes foram agrupados com base nessa métrica através do algoritmo UPGMA. Os resultados obtidos tiveram papel importante para evidenciar o potencial da técnica de microarranjo [MBEB98].

Também em 1998, Tamayo e colaboradores desenvolveram o programa GeneCluster, que utiliza o algoritmo SOM. No entanto, consideraram o número de grupos como correpondente à dimensão da matriz, por exemplo, uma matrix definida com a dimensão 2x2 corresponde à formação de 4 grupos. Outra implementação do algoritmo SOM para agrupamento de dados de expressão foi desenvolvido por Toronen e colaboradores.

Embora pioneiras, as técnicas de agrupamento hierárquico, k-médias e SOM continuam sendo utilizadas com freqüência na mineração de dados de microarranjo. Outras alternativas envolvem *simulated annealing* [UAL99] e técnicas baseadas na teoria de grafos: HCS [HS00] e CAST [ABDY99]. Recentemente vem ocorrendo grandes avanços na melhoria das técnicas de agrupamento aplicadas em dados de expressão. Exemplos mais proeminentes incluem a técnica de agrupamento bidimensional [MO04] e *gene shaving* [THB00].

Neste capítulo são apresentadas as etapas do processo de análise de agrupamento, desde a preparação dos dados até a interpretação e validação dos resultados obtidos. São descritas as medidas de distâncias mais usuais, os algoritmos de agrupamento hierárquico, k-médias, SOM e SAMBA, de acordo com suas abordagens unidimensional e bidimensional. Também são descritas as técnicas de validação estatística e biológicas de enriquecimento funcional e identificação de fatores de transcrição.

3.1 Etapas do processo de agrupamento de dados

O processo de agrupamento envolve diversas etapas que vão desde o pré-processamento dos dados, até a interpretação dos grupos obtidos.

A Figura 3.1 apresenta as etapas do processo de agrupamento. Cada uma das etapas são descritas na seqüência. A seção 3.2 apresenta a etapa de pré-processamento dos dados, a seção 3.3 apresenta a etapa de definição de medida de similaridade (não representada na figura), a seção 3.4 a etapa de definição do algoritmo de agrupamento, a seção 3.5 a etapa de validação dos resultados de agrupamento, a seção 3.6 apresenta técnicas de validação estatística, a seção 3.7 apresenta técnicas de validação biológica e a seção 3.8 apresenta a etapa de interpretação dos resultados de agrupamento.

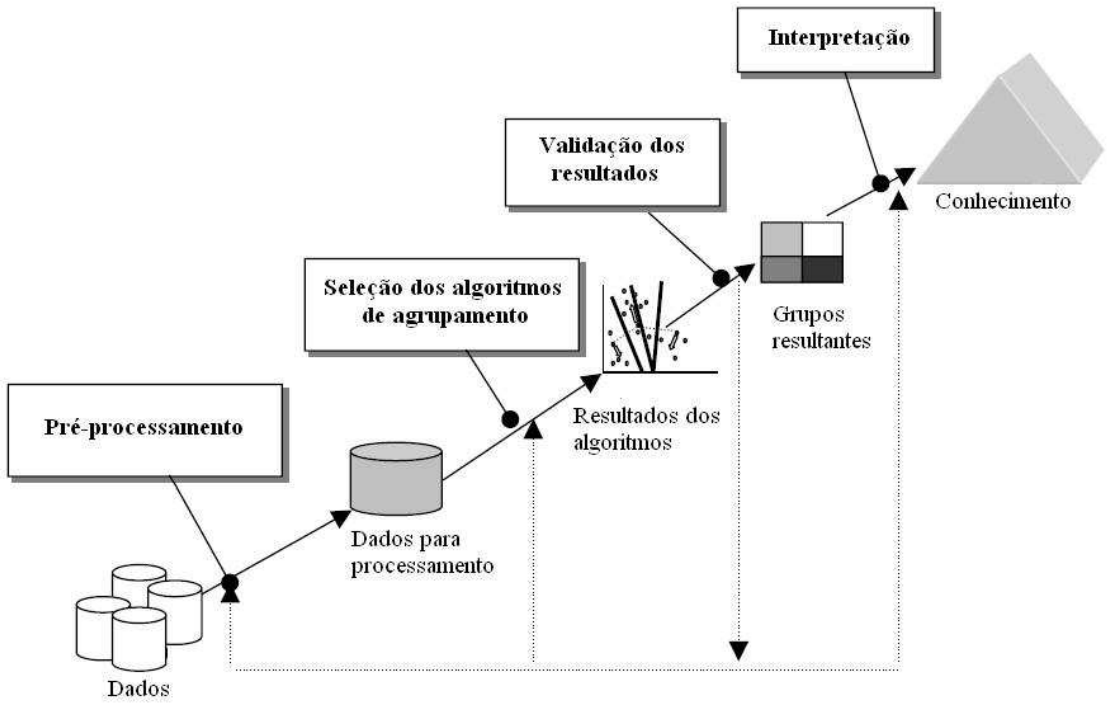


Figura 3.1: Etapas do processo de agrupamento de dados. Imagem adaptada de [MHV01].

3.2 Pré-processamento dos dados

Nesta fase, é comum recorrer às técnicas de extração e seleção de características. No caso de dados de expressão gênica, um recurso comumente utilizado são os filtros de dados, que são úteis para minimizar a quantidade de dados redundantes, ruidosos ou irrelevantes, que podem influenciar na tarefa de agrupamento.

O propósito da utilização das técnicas de extração e seleção de características é o melhoramento da aprendizagem indutiva, em termos de velocidade da aprendizagem, capacidade de generalização ou simplicidade da representação. Elas também facilitam o entendimento dos resultados obtidos diminuindo o volume de armazenamento, reduzindo o ruído gerado por características irrelevantes ou redundantes e eliminando o conhecimento inútil.

Seleção de características é uma solução computacional que é motivada por uma certa definição de relevância. Contudo, a relevância de uma característica pode ter várias definições, dependendo do objetivo do problema. No caso de dados de expressão gênica, a seleção pode identificar grupos de genes pouco informativos quando considerados individualmente, mas que são muito significativos quando considerados em conjunto.

Além disso, a seleção de características é geralmente utilizada como pré-processamento de dados da tarefa de classificação [LY05]. Entretanto, para a tarefa

de agrupamento é uma necessidade *ad-hoc*, podendo envolver um processo de tentativa e erro, onde vários subconjuntos de características são selecionados e as saídas avaliadas utilizando um índice matemático, por exemplo. [AKJF99].

3.3 Definição de medida de similaridade

A medida de similaridade ou proximidade é um passo fundamental para a definição de um agrupamento. A definição da medida a ser utilizada é um passo que ocorre entre as etapas de pré-processamento dos dados e definição do algoritmo de agrupamento.

A medida de similaridade calcula o quanto um elemento é similar a outro e, assim, ajuda a determinar se ambos devem estar contidos em um mesmo grupo ou não.

Na realidade, as medidas podem se referir à similaridade ou dissimilaridade. As medidas de dissimilaridade referem-se às distâncias, que correspondem às diferenças entre os valores de cada atributo dos elementos. Neste caso, considera-se que quanto menor for a distância entre um par de elementos maior é a similaridade entre eles e vice-versa. Nos dados de expressão gênica, elementos são usualmente genes e os atributos de cada gene correspondem à sua expressão nas diferentes condições experimentais.

Nesta seção são apresentadas as medidas de distância Euclidiana e o coeficiente de correlação de Person, que são os mais utilizados nos trabalhos de agrupamento de dados de expressão gênica. Em [Dra03] são apresentadas outras medidas de distância usuais neste contexto. Também é apresentada uma seção de observações que auxiliam na escolha da medida de distância e uma seção de comparação de diferentes medidas.

Maiores detalhes destas e outras medidas também podem ser obtidos em [BSEL01].

Distância Euclidiana

É simplesmente a distância geométrica dos objetos x e y em um espaço multidimensional, dada pela equação 3.1:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3.1}$$

em que:

$$\begin{array}{lcl} d(x,y) & = & \text{distância do vetor } x \text{ para o vetor } y; \\ x_i, y_i & = & \text{elemento da dimensão ou atributo } i \text{ dos vetores } x \text{ e } y; \\ n & = & \text{número total de dimensões ou atributos;} \end{array}$$

Quando utilizada a distância Euclidiana, é necessário que os dados tenham sido normalizados adequadamente. Diferente das funções de distância baseadas em correlação, a distância Euclidiana é sensível à magnitude das diferenças dos valores dos dados (ver exemplo da Figura 2.4 do Capítulo 2), enquanto o coeficiente de correlação não apresenta este comportamento. É apropriada para conjuntos de dados que possuem grupos compactos ou isolados [AKJF99].

Coeficiente de correlação de Pearson

O coeficiente de correlação é freqüentemente descrito como uma medida da forma, no sentido de que é insensível a diferenças na magnitude dos atributos, conforme a fórmula 3.2:

$$S_{ij} = \frac{\sum_{k=1}^{d}(x_{ik} - \overline{x_i})(x_{jk} - \overline{x_j})}{(\sum_{k=1}^{d}(x_{ik} - \overline{x_i})^2 \sum_{l=1}^{d}(x_{jl} - \overline{x_j})^2)^{1/2}}, \tag{3.2}$$

em que $\overline{x_i} = \sum_{k=1}^{d} \frac{x_{ik}}{d}$. Os valores dessa medida estão no intervalo [-1,1], sendo:

- $S_{ij} = 1$, significa uma correlação perfeita positiva entre as duas variáveis.
- $S_{ij} = -1$, significa uma correlação negativa perfeita entre as duas variáveis, ou seja, se uma aumenta, a outra sempre diminui.
- $S_{ij} = 0$, significa que as duas variáveis não dependem linearmente uma da outra. No entanto, pode existir uma outra dependência que seja "não linear". Assim, o resultado $S_{ij} = 0$ deve ser investigado por outros meios.

O coeficiente de correlação é sensível à *outliers* (pontos distantes do padrão) e é menos intuitivo do que a distância Euclidiana, por exemplo. Deve ser utilizado quando se pretende analisar a semelhança no comportamento em detrimento da semelhança nos valores.

A distância Euclidiana é mais apropriada para dados de razão logarítmica e o coeficiente de correlação de Pearson é mais indicado para valores absolutos [D'H05].

3.4 Definição do algoritmo de agrupamento

Esta etapa consiste da aplicação de um algoritmo de agrupamento apropriado para agrupar dados de acordo com um objetivo específico. Existem inúmeros algoritmos de agrupamento que podem ser aplicados nesta etapa. Alguns deles são apresentados a seguir de acordo com as abordagens unidimensional e bidimensional.

A abordagem de agrupamento unidimensional considera instâncias (genes) e atributos(condições) da matriz de dados separadamente, enquanto a abordagem de agrupamento bidimensional possibilita o agrupamento das duas dimensões simultaneamente.

3.4.1 Abordagem de agrupamento unidimensional

Embora existam diferentes classificações, as técnicas de agrupamento unidimensional podem ser divididas basicamente em duas categorias: hierárquicas e não-hierárquicas.

Jain e colaboradores classificam os algoritmos de agrupamento nas categorias hierárquica e particional [AKJF99]. Aldenderfer e Blashfield classificam de forma mais detalhada: hierárquicos aglomerativos, hierárquicos divisivos, particionamento iterativo, busca em profundidade, fator-analítico, amontoamento e baseados na teoria de grafos [AB84].

A seguir, são apresentadas as categorias de agrupamento baseados na taxonomia de Jain. Os algoritmos apresentados são o hierárquico, k-médias e SOM, comumente utilizados para o agrupamento de dados de expressão gênica. Neste trabalho foram utilizados o k-médias e SOM.

Técnicas de agrupamento hierárquico

As técnicas hierárquicas podem ser classificadas de acordo com a forma com que a decomposição hierárquica é realizada: *bottom-up* ou *top-down*, e são conhecidas como:

1. Agrupamento hierárquico aglomerativo: nesta estratégia *bottom-up* a idéia é juntar os objetos em grupos cada vez maiores, incluindo não só novos objetos, mas também os grupos já formados. Novos objetos são adicionados ao grupo utilizando métodos de determinação de distância entre grupos (*linkage metrics*). O processo continua até que todos os objetos sejam agrupados, conforme mostrado na Figura 3.2. As técnicas hierárquicas, em sua maioria, pertencem a esta categoria, diferindo entre si apenas pela definição da semelhança entre os grupos.

2. Agrupamento hierárquico divisivo: esta estratégia *top-down* trabalha de forma oposta ao agrupamento aglomerativo. Começa com um único grupo, conforme mostrado na Figura 3.2, e subdividide os grupos em grupos menores até que cada objeto forme o próprio grupo ou até que o algoritmo satisfaça determinadas condições de conclusão, tais como a obtenção de um número desejado de grupos ou, ainda, que a distância entre dois grupos próximos ultrapasse um determinado limiar.

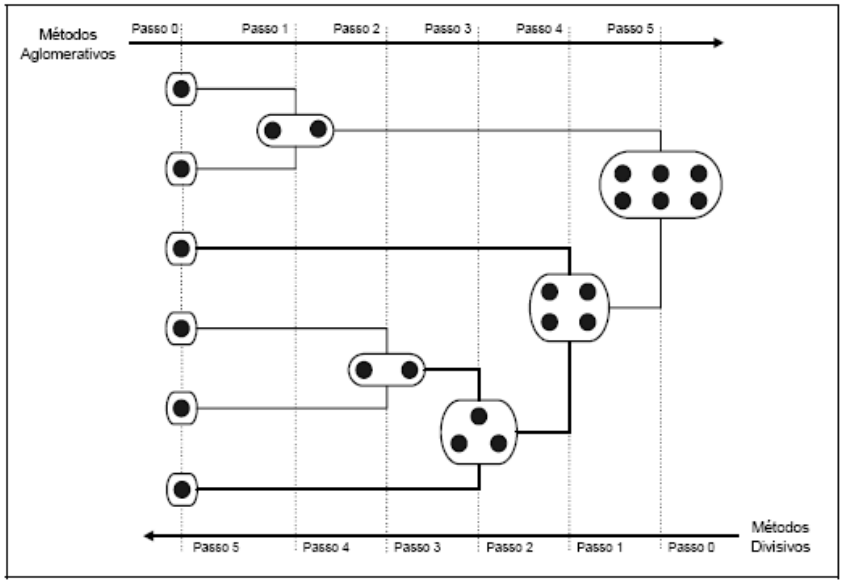


Figura 3.2: Esquema dos métodos de agrupamento aglomerativo e divisivo [And04].

Os métodos de determinação de distância entre grupos correspondem a medidas do grau de similaridade entre dois grupos. Os principais são:

(a) Ligação simples (*single linkage*) : a distância entre dois grupos é definida como a menor distância entre os objetos dos dois grupos, conforme a fórmula 3.3:

$$d(c_1, c_2) = \min \left\{ \begin{array}{l} d(x, y) \\ x \in c_1 \\ y \in c_2 \end{array} \right\} \tag{3.3}$$

(b) Ligação média (*average linkage*) : a distância entre dois grupos é definida como a distância média entre os objetos dos dois grupos, conforme a fórmula 3.4:

$$d(c_1, c_2) = \frac{1}{n_1 n_2} \sum_{x \in c_1, y \in c_2} d(x, y) \tag{3.4}$$

(c) Ligação completa (*complete linkage ou farthest neighbor*) : a distância entre dois grupos é definida como a maior distância entre os objetos dos dois grupos, conforme a fórmula 3.5:

$$d(c_1, c_2) = \max \left\{ \begin{array}{l} d(x, y) \\ x \in c_1 \\ y \in c_2 \end{array} \right\} \tag{3.5}$$

(d) Ligação centróide (*centroid linkage*): a distância entre dois grupos é definida como a distância entre o centro dos dois grupos. , conforme a fórmula 3.6:

$$d(c_1, c_2) = d(vc_1, vc_2), \tag{3.6}$$

onde:

$$vc_1 = \frac{1}{n_1} \sum_{x \in c_1} x, vc_2 \frac{1}{n_2} \sum_{y \in c_2} y \tag{3.7}$$

(e) Variância mínima (*minimum variance clustering ou Ward's method of sum-of-squares method*)

O método de variância mínima é uma variação dos anteriores, que busca otimizar a mínima variância entre os agrupamentos, juntando os elementos cuja soma dos quadrados entre eles seja mínima ou que o erro desta soma (denominada ESS (*Error Sum of Squares*)) seja mínimo.

(f) Agrupamento pareado igualmente ponderado (*unweighted pair-group method, UPGM*)

Neste método, a distância entre dois grupos é calculada como a distância média entre todos os pares dos objetos nos dois grupos diferentes. Este método é eficiente quando os objetos naturalmente assumem a forma de grupos distintos, entretanto, ele executa igualmente bem com a forma alongada.

(g) Agrupamento pareado proporcionalmente ponderado (*weighted pair-group method, WPGM*).

Este método é idêntico ao UPGM, exceto pelo fato de que o número dos objetos contidos nos grupos seja usado como peso. Assim, este método (melhor que o método anterior) deve ser usado quando existir a suspeita de que o tamanho dos grupos serão extremamente desiguais.

A aplicação de diferentes métodos de determinação de distância entre grupos pode resultar em agrupamentos bem diferentes.

O resultado do agrupamento hierárquico é tipicamente representado na forma de um dendograma, conforme Figura 3.3:

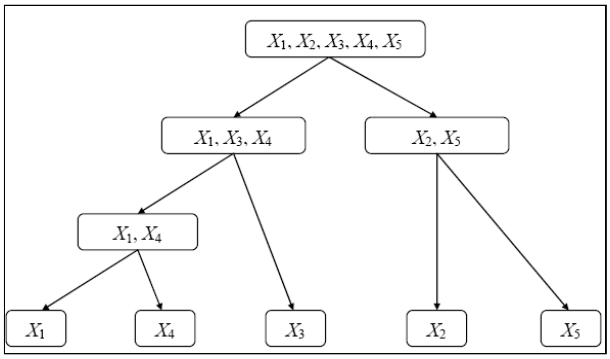


Figura 3.3: Exemplo de um dendograma do agrupamento hierárquico.

O dendograma é um diagrama em forma de árvore, muito utilizada na área de Filogenia. A forma geral é apresentada com uma raiz no topo, com uma escala de distâncias indicando em que nível dois grupos tornam-se um só.

As vantagens do agrupamento hierárquico são sua simplicidade e o fato de produzir resultados mais fáceis de serem visualizados e interpretados do que os gerados por outros algoritmos, como o k-médias, por exemplo.

Berkhin aponta como vantagens dos algoritmos de agrupamento hierárquico a facilidade em lidar com qualquer medida de similaridade utilizada e a sua conseqüente aplicabilidade a qualquer tipo de atributo (numérico ou categórico). A desvantagem corresponde à sua inabilidade em executar ajustes após uma fusão ou divisão terem sido executadas (*backtracking*). Este aspecto está relacionado ao fato dos algoritmos hierárquicos serem apenas algoritmos construtivos, não permitindo o refinamento de soluções obtidas durante a sua execução, o que normalmente fornece um caráter guloso à técnica hierárquica tradicional [Ber02].

Outra desvantagem dos algoritmos hierárquicos está relacionada à imprecisão do critério de parada. À medida que os grupos crescem, o vetor que representa o grupo pode não representar mais nenhum dos elementos do grupo. No caso de dados de expressão gênica, pode tornar os padrões de expressão dos genes menos relevantes. Algumas técnicas híbridas vêm sendo criadas para tentar descobrir o momento certo para o algoritmo parar de juntar elementos.

A comparação das principais características dos algoritmos de agrupamento hierárquico pode ser consultada em [MHV01]. Neste material é apresentada uma tabela com os algoritmos hierárquicos mais comuns, o tipo de dados adequado, a complexidade do algoritmo, a geometria dos dados, o tratamento de ruídos (*outliers*), os parâmetros de entrada necessários e o critério de agrupamento.

Técnicas de agrupamento particional

As técnicas de particionamento criam agrupamentos através de diversas iterações. Segundo Aldenderfer e Blashfield, a maior vantagem dessas técnicas está nas diversas passadas (iterações) no conjunto de dados, podendo corrigir eventuais problemas de alocação inadequada (muito comum nos algoritmos hierárquicos) [AB84]. Porém a desvantagem é que elas são mais demoradas do que os hierárquicos. O maior problema desses algoritmos está no fato de o usuário ter que especificar o número de agrupamentos desejado. Isto porque não há maneira de prever se o número de agrupamentos é o mais adequado para determinado conjunto de objetos.

Uma das maneiras sugeridas para se chegar a um número de partições mais adequado é executar o método diversas vezes, testando diferentes configurações possíveis de agrupamentos [AB84]. Outra alternativa sugerida, consiste em aplicar algoritmos hierárquicos, a fim de se identificar a melhor quantidade de agrupamentos e, em seguida, utilizar essa quantidade como parâmetro inicial para um algoritmo de particionamento. Everitt [BSEL01] também cita outros critérios que podem auxiliar na identificação da melhor quantidade de agrupamentos para cada conjunto de dados.

As técnicas de agrupamento hierárquico são interessantes quando se deseja analisar relações de abrangência ou especificidade entre os objetos, mas os usuários devem percorrer toda a estrutura para compreender as inúmeras relações entre eles. Já os grupos de métodos particionais podem ser visualizados e compreendidos mais facilmente, desde que o número total de agrupamentos não seja muito grande.

k-médias

O algoritmo mais conhecido dessa categoria é o k-médias (*k-means*). O usuário indica o número de grupos desejado e o algoritmo de agrupamento cria (de forma aleatória ou por outro processo) um conjunto inicial de grupos. A seguir, o centróide de cada um desses grupos é calculado e o algoritmo analisa a distância ou similaridade desses centróides com todos os elementos a serem agrupados. Em seguida cada objeto é alocado ao grupo cujo centróide esteja mais próximo e, ao ser incluído, este centróide é recalculado para representar esse novo objeto. O processo é repetido até que os centróides não mudem mais de posição [AKJF99].

O algoritmo é sensível à escolha inicial dos centróides e da sua forma de atualização. Dependendo da escolha dos centróides o algoritmo pode convergir para um ótimo local [FACFLFC05].

A Figura 3.4 ilustra os passos do processamento do algoritmo k-médias. Inicial-

mente, são escolhidos 3 objetos para representar o centróide de cada um dos 3 grupos que deverão ser formados (os centros dos grupos estão marcados com o símbolo "+" em todas as figuras) e os objetos remanescentes são agrupados de acordo com suas distâncias em relação aos centróides, como mostra a Figura 3.4 (a). Após a realocação dos objetos, formando nova composição de grupos, os novos centróides são calculados, iterativamente, conforme mostrado na Figura 3.4 (b). Quando não há mais possibilidade de realocar qualquer objeto, o processo é finalizado, conforme visto na Figura 3.4 (c) [And04].

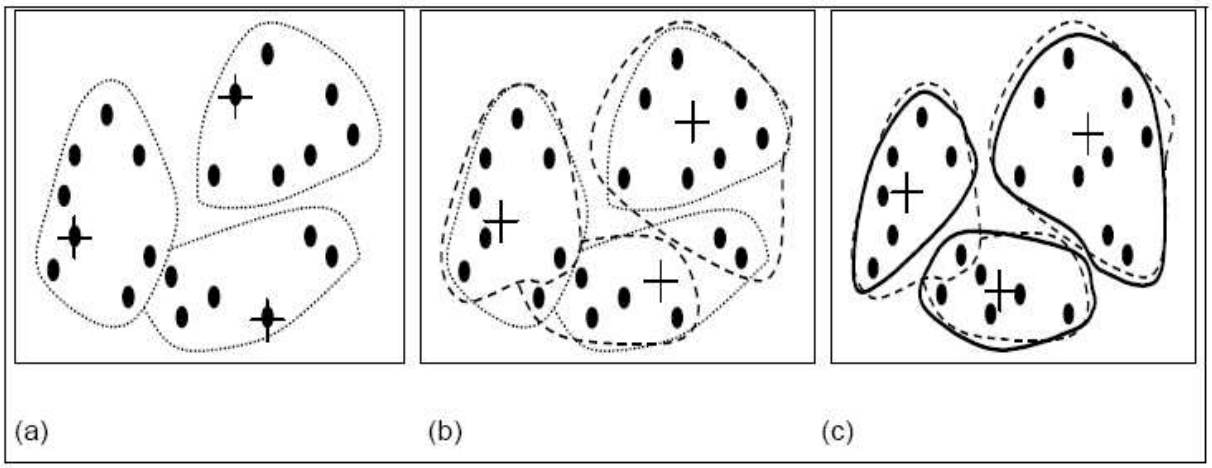


Figura 3.4: Funcionamento do algoritmo k-médias [And04].

A maior vantagem do k-médias está no fato dele fazer diversas passadas (iterações) no conjunto de dados, podendo corrigir eventuais problemas de alocação inadequada (comum nos algoritmos hierárquicos). A maior desvantagem está na necessidade da definição do número de grupos desejado. Isto porque não há uma maneira de prever se o número de grupos é adequado para determinado conjunto de objetos. Uma maneira adequada de descobrir a quantidade de grupos mais apropriada é executando o algoritmo várias vezes [AB84].

Existem vários critérios para a determinação do número de grupos e quase todos funcionam da seguinte maneira: realizar o agrupamento dos dados considerando 2 grupos e calcular o valor de uma função proposta que tenha o número de grupos como um de seus parâmetros, realizar o agrupamento dos dados considerando 3 grupos e calcular o valor da mesma função, repetir este procedimento até atingir um número máximo de grupos estabelecido. O agrupamento que ocasionar o valor máximo (ou, em alguns casos, mínimo) da função, deve ser considerado como o melhor agrupamento possível para a base de dados [And04].

SOM (Self Organizing Maps)

SOM [Koh01] é o algoritmo mais popular de rede neural artificial, baseado em aprendizado competitivo e não supervisionado, freqüentemente utilizado para tarefas de agrupamento e vizualização.

Nesse tipo de rede, os neurônios são organizados em um arranjo unidimensional ou bidimensional. Cada neurônio no arranjo está conectado a todas as entradas da rede. Esta rede geralmente utiliza uma única camada computacional. A cada padrão de entrada apresentado à rede, os neurônios computam seus valores de ativação, ativando uma região diferente do arranjo. Para cada padrão de entrada, os neurônios de saída da rede competem entre si para serem ativados. O neurônio com maior valor de ativação é o vencedor da competição. Em seguida, é determinada a localização espacial de uma vizinhança topológica de neurônios excitados, centrada no neurônio vencedor.

O próximo passo consiste de uma adaptação dos pesos. Os ajustes dos pesos são tais que a resposta do neurônio vencedor à aplicação subseqüente de um padrão de entrada similar é melhorada. O algoritmo procura fazer com que os neurônios vizinhos no arranjo apresentem vetores de pesos que retratem as relações de vizinhança entre os dados. Assim, durante a execução do algoritmo, os vetores de entrada direcionam o movimento dos vetores de peso, promovendo uma organização topológica dos neurônios da rede. Ainda durante o treinamento, a região de vizinhança dos neurônios é gradativamente reduzida.

A primeira aplicação para a análise de dados de expressão gênica foi feita por Tamayo e colaboradores no ano de 1999 [PTG99]. A aplicação comum na comunidade de bioinformática é que cada unidade do mapa gerado pela rede SOM é considerada como um grupo separado, sendo que na realidade, várias unidades vizinhas podem modelar um único grupo [FACFLFC05].

A comparação das principais características dos algoritmos de agrupamento particional pode ser consultada em [MHV01]. Neste material é apresentada uma tabela com os algoritmos particionais mais comuns, o tipo de dados adequado, a complexidade do algoritmo, a geometria dos dados, o tratamento de ruídos (*outliers*), os parâmetros de entrada necessários e o critério de agrupamento.

3.4.2 Abordagem de agrupamento bidimensional

Um grande número de abordagens de agrupamento unidimensional de dados tem sido proposto para a análise de dados de expressão gênica. No entanto, os resultados

dessas aplicações são limitados. Essa limitação é devido à incapacidade desses algoritmos de identificar padrões de expressão gênica viáveis somente num subconjunto de condições experimentais. Por esta razão, são propostos algoritmos que agrupam genes e condições simultaneamente, onde os genes demonstram atividades altamente correlacionadas para todas as condições experimentais selecionadas.

O conceito de agrupamento bidimensional, ou, do inglês, biclustering, foi introduzido por [Har72]. É também conhecido como agrupamento simultâneo (*simultaneous clustering*), agrupamento em blocos (*block clustering*) e co-agrupamento (*co-clustering*) [Ber02], sendo uma abordagem utilizada há muito tempo na estatística. Cheng e Church foram os primeiros a aplicar a técnica de agrupamento bidimensional em dados de expressão gênica [CC00]. Algumas implementações do agrupamento bidimensional podem ser vistas em CTWC (*Coupled Two-Way Clustering*) [GGD00], *Spectral biclustering of microarray data* [YKG03], SAMBA (*Statistical Algorithmic Method for Biclustering Analysis*) [SMKT+05] e GEMS (Gene Expression Mining Server) [WK05].

Diferença entre agrupamento unidimensional e bidimensional

Agrupamento unidimensional pode ser aplicado a instâncias e a atributos da matriz de dados, separadamente. Agrupamento bidimensional, no entanto, possibilita o agrupamento das duas dimensões simultaneamente. Isto significa que o agrupamento unidimensional obtém um modelo global enquanto que o agrupamento bidimensional resulta em um modelo local [MO04]. Quando algoritmos de agrupamento unidimensional são utilizados, cada grupo de genes é definido considerando todas as condições. Da mesma forma, cada grupo de condições é caracterizado pela atividade de todos os genes.

A diferença entre agrupamento unidimensional e bidimensional é ilustrada na Figura 3.5.

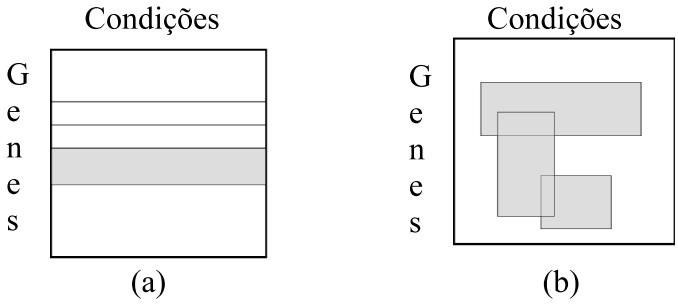


Figura 3.5: (a) Grupo unidimensional: grupo de genes considerando todas as condições. (b) Grupos bidimensionais: subgrupos de genes e subgrupos de condições.

Diferente da abordagem de agrupamento unidimensional, o agrupamento bidimensional pode identificar grupos com as seguintes restrições:

1. Um agrupamento de genes pode ser definido com somente um subconjunto de condições;

2. Um agrupamento de condições pode ser definido com somente um subconjunto de genes;

3. Um agrupamento pode não ser exclusivo e/ou exaustivo: Um gene/condição pode pertencer a mais de um grupo ou não pertencer a nenhum grupo [MO04].

Algoritmo SAMBA de agrupamento bidimensional

O algoritmo SAMBA utiliza modelagem probabilística dos dados e técnica de teoria de grafos para identificar subconjuntos de genes que respondem coordenadamente a um subconjunto de condições.

O algoritmo SAMBA inicia com a formação do grafo bipartido a parti da matriz de dados. Um grafo é dito bipartido quando seu conjunto de vértices V puder ser particionado em dois subconjuntos V_1 e V_2, tais que toda aresta de G une um vértice de V_1 a outro de V_2.

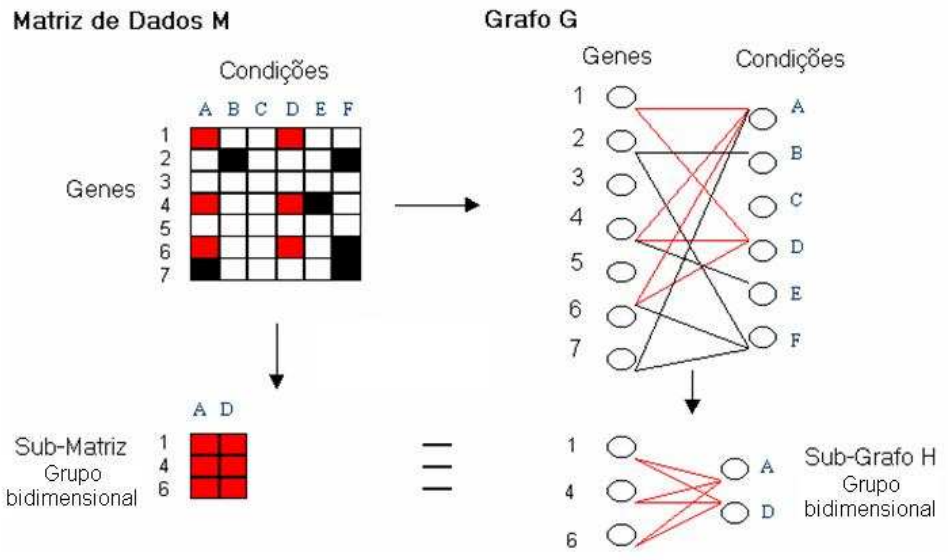


Figura 3.6: Algoritmo SAMBA: da matriz de dados ao grafo bipartido.

A Figura 3.6 ilustra um grafo bipartido G formado por uma coluna de vértices que

corresponde aos genes e outra coluna que corresponde às condições. Este grafo corresponde à estrutura da matriz de dados M.

As arestas indicam em quais condições o gene se expressa. No exemplo da figura, o gene 1 se expressa nas condições A e D, o gene 2 nas condições B e F, o gene 3 em nenhuma condição, o gene 4 nas condições A, D e E, e assim sucessivamente com os demais genes do conjunto de dados.

Mais precisamente, a expressão da matriz é transformada em um grafo bipartido $G = (U, V, E)$, onde U corresponde às condições, V aos genes e $(u, v) \in E$ se v responde à condição u. O peso de cada subgrafo $H = (U', V', E')$ é a soma dos pesos dos pares de arestas (gene-condição).

O peso de uma aresta (u, v) é o $log\frac{p_c}{p_{u,v}}$, onde $p_{u,v}$ é a fração do grafo bipartido idêntica a G que contém a aresta (u, v), e p_c resulta de um modelo alternativo que assume que cada aresta em um grupo verdadeiro ocorre com uma probabilidade constante. Da mesma forma, o peso de cada não-aresta (u, v) corresponde ao $log\frac{1-p_c}{1-p_{u,v}}$, e o peso de E é dado pela fórmula 3.8:

$$\sum_{(u,v)\in E'} log\frac{p_c}{p_{u,v}} + \sum_{(u,v)\in (U'V')} log\frac{1-p_c}{1-p_{u,v}} \tag{3.8}$$

Sob este esquema de escore, o peso de um subgrafo é a razão log-probabilidade de um grupo, só então procura-se identificar os k subgrafos de maior peso de G. Este problema é *NP-hard*, SAMBA emprega uma busca heurística para cada subgrafo. Mais informações do algoritmo SAMBA e outros algoritmos bidimensionais podem ser consultadas em [Ami04, ATS04, HLTH05].

3.4.3 Critério básico para a seleção de técnicas de agrupamento

Existem inúmeras técnicas de agrupamento que podem ser utilizadas para a análise de dados de expressão gênica, mas escolher "a melhor" delas para um problema específico pode representar um desafio. Vantagens e limitações podem depender de fatores como a natureza estatística dos dados, procedimentos de pré-processamento, número de características, etc. Além disso, não é incomum observar resultados inconsistentes quando diferentes algoritmos de agrupamento são utilizados num mesmo conjunto de dados. Para fazer uma escolha mais apropriada é importante entender bem o domínio do problema e as opções de agrupamento disponíveis.

Vários algoritmos indiretamente assumem que a estrutura dos dados exibe características particulares. Por exemplo, o algoritmo k-médias assume que a forma dos grupos

é esférica e o agrupamento hierárquico de ligação simples assume que os grupos são bem separados. Infelizmente, este não há tipo de conhecimento nos dados de expressão gênica [D'H05].

A necessidade específica do usuário também pode influenciar na decisão da seleção do algoritmo de agrupamento. Por exemplo, um usuário pode ter interesse em observar relações diretas dos grupos. Neste caso a abordagem de agrupamento hierárquico pode representar uma solução básica. Mas em alguns estudos pode ser difícil de visualizar o resultado do agrupamento hierárquico por causa da quantidade de genes e condições envolvidos.

Em geral, a aplicação de duas ou mais técnicas de agrupamento pode oferecer uma base para melhorar a confiança dos resultados. Um usuário pode ter maior segurança num agrupamento se muitos resultados similares forem obtidos de diferentes técnicas [AB].

3.5 Validação dos resultados de agrupamento

Após o processo de análise de agrupamentos inicia-se uma tarefa importante, que é a validação dos grupos resultantes. Pelo fato de a tarefa de agrupamento ser de origem exploratória, isto é, devido ao seu objetivo ser o de conhecer as características de um conjunto de dados, torna-se difícil avaliar se determinado resultado é correto ou não. Isto porque, para avaliar se um determinado resultado está correto ou não, deve-se compará-lo com um resultado conhecido. Por esta razão a análise de agrupamento é altamente delicada podendo ser subjetiva.

Os resultados dos agrupamentos são geralmente aceitos sem objeções ou suposições de resultados contraditórios, tornando o trabalho totalmente contra-produtivo, uma vez que a idéia de aprendizagem não-supervisionada seja justamente surpreender com a identificação de padrões inesperados para a geração de novas hipóteses.

A maneira mais indicada para a validação de agrupamentos é considerar que uma estrutura de agrupamento é válida se não ocorreu por acaso, ou se é "rara" em algum sentido, já que qualquer algoritmo de agrupamento encontrará grupos, independente se existe ou não similaridade nos dados. Entretanto, se essa similaridade existe, alguns algoritmos podem encontrar grupos mais adequados que outros [FACFLFC05].

Não há um método de validação de agrupamento universal [SS01b], mas o uso de técnicas de validação aplicadas em diferentes etapas do processo de análise de agrupamento ajuda a melhorar a qualidade dos resultados e aumentar a confiança do resultado final [JHK05]. Um estudo sobre validação de agrupamentos pode ser encontrado em [MHV01].

Muitas técnicas de validação de agrupamento têm sido propostas, embora pouca

atenção tenha sido destinada às aplicações da biologia [JHK05]. São poucos os trabalhos de agrupamento de dados de expressão gênica que aplicaram técnicas de validação biológica. Estas técnicas ainda não estão definidas, poucas ferramentas estão disponíveis e, além disso, buscar coerência biológica em grupos de dados não é uma tarefa trivial. Os autores geralmente utilizam os índices de homogeneidade e separação como métricas para avaliar a qualidade dos grupos.

A proposta deste trabalho consiste na aplicação das técnicas de validação estatística e biológica nos resultados dos agrupamentos, portanto, na seqüência são descritas ambas as técnicas de validação. A validação estatística está dividida nas abordagens de agrupamento unidimensional e bidimensional.

3.6 Validação estatística

A validação dos resultados de um agrupamento, em geral, é feita com base em índices estatísticos, que julgam, de uma maneira quantitativa, o mérito das estruturas encontradas. Um índice quantifica alguma informação a respeito da qualidade de um agrupamento. A maneira pela qual um índice é aplicado é dada pelo critério de validação. Assim, um critério de validação expressa a estratégia utilizada para validar uma estrutura de agrupamento, enquanto um índice é uma estatística pela qual a validade é testada.

Os resultados de agrupamento podem ser avaliados baseados em critérios externos, internos e relativos.

- Critérios internos: Segundo Halkidi e colaboradores, as técnicas de validação baseadas em critérios internos utilizam os próprios dados para realizar a validação dos resultados [MHV01]. Nos casos em que as classes são desconhecidas ou existir a possibilidade de mais de uma classe como resposta, os critérios de avaliação internos são mais apropriados. Técnicas de validação interna não utilizam o conhecimento das classes, mas a informação dos próprios grupos. Baseiam-se nas várias propriedades estatísticas dos grupos.

- Critérios externos: as técnicas de validação baseadas em critérios externos avaliam os grupos baseados no conhecimento prévio das classes dos grupos. Essa estrutura geralmente reflete alguma intuição sobre a estrutura de agrupamentos do conjunto de dados e deve ser criada por algum especialista.

- Critérios relativos: neste caso, os agrupamentos são avaliados e validados comparando a estrutura dos agrupamentos resultantes com outras estruturas geradas pelo

mesmo algoritmo mas executado com diferentes parâmetros de entrada. O objetivo, neste caso é identificar os melhores valores de parâmetro de entrada para o conjunto de dados.

Nos programas de agrupamento, cuja solução não é conhecida, usualmente avalia-se a qualidade dos grupos através dos índices de homogeneidade e separação. A seguir são descritas as técnicas de validação estatística, utilizadas neste trabalho.

3.6.1 Homogeneidade

A homogeneidade corresponde à minimização das distâncias intragrupos. Nos dados de expressão gênica, a homogeneidade é avaliada pela média do vetor de expressão dos genes do mesmo grupo. Mais precisamente, se $cl(u)$ é um grupo de u, $F(X)$ e $F(u)$ correspondem à identidade do grupo X e do elemento u, respectivamente. S é a função de similaridade, dada pela fórmula 3.9:

$$H_{Ave} = \frac{1}{N} \sum_{u \in N} S(F(u), F(cl(u)) \tag{3.9}$$

3.6.2 Separação

A separação corresponde à maximização das distâncias intergrupos. É avaliada pela média de similaridade entre os vetores de expressão dos grupos, dada pela fórmula 3.10:

$$Sep_{Ave} \frac{1}{\sum_{i \neq j} |X_i||X_j|} \sum_{i \neq j} |X_i||X_j| S(F(X_i), F(X_j)) \tag{3.10}$$

As duas medidas são inerentemente conflitantes, pois a melhora de uma corresponde à piora da outra. A solução melhora se H_{Ave} aumenta e se Sep_{Ave} diminui. Para os cálculos acima, os elementos sozinhos (*singletons*) são considerados como um grupo [SS01b].

3.6.3 Índice C

O índice C é a medida de quão próximos estão os itens de um grupo. Ele é definido pela fórmula 3.11:

$$C = \frac{S - S_{min}}{S_{max} - S_{min}}, \tag{3.11}$$

onde S é a soma das distâncias de todos os pares de genes de um mesmo grupo (sobre todos os grupos). Dado l sendo um número desses pares, então S_{min} é a soma das l menores distâncias entre todos os pares de genes e S_{max} é a soma das l maiores distâncias. É fácil perceber que o numerador da fórmula acima será menor para pares de genes com a menor distância. Portanto, um bom agrupamento é indicado pelo menor valor de C [HS76].

O índice C tem se mostrado eficiente para estimar a qualidade de diferentes tipos de aplicações de agrupamento. No entanto, é importante ressaltar que um grupo com baixo valor de C não significa ser um grupo significante biologicamente e que dependendo da medida de distância podem fornecer diferentes resultados [NBC05].

3.6.4 Índice Dunn

Este índice é baseado na idéia da identificação de grupos compactos e bem separados [Dun74]. Para qualquer partição de grupos onde c_i representa o grupo i de cada partição, o índice Dunn, D, é calculado com a fórmula 3.12:

$$D = min_{1\leq i\leq n}\{min_{1\leq j\leq n\ i\neq j}\{\frac{d(c_i,c_j)}{max_{1\leq k\leq n}\{d'(c_k)\}}\}\}, \tag{3.12}$$

onde $d(c_i, c_j)$ é a distância entre os grupos c_i e c_j (distância intergrupo); $d'(ck)$ é a distância intragrupo do grupo c_k e n é o número de grupos. O principal objetivo da medida é maximizar as distâncias intergrupo e minimizar as distâncias intragrupos. Portanto, o número do grupo que maximiza D é considerado o número ótimo de grupos.

3.6.5 Índice Davies-Bouldin

O índice Davies-Bouldin é baseado na idéia da identificação de grupos compactos e bem separados, assim como o índice Dunn. O índice Davies-Bouldin é uma função da razão da soma intragrupo para a separação entre grupos [DB79]. De acordo com este índice, o melhor agrupamento minimiza a equação 3.13:

$$DB = \frac{1}{n}\sum_{i-1}^{n} max_{i\neq j}\{\frac{S_n(Q_i)+S_n(Q_j)}{S(Q_i,Q_j)}\}, \tag{3.13}$$

onde n é o número de grupos, S_n é a distância entre os objetos dos grupos i e j e $S(Q_i, Q_j)$ é a distância entre os centróides dos respectivos grupos. Assim, a razão é pequena se os grupos são compactos e distantes dos outros grupos. Conseqüentemente, o melhor agrupamento será indicado pelo índice Davies-Bouldin com o menor valor.

3.6.6 Silhueta

A técnica Silhueta calcula a largura da silhueta de cada objeto, a largura da silhueta média para cada grupo é o total da largura da silhueta média de todo o agrupamento [Rou87]. Usando esta abordagem cada grupo pode ser representado por uma silhueta, que é baseada na comparação da sua compactação e separação. O conceito desta técnica foi desenvolvido para determinar o número correto de grupos considerando-se que diversos agrupamentos diferentes tenham sido obtidos.

Basicamente, considerando um objeto i pertencente ao grupo A. Então a dissimilaridade média de i em relação a todos os outros objetos de A pode ser denotada por $a(i)$. Considerando um grupo diferente C. É então calculada a dissimilaridade média de i em relação a todos os objetos de C, que será denotada por $d(i, C)$. Após calcular $d(i, C)$ para todos os grupos $C \neq A$, seleciona-se a menor delas, aqui denotada por $b(i)$, conforme a fórmula 3.14:

$$b(i) = \min d(i, C), C \neq A \tag{3.14}$$

Este número representa a dissimilaridade de i em relação ao seu grupo vizinho [Hru01]. Assim define-se a silhueta $s(i)$ como:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{3.15}$$

Ela é seguida da fórmula $(-1 \leq s(i) \leq 1)$. Se o valor da silhueta é próximo a 1, significa que os objetos foram bem agrupados. Se o valor da silhueta é 0, significa que um outro grupo próximo deve ser a melhor opção. A função objetivo representa a média de $s(i)$ para $i = 1, 2, ..., N$. Neste caso, o melhor número de grupos ocorre quando o valor de $s(i)$ médio for máximo.

3.6.7 Índice Isolamento

Esta técnica é baseada na afirmação de que a vizinhança num espaço de características ocorre naturalmente num mesmo grupo [PF99]. O isolamento de cada grupo é medido usando a regra do k-vizinho mais próximo (KNN), onde a regra a de cada exemplo a é definida como a proporção de seus k vizinhos mais próximos que foram atribuídos ao mesmo grupo que a. Calculando a média sobre todos os n exemplos nos dados, a homogeneidade de um grupo pode ser calculada com a fórmula 3.16:

$$l_k = \frac{1}{n}\sum_{i=1}^{n} v_k(x_i) \tag{3.16}$$

O maior valor para esta medida indica grupos mais separados. Os autores reconhecem que, quando este índice recompensa os grupos com regiões de exemplos bem conectados para o mesmo grupo, ela não penalisa os grupos onde os conjuntos bem-separados são unidos, desde que somente um local limitado seja considerado para cada ponto.

3.6.8 Validação estatística para a abordagem de agrupamento bidimensional

Estes indices de validação estatística são comumente aplicados nos resultados dos agrupamentos unidimensionais. Para os resultados do agrupamento bidimensional ainda são poucos os recursos de validação. Tanay e colaboradores desenvolveram um método de validação estatística específico para o algoritmo bidimensional SAMBA. A idéia é aribuir pesos à arestas do grafo bipartido e extrair somente as conexões de maior peso, que correspondem aos grupos mais significativos. As arestas ligam duas colunas de vértices, uma que corresponde aos genes e a outra às condições. Estes pesos dependem do nível da expressão dos genes em cada uma das condições [ATS02].

Prelic e colaboradores desenvolveram um trabalho que propõem uma metolodologia para a comparação e validação adequada para o contexto de agrupamento bidimensional de dados. Utilizam as técnicas mais proeminentes de agrupamento bidimensional de dados de expressão gênica, incluive o SAMBA. O modelo proposto busca a definição do número ótimo de grupos através do algoritmo dividir e conquistar. A relevância biológica dos grupos bidimensionais obtidos foram analisadas com auxílio das anotações funcionais dos genes do *Gene Ontology* [APZ06].

3.7 Validação biológica

As técnicas de validação biológicas aplicadas nos resultados de agrupamentos de dados buscam identificar associações dos genes dos grupos com informações biológicas conhecidas, enriquecendo os agrupamentos com coerência biológica e não se limitando a somente cálculos estatísticos. Uma analogia a este procedimento de conciliação de duas diferentes abordagens de validação de agrupamentos, seria tornar o abstrato (resultados estatísticos) em um problema real (resultados biológicos) e, dessa forma, estabelecer novos desafios através do trabalho em conjunto das duas abordagens de validação.

A ontologia dos genes tem sido o recurso mais utilizado para o processo de enriquecimento biológico dos agrupamentos. Neste trabalho, além da utilização da ontologia gênica, também foi utilizado o recurso de identificação de fatores de transcrição. Ambos são descritos com mais detalhes a seguir.

3.7.1 Enriquecimento funcional dos genes

A crescente massa de dados proveniente de modernas técnicas de seqüenciamento, principalmente de genomas inteiros, trouxe a limitação dos recursos de interpretação dessa grande quantidade de dados. Desta limitação surgiu a necessidade da criação de sistemas que pudessem transcrever o conhecimento dos especialistas do domínio em informações biológicas representadas por meios computacionais. Estes sistemas poderiam ter um papel crucial no processamento de informações e interação com estes especialistas.

O uso de ontologias foi então adotado como forma de organização deste conhecimento e recurso de disponibilização dessas informações para pesquisadores e aplicativos computacionais.

A palavra ontologia tem ganhado popularidade, principalmente nas áreas que estudam o compartilhamento de conhecimentos. O exemplo mais comum do uso de ontologia na biologia molecular é o uso de comparação de seqüências para inferir a função de uma nova seqüência de proteínas, ou até a descoberta de novas vias metabólicas. A causa disto é que se uma seqüência de função desconhecida é altamente similar a uma seqüência de função conhecida, então é provável que a nova seqüência também tenha a mesma função. Então, ao invés de usar uma regra, lei ou equação para encontrar a função da proteína, um biólogo usa o conhecimento de que uma seqüência similar tem uma função conhecida, para fazer um julgamento sobre a função da nova seqüência.

O problema é que nem sempre essa comparação entre seqüências é possível. Os sistemas de nomenclatura utilizados são divergentes, tornando a interoperabilidade entre bancos de dados genômicos limitada. Este obstáculo que incentivou a formação do consórcio GO (*Gene Ontology*) [Con01, MAS00].

O GO consiste na organização de um vocabulário controlado, estruturado, precisamente definido e comum para descrever o papel dos genes e proteínas/RNAs de qualquer organismo. Ele é estruturado em três ontologias independentes - processo biológico, função molecular e componente celular.

A categoria que descreve o processo biológico refere-se ao objetivo biológico no qual o gene, proteína ou RNA contribui. A função molecular é a atividade ou tarefa que a proteína/RNA realiza. Esta ontologia descreve somente o que é feito, sem especificar

aonde ou quando o evento realmente acontece. Componente celular refere-se ao lugar na célula onde a proteína/RNA se encontra.

O projeto GO disponibiliza, na sua página na internet, as bases dos termos que fazem parte das ontologias. Estes termos estão conectados na forma de grafos acíclicos direcionados (DAGs), representados em redes hierárquicas. Esta representação foi a escolhida por causa da propriedade da conexão múltipla entre nós pais e filhos.

Cada termo da ontologia tem um identificador único, que permite referência cruzada entre os bancos de dados e os termos do consórcio. A sintaxe do identificador é GO:nnnnnnn, onde n representa um número seqüencial preenchido com zeros à direita.

Para a tarefa de agrupamentos de dados de expressão gênica, a idéia é enriquecer os grupos de genes com as ontologias de funções, processos biológicos e componente celular. Genes de um mesmo grupo significantemente enriquecido com a função de replicação do DNA, por exemplo, fornece um alto indício de coerência biológica do grupo.

São inúmeras as ferramentas disponíveis para o trabalho com ontologias de dados de expressão gênica: *GOToolBox* [DMJ04], *CLENCH* [SF04], *GOstat* [BS04], *FatiGo* [FASD04] e *GOTM* [BZS04] são alguns exemplos. Mais informações sobre ferramentas, análises de ontologia de dados de expressão gênica, problemas e limitação desta tarefa podem ser consultadas no trabalho de revisão dos autores Khatri e Draghici [KD05]. Sevilla e colaboradores desenvolveram um trabalho com o objetivo de confirmar a correlação entre dados de expressão e a similaridade semântica do GO. Os resultados obtidos comprovaram a existência dessa correlação, inclusive nos três níveis de ontologia do GO. Eles ainda sugerem que a similaridade semântica pode ser usada para melhorar os algoritmos de agrupamento através do desenvolvimento de uma ferramenta de busca semântica [JLSR05].

3.7.2 Identificação de fatores de transcrição

TFs (Fatores de transcrição) são proteínas capazes de regular a transcrição gênica. Essas proteínas controlam, quando, onde e como os genes serão transcritos, sendo a base para o controle da expressão gênica. Sem os fatores de transcrição, a maioria dos genes não seria capaz de ser expresso nas células e não haveria, portanto, a chance de as células eucariotas se diferenciarem.

As técnicas de validação de agrupamento baseadas na identificação de fatores de transcrição consistem na associação de fatores conhecidos, geralmente disponíveis em bancos de dados específicos, com os genes pertencentes a um grupo [RES]. A idéia é que genes com expressão co-regulada sobre várias condições são regulados pelos mesmos

fatores de transcrição e ainda é esperado que compartilhem elementos reguladores comuns em suas regiões promotoras (Mais informações sobre fatores de transcrição e elementos promotores são disponibilizadas no apêndice de Fundamentos de Biologia).

A ferramenta PRIMA [SMKT+05] utilizada neste trabalho é um exemplo de ferramenta de identificação de fatores de transcrição para a análise de agrupamentos de dados de expressão. Ela identifica fatores de transcrição cujos sítios de ligação são super-representados em um dado grupo de promotores.

3.8 Interpretação dos resultados de agrupamento

A interpretação dos resultados e o estabelecimento de medidas de avaliação de desempenho são as últimas, mas não menos importantes etapas da análise. A participação de um especialista é fundamental, já que parte do esforço nesta etapa do trabalho depende também de uma análise subjetiva [MHV01].

A interpretação dos agrupamentos deste trabalho foi beneficiada pelo uso das ferramentas de validação estatística e biológica. De acordo com os índices estatísticos, a melhor solução de agrupamento corresponde ao maior ou menor valor obtido, dependendo do índice. De acordo com as ferramentas de validação biológica o melhor agrupamento é aquele enriquecido com o maior número de funcões biológicas e com maior número de fatores de transcrição. Portanto, com o auxílio destas duas diferentes abordagens de validação, a melhor solução de agrupamento seria aquela indicada pela maioria dos índices estatísticos e corroboradas pela indicação de ambas as ferramentas de validação biológica.

3.9 Conclusão do capítulo

Neste capítulo foram descritos os conceitos da técnica de agrupamento de dados, organizados de acordo com as etapas de realização de um agrupamento: pré-processamento dos dados, definição de medidas de similaridade e algoritmos de agrupamento, separados nas abordagens unidimensional e bidimensional de dados. Na etapa de validação foram apresentadas as técnicas de validação estatística e biológica e por último foi apresentada a etapa de interpretação dos resultados de agrupamento de dados.

O próximo capítulo apresenta a metodologia adotada para a realização do trabalho.

Capítulo 4

Metodologia

A metodologia do trabalho é ilustrada na Figura 4.1. Os experimentos foram realizados com a utilização de duas bases de dados e com a combinação de diferentes etapas do trabalho, com o objetivo de identificar a metodologia mais qualificada para o problema.

Os algoritmos utilizados foram o k-médias e SOM, de agrupamento unidimensional e o algoritmo SAMBA de agrupamento bidimensional. Os números (4.1, 4.2, etc.) indicados na figura representam as subseções deste capítulo, onde são descritas cada uma das etapas da metodologia.

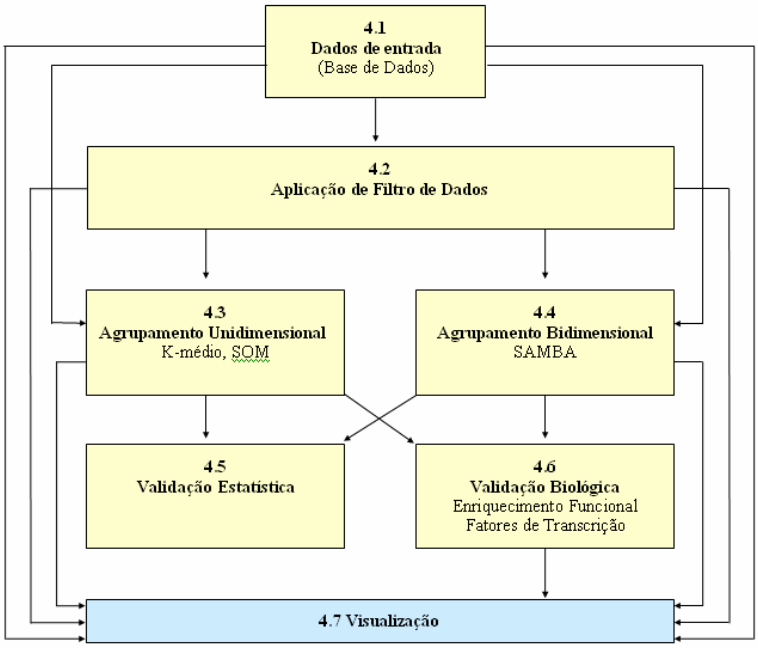


Figura 4.1: Esquema da metodologia do trabalho.

4.1 Definição da base de dados

A qualidade dos resultados dos agrupamentos é inerente à qualidade dos dados de entrada, por isso a definição da base de dados é um passo fundamental.

Neste trabalho foram utilizadas duas bases de dados, definidas considerando o organismo estudado, o preparo e a qualidade do experimento. Ambas, contém dados da levedura *Saccharomyces cerevisiae*, sugerido como modelo por se tratar de um organismo muito bem estudado, com grande variedade de materiais biológicos e de literatura, viabilizando a aplicação de diferentes estratégias computacionais de análise de dados e facilitando a validação dos resultados.

4.1.1 Base de dados CCSc

Assim chamada neste trabalho por se tratar de uma base de dados de genes possivelmente envolvidos no ciclo celular de *Saccharomyces cerevisiae* [PTSea98, Uni06b]. Ela é resultante da técnica de microarranjo de DNA. Contém 799 genes de *S. cerevisiae* submetidos a 77 condições experimentais de ciclo celular, conforme mostrado na Tabela 4.1. Todos os valores da base de dados foram medidos contra uma amostra referência no tempo 0. Cada valor representa a razão de fluorescência entre Cy5/Cy3 e foram normalizados com logaritmo na base 2. A base de dados contém 5.6% de valores faltantes.

Os 799 genes foram selecionados através de algoritmos de correlação. São genes que responderam a um critério de regulação de ciclo celular. Outras análises deste conjunto de dados ainda revelaram a presença de elementos promotores[1] novos e conhecidos, sendo os conhecidos, em sua maioria, envolvidos na regulação do ciclo celular. A descrição completa do conjunto de dados está disponível no endereço [Uni06b].

Além dos valores de expressão, a base de dados CCSc contém uma coluna que armazena a anotação funcional do gene e a fase do ciclo celular que possivelmente o gene esteja envolvido. Do total de 799 genes, 300 estão anotados como envolvidos na fase G1 do ciclo celular, 71 envolvidos na fase S, 121 na fase G2, 195 na fase M (G2/M), e os 112 restantes na fase M (M/G1). Estas informações são bastante úteis para a avaliação preliminar da qualidade dos grupos. Os programas de agrupamento direcionados para a análise de dados de expressão geralmente incorporam a funcionalidade desta coluna de anotação, disponibilizando essas informações junto com os resultados dos agrupamentos, permitindo facilmente verificar se genes com funções relacionadas foram alocados em um mesmo grupo.

[1] Pequenas seqüências que sinalizam onde a síntese do RNA deve ser iniciada.

Tabela 4.1: Estrutura da base de dados CCSc.

CONDIÇÕES		**TEMPO**
Ciclo celular	CLN3	1, 2
	CLB2	2, 1
	Fator Alpha	0, 7, 14, 21, 28, 35, 42, 49, 56, 63, 70, 77, 84, 91, 98, 105, 112, 119
	CDC15	10m, 30m, 50m, 70m, 80m, 90m, 100m, 110m, 120m, 130m, 140m, 150m, 160m, 170m, 180m, 190m, 200m, 210m, 220m, 230m, 240m, 250m, 270m, 290m
	CDC28	0m, 10m, 20m, 30m, 40m, 50m, 60m, 70m, 80m, 90m, 100m, 110m, 120m, 130m, 140m, 150m, 160m
	Elutriação	0m, 30m, 60m, 90m, 120m, 150m, 180m, 210m, 240m, 270m, 300m, 330m, 360m, 390m
TOTAL		77 condições experimentais

Esta base de dados foi escolhida devido à característica de todos os seus genes serem regulados durante o processo de ciclo celular [PTSea98]. Baseado nas análises conduzidas por Spellman e colaboradores esperava-se obter 4 grupos, correspondentes às fases G1, S, G2 e M do ciclo celular.

O controle do ciclo é feito por diversos produtos gênicos, que são, por sua vez, regulados por fatores extracelulares, sejam eles nutrientes ou fatores de crescimento, que fazem com que a divisão celular ocorra coordenadamente com as necessidades do organismo como um todo. As condições CLN3, CLB2, fator alpha, CDC15, CDC28 e elutriação, correspondem à proteínas envolvidas na regulação do ciclo celular de *Saccharomyces cerevisiae*.

Mais detalhes do organismo *Saccharomyces cerevisiae* e do ciclo celular estão disponíveis no Apêndice B.

Na Tabela 4.2 abaixo as condições da base de dados CCSc estão numeradas, porque as imagens das análises dos agrupamentos referenciam as condições através destes números.

Tabela 4.2: Condições numeradas da base de dados CCSc.

Rótulo	Condição
1	cln3-1
2	cln3-2
3	clb2-2
4	clb2-1
5	alpha0
6	alpha7
7	alpha14
8	alpha21
9	alpha28
10	alpha35
11	alpha42
12	alpha49
13	alpha56
14	alpha63
15	alpha70
16	alpha77
17	alpha84
18	alpha91
19	alpha98
20	alpha105
21	alpha112
22	alpha119
23	cdc15_10
24	cdc15_30
25	cdc15_50
26	cdc15_70
27	cdc15_80
28	cdc15_90
29	cdc15_100
30	cdc15_110

Continua na próxima página

Tabela 4.2 – continuação da página anterior

31	cdc15_120
32	cdc15_130
33	cdc15_140
34	cdc15_150
35	cdc15_160
36	cdc15_170
37	cdc15_180
38	cdc15_190
39	cdc15_200
40	cdc15_210
41	cdc15_220
42	cdc15_230
43	cdc15_240
44	cdc15_250
45	cdc15_270
46	cdc15_290
47	cdc28_0
48	cdc28_10
49	cdc28_20
50	cdc28_30
51	cdc28_40
52	cdc28_50
53	cdc28_60
54	cdc28_70
55	cdc28_80
56	cdc28_90
57	cdc28_100
58	cdc28_110
59	cdc28_120
60	cdc28_130
61	cdc28_140
62	cdc28_150
63	cdc28_160

Continua na próxima página

Tabela 4.2 – continuação da página anterior

64	elu0
65	elu30
66	elu60
67	elu90
68	elu120
69	elu150
70	elu180
71	elu210
72	elu240
73	elu270
74	elu300
75	elu330
76	elu360
77	elu390

4.1.2 Base de dados GSc

Assim chamada neste trabalho por se tratar de uma base de dados do genoma completo de *Saccharomyces cerevisiae* [MBEB98, Eis06b].

A base de dados GSc também é resultante da técnica de microrranjo de DNA. Contém 6621 ORFs[2] de *S. cerevisiae* submetidas a 80 condições experimentais sincronizados em cinco situações distintas de ciclo de divisão celular e duas condições correspondentes a respostas a diferentes estresses ambientais, conforme mostrado na Tabela 4.3. Todos os valores da base de dados foram medidos contra uma amostra referência no tempo 0. Cada valor representa a razão de fluorescência entre Cy5/Cy3 e foram normalizados com logaritmo na base 2. A base de dados contém 4.32% de valores faltantes.

O ciclo celular compreende os processos que ocorrem desde a formação de uma célula até sua divisão em duas células-filhas, iguais entre si. Serve tanto para manter a vida, no caso dos organismos pluricelulares, como para gerar a vida, no caso dos organismos unicelulares. O ciclo celular é dividido em quatro fases distintas: G1, S, G2 e M.

[2]ORF (*Open Read Frames*) corresponde a um quadro de leitura que inicia com um códon de início até terminar com um códon de parada. Embora seja comum o uso dos termos ORF e gene indistintamente, toda região codificadora de um gene é uma ORF, mas nem toda ORF é um gene.

Tabela 4.3: Estrutura da base de dados GSc.

CONDIÇÕES		TEMPO
Ciclo celular	Fator Alpha	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18
	CDC15	10m, 30m, 50m, 70m, 80m, 90m, 100m, 110m, 120m, 120m, 130m, 140m, 150m, 160m, 160m, 170m, 180, 190m, 200m, 210m, 220m, 240m, 250m, 270m, 290m
	Elutriação	0.0hrs, 0.5hrs, 1.0hrs, 1.5hrs, 2.0hrs, 2.5hrs, 3.0hrs, 3.5hrs, 4.0hrs, 4.5hrs, 5.0hrs, 5.5hrs, 6.0hrs, 6.5hrs
	CLN3	30m, 40m
	CLB5	40m
Esporulação		0, 30m, 2h, 5h, 7h, 9h, 11h, 2h (v.5h), 7h (v.5h), 11h (v.5h), ndt80-Early, ndt80-Middle, ndt80-Over
Mudança diáuxica		19.0g/L, 18.7g/L, 17.6g/L, 14.0g/L, 7.5g/L, 0.2g/L, 0g/L
TOTAL		80 condições experimentais

O controle do ciclo celular é feito por diversos produtos gênicos, que são, por sua vez, regulados por fatores extracelulares, sejam eles nutrientes ou fatores de crescimento, que fazem com que a divisão celular ocorra coordenadamente com as necessidades do organismo como um todo.

O fator alpha, CDC15, elutriação, CLN3 e CLB5 da base de dados GSc, correspondem à proteínas envolvidas na regulação do ciclo celular de *Saccharomyces cerevisiae*. Esporulação corresponde à geração de novos esporos[3] e é induzida por limitação nutricional e neste trabalho medida em uma série de tempos diferentes. A mudança diáuxica refere-se à indução dos genes em meios de cultura com diferentes concentrações de glicose.

Na Tabela 4.4 abaixo as condições da base de dados GSc estão numeradas, porque as imagens das análises dos agrupamentos referenciam as condições através destes números.

[3]Esporo corresponde à unidade reprodutiva produzida pelas plantas, protozoários e bactérias.

Tabela 4.4: Condições numeradas da base de dados GSc.

Rótulo	Condição
1	Cell-cycle Alpha-Factor 1
2	Cell-cycle Alpha-Factor 2
3	Cell-cycle Alpha-Factor 3
4	Cell-cycle Alpha-Factor 4
5	Cell-cycle Alpha-Factor 5
6	Cell-cycle Alpha-Factor 6
7	Cell-cycle Alpha-Factor 7
8	Cell-cycle Alpha-Factor 8
9	Cell-cycle Alpha-Factor 9
10	Cell-cycle Alpha-Factor 10
11	Cell-cycle Alpha-Factor 11
12	Cell-cycle Alpha-Factor 12
13	Cell-cycle Alpha-Factor 13
14	Cell-cycle Alpha-Factor 14
15	Cell-cycle Alpha-Factor 15
16	Cell-cycle Alpha-Factor 16
17	Cell-cycle Alpha-Factor 17
18	Cell-cycle Alpha-Factor 18
19	Cell-cycle cdc15 10m
20	Cell-cycle cdc15 30m
21	Cell-cycle cdc15 50m
22	Cell-cycle cdc15 70m
23	Cell-cycle cdc15 80m
24	Cell-cycle cdc15 90m
25	Cell-cycle cdc15 100m
26	Cell-cycle cdc15 110m
27	Cell-cycle cdc15 120m
28	Cell-cycle cdc15 120m
29	Cell-cycle cdc15 130m
30	Cell-cycle cdc15 140m
31	Cell-cycle cdc15 150m

Continua na próxima página

Tabela 4.4 – continuação da página anterior

32	Cell-cycle cdc15 160m
33	Cell-cycle cdc15 160m
34	Cell-cycle cdc15 170m
35	Cell-cycle cdc15 180m
36	Cell-cycle cdc15 190m
37	Cell-cycle cdc15 200m
38	Cell-cycle cdc15 210m
39	Cell-cycle cdc15 220m
40	Cell-cycle cdc15 240m
41	Cell-cycle cdc15 250m
42	Cell-cycle cdc15 270m
43	Cell-cycle cdc15 290m
44	Cell-cycle Elutriation 0.0hrs
45	Cell-cycle Elutriation 0.5hrs
46	Cell-cycle Elutriation 1.0hrs
47	Cell-cycle Elutriation 1.5hrs
48	Cell-cycle Elutriation 2.0hrs
49	Cell-cycle Elutriation 2.5hrs
50	Cell-cycle Elutriation 3.0hrs
51	Cell-cycle Elutriation 3.5hrs
52	Cell-cycle Elutriation 4.0hrs
53	Cell-cycle Elutriation 4.5hrs
54	Cell-cycle Elutriation 5.0hrs
55	Cell-cycle Elutriation 5.5hrs
56	Cell-cycle Elutriation 6.0hrs
57	Cell-cycle Elutriation 6.5hrs
58	Cell-cycle CLN3 induction 30m
59	Cell-cycle CLN3 induction 40m
60	Cell-cycle CLB5 induction 40m
61	Sporulation 0
62	Sporulation 30m
63	Sporulation 2h
64	Sporulation 5h

Continua na próxima página

Tabela 4.4 – continuação da página anterior

65	Sporulation 7h
66	Sporulation 9h
67	Sporulation 11h
68	Sporulation 2h (v. 5h)
69	Sporulation 7h (v. 5h)
70	Sporulation 11h (v. 5h)
71	Sporulation ndt80- Early
72	Sporulation ndt80- Middle
73	Sporulation ndt80over
74	Diauxic Shift 19.0g/L
75	Diauxic Shift 18.7g/L
76	Diauxic Shift 17.6g/L
77	Diauxic Shift 14.0g/L
78	Diauxic Shift 7.5g/L
79	Diauxic Shift 0.2g/L
80	Diauxic Shift 0g/L

4.2 Aplicação de filtros de dados

A maioria das bases de dados contém dados incompletos ou com ruídos. Nos dados resultantes da técnica de microarranjo e DNA, estas características indesejáveis aumentam com a influência de manchas espúrias (*background*) e de variações do processo de hibridação dos valores de cada *spot*, mas que podem ser minimizadas com a normalização dos dados, processo descrito em detalhes no Capítulo 1. As duas bases de dados utilizadas neste trabalho foram disponibilizadas já normalizadas.

A presença de dados faltantes na base é tema de discussão. Muitos algoritmos requerem a matriz de dados completa. Os algoritmos hierárquico e k-médias, por exemplo, não são robustos para trabalhar com dados faltantes, podendo ter sua eficácia comprometida [TCS+01]. Alguns trabalhos sugerem que os genes que apresentam valores faltantes sejam eliminados, mas também é comum que nada seja feito com eles. Uma outra opção freqüentemente adotada é a substituição desses dados pela média dos valores dos registros ou colunas da base de dados, dependendo do problema.

Neste trabalho os dados faltantes foram substituídos pelo cálculo da média das

condições (colunas), conforme sugerido pelos autores do programa Machaon CVE (*Clustering and Validation Environment*) [NBC05, Bol06] de validação estatística. Um estudo comparativo de vários métodos de estimação de valores faltantes nos dados de microarranjo de DNA são encontrados em [TCS+01].

A tentativa de diminuir a presença de dados ruidosos e/ou pouco representativos da base de dados foi feita com a aplicação de filtros de dados, descritos na seqüência.

Os agrupamentos foram aplicados nas bases de dados com e sem a aplicação de filtros de dados. O programa utilizado para a aplicação de filtros foi o *Cluster* v.3.0 [MJLDHM04, Eis06a][4]. Este programa permite gravar a base de dados resultante da aplicação de filtros, podendo, dessa forma, ser aplicada em outro programa, assim como foi feito neste trabalho.

Para todas as bases de dados, foi utilizada a opção de filtro de dados que considera somente os genes que apresentam pelo menos 1 valor de expressão $>= 1,5$ em todas as condições.

Os valores de expressão variam de acordo com a intensidade do sinal. A ausência de sinal de expressão é representada pelo valor 1 ou pela cor preta. Sinais baixos são representados pelo intervalo de 1 a 28 ou pela cor verde. O aumento da expressão dos genes é representado por valores maiores que 2.8 ou pela cor vermelha.

4.3 Aplicação de algoritmos de agrupamento unidimensional

O programa utilizado para a aplicação dos algoritmos k-médias e SOM foi o Expander v.2.0 (*Expression Analyzer and Displayer*) [SMKT+05, ea06].

A medida de distância utilizada em ambos os algoritmos foi a distância Euclidiana, medida padrão do programa Expander. O algoritmo k-médias foi aplicado com diferentes valores atribuídos à k (número desejado de grupos) e o SOM foi aplicado com diferentes dimensões da matriz.

Para o agrupamento da base de dados CCSc foram adotados os valores de $k = 2, 4, 5, 8, 10$. As dimensões da matriz do SOM foram definidas como 2x2, 5x1, 2x3, 2x4 e 2x5.

Para o agrupamento da base de dados GSc foram adotados os valores de $k =$

[4]Para utilizar o programa é necessário adaptar as configurações regionais do computador ao formato dos valores numéricos da base de dados, ou o contrário. O formato dos valores numéricos dos EUA, por exemplo, o ponto representa o símbolo decimal e a vírgula o símbol o de agrupamento de dígitos (123,456,789.00), diferente do padrão adotado no Brasil.

$5, 10, 20, 30, 50$. As dimensões da matriz do SOM foram definidas como 5x1, 2x5, 5x5 e 5x10.

Os valores para o algoritmo k-médias e SOM foram definidos inicialmente num intervalo entre 2 até 200 para a base de dados CCSc e de 5 até 1000 para a base de dados GSc, no entanto, um número muito grande de grupos dificulta o processo de análise dos agrupamentos. Por esta razão foram definidos valores próximos a 4 para a base CCSc, baseado na expectativa da formação de 4 grupos, correspondentes às 4 fases do ciclo celular. Para a base GSc os valores foram definidos baseados na relação dos atributos da base. Esta definição depende do objetivo do problema, pois sabe-se que quanto maior o número de grupos, mais homogêneos os elementos pertencentes ao mesmo grupo. Como neste trabalho o objetivo principal é o processo de validação de agrupamento, os valores foram limitados a no máximo 50 grupos, conforme descrito acima.

4.4 Aplicação do algoritmo de agrupamento bidimensional

Para a aplicação do algoritmo de agrupamento bidimensional também foi utilizado o programa Expander.

4.5 Validação estatística

Para a validação estatística foi utilizado o programa Machaon CVE, que avalia a qualidade dos agrupamentos obtidos através de diferentes índices estatísticos: C, Davies-Bouldin, Dunn, Silhueta e Isolamento. Todos esses índices foram aplicados com distância Euclidiana. O trabalho de [NBC05] mostra que diferentes medidas de distância aplicada nesses índices não interfere significativamente nos resultados. Os índices Davies-Bouldin e Dunn, além da distância Euclidiana foram aplicados com a distância completa intergrupo e o diâmetro completo intragrupo. O tamanho da vizinhança do índice de Isolamento foi definido com o valor de 0,1. Além desses índices, também foram utilizados os índices de homogeneidade e separação para a avaliação da qualidade dos agrupamentos. Estes índices são fornecidos junto com o resultado dos agrupamentos do programa Expander.

Os arquivos com o resultado dos agrupamentos foram adaptados ao formato do programa Machaon CVE. Cada arquivo deve conter todo o conteúdo da base de dados e a última coluna com o número do grupo que o gene pertence.

4.6 Validação biológica

A validação biológica foi feita com o auxílio das ferramentas TANGO (*Tool for Analysis of GO Enrichment*) e PRIMA (*Promoter Integration in Microarray Analysis*), disponíveis no programa Expander.

A ferramenta TANGO auxilia o processo de validação biológica porque enriquece os resultados dos agrupamentos através da associação dos grupos com funções biológicas. As funções dos genes são determinadas de acordo com as definições do projeto GO e são adaptadas ao programa Expander através de arquivos disponibilizados no endereço do projeto na internet [Con01]. Esta adaptação permite que os arquivos sejam facilmente atualizados, acompanhando a dinâmica das informações das funções dos genes.

A ferramenta TANGO foi aplicada nos resultados de todos os agrupamentos. Os parâmetros utilizados foram dois níveis do GO: Processo e Função. O número de iterações e valor de significância (*p-value*) foram mantidos com o valor padrão de 1.000 e 0,05 respectivamente. Isto significa que, quanto maior o número de iterações, maior a resolução dos valores de significância corrigidos e maior o tempo de execução do algoritmo. Os valores de significância corrigidos estarão num intervalo entre 1/iterações e 1. A classe funcional é considerada significantemente enriquecida se o seu valor de significância corrigido estiver abaixo do limiar. As funções biológicas identificadas em cada grupo são apresentadas na forma de histogramas.

PRIMA é outra ferramenta útil para auxiliar a análise dos resultados dos agrupamentos de dados de expressão. Ela identifica fatores de transcrição cujos sítios de ligação são super-representados em um dado grupo de promotores. A idéia é que genes com expressão co-regulada sobre várias condições são regulados pelos mesmos fatores de transcrição e ainda é esperado que compartilhem elementos reguladores comuns em suas regiões promotoras.

Assim como a ferramenta TANGO, a ferramenta PRIMA também foi aplicada nos resultados dos agrupamentos. Foi utilizado o valor de significância (*p-value*) padrão: 1.0E-4.

PRIMA trabalha com um conjunto de promotores de genes humanos conhecidos, denominado 13K e um modelo de matriz de pesos (PWM) para a modelagem de sítios de ligação reconhecidos por TFs. O programa obtém essa matriz de pesos do banco de dados TRANSFAC [TRA06].

Utilizando as seqüências do genoma humano e modelos de sítios de ligação (BSs) reconhecidos por TFs, PRIMA identifica TFs cujos BSs são significantemente representados num dado conjunto de promotores.

A entrada do programa são dois conjuntos de genes: um deles que corresponde aos genes dos grupos e o outro corresponde ao arquivo de genes conhecidos (13K). Para cada PWM, P é calculado da seguinte forma:

1. O programa inicia calculando um limiar de similaridade $T(P)$. Em seguida, são procurados os promotores com escore de similaridade acima deste limiar, considerados como *hits* de P.

2. O próximo passo é a realização de um teste estatístico que verifica quais os *hits* de P estão altamente representados nos grupos de genes em relação ao arquivo de promotores do programa PRIMA.

Os TFs identificados em cada grupo são apresentados na forma de histogramas e representados pela nomenclatura padrão, por exemplo: MBP1, STB1, etc. A descrição de cada fator, a função que desempenham e o organismo em que foram identificados podem ser consultados no banco de dados TRANSFAC.

Mais detalhes do algoritmo são descritos em [RES].

A idéia para a análise biológica dos dados da base CCSc, adotada como modelo experimental é explorar os resultados de todos os agrupamentos detalhadamente, conciliando o conhecimento prévio dos dados (através das anotações funcionais dos genes e anotação da fase do ciclo celular que cada gene possivelmente está envolvido), com os recursos de visualização do programa Expander. Desta forma confirmar se a metodologia é a mais indicada para a análise da base de dados GSc. As etapas desta análise são descritas a seguir:

1. Análise do gráfico do perfil médio da expressão dos genes dos grupos: para visualizar o comportamento dos genes em cada grupo e as variações dos sinais de expressão através do desvio padrão.

2. Análise dos *heat maps*: para visualizar a expressão dos genes em cada uma das condições do microarranjo e os padrões de expressão do grupo.

3. Consulta às informações de anotação funcional presentes na base de dados: estas informações permitiram que os genes fossem analisados de acordo com a fase do ciclo celular que estão envolvidos e, conseqüentemente analisar se um grupo foi formado por genes da fase G1, por exemplo.

4. Análise dos resultados obtidos com as técnicas de validação biológica de enriquecimento funcional e identificação de fatores de transcrição.

5. Consulta aos bancos de dados de ontologia e de fatores de transcrição: para confirmar se as funções e fatores identificadas pelos processos do item anterior confirmam a análise do 3^o passo, de associação dos grupos com as fases do ciclo celular.

4.7 Visualização

A visualização dos resultados é um instrumento valioso de apoio ao processo de mineração de dados. Foi adotada em todas as fases deste trabalho para o acompanhamento do comportamento dos dados e para a visualização dos resultados parciais. A visualização permite ao usuário adquirir percepções dos dados, confirmar expectativas, desenvolver novas idéias podendo, inclusive, sugerir novas hipóteses.

4.7.1 Visualização dos dados de entrada

Os dados das bases, com e sem a aplicação de filtros, foram visualizados através das ferramentas disponíveis no programa Expander. A primeira delas consiste num gráfico *box-plot*, onde o eixo X corresponde às condições e o eixo Y aos valores de expressão dos genes. O gráfico apresenta o perfil de cada umas das condições, qual o valor de expressão máximo e mínimo, a mediana e os intervalos *inter-quartis*, conforme ilustrado na Figura 4.2.

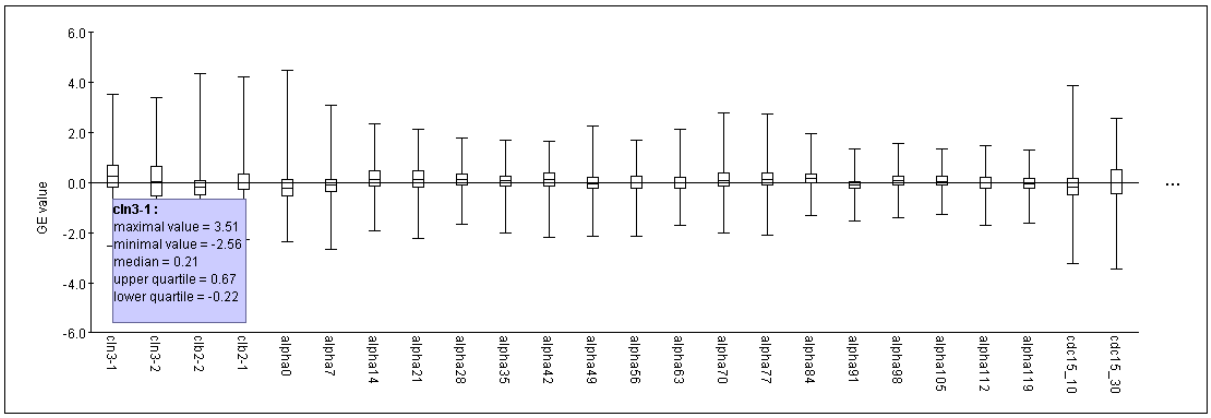


Figura 4.2: Gráfico do perfil de algumas condições da base de dados CCSc (imagem adaptada da captura da tela do programa Expander).

A outra opção são os *heat maps*, freqüentemente utilizados em biologia molecular para a visualização de dados de microarranjo de DNA. Um *heat map* é uma representação gráfica dos dados que são apresentados em cores distribuídos em um mapa bidimensional.

É muito útil para a visualização de resultados de agrupamentos. As cores do *heat map* correspondem aos níveis de expressão dos genes sobre as condições. A cor vermelha indica maior expressão (ou expressão induzida), a cor verde pouca expressão (ou expressão reprimida) e a cor preta, representa a ausência ou baixa qualidade do sinal de expressão.

Os programas geralmente implementam *heat maps* interativos. No caso específico de dados biológicos, os programas geralmente permitem o usuário visualizar o nível de expressão de cada gene e sua anotação funcional clicando sobre uma coordenada do *heat map*, conforme ilustrado na Figura 4.3.

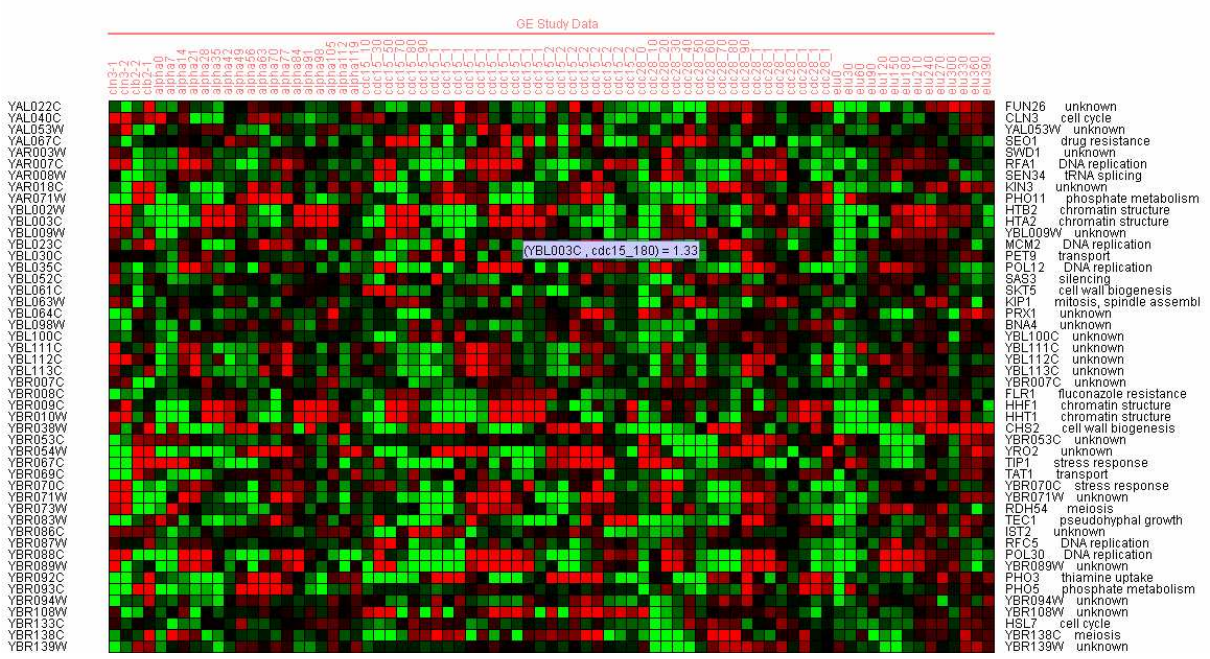


Figura 4.3: Um trecho do *heat map* da base de dados CCSc (captura da tela do programa Expander).

4.7.2 Visualização dos agrupamentos unidimensionais

Os resultados dos agrupamentos dos algoritmos k-médias e SOM foram visualizados através dos *heat maps* e dos gráficos do perfil da expressão dos genes. São dois tipos de gráficos, o primeiro corresponde ao perfil médio da expressão de um grupo e o segundo corresponde ao perfil individual da expressão de cada gene.

A Figura 4.4 ilustra o gráfico do perfil médio da expressão de um grupo. Este gráfico permite a comparação do perfil da expressão de diferentes grupos.

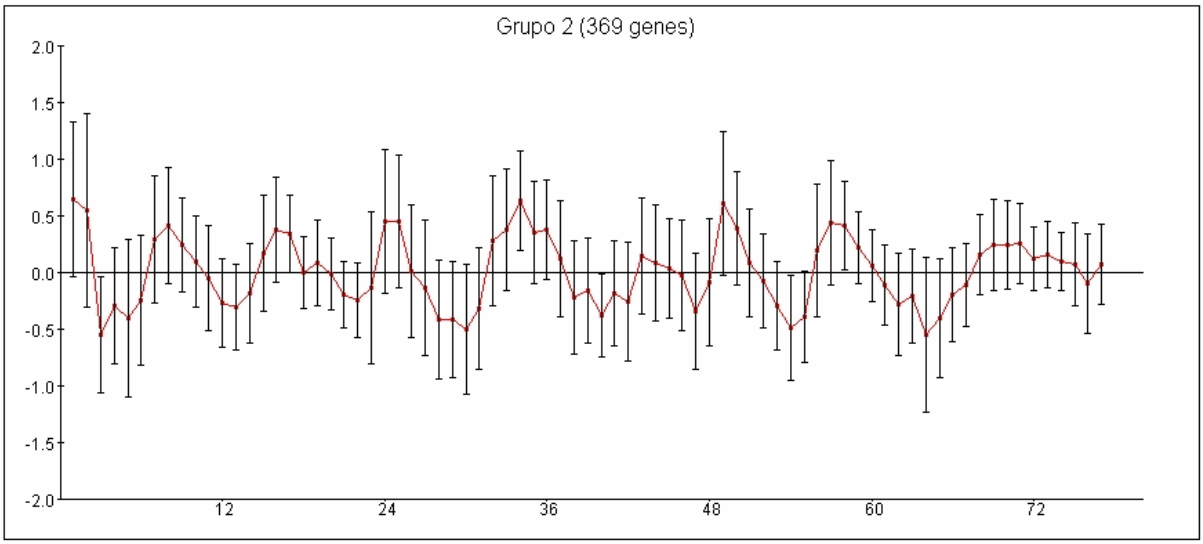


Figura 4.4: Perfil médio da expressão do grupo.

O eixo X corresponde às condições, numerada de acordo com as Tabelas X e Y apresentadas anteriormente. O eixo Y aos níveis de expressão. A linha vermelha indica o perfil médio da expressão dos genes 369 genes do grupo 2. Na condição 12, por exemplo, os genes apresentam expressão reprimida. Na condição 36 apresentam expressão induzida. As barras verticais indicam o desvio padrão identificado em cada uma das condições.

Da mesma forma, também é possível visualizar o perfil individual da expressão dos genes, conforme ilustrado na Figura 4.5.

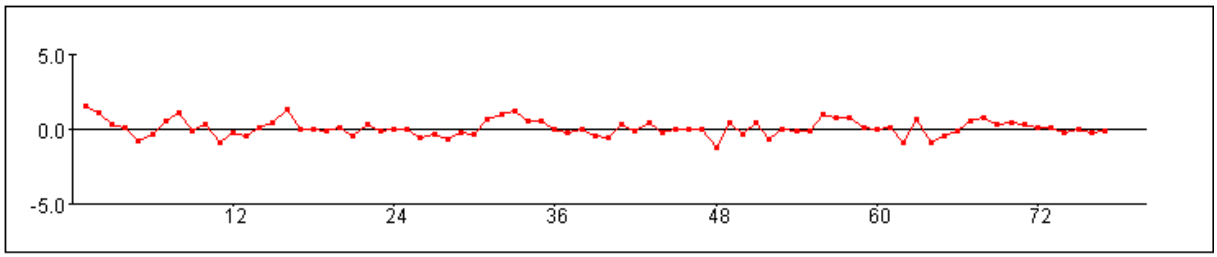


Figura 4.5: Perfil individual da expressão de um gene.

Através deste gráfico é possível verificar se um gene tem o mesmo perfil de expressão dos demais genes do grupo. Também é útil para comparar o perfil da expressão individual dos genes com o gráfico do perfil médio do grupo e, se for o caso, identificar em quais condições eles não demonstram conformidade.

4.7.3 Visualização dos agrupamentos bidimensionais

Da mesma forma que os agrupamentos unidimensionais, os resultados dos agrupamentos bidimensionais também foram visualizados através dos *heat maps*.

4.7.4 Visualização dos resultados de validação biológica

Os resultados dos processos de validação biológica de enriquecimento funcional e identificação de fatores de transcrição foram visualizados através de histogramas. Cada barra do histograma representa uma função ou fator. Cada função ou fator é representada por uma cor. Sobre cada barra do histograma é apresentado o percentual da freqüência de cada função ou fator no grupo, conforme ilustrado na Figura 4.6.

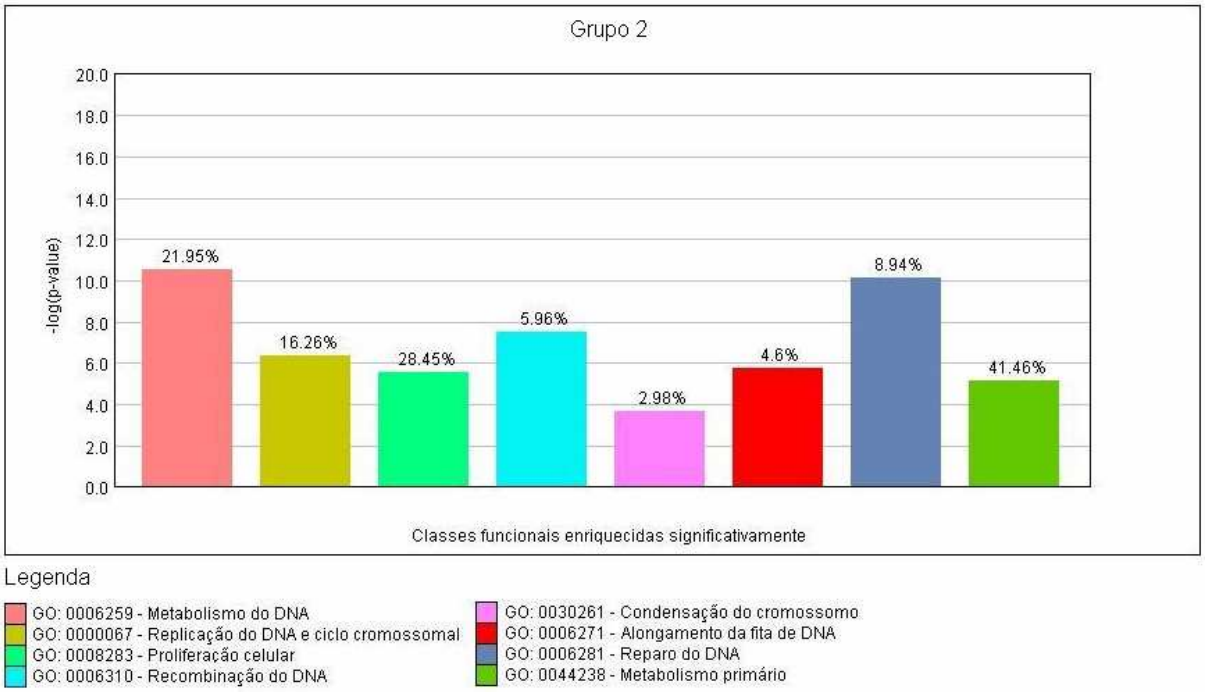


Figura 4.6: Histograma do resultado do processo de validação biológica por enriquecimento funcional do grupo 2.

4.8 Conclusão do capítulo

Neste capítulo foram descritas todas as etapas da metodologia do trabalho, desde a definição da base de dados até o processo de visualização dos resultados.

No próximo capítulo são apresentados os resultados e discussões das técnicas de agrupamento de dados e de validação estatística e biológica.

Capítulo 5

Resultados e Discussões

Este capítulo apresenta os resultados e discussões obtidos dos diferentes algoritmos de agrupamento de dados aplicados em dados de expressão gênica.

A avaliação dos resultados de um agrupamento não é uma tarefa trivial em decorrência da sua característica de aprendizagem não-supervisionada, ou seja, de não haver conhecimento *a priori* dos dados. Nos trabalhos publicados de agrupamento de dados de expressão, usualmente são aplicadas técnicas de validação estatística ou biológica que medem a qualidade da solução de um agrupamento, embora poucos trabalhos tenham relacionado estas duas opções.

Neste trabalho os resultados dos agrupamentos foram analisados com o auxílio da combinação das técnicas de validação estatística e biológica. O objetivo do trabalho é, além de minimizar a subjetividade da escolha da melhor solução de agrupamento, identificar se a resposta de ambas as técnicas de validação coincidem, se elas se complementam ou se não é possível observar nenhuma associação entre elas.

O processo de análise de agrupamentos foi o mesmo para os diferentes algoritmos e para as duas bases de dados. Por esta razão e para não tornar o capítulo muito extenso, são apresentados aqui os resultados e discussões do agrupamento do algoritmo k-médias somente quando $k = 4$, da base de dados CCSc. A análise dos demais agrupamentos são apresentados no Apêndice A.

A idéia do processo de análise adotado neste trabalho foi iniciar com resultados da base CCSc, que refere-se somente ao processo biológico de ciclo celular de *S. cerevisiae* e aplicar o procedimento na segunda base de dados GSc que corresponde ao genoma completo do mesmo organismo, a fim de reforçar o procedimento do trabalho numa base onde não há conhecimento prévio do comportamento dos dados.

5.1 Base de dados CCSc

5.1.1 Agrupamento k-médias k = 4

Os 799 genes da base de dados CCSc foram agrupados em 4 grupos, conforme mostrado na Tabela 5.1.

Tabela 5.1: Agrupamento $k = 4$.

Grupos	Quantidade de genes	Homogeneidade
1	217	0,51
2	1	1,0
3	304	0,604
4	277	0,344

5.1.2 Validação estatística do agrupamento k = 4

A Tabela 5.2 a seguir apresenta os resultados das técnicas de validação estatística aplicadas no agrupamento $k = 4$. Este agrupamento foi indicado como a melhor solução somente pelo índice de homogeneidade, conforme indicado em negrito na tabela.

Tabela 5.2: Validação estatística do agrupamento $k = 4$

k	Homogeneidade	Separação	C	D. Bouldin	Dunn	Silhueta	Isolamento
4	**0,491**	-0,049	0,287	1,765	0,702	0,087	0,526

5.1.3 Validação biológica do agrupamento k = 4

A Tabela 5.3 a seguir apresenta a quantidade de funcões biológicas (TANGO) e fatores de transcrição (PRIMA) identificados nos agrupamentos do algoritmo k-médias. Embora o índice estatístico de homogeneidade tenha indicado o agrupamento $k = 4$ como a melhor solução de agrupamento, as técnicas de validação biológica indicaram o agrupamento $k = 10$.

No Apêndice A são apresentadas as análises detalhadas da significância biológica de todos estes agrupamentos do k-médias. A seguir é apresentada a análise do agrupamento $k = 4$, conforme estabelecido para ser apresentado neste capítulo.

Tabela 5.3: Validação biológica dos agrupamentos k-médias

k	TANGO	PRIMA
2	9	3
4	12	3
5	8	3
8	10	5
10	**17**	**5**

5.1.4 Significância biológica do agrupamento k = 4

A seguir são apresentadas as análises de cada um dos 4 grupos. Estas análises foram baseadas na informação da funcionalidade dos genes contida na base de dados, com o auxílio das diferentes ferramentas de visualização e com as informações disponíveis na literatura. Em seguida, são apresentadas as funções biológicas e os fatores de transcrição associados aos grupos.

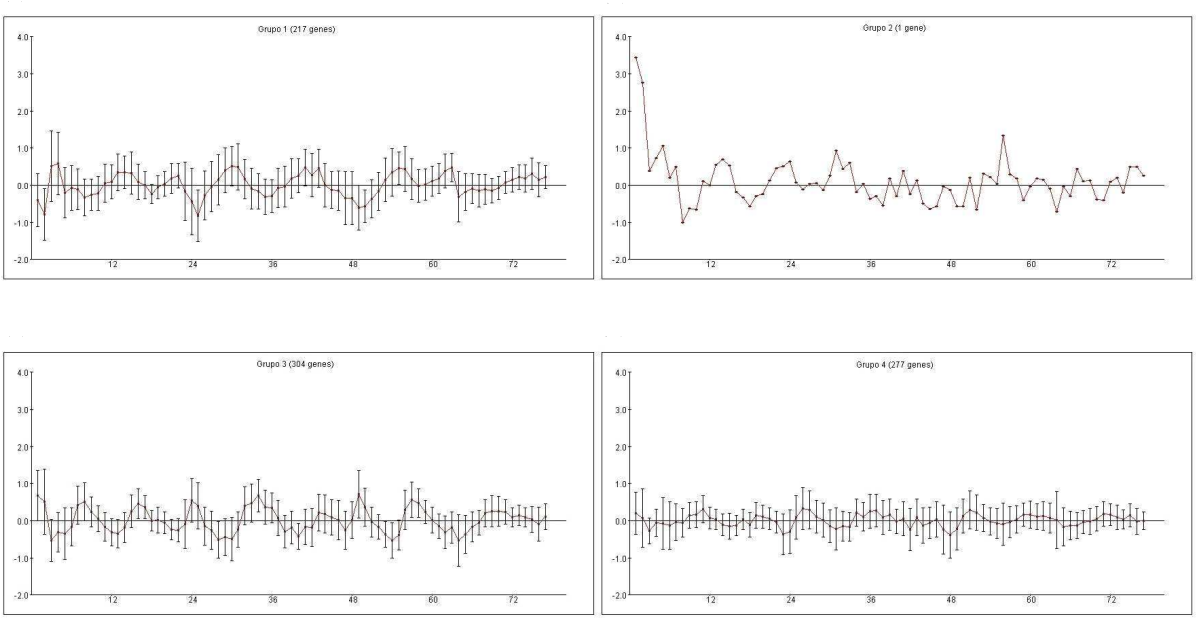


Figura 5.1: Perfil médio da expressão dos genes dos 4 grupos ($k = 4$).

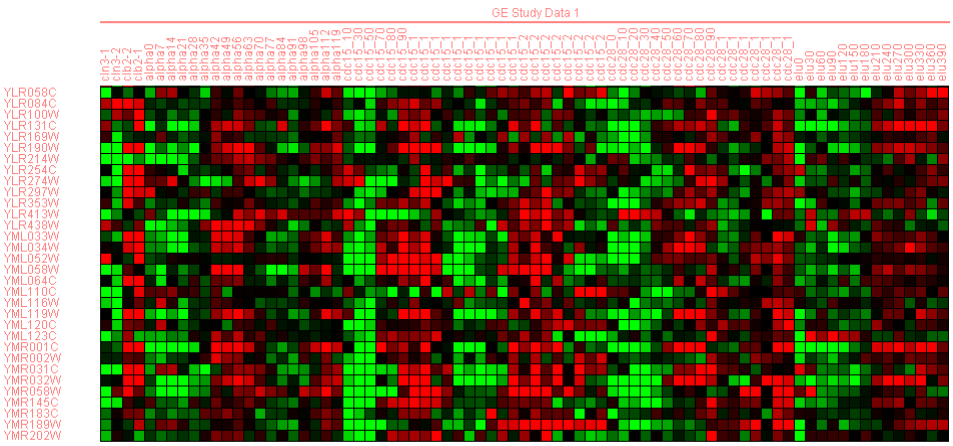
Quando $k = 4$, esperava-se obter 4 grupos representantes das 4 fases do ciclo celular. No entanto, um dos grupos foi formado por somente 1 gene, comprometendo a distribuição do restante dos genes nos 3 grupos. A Figura 5.1 mostra o perfil médio da expressão dos genes dos 4 grupos.

O grupo 1 foi formado por genes envolvidos na fase M (G2/M e M/G1). O grupo 2 foi formado por somente o gene CLN3. O grupo 3 foi formado por genes da fase G1 do

ciclo. O grupo 4 foi formado por genes envolvidos nas fases S e G2 do ciclo. A formação de 4 grupos correspondentes às 4 fases do ciclo celular foi prejudicada pela formação do grupo 2 com 1 gene.

A Figura 5.2 (a) ilustra o *heat map* de alguns genes pertencentes ao grupo 1. A imagem mostra o comportamento de co-regulação desses genes em todas as condições experimentais através das colunas verdes e vermelhas bem definidas. A Figura 5.1 (b) ilustra o *heat map* de alguns genes do grupo 4. Diferente do perfil da expressão dos genes do grupo 1, a imagem do grupo 4 mostra que não há um perfil bem definido de co-regulação dos genes, corroborando ser o grupo de menor homogeneidade, conforme informações da Tabela 5.1.

a)



b)

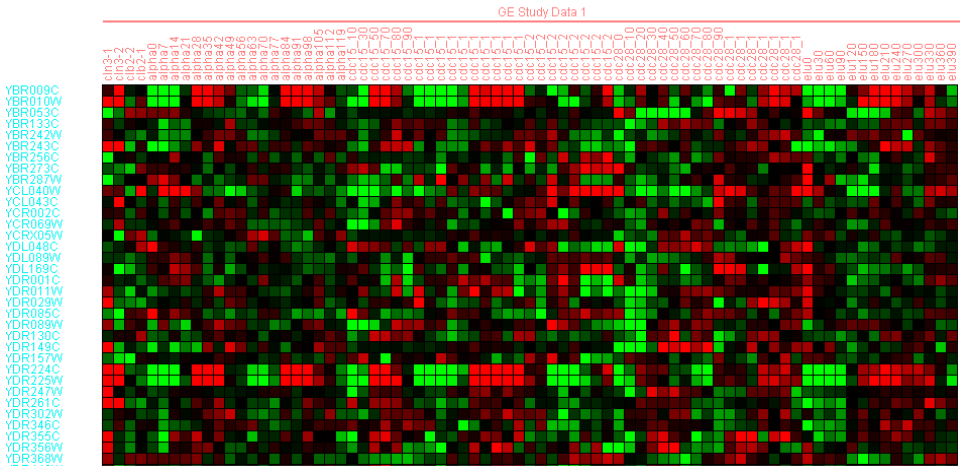


Figura 5.2: (a) *Heat map* de alguns genes pertencentes ao grupo 1. (b) *Heat map* de alguns genes pertencentes ao grupo 4 (captura da tela do programa Expander).

A validação por enriquecimento funcional do agrupamento quando $k = 4$ identifi-

cou um total de 12 funções, mas somente em 3 grupos, conforme ilustrado na Figura 5.3. No grupo 1, com 217 genes foi identificada a função de localização e transporte de cálcio. No grupo 3, com 304 genes, foram identificadas as funções de metabolismo do DNA, replicação do DNA e ciclo cromossomal, proliferação celular, recombinação do DNA, condensação do cromossomo, alongamento da fita de DNA, reparo do DNA, metabolismo primário e replicação do DNA.

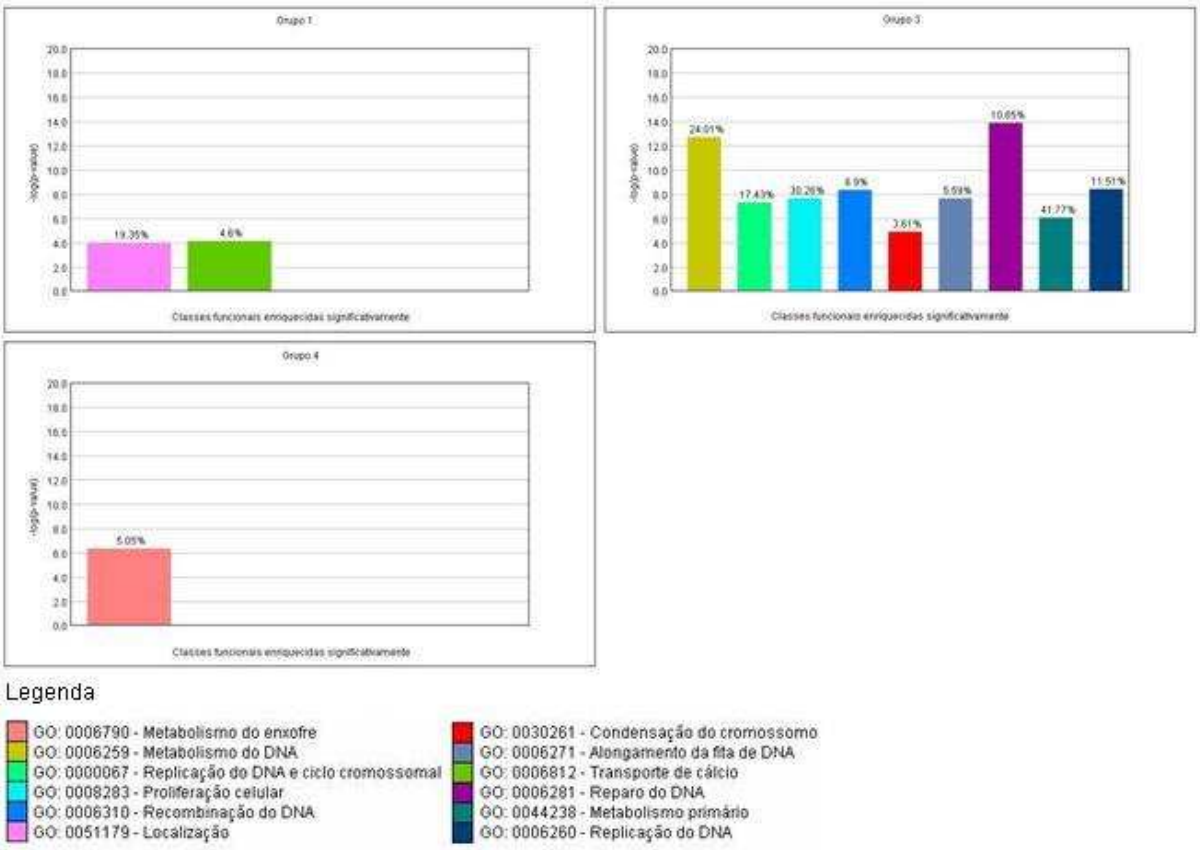


Figura 5.3: Enriquecimento funcional do agrupamento quando $k = 4$.

A Figura 5.4 apresenta o resultado da técnica de validação biológica de identificação de fatores de transcrição. Foi identificado um total de 3 fatores de transcrição, mas somente nos grupos 1 e 3.

No grupo 1, representativo da fase M, foi identificado o fator MCM1. No grupo 3, representativo da fase G1, foram identificados os fatores MBP1 e STB1.

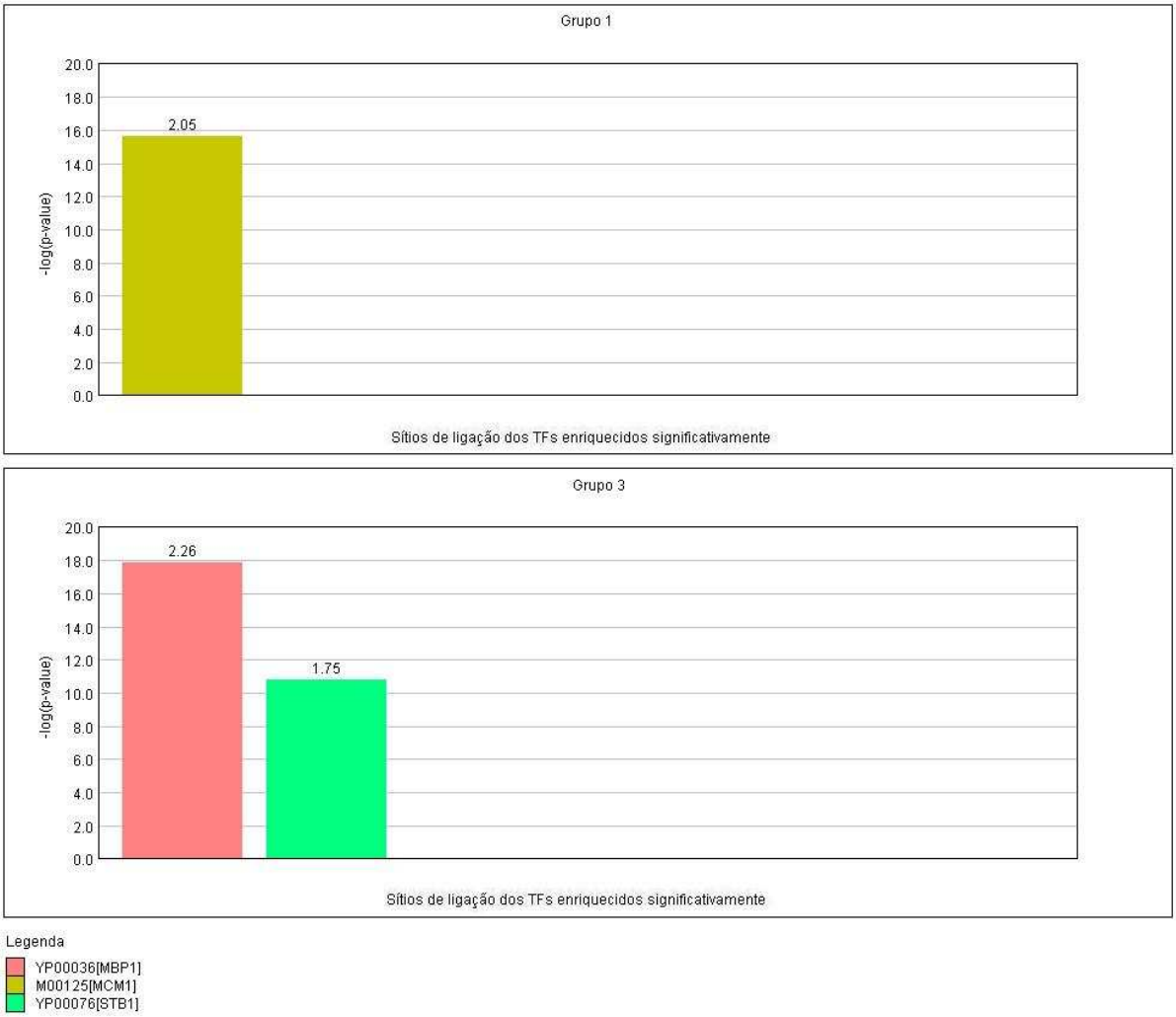


Figura 5.4: Identificação de fatores de transcrição no agrupamento quando $k = 4$.

5.2 Conclusão do capítulo

Neste capítulo foram apresentados em detalhes os resultados e as discussões obtidos do agrupamento do algoritmo k-médias quando $k = 4$ da base de dados CCSc. Para evitar tornar este capítulo muito extenso, os demais resultados dos agrupamentos dos diferentes algoritmos e bases de dados são disponibilizados no Apêndice A. Ainda no Apêndice A, no final da seção de cada base de dados, é apresentada uma conclusão com a comparação da performance dos algoritmos de agrupamento utilizados, do uso ou não de filtros de dados e a comparação dos resultados das técnicas de validação estatísticas e biológicas.

O próximo capítulo apresenta a conclusão geral do trabalho e as perspectivas de trabalhos futuros.

Capítulo 6

Conclusão

Técnicas de agrupamento têm sido adotadas como ferramenta padrão na análise de dados de expressão gênica.

Neste trabalho foram utilizados os algoritmos k-médias e SOM da abordagem de agrupamento unidimensional, o algoritmo SAMBA da abordagem bidimensional e técnicas de validação estatística e biológica para identificar a melhor solução de agrupamento.

Um resultado bastante útil foi a possibilidade de inferir funções biológicas para genes com função até então desconhecida, mas agrupados com genes de função conhecida. É bem possível que genes pertencentes a um mesmo grupo estejam envolvidos em um mesmo processo biológico porque eles devem ter demonstrado expressão co-regulada no microarranjo.

Outro resultado interessante e inesperado foi a possibilidade de associar funções biológicas e fatores de transcrição às fases G1, M, G2 e S do ciclo celular. Esta associação foi possível graças ao trabalho em conjunto com o algoritmo de agrupamento bidimensional.

A maioria dos trabalhos de agrupamento de dados de expressão gênica utiliza índices estatísticos para a definição da melhor solução de agrupamento. O problema é que estes índices podem não refletir um significado biológico. Por esta razão, o objetivo deste trabalho foi aferir a significância destes índices com a análise biológica detalhada dos grupos em conjunto com a aplicação de técnicas de validação biológica.

A aplicação dessas duas diferentes abordagens de validação em diversos agrupamentos levou à conclusão de que os 7 índices estatísticos dificilmente indicam a mesma solução de agrupamento. O índice C foi o índice que melhor corroborou a validação biológica, partindo do princípio de que a melhor solução, neste caso, é aquela em que foi associada maior quantidade de funções biológicas e fatores de transcrição.

Quanto aos algoritmos de agrupamento, os resultados obtidos revelaram a eficácia

do algoritmo SOM com relação ao k-médias, que demonstrou ser sensível a ruídos. A aplicação de filtros nas bases dados demonstrou ser um recurso eficiente. Garantiu melhor desempenho dos algoritmos, melhores resultados do k-médias, além de que a quantidade reduzida de dados facilitar a análise dos agrupamentos sem comprometer a significância biológica dos grupos, embora a redução de quantidade de genes resulte na redução da quantidade de funções biológicas e fatores de transcrição.

O algoritmo da abordagem de agrupamento bidimensional apresentou resultados satisfatórios, condizentes com a questão biológica fundamental de que um gene possa participar de diferentes processos biológicos e, portanto, pertencer a diferentes grupos. Esta abordagem, ainda pouco utilizada na análise de dados de expressão gênica, potencializa a identificação de estruturas nos dados, não perceptíveis pelas abordagens tradicionais de agrupamento, por isso revelou-se como uma boa opção para o trabalho em conjunto com os algoritmos da abordagem unidimensional. A aplicação de filtros de dados não apresentou vantagens com o algoritmo bidimensional.

A perspectiva de trabalho futuro é baseada no desenvolvimento de técnicas de validação de agrupamento associadas à técnicas de validação estatística. Um trabalho interessante seria uma ferramenta que transformasse a significância biológica e estatística em um único índice, minimizando a necessidade da análise subjetiva dos dados, idéia também válida para a aplicação nos resultados do agrupamento bidimensional.

Para a área biológica, uma sugestão de trabalho futuro consiste na comprovação das funções biológicas e fatores de transcrição que foram associadas a cada fase do ciclo celular de *S. cerevisiae*.

As conclusões dos algoritmos de agrupamento, aplicação de filtros de dados e técnicas de validação resultantes deste trabalho serão empregadas nas análises dos dados de microarranjo do IBMP. A utilização do organismo modelo *S. cerevisiae* foi muito importante para a validação de todo o processo de análise dos resultados e será bastante útil para ser empregado nas bases de dados do IBMP, por serem, na maioria, de organismos pouco conhecidos, como o *Trypanossoma cruzi*, por exemplo.

Referências Bibliográficas

[AB] F. Azuaje and N. Bolshakova. Clustering genomic expression data: Design and evaluation principles.

[AB84] M. S. Aldenderfer and R. K. Blashfield. *Cluster Analysis*. CA: Sage, 1984.

[ABDY99] R. Shamir A. Ben-Dor and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3/4):281–297, 1999.

[ABE06] ABEM. Arquivos brasileiros de endocrinologia e metabologia. disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0004-27302002000400006>, acesso em: 09/03/2006.

[AKJF99] M. N. Murty A. K. Jain and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, 1999.

[Alb04] B. Alberts. *Biologia Molecular da Célula*. Porto Alegre: Artmed, 4 edition, 2004.

[Ami04] N. Amit. The bicluster graph editing problem. Master's thesis, 2004.

[And04] L. P. Andrade. *Procedimento Interativo de Agrupamento de Dados*. PhD thesis, Universidade Federal do Rio de Janeiro, COPPE, 2004.

[APZ06] P. Zimmermann A. Wille P. Bühlmann W. Gruissem L. Hennig L. Thiele A. Prelić, S. Bleuler and E. Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129, 2006.

[ATS02] R. Sharan A. Tanay and R. Shamir. Discovering statistically significant biclusters in gene expression data, 2002.

[ATS04] R. Sharan A. Tanay and R. Shamir. Biclustering algorithms: A survey, 2004.

[Ber02] P. Berkhin. Survey of clustering data mining techniques. Technical report, Accure Software, San Jose, CA, 2002.

[BJB+00] S. Brenner, M. Johnson, J. Bridgham, G. Golda andD.H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, R. Roth, D. George, S. Eletr, G. Albrecht, E. Vermaas, S.R. Williams, K. Moon, T. Burcham, M. Pallas, R.B. DuBridge, J. Kirchner, K. Fearon, J. Mao, and K. Corcoran. Gene expression analysis by massively parallel signature sequencing (mpss) on microbead arrays. *Nature Biotechnology*, pages 630–634, 2000.

[Bol06] N. Bolshakova. Machaon clustering and validation environment. disponível em: <https://www.cs.tcd.ie/Nadia.Bolshakova/Machaon.html>, acesso em: 21/07/2006.

[BS04] T. BeiBbarth and T. P. Speed. Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465, 2004.

[BSEL01] S. Landau B. S. Everitt and M. Leese. *Cluster Analysis*. Arnold Publishers, May 2001.

[BZS04] S. Kirov B. Zhang, D. Schmoyer and J. Snoddy. Gotree machine (gotm): a web-based platform for interpreting sets of interesting genes using gene ontology hierarchies. *BMC Bioinformatics*, 5:16, 2004.

[CAHR00] C. Rosenow C. A. Harrington and J. Retief. Monitoring gene expression using dna microarrays. *Curr. Opin. Microbol.*, 3:285–291, 2000.

[CC00] Y. Cheng and G. M. Church. Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol*, 8:93–103, 2000.

[Con01] The Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Research*, 11:1425–1433, 2001.

[DB79] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 1:224–227, 1979.

[D'H05] P. D'Haeseleer. How does gene expression clustering work? *Nature Biotechnology*, 23(12):1499–1501, 2005.

[DMJ04] E. Remy P. Mouren D. Thieffry D. Martin, C. Brun and B. Jacq. Gotoolbox: functional analysis of gene datasets based on gene ontology. *Genome Biology*, 5(12), 2004.

[Dra03] S. Draghici. *Data Analysis Tools for DNA Microarrays*. Chapman & Hall/CRC, June 2003.

[Dun74] J. Dunn. Well separated clusters and optimal fuzzy partitions. *J. Cybernetics*, 4:95–104, 1974.

[ea06] R. Shamir et al. Expander. disponível em: <http://www.cs.tau.ac.il/ rshamir/expander/expander.html>, acesso em: 21/07/2006.

[Eis06a] M. Eisen. Cluster v.3.0. disponível em: <http://bonsai.ims.u-tokyo.ac.jp/m̃dehoon/software/cluster/software.htm>, acesso em: 09/03/2006.

[Eis06b] M. Eisen. Eisen lab. disponível em: <http://rana.lbl.gov/EisenData.htm>, acesso em: 21/07/2006.

[FACFLFC05] K. Faceli and M. C. P. Souto A. C. F. L. F. Carvalho. Algoritmos de agrupamento de dados. *Instituto de Ciências Matemáticas e de Computação*, 2005.

[FASD04] R. Diaz-Uriarte F. Al-Shahrour and J. Dopazo. Fatigo: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics*, 20(4):578–580, 2004.

[Gen06a] Genetique. Genetique et biotech. disponível em: <http://www.colloutao.qc.ca/bio/Imagebiologie/Imagegenetique/biotech.htm>, acesso em: 09/03/2006.

[Gen06b] Genome. The human genome. Disponível em: <http://genome.wellcome.ac.uk/doc_WTD020745.html>, Acesso em: 10 abr. 2006.

[GGD] E. Levine G. Getz and E. Domany. Coupled two-way clustering analysis of gene microarray data.

[GGD00] E. Levine G. Getz and E. Domany. Coupled two-way clustering analysis of gene microarray data. *Proc Natl Acad Sci U S A*, 97(22):12079–12084, October 2000.

[GO06] GO. Gene ontology. disponível em: <http://www.geneontology.org>, acesso em: 15/07/2006.

[Har72] J. A. Hartigan. Direct clustering of a data matrix. *Journal of American Statistical Association*, 67(337):123–129, 1972.

[HLTH05] W. J. Krzanowski H. L. Turner, T. C. Bailey and C. A. Hemingway. Biclustering models for structured microarray data. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 2(4):316–329, 2005.

[Hru01] E. R. Hruschka. Algoritmos genéticos de agrupamento para extração de regras de redes neurais. Master's thesis, Universidade Federal do Rio de Janeiro, COPPE, 2001.

[HS76] L. Hubert and J. Schultz. Quadratic assignment as a general data-analysis strategy. *British Journal of Mathematical and Statistical Psychologie*, 29:190–241, 1976.

[HS00] E. Hartuv and R. Shamir. A clustering algorithm based on graph connectivity. *Information Processing Letters*, 76(4–6):175–181, 2000.

[JHK05] J. D. Knowles J. Handl and D. B. Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212, 2005.

[JLSR05] A. Podhorski E. Guruceaga J. M. Mato L. A. Martinez-Cruz F. J. Corrales J. L. Sevilla, V. Segura and A. Rubio. Correlation between gene expression and go semantic similarity. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 2(4):330–338, 2005.

[KBR06] KBRIN. Kentucky biomedical research infrastructure network. Disponível em: <http://www.kbrin.louisville.edu/archives/fellows/dobbins.html>, Acesso em: 2006.

[KD05] P. Khatri and S. Draghici. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–3595, 2005.

[Koh01] T. Kohonen. *Self-Organizing Maps*, volume 30 of Springer Series in Information Sciences. Springer, Berlin, 3rd edition, 2001.

[LY05] H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans. Knowl. Data Eng.*, 17(4):491–502, 2005.

[MAS00] J. A. Blake D. Botstein H. Butler J. M. Cherry-A. P. Davis K. Dolinski S. S. Dwight J. T. Eppig M. A. Harris D. P. Hill L. Issel-Tarver A. Kasarskis S. Lewis J. C. Matese J. E. Richardson M. Ringwald G. M Rubin M. Ashburner, C. A. Ball and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25(1):25–29, May 2000.

[MBEB98] P. O. Brown M. B. Eisen, P. T. Spellman and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–14868, December 1998.

[MHV01] Y. Batistakis M. Halkidi and M. Vazirgiannis. On clustering validation techniques. *J. Intell. Inf. Syst.*, 17(2-3):107–145, 2001.

[MJLDHM04] J. Nolan M. J. L. De Hoon, S. Imoto and S. Miyano. Open source clustering software. *Bioinformatics*, 20(9):1453–1454, 2004.

[MO04] S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 1(1):24–45, 2004.

[MS94] J. Meidanis and J. C. Setubal. Uma introducao a biologia computacional. *Recife, PE, Brasil: Universidade Federal de Pernambuco*, 1994.

[MSB95] Davis R. W. M. Schena, D. Shalon and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, October 1995.

[Nat06] Nature. Nature genetics. disponível em: <http://www.nature.com/cgi-taf/dynapage.taf?file=/ng/journal/v21/n1s/ index.html e http://www.nature.com/cgi-taf/dynapage.taf?file=/ng/journal/ v32/n4s/index.html>, acesso em: 2006.

[NBC05] F. Azuaje N. Bolshakova and P. Cunningham. An integrated tool for microarray data clustering and cluster validity assessment . *Bioinformatics*, 21(4):451–455, 2005.

[ol06] Master of life. Dna - master of life. disponível em: <http://www.il.mahidol.ac.th/course/dna/chapter/images/editable/expression_dogma.jpg>, acesso em: 21/07/2006.

[PF99] E. J. Pauwels and G. Frederix. *Finding salient regions in images: nonparametric clustering for image segmentation and grouping*, volume 75. 1999.

[PTG99] J. Mesirov Q. Zhu S. Kitareewan E. Dmitrovsky E. S. Lander P. Tamayo, D. Slonim and T. R. Golub. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*, 96(6):2907–2912, 1999.

[PTSea98] G. Sherlock P. T. Spellman and et al. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization, 1998.

[RES] R. Sharan R. Shamir R. Elkon, C. Linhart and Y. Shiloh. Genome-wide in-silico identification of transcriptional regulators controlling cell cycle in human cells.

[Rou87] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comp App. Math*, 20:53–65, 1987.

[SF04] N. H. Shah and N. V. Fedoroff. Clench: a program for calculating cluster enrichment using the gene ontology. *Bioinformatics*, 20(7):1196–1197, 2004.

[Slo02] D. Slonim. From patterns to pathways: gene expression data analysis comes of age. *Nature Genetics*, 32:502 – 508, 2002.

[SMKT+05] R. Shamir, A. Maron-Katz, A. Tanay, C. Linhart, I. Steinfeld, R. Sharan, Y. Shiloh, and R. Elkon. Expander - an integrative program suite for microarray data analysis. *BMC Bioinformatics*, 6:232, 2005.

[SS01a] R. Shamir and R. Sharan. Algorithmic approaches to clustering gene expression data. In T. Jiang T. Smith Y. Xu and M. Q. Zhang, editors, *Current Topics in Computational Biology*. MIT press, 2001.

[SS01b] R. Shamir and R. Sharan. Algorithmic approaches to clustering gene expression data. In T. Jiang, T. Smith, Y. Xu, and M. Q. Zhang, editors, *Current Topics in Computational Biology*. MIT press, 2001. To appear.

[STC99] M. J. Campbell R. J. Cho S. Tavazoie, J. D. Hughes and G. M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22(3):281–285, July 1999.

[TCS+01] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, June 2001.

[THB00] M. B. Eisen A. Alizadeh R. Levy L. Staudt W. C. Chan D. Botstein T. Hastie, R. Tibshirani and P. Brown. 'gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology*, 1:1–21, 2000.

[TRA06] TRANSFAC. Transfac. Disponível em: <http://www.Gene-regulation.com/pub/databases.html#transfac>, Acesso em: 15/07/2006.

[UAL99] D. A. Notterman K. Gish S. Ybarra D. Mack U. Alon, N. Barkai and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A*, 96(12):6745–6750, June 1999.

[Uni06a] Stanford University. Saccharomyces genome database. disponível em: <http://genome-www.stanford.edu/Saccharomyces/>, acesso em: 09/03/2006.

[Uni06b] Stanford University. The yeast cell cycle analysis project. Disponível em: <http://cellcycle-www.stanford.edu>, Acesso em: 21/07/2006.

[VEVK95] B. Vogelstein V. E. Velculescu, L. Zhang and K. W. Kinzler. Serial analysis of gene expression. *Science*, 270:484–487, 1995.

[VEVW95] B. Vogelstein V. E. Velculescu, L. Zhang and Kinzler K. W. Serial analysis of gene expression. *Science*, 270(5235):484–487, October 1995.

[WK05] C. J. Wu and S. Kasif. Gems: a web server for biclustering analysis of expression data. *Nucleic Acids Research*, 33(Web Server issue):W596-9, 2005.

[WMFV99] S. J. Walker W. M. Freeman and K. E. Vrana. Quantitative rt-pcr: pitfalls and potential. *Biotechniques*, 26(1):112–22, 124–5, Jan 1999.

[YKG03] J. T. Chang Y. Kluger, R. Basri and M. Gerstein. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res*, 13(4):703–716, April 2003.

[Zah03] A. Zaha. *Biologia Molecular Básica*. Mercado Aberto, 3 edition, 2003.

Apêndice A

Resultados

No capítulo Resultados e Discussões foi apresentada somente a análise do agrupamento $k = 4$ para evitar tornar o capítulo muito extenso. Neste material são apresentadas as análises de todos os agrupamentos do algoritmo k-médias, SOM e SAMBA.

Os resultados são apresentados organizados por bases de dados e pelas opções das base de dados com e sem a aplicação de filtros. No final da análise de cada base de dados é apresentada uma conclusão, contendo a melhor solução de agrupamento de acordo com as técnicas de validação estatística e biológica, o algoritmo de agrupamento que apresentou melhores resultados e uma discussão das vantagens obtidas com a utilização dos filtros de dados.

A.1 Base de dados CCSc

A.1.1 Agrupamento k-médias

Com a aplicação do algoritmo k-médias nesta base de dados esperava-se a identificação de 4 grupos, correspondentes às 4 fases do ciclo celular. Por esta razão, neste capítulo é apresentada a análise detalhada do agrupamento $k = 4$. Os demais agrupamentos, obtidos quando *k = 2, 5, 8 e 10* são apresentados no Apêndice A.

k = 2

Os 799 genes da base de dados CCSc foram agrupados em 2 grupos, conforme mostrado na Tabela A.1.

Tabela A.1: Agrupamento $k = 2$.

Grupos	Quantidade de genes	Homogeneidade
1	430	0,332
2	369	0,565

k = 4

Os 799 genes da base de dados CCSc foram agrupados em 4 grupos, conforme mostrado na Tabela A.2.

Tabela A.2: Agrupamento $k = 4$.

Grupos	Quantidade de genes	Homogeneidade
1	217	0,51
2	1	1,0
3	304	0,604
4	277	0,344

k = 5

Os 799 genes da base de dados CCSc foram agrupados em 5 grupos, conforme mostrado na Tabela A.3.

Tabela A.3: Agrupamento $k = 5$.

Grupos	Quantidade de genes	Homogeneidade
1	217	0,509
2	1	1,0
Continua na próxima página		

Tabela A.3 – continuação da página anterior

3	475	0,318
4	4	0,609
5	102	0,721

k = 8

Os 799 genes foram agrupados em 8 grupos, conforme mostrado na Tabela A.4.

Tabela A.4: Agrupamento $k = 8$.

Grupos	Quantidade de genes	Homogeneidade
1	163	0,483
2	1	1,0
3	393	0,33
4	2	0,816
5	77	0,747
6	115	0,556
7	42	0,707
8	6	0,835

k = 10

Os 799 genes foram agrupados em 10 grupos, conforme mostrado na Tabela A.5.

Tabela A.5: Agrupamento $k = 10$.

Grupos	Quantidade de genes	Homogeneidade
1	163	0,483
2	1	1,0
3	375	0,314
4	2	0,816

Continua na próxima página

Tabela A.5 – continuação da página anterior

5	75	0,756
6	115	0,556
7	42	0,707
8	6	0,835
9	3	0,867
10	17	0,829

A.1.2 Agrupamento SOM

O algoritmo SOM foi aplicado com as dimensões da matriz definidas por 2x2, 5x1, 2x3, 2x4 e 2x5. Os resultados obtidos foram os seguintes:

SOM = 2x2

A quantidade de genes e a taxa de homogeneidade de cada grupo são apresentadas na Tabela A.6.

Tabela A.6: Agrupamento SOM = 2x2.

Grupos	Quantidade de genes	Homogeneidade
1	246	0,667
2	139	0,494
3	212	0,514
4	202	0,549

SOM = 5x1

A quantidade de genes e a taxa de homogeneidade de cada grupo são apresentadas na Tabela A.7.

Tabela A.7: Agrupamento SOM = 5x1.

Grupos	Quantidade de genes	Homogeneidade
1	209	0,479
2	180	0,566
3	114	0,533
4	251	0,622
5	45	0,778

SOM = 2x3

A quantidade de genes e a taxa de homogeneidade de cada grupo são apresentadas na Tabela A.8.

Tabela A.8: Agrupamento SOM = 2x3.

Grupos	Quantidade de genes	Homogeneidade
1	158	0,443
2	51	0,63
3	254	0,651
4	63	0,705
5	211	0,453
6	62	0,737

SOM = 2x4

A quantidade de genes e a taxa de homogeneidade de cada grupo são apresentadas na Tabela A.9.

Tabela A.9: Agrupamento SOM = 2x4.

Grupos	Quantidade de genes	Homogeneidade
1	43	0.673
2	87	0,571
3	219	0,581
4	59	0,815
5	35	0,797
6	162	0,48
7	162	0,526
8	32	0,791

SOM = 2x5

A quantidade de genes e a taxa de homogeneidade de cada grupo são apresentados na Tabela A.10.

Tabela A.10: Agrupamento SOM = 2x5.

Grupos	Quantidade de genes	Homogeneidade
1	53	0,699
2	159	0,647
3	75	0,429
4	76	0,508
5	41	0,674
6	50	0,831
7	29	0,8
8	135	0,567
9	143	0,527
10	38	0,79

A.1.3 Agrupamento SAMBA

A aplicação do algoritmo de agrupamento bidimensional na base de dados CCSc resultou em 50 grupos, conforme apresentado na Tabela A.11. O resultado do agrupamento é apresentado com um escore, o número de condições e genes atribuídos a cada grupo.

Tabela A.11: Agrupamento SAMBA.

Grupo	Escore	Condições	Genes
1	385,972	15	42
2	236,846	4	56
3	316,706	9	44
4	198,045	12	26
5	173,463	5	41
6	368,667	11	57
7	175,374	5	40
8	238,274	9	39
9	230,267	5	75
10	307,388	9	57
11	351,838	10	50
12	235,265	8	41
13	189,242	5	54
14	331,713	13	43
15	299,126	19	27
16	253,264	7	51
17	243,572	9	43
18	145,216	5	41
19	227,665	7	42
20	345,258	12	44
21	250,176	8	47
22	183,727	7	38
23	130,062	7	24
Continua na próxima página			

Tabela A.11 – continuação da página anterior

24	99,1136	7	19
25	63,585	5	18
26	167,818	14	21
27	152,237	10	22
28	124,676	10	17
29	115,122	7	17
30	135,142	9	18
31	340,747	14	39
32	133,47	5	40
33	163,942	10	26
34	187,246	7	36
35	99,6389	15	9
36	184,877	9	32
37	213,228	9	40
38	161,168	7	37
39	215,551	8	41
40	62,167	4	15
41	72,2183	7	12
42	123,173	9	21
43	119,11	6	20
44	215,642	8	42
45	127,966	14	15
46	206,026	4	49
47	186,479	8	29
48	74,6247	5	22
49	267,666	11	42
50	249,156	15	25

A.1.4 Validação estatística k-médias

A Tabela A.12 a seguir contém os valores obtidos das técnicas de validação estatística aplicadas nos resultados do algoritmo k-médias.

Tabela A.12: Validação estatística dos agrupamentos k-médias.

k	Homogeneidade	Separação	C	D. Bouldin	Dunn	Silhueta	Isolamento
2	0,431	**-0,094**	0,341	1,761	**1,088**	**0,143**	**0,863**
4	**0,491**	-0,049	0,287	1,765	0,702	0,087	0,526
5	0,364	-0,017	0,257	1,749	0,672	0,101	0,392
8	0,38	0,006	0,236	**1,72**	0,605	0,058	0,365
10	0,372	0,012	**0,206**	1,768	0,579	0,06	0,352

Nos índices de separação, C e Davies Bouldin, o menor valor do índice corresponde ao melhor agrupamento. Para os índices de homogeneidade, Dunn, Silhueta e Isolamento, o melhor agrupamento corresponde ao maior valor do índice.

Os índices de homogeneidade e separação foram obtidos do programa Expander. Segundo o índice de homogeneidade o melhor agrupamento foi $k = 4$ e de acordo com o índice de separação, o Dunn, Silhueta, e de Isolamento, $k = 2$. O melhor agrupamento para o índice C foi $k = 10$. O índice Davies Bouldin identificou o melhor agrupamento $k = 8$.

No contexto biológico, o agrupamento mais significativo seria $k = 4$, por causa das 4 fases do ciclo celular. Situações como esta, evidenciam a necessidade da utilização de outros recursos que enriqueçam a análise de uma solução de agrupamento. Os resultados das técnicas de validação biológica são apresentados após os resultados das técnicas de validação estatística.

A.1.5 Validação estatística SOM

A Tabela A.13 a seguir contém os valores obtidos das técnicas de validação estatística aplicadas nos agrupamentos resultantes do algoritmo SOM.

Tabela A.13: Validação estatística dos agrupamentos SOM.

k	Homogeneidade	Separação	C	D. Bouldin	Dunn	Silhueta	Isolamento
2x2	0,576	**-0,045**	0,231	1,787	0,865	**0,101**	**0,672**
5x1	0,554	-0,031	0,188	**1,663**	0,828	0,094	0,579
2x3	0,557	-0,024	0,164	1,759	**0,882**	0,1	0,516
2x4	0,557	-0,004	**0,142**	1,696	0,755	0,083	0,499
2x5	**0,588**	0,008	0,144	1,814	0,733	0,065	0,461

Todas as dimensões da matriz definidas para o SOM foram indicadas como melhor agrupamento por um dos índices de validação estatística.

A matriz 2x2, no entanto, foi indicada como melhor agrupamento por três índices diferentes: separação, Silhueta e Isolamento, resultado esperado conforme as quatro fases do ciclo celular.

A.1.6 Validação estatística SAMBA

Os índices de validação estatística não foram aplicados nos resultados do agrupamento bidimensional porque, além de não serem adequados para esta abordagem (conforme descrito na seção 3.6.8), o propósito da utilização do algoritmo bidimensional é comparar com os resultados dos agrupamentos unidimensionais.

O escore atribuído a cada grupo bidimensional é dependente do tamanho do grupo e, por isso, não é recomendada sua utilização para comparar a qualidade de dois grupos diferentes [SMKT$^+$05].

A.1.7 Validação biológica k-médias

k = 2

De acordo com a maioria das técnicas de validação estatística, o melhor agrupamento foi $k = 2$.

A Figura A.1 mostra o perfil médio da expressão dos genes dos 2 grupos, indicado por uma linha vermelha. O eixo X corresponde às condições, representadas por números (de 1 a 77) e o eixo Y corresponde aos níveis de expressão. As barras pretas na vertical indicam o desvio padrão identificado em cada condição, sendo menor quando é atribuído

um valor maior de k, devido à maior homogeneidade entre os elementos do grupo (ver o perfil médio da expressão dos genes dos agrupamentos quando $k = 4, 5, 8\ e\ 10$).

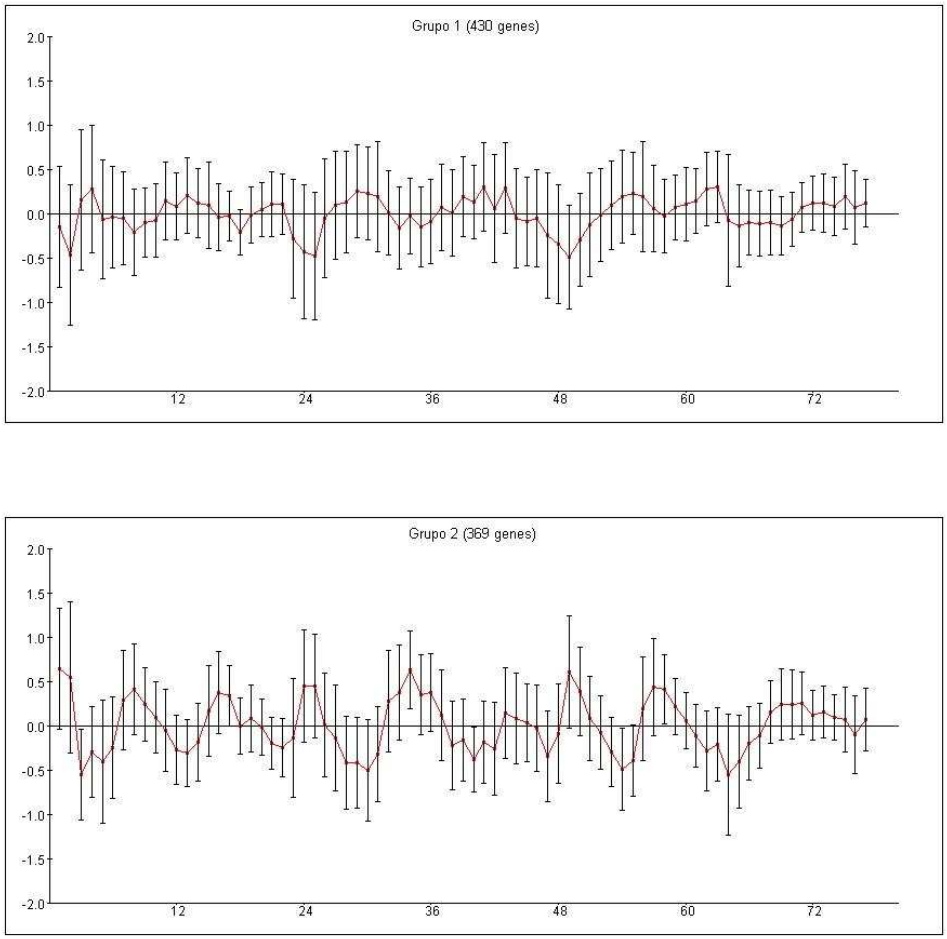


Figura A.1: Perfil médio da expressão dos genes dos 2 grupos ($k = 2$).

O grupo 1 foi formado por genes envolvidos nas fases G2 e M. O grupo 2 foi formado genes envolvidos nas fases G1 e S. Os dois grupos resultantes são potencialmente significativos, considerando a ordem que ocorre as etapas do ciclo celular.

A Tabela A.14 apresenta as funções biológicas e fatores de transcrição associados a cada grupo do agrupamento $k = 2$.

Tabela A.14: Validação biológica do agrupamento $k = 2$.

Grupos	TANGO	PRIMA
1	GO:0051179 - Localização	AZF1
2	GO:0006259 - Metabolismo do DNA	MBP1
	GO:0000067 - Replicação do DNA e ciclo cromossomal	STB1
	GO:0008283 - Proliferação celular	
	GO:0006310 - Recombinação do DNA	
	GO:0030261 - Condensação do cromossomo	
	GO:0006271 - Alongamento da fita do DNA	
	GO:0006281 - Reparo do DNA	
	GO:0044238 - Metabolismo primário	

O grupo 2 foi o grupo que mais refletiu funções biológicas, porque as fases G1 e S são as fases de maior atividade do ciclo além de serem as mais demoradas. A fase G1 dura 12 horas e a fase S dura 7 a 8 horas, G2 (3 a 4 horas) e M (1 a 2 horas). Na G1 ocorre o estímulo do crescimento da célula e progressão do ciclo para a fase S. A fase S é caracterizada pela replicação (síntese) do DNA.

Informações mais detalhadas das funções identificadas nos grupos, são encontradas no site do *Gene Ontology* [GO06]. Informações dos fatores de transcrição podem ser encontradas no site do banco de dados TRANSFAC [TRA06].

k = 4

Os resultados e discussões do agrupamento $k = 4$ foram apresentados no Capítulo 5. Este agrupamento foi utilizado como modelo para descrever o processo de análise adotado neste trabalho. Por isso neste material os resultados desse agrupamento não serão discutidos, a não ser quando comparados com os resultados de outros agrupamentos.

k = 5

Tendo em vista os resultados obtidos do agrupamento $k = 4$, esperava-se que quando $k = 5$ fossem obtidos 5 grupos, um deles contendo o gene CLN3 e os demais correspondentes às 4 fases do ciclo celular.

O grupo 1 foi formado por genes da fase M (G2/M e M/G1), o grupo 2 pelo gene CLN3, o grupo 3 por genes das fases G1, S, G2 e M. O grupo 4 foi formado por genes das fases G1, S e M (M/G1) e o grupo 5 por genes da fase G1.

A Tabela A.15 apresenta as funções biológicas e fatores de transcrição associados a cada grupo do agrupamento $k = 5$.

Tabela A.15: Validação biológica do agrupamento $k = 5$.

Grupos	TANGO	PRIMA
1	GO:0006812 - Transporte de cálcio GO:0051179 - Localização	MCM1
3	GO:0005200 - Constituinte estrutural do citoesqueleto	
5	GO:0000082 - Transição da fase G1/S do ciclo celular mitótico GO:0006260 - Replicação do DNA GO:0006974 - Resposta à estímulos de danos ao DNA GO:0006259 - Metabolismo do DNA GO:0008283 - Proliferação celular	MBP1 STB1

A tentativa da definição de $k = 5$ não adicionou nenhuma informação, se comparada aos resultados obtidos do agrupamento $k = 4$.

k = 8

A Figura A.2 ilustra o perfil médio da expressão dos genes dos 8 grupos. Conforme a quantidade de grupos aumenta, o desvio padrão dos perfis de expressão diminui.

O agrupamento $k = 8$ formou grupos de baixa homogeneidade e pouco representativos. O grupo 1 foi formado por genes das fases M (M/G1 e G2/M). O agrupamento manteve o grupo 2 com o gene CLN3, o grupo 3 foi formado por genes das fases G1, S e G2. O grupo 4 por genes da fase S, os grupos 5 e 6 por genes da fase G1, os grupos 7 e 8 por genes da fase M (G2/M).

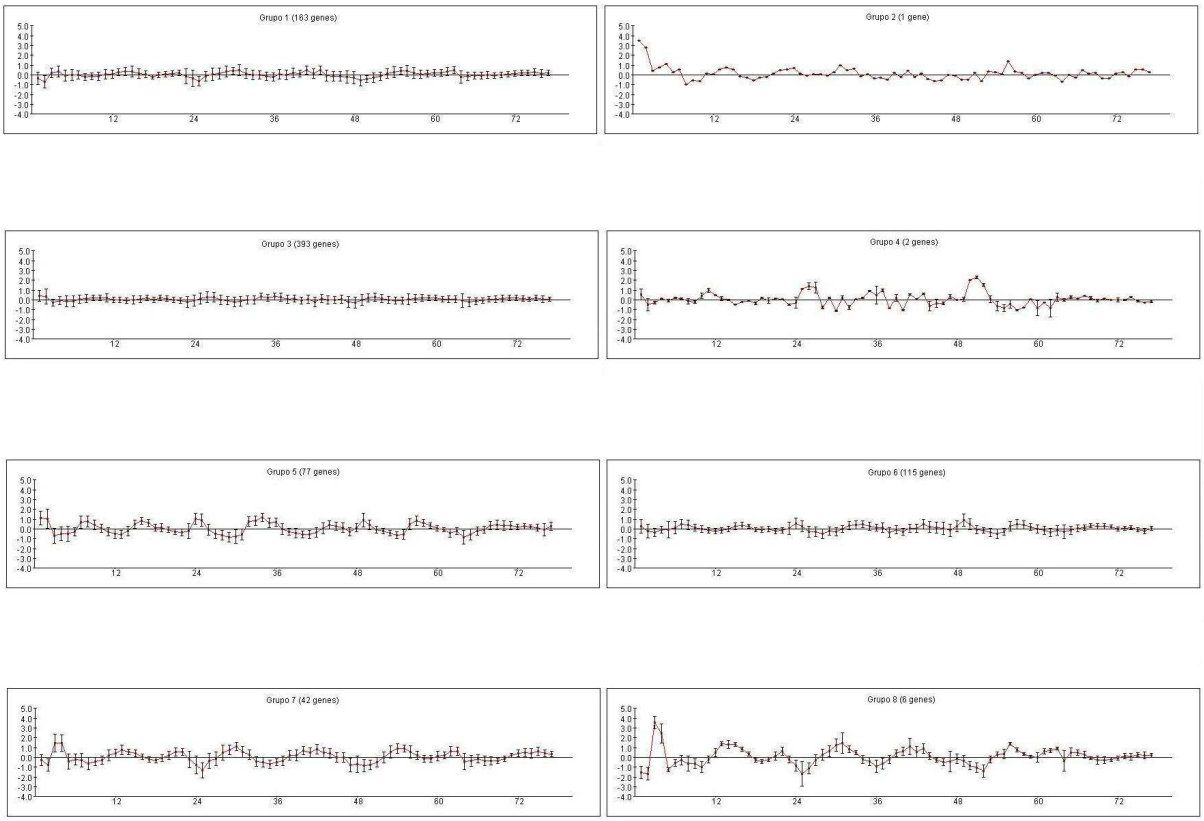


Figura A.2: Perfil médio da expressão dos genes dos 8 grupos ($k = 8$).

A Tabela A.16 apresenta as funções biológicas e fatores de transcrição associados a cada grupo do agrupamento $k = 8$.

Tabela A.16: Validação biológica do agrupamento $k = 8$.

Grupos	TANGO	PRIMA
1	GO:0015082 - Atividade de transporte de cálcio inorgânico GO:0051179 - Localização GO:0006812 - Transporte de cálcio	MCM1
3		ABF1
5	GO:0006259 - Metabolismo do DNA GO:0008283 - Proliferação celular GO:0006974 - Resposta à estímulos de danos ao DNA GO:0006312 - Recombinação mitótica	MBP1 STB1
6	GO:0006259 - Metabolismo do DNA GO:0000067 - Replicação do DNA e ciclo cromossomal	MBP1
Continua na próxima página		

Tabela A.16 – continuação da página anterior

	GO:0030261 - Condensação do cromossomo GO:0006281 - Reparo do DNA	
7		MCM1
8		PHO4

Embora este agrupamento não tenha sido enriquecido com novas funções biológicas, foram identificados dois diferentes fatores de transcrição nos grupos: ABF1 e PHO4, conforme ilustrado na Tabela A.16.

Nos grupos 1 e 7, representativos da fase M, foi identificado o fator de transcrição MCM1, no grupo 3 das fases G1, S e G2, o fator ABF1, no grupo 5 da fase G1, os fatores MBP1 e STB1, no grupo 6, também da fase G1, o fator MBP1 e no grupo 8 representativo da fase M (G2/M), o fator PHO4.

k = 10

Este agrupamento foi identificado como o melhor pelo índice estatístico C. As técnicas de validação biológica corroboram esta indicação, conforme é apresentado nas análises a seguir.

A Figura A.3 ilustra o perfil médio da expressão dos genes dos 10 grupos.

Os dois primeiros grupos são os mesmos que no agrupamento $k = 8$. O grupo 1 formado por genes envolvidos na fase M (M/G1 e G2/M) e o grupo 2 contendo somente o gene CLN3. Também da mesma forma que o agrupamento $k = 8$, o grupo 3 apresentou baixa homogeneidade, com genes envolvidos em todas as fases do ciclo celular. O grupo 4 foi formado por somente 2 genes, ambos da fase S. O grupo 5 e 6 são representativos da fase G1, ambos de alta homogeneidade. O grupo 7 e 8 são também representativos da fase M (G2/M). O grupo 9 foi formado por 3 genes, sendo dois deles pertencentes à fase G1 e um à fase S e finalmente o grupo 10, contendo 17 genes, todos eles envolvidos na fase S do ciclo celular.

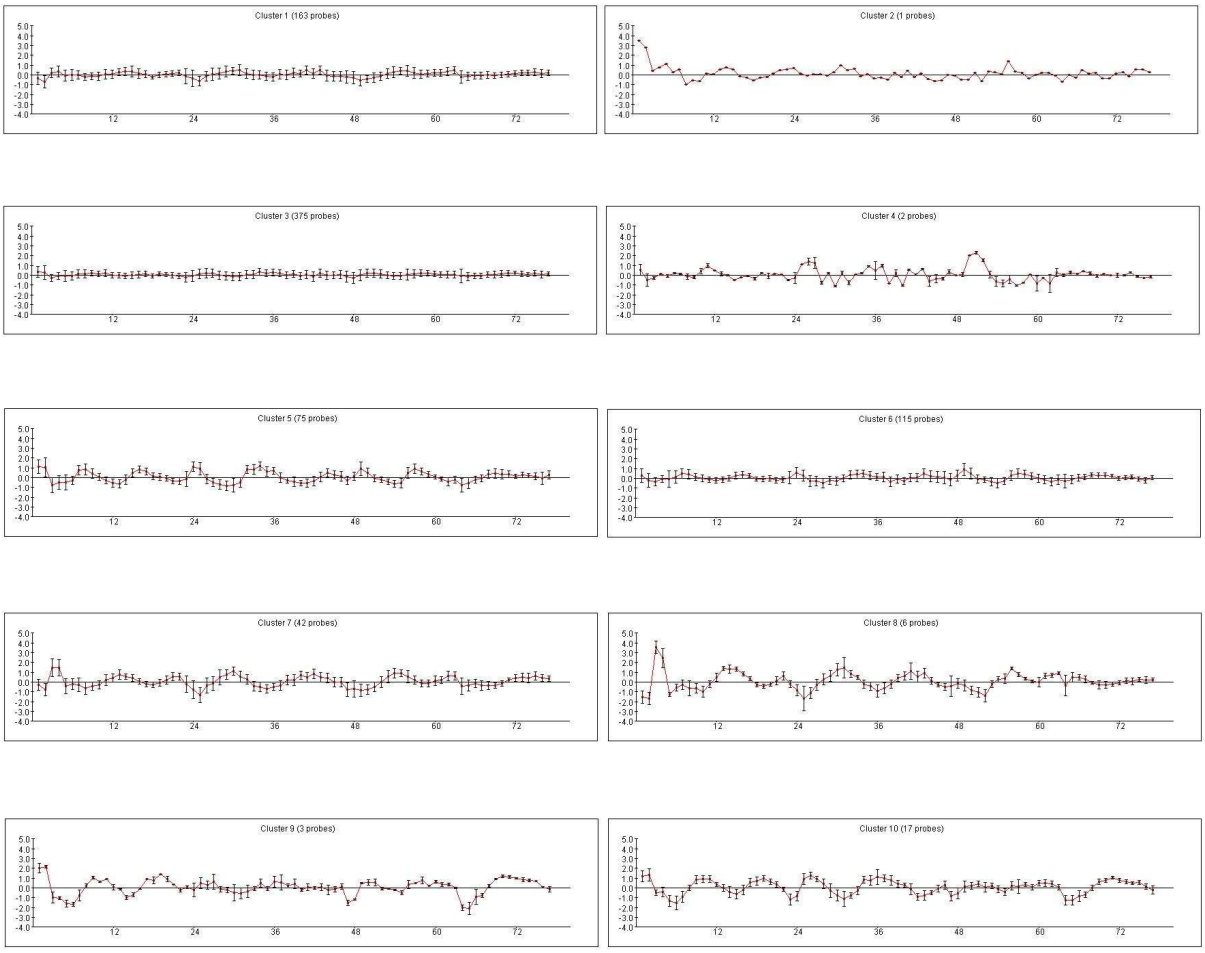


Figura A.3: Perfil da expressão dos genes dos 10 grupos ($k = 10$).

Este foi o único agrupamento que criou um grupo representativo da fase S, embora esses genes demonstrem alta co-regulação em todas as condições, conforme ilustrado na Figura A.4.

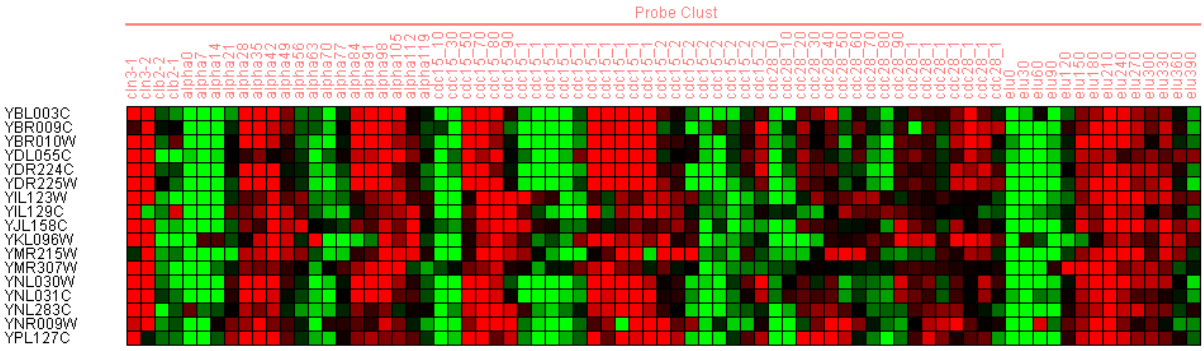


Figura A.4: *Heat map* do grupo 10 quando $k = 10$ (captura da tela do programa Expander).

A Tabela A.17 apresenta as funções biológicas e fatores de transcrição associados a cada grupo do agrupamento $k = 10$.

Tabela A.17: Validação biológica do agrupamento $k = 10$.

Grupos	TANGO	PRIMA
1	GO:0051179 - Localização GO:0006812 - Transporte de cálcio	MCM1
3	GO:0046467 - Biosíntese da membrana lipídica GO:0016192 - Transporte mediado-vesículo GO:0043170 - Metabolismo de macromolécula GO:0044249 - Biosíntese celular	ABF1
5	GO:0006259 - Metabolismo do DNA GO:0008283 - Proliferação celular GO:0006974 - Resposta à estímulos de danos ao DNA GO:0006312 - Recombinação mitótica GO:0006260 - Replicação do DNA	MBP1
6	GO:0006259 - Metabolismo do DNA GO:0000067 - Replicação do DNA e ciclo cromossomal GO:0030261 - Condensação do cromossomo	MBP1
Continua na próxima página		

Tabela A.17 – continuação da página anterior

	GO:0006281 - Reparo do DNA	
7		MCM1
8		PHO4
10	GO:0006333 - União ou separação da cromatina GO:0003677 - Ligação do DNA GO:0016043 - Organização celular e biogênese	HAC1

A.1.8 Validação biológica SOM

SOM = 2x2

O perfil da expressão dos genes de cada grupo é ilustrado na Figura A.5.

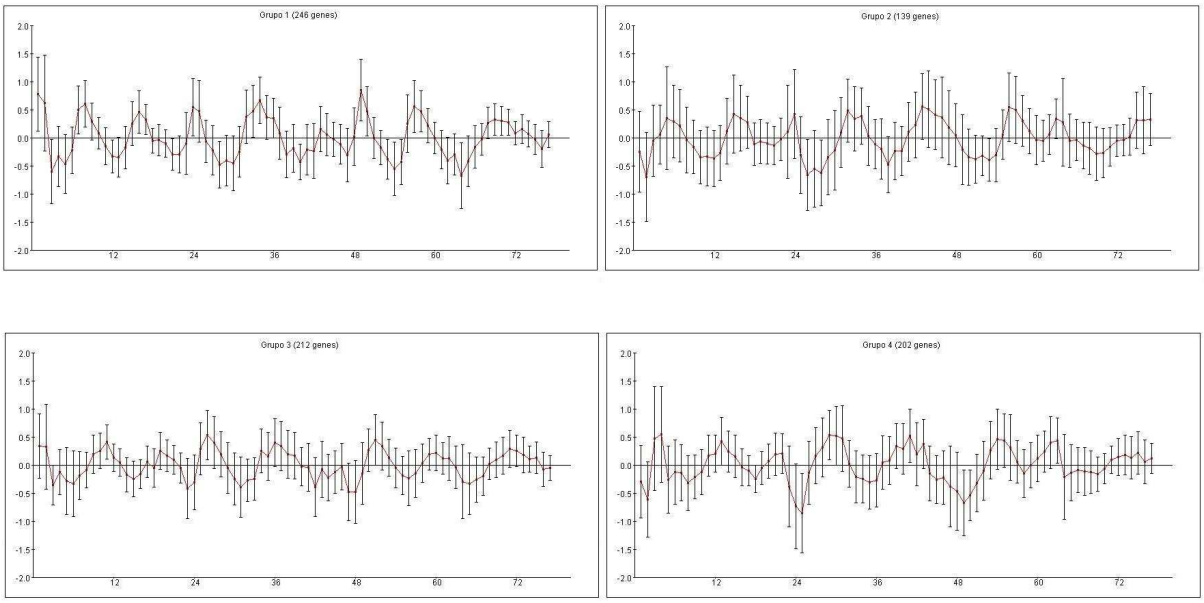


Figura A.5: Perfil médio da expressão dos genes dos 4 grupos (SOM = 2x2).

O grupo 1 foi formado por 246 genes envolvidos na fase G1 do ciclo celular O grupo 2 foi formado por 139 genes envolvidos na fase M, o grupo 3 foi formado por 212 genes envolvidos na fase S e finalmente o grupo 4, foi formado por genes da fase G2. Cada grupo é altamente representativo das 4 fases do ciclo celular, indicando o SOM como uma alternativa de algoritmo melhor do que o k-médias, do ponto de vista biológico.

A Tabela A.18 apresenta as funções biológicas e fatores de transcrição associados a cada grupo do agrupamento SOM = 2x2.

Tabela A.18: Validação biológica do agrupamento SOM = 2x2.

Grupos	TANGO	PRIMA
1	GO:0006259 - Metabolismo de DNA	MBP1
	GO:0007049 - Ciclo celular	STB1
	GO:0000067 - Replicação do DNA e ciclo cromossomal	
	GO:0006310 - Recombinação do DNA	
	GO:0030261 - Condensação do cromossomo	
	GO:0006271 - Alongamento da fita de DNA	
	GO:0006281 - Reparo do DNA	
	GO:0006260 - Replicação do DNA	
2	GO:0000749 - Resposta ao feromônio	MSN4
	GO:0007154 - Comunicação celular	
3	GO:0005200 - Metabolismo de enxofre	ABF1
	GO:0016043 - Organização celular e biogênse	
4	GO:0006812 - Transporte de cálcio	MCM1
	GO:0051179 - Localização	
	GO:0008324 - Atividade de transporte de cálcio	

Com o resultado deste agrupamento, sugere-se que o fator ABF1 seja específico dos genes da fase S. Além disso, de acordo com o banco de dados TRANSFAC, este fator é um ativador da replicação do DNA e da transcrição em levedura, atividades características da fase S do ciclo celular.

SOM = 5x1

A Figura A.6 ilustra o perfil da expressão dos genes dos 5 grupos. O resultado é muito semelhante do SOM = 2x2, inclusive as funções e os fatores de transcrição identificados.

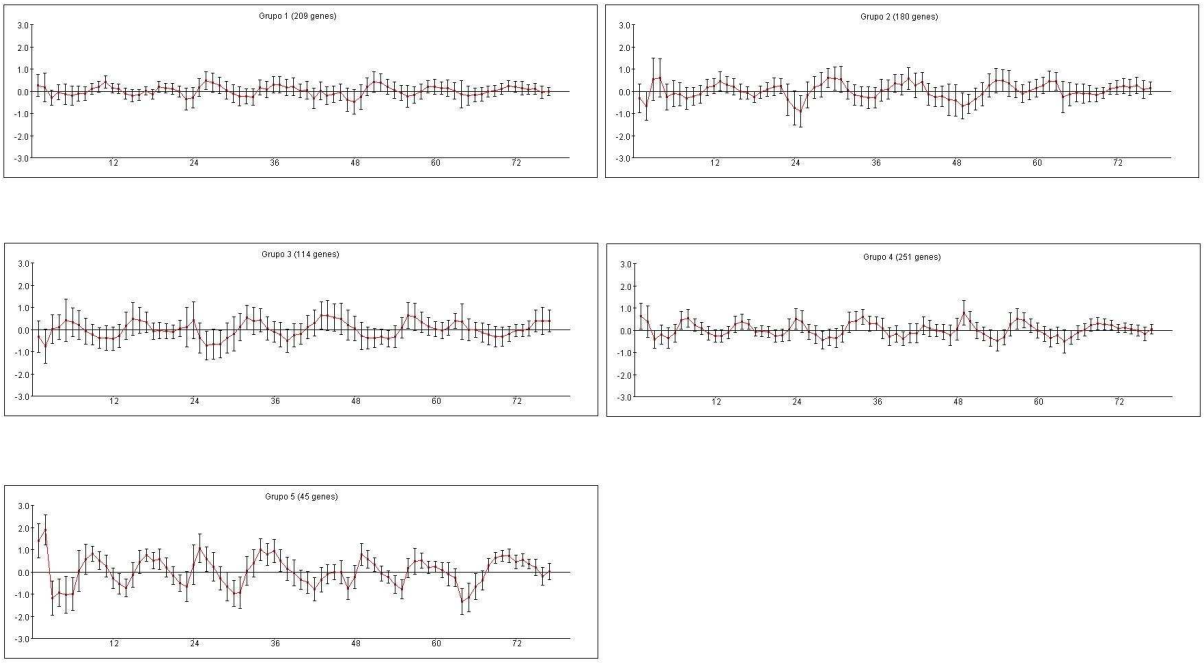


Figura A.6: Perfil médio da expressão dos genes dos 4 grupos (SOM = 5x1).

A Tabela A.19 apresenta as funções biológicas e fatores de transcrição associados a cada grupo do agrupamento SOM = 5x1.

Tabela A.19: Validação biológica do agrupamento SOM = 5x1.

Grupos	TANGO	PRIMA
1	GO:0006790 - Metabolismo de enxofre GO:0006520 - Metabolismo de aminoácido	GCN4
2	GO:0006812 - Transporte de cálcio GO:0051179 - Localização	MCM1
3	GO:0000749 - Resposta ao feromônio GO:0007154 - Comunicação celular	MSN4

Continua na próxima página

Tabela A.19 – continuação da página anterior

4	GO:0006259 - Metabolismo do DNA GO:0000067 - Replicação do DNA e ciclo cromossomal GO:0006310 - Recombinação do DNA GO:0030261 - Condensação do cromossomo GO:0006271 - Alongamento da fita de DNA GO:0004519 - Atividade de endonuclease GO:0006281 - Reparo do DNA	MBP1
5	GO:0006333 - União ou separação da cromatina GO:0003677 - Ligação do DNA	STB1

Esta diferente distribuição dos genes nos grupos reafirma a hipótese de que os fatores MBP1 e STB1 são fatores de transcrição dos genes da fase G1, o fator MCM1 da fase M (G2/M) e o fator MSN4 também da fase M mas de (M/G1). No grupo 1, representativo da fase S, foi identificado o fator GCN4, fator não identificado nos agrupamentos anteriores.

SOM = 2x3

A Figura A.7 ilustra o perfil médio da expressão dos genes dos 6 grupos. O grupo 1 contém os genes envolvidos na fase M (G2/M e M/G1), no grupo 2 os genes das fases M (M/G1) e G1, no grupo 3 os genes da fase G1, no grupo 4 os genes da fase M (G2/M), no grupo 5 os genes das fases S e G2 e no grupo 6 os genes das fases G1, S e G2.

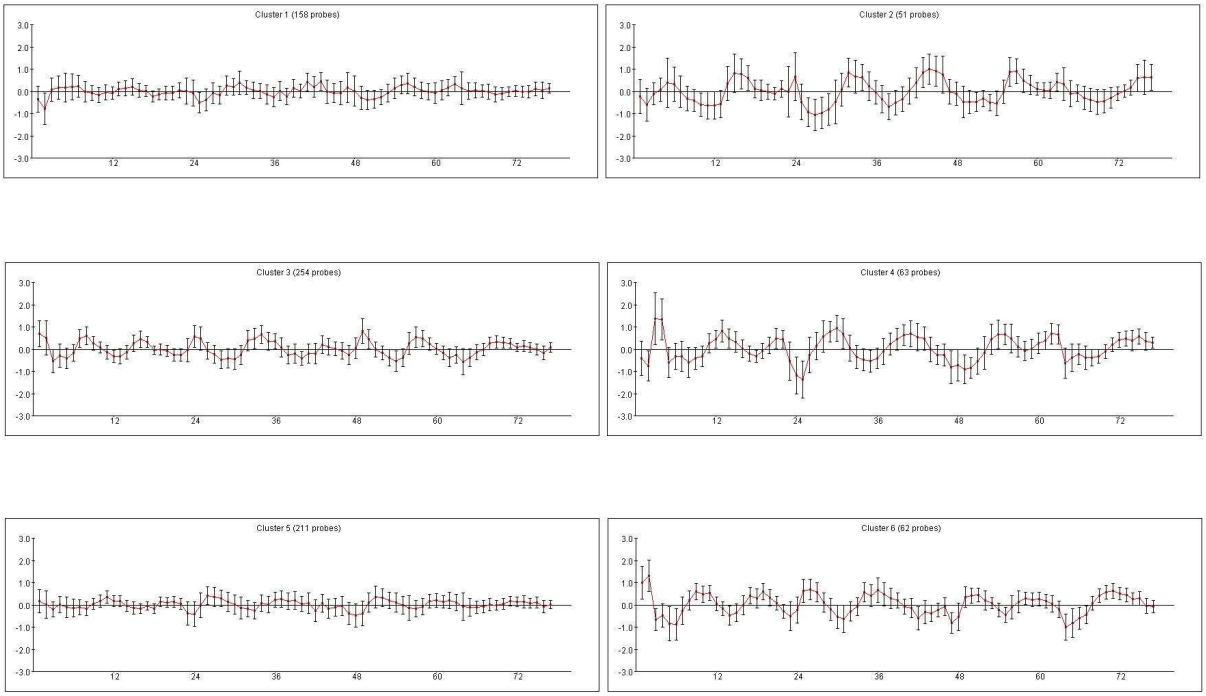


Figura A.7: Perfil da expressão dos genes dos 6 grupos (SOM = 2x3).

Somente 4 grupos foram enriquecidos funcionalmente. A Tabela A.20 apresenta as funções biológicas e fatores de transcrição associados a cada grupo do agrupamento SOM = 2x3.

Tabela A.20: Validação biológica do agrupamento SOM = 2x3.

Grupos	TANGO	PRIMA
1	GO:0006790 - Transporte de açúcar GO:0006520 - Localização	MCM1
2		ACE2 SWI5
3	GO:0006259 - Metabolismo de DNA GO:0007049 - Ciclo celular GO:0000067 - Replicação do DNA e ciclo cromossomal GO:0006310 - Recombinação do DNA GO:0030261 - Condensação do cromossomo GO:0006271 - Alongamento da fita de DNA GO:0006281 - Reparo do DNA	MBP1 STB1
Continua na próxima página		

Tabela A.20 – continuação da página anterior

	GO:0006260 - Replicação do DNA	
4		MCM1
5	GO:0006790 - Metabolismo de enxofre GO:0006520 - Metabolismo de aminoácidos	BAS1 GCN4
6	GO:0006333 - União ou separação da cromatina GO:0016043 - Organização celular e biogênese GO:0045229 - Org. da estrutura de encapsulamento externo e biogênese GO:0009101 - Biosíntese de glicoproteína	STB1

SOM = 2x4

A Figura A.8 ilustra o perfil médio da expressão dos genes dos 8 grupos quando SOM = 2x4.

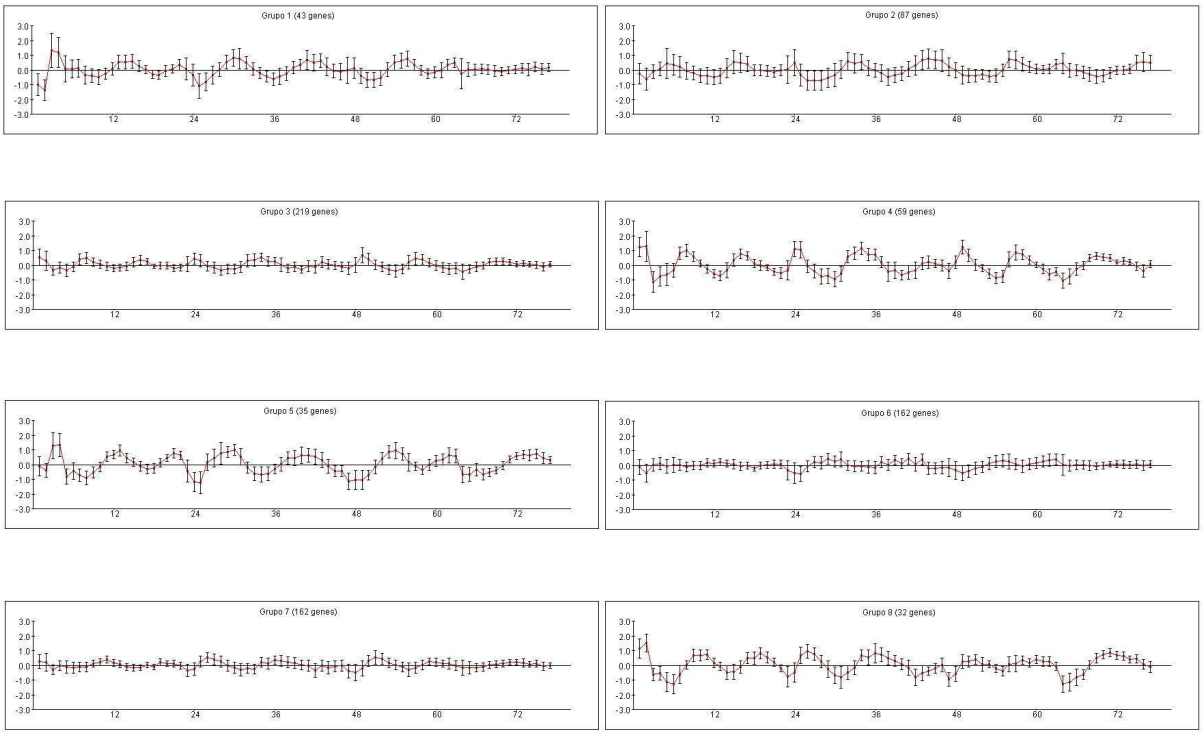


Figura A.8: Perfil médio da expressão dos genes dos 8 grupos (SOM = 2x4).

O grupo 1 é representativo da fase M (G2/ M) , o grupo 2 das fases M (M/G1) e G1, os grupos 3 e 4 da fase G1, os grupos 5 da fase M (G2/M), o grupo e 6 das fases G2 e M (G2/M), o grupo 7 das fases S e G2 e o grupo 8 é representativo das fases G1, S e G2.

O resultado deste agrupamento é muito semelhante ao resultado do agrupamento anterior, inclusive pelas funções identificadas nos grupos. A Tabela A.21 apresenta as funções biológicas e fatores de transcrição associados a cada grupo do agrupamento SOM = 2x4.

Tabela A.21: Validação biológica do agrupamento SOM = 2x4.

Grupos	TANGO	PRIMA
1	GO:0008026 - Atividade da helicase dependente de ATP	MCM1
2	GO:0000746 - Conjugação	MSN4 SWI5
3	GO:0006259 - Metabolismo de DNA GO:0006310 - Recombinação do DNA GO:0006281 - Reparo do DNA	MBP1
4	GO:0008283 - Proliferação celular GO:0006974 - Resposta à estímulos de danos ao DNA GO:0006260 - Replicação do DNA	MBP1 STB1
5		MCM1
6	GO:0015082 - Atividade de transporte de cálcio inorgânico GO:0051179 - Localização	
7	GO:0006790 - Metabolismo de enxofre GO:0006520 - Metabolismo de aminoácido	
8	GO:0006333 - União ou separação da cromatina GO:0016043 - Organização celular e biogênse	

SOM = 2x5

A Figura A.9 ilustra o perfil médio da expressão dos genes dos 10 grupos.

Os grupos 1 e 2 foram formados por genes envolvidos na fase G1 do ciclo celular, o grupo 3 por genes das fases G1 e S, os grupos 4 e 5 por genes da fase M (G2/M e M/G1), o grupo 6 por genes da fase G1, os grupos 7 e 8 por genes da fase S e G2, o grupo 9 por genes das fases G2 e M (G2/M) e o grupo 10 por genes da fase M (G2/M).

A Tabela A.22 apresenta as funções biológicas e fatores de transcrição associados a cada grupo do agrupamento SOM = 2x5.

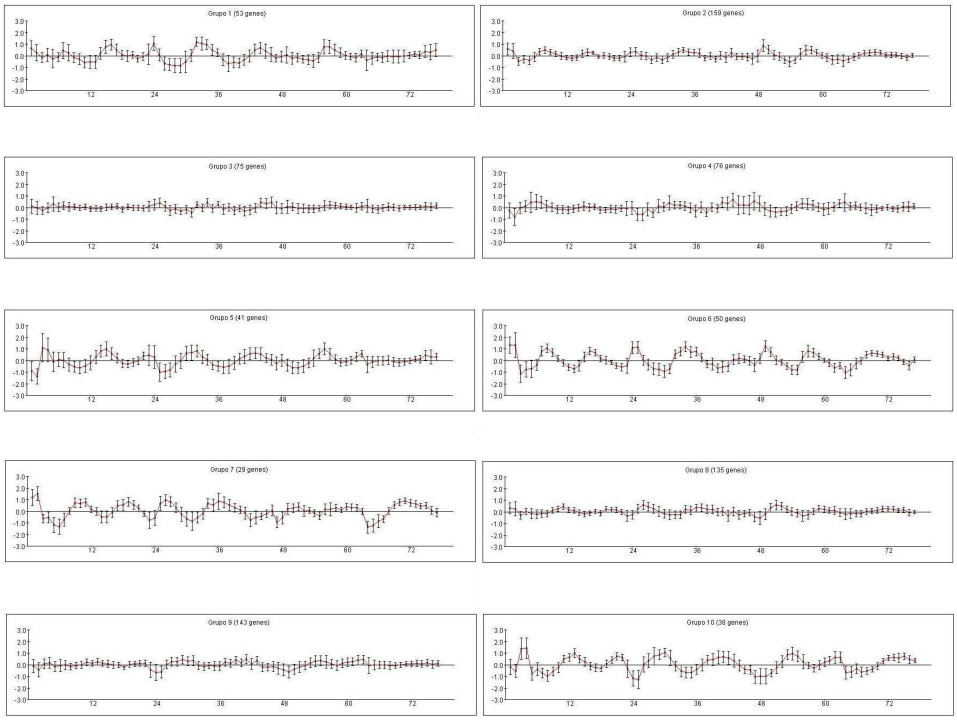


Figura A.9: Perfil médio da expressão dos genes dos 10 grupos (SOM = 2x5).

Tabela A.22: Validação biológica do agrupamento SOM = 2x5.

Grupos	TANGO	PRIMA
1	GO:0000722 - Manutenção dos telômeros	SWI5
2	GO:0006259 - Metabolismo de DNA GO:0006310 - Recombinação do DNA GO:0006281 - Reparo do DNA	MBP1
4	GO:0019236 - Resposta ao feromônio GO:0051119 - Atividade de transporte de açúcar	
5	GO:0006267 - Formação e manutenção do complexo pré-replicativo	MCM1
6	GO:0008283 - Proliferação celular GO:0006260 - Replicação do DNA	MBP1 STB1
7	GO:0045229 - Org. da estrutura de encapsulamento externo e biogênese GO:0016043 - Organização celular e biogênese GO:0006333 - União ou separação da cromatina	

Continua na próxima página

Tabela A.22 – continuação da página anterior

8	GO:0006790 - Metabolismo de enxofre	MET31
9	GO:0005386 - Atividade de transporte GO:0006812 - Transporte de cálcio GO:0051179 - Localização GO:0015082 - Atividade de transporte de cálcio inorgânico	BAS1
10		MCM1

A.1.9 Validação biológica SAMBA

Nesta seção é apresentado o resultado da aplicação das técnicas de validação biológica no agrupamento do algoritmo bidimensional. Foi adotado o mesmo processo utilizado na análise dos agrupamentos unidimensionais, mas aqui não são descritos todos os 50 grupos bidimensionais. O comportamento de alguns deles são descritos a seguir junto com seus respectivos *heat maps*.

A Figura A.10 ilustra o *heat map* do grupo 4. As linhas correspondem aos genes e as colunas às condições experimentais do microarranjo. As cores correspondem aos níveis de expressão dos genes sobre as condições. Lembrando que a cor vermelha indica maior expressão, a cor verde pouca expressão e a cor preta, representa a ausência ou baixa qualidade do sinal de expressão, conforme a legenda da figura. As últimas colunas da figura correspondem ao nome e anotação funcional dos genes já caracterizados senão, é repetida a identificação do gene na base de dados seguido da descrição "*unknown*".

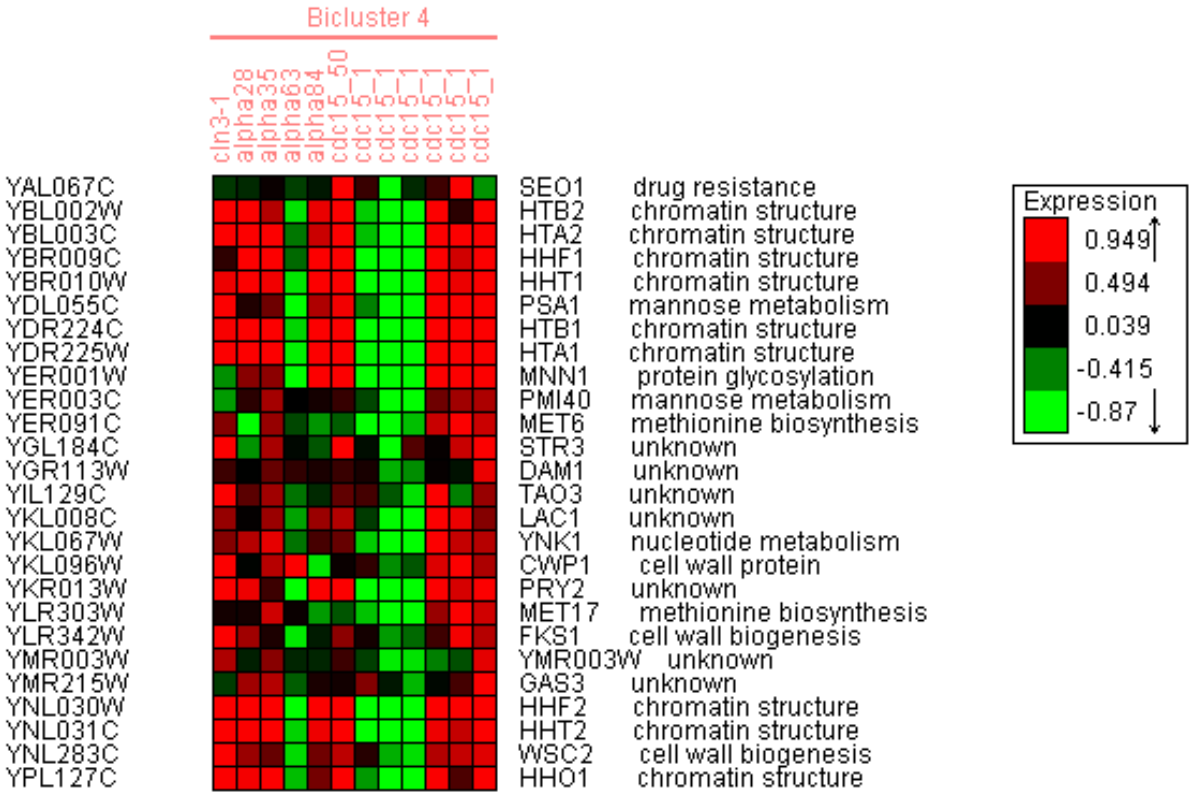


Figura A.10: *Heat map* do grupo 4.

A imagem do grupo 4 da Figura A.10 reflete a característica da abordagem de agrupamento bidimensional, onde os genes são agrupados somente nas condições em que são co-regulados. Este grupo foi formado por 26 genes que apresentaram expressão co-regulada em 12 das 77 condições do microarranjo e a maioria deles estão envolvidos na função de estruturação da cromatina.

A Figura A.11 ilustra o *heat map* do grupo 27.

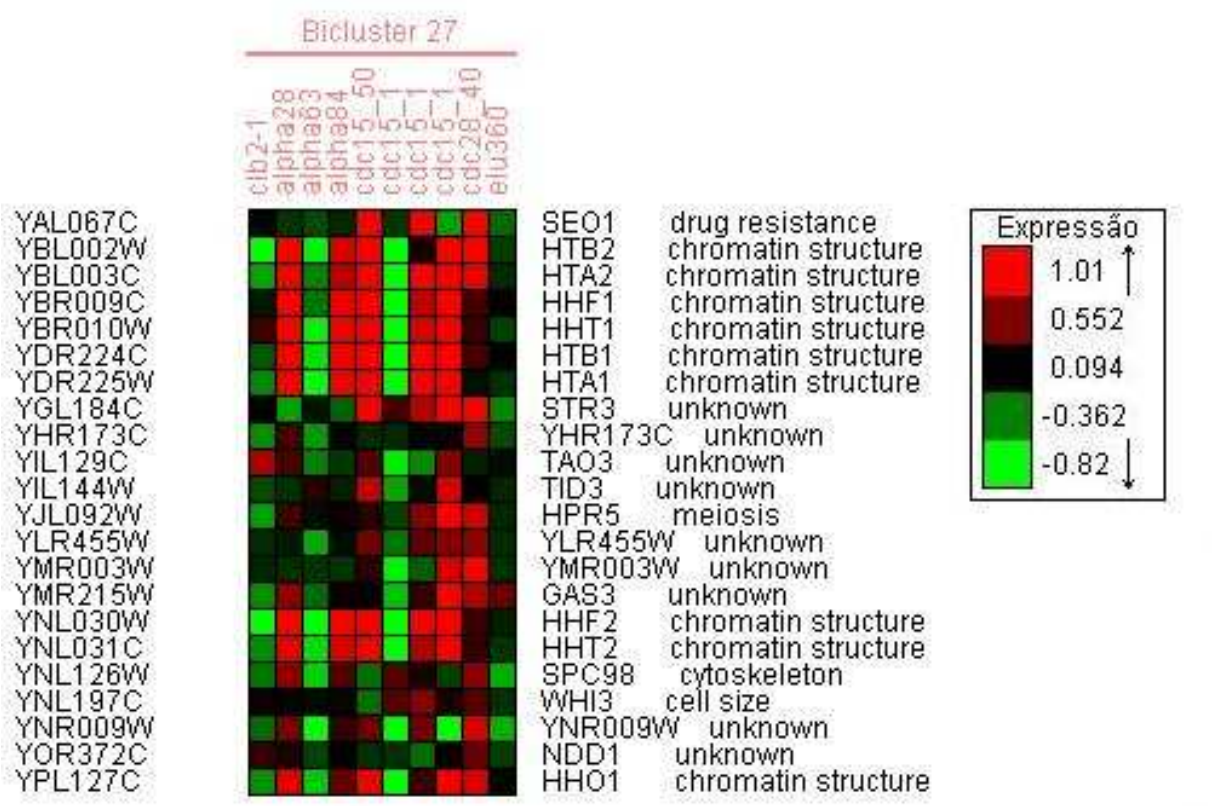


Figura A.11: *Heat map* do grupo 27.

O grupo 27 foi formado por 22 genes em 10 condições. Embora esses dois grupos sejam muito semelhantes, os genes do grupo 4 são co-regulados nas condições de CLN3, alpha e cdc15, em diferentes tempos e os genes do grupo 27 são co-regulados nas condições clb2, alpha, cdc15 e também cdc28 e elutriação.

No grupo 4, os genes com a função de estruturação da cromatina são co-regulados com genes envolvidos nas funções de metabolismo de manose e biosíntese da metionina. Então é possível que estas diferentes funções sejam executadas de forma coordenada, mas

somente nestas condições do grupo 4. O grupo 27, além dos genes envolvidos na função de estruturação da cromatina, o grupo também foi formado por genes envolvidos nas funções de meiose e citoesqueleto. Da mesma forma que no grupo 4, a co-regulação dos genes deste grupo deve ocorrer somente nestas condições. Portanto, os diferentes subgrupos de condições dos grupos, possibilitaram a identificação do envolvimento de genes de diferentes funções.

Ainda da análise dos grupos 4 e 27, é possível inferir que os genes com função desconhecida "*unknown*", mas presentes em ambos os grupos, também estejam envolvidos na função de estruturação da cromatina, como por exemplo, os genes YGL184C, YIL129C e YMR003W.

Os grupos 1, 3, 32 e 33 também foram formados por genes envolvidos na função de estruturação da cromatina, mas também em diferentes condições, podendo revelar associações com outras funções, além das identificadas nos grupo 4 e 27.

Essas diferentes combinações de genes e condições também podem revelar relações entre diferentes fases do ciclo celular. Ou seja, nos grupos 4 e 27 quase a totalidade dos genes estão envolvidos na fase S do ciclo. Já no grupo 1, os genes estão envolvidos nas fases G1 e S e no grupo 32 nas fases S e G2, sugerindo a hipótese de que os genes desses grupos estão envolvidos na transição entre essas fases.

Uma outra observação é a vantagem do agrupamento bidimensional com relação ao k-médias. Na Figura 5.2 (b) do grupo 4 do algoritmo k-médias, os genes YBR009C, YBR010W, YDR224C e YDR225W demonstram co-regulação entre eles e não com os demais genes do grupo. No agrupamento bidimensional estes genes foram encontrados nos grupos 1, 3, 4 e 27. A Figura A.11, do grupo bidimensional 4 mostra a co-regulação destes genes em 12 condições, junto com os genes YDL055C e YNL030W, que demonstram o mesmo perfil de expressão nestas condições, embora na solução do algoritmo k-médias estes genes não tenham sido alocados no mesmo grupo. Este exemplo evidencia a incapacidade do algoritmo k-médias em agrupar genes co-regulados em somente algumas condições.

Um outro exemplo, ainda na Figura 5.2 (b), é observado com os genes YBR273C, YDL169C e YDR368W das linhas 8, 17 e 34 do *heat map*, respectivamente. Conforme o *heat map*, estes genes não demonstram perfil de expressão bem definidos quando comparados com os demais elementos do grupo. No agrupamento bidimensional estes genes foram encontrados nos grupos 2, 46 e 47, demonstrando alta co-regulação entre eles e com os demais genes do grupo.

A Figura A.12 ilustra o *heat map* do grupo 47 do agrupamento bidimensional.

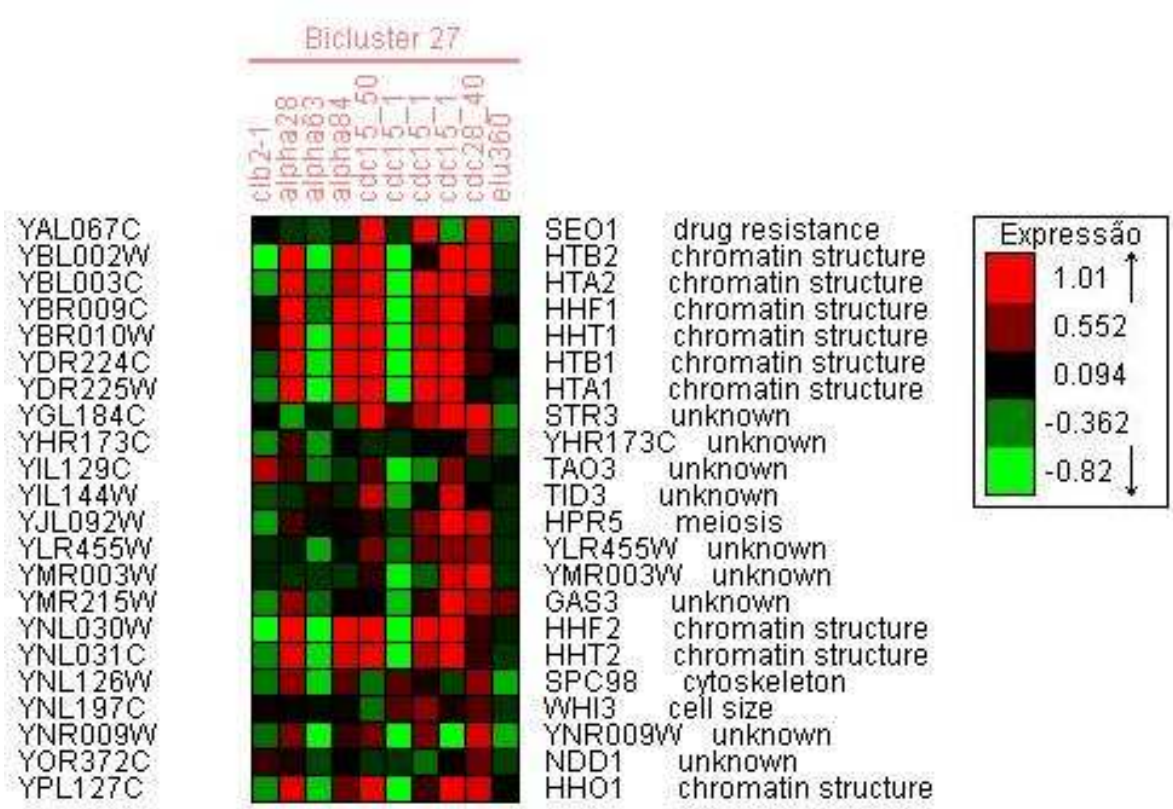


Figura A.12: *Heat map* do grupo 47.

A Tabela A.23 apresenta as funções biológicas e fatores de transcrição associados a cada grupo do agrupamento SAMBA.

Tabela A.23: Validação biológica do agrupamento SAMBA.

Grupos	TANGO	PRIMA
1	GO:0006333 - União ou separação da cromatina	
2	GO:0000746 - Conjugação	
3	GO:0006333 - União ou separação da cromatina	STB1
4	GO:0006333 - União ou separação da cromatina	
5	GO:0006790 - Metabolismo de enxofre	
6		ACE2 SWI5
8		SWI5
11	GO:0008283 - Proliferação celular	STB1 ACE2
12	GO:0008283 - Proliferação celular	MBP1 STB1
14	GO:0008283 - Proliferação celular	MCM1
15		ACE2 SWI5
16	GO:0008026 - Atividade de helicase	MCM1
17		STB1
19	GO:0008026 - Manutenção e formação do complexo pré-replicativo	MCM1
20	GO:0008283 - Proliferação celular GO:0006281 - Reparo do DNA GO:0006260 - Replicação do DNA	MBP1 STB1
21		MCM1
22		MCM1
23	GO:0019236 - Resposta ao feromônio	MCM1
26		MBP1
27	GO:0006333 - União ou separação da cromatina	

Continua na próxima página

Tabela A.23 – continuação da página anterior

	GO:0003676 - Ligação de ácido nucléico	
28		MCM1
31		MBP1 STB1
32	GO:0006333 - União ou separação da cromatina	
37		MCM1
39	GO:0008026 - Atividade de helicase	
42		MCM1
44		MSN4
47		MSN4 ACE2
49	GO:0008283 - Proliferação celular	MBP1
50		MBP1 STB1

Não foram identificadas todas as funções identificadas nos agrupamentos anteriores, embora este resultado do agrupamento bidimensional auxilie na confirmação de qual fase realmente a função está envolvida. A função de atividade de helicase, por exemplo, foi identificada num grupo bastante representativo da fase M, mais especificamente da transição de M para G1 e não foi identificada na fase G2/M, sugerindo que é uma função específica dessa transição.

Da mesma forma, a validação por identificação de fatores de transcrição não adicionou informação adicional nos resultados do agrupamento bidimensional, embora esse resultado seja útil para corroborar a hipótese de associação de um fator com uma fase específica do ciclo, já que os grupos bidimensionais são bem mais específicos.

A.2 Base de dados CCSc - com a aplicação de filtros de dados

A.2.1 Agrupamento k-médias

Da mesma forma que com a base de dados sem a aplicação de filtros, com a aplicação do algoritmo k-médias nesta base de dados filtrada esperava-se a identificação de 4 grupos, correspondentes às 4 fases do ciclo.

O algoritmo k-médias também foi aplicado com o valor de $k = 2, 4, 5, 8$ *e* 10. Os resultados obtidos foram os seguintes:

k = 2

Os 372 genes da base de dados CCSc foram agrupados em 2 grupos, conforme mostrado na Tabela A.24.

Tabela A.24: Agrupamento $k = 2$.

Grupos	Quantidade de genes	Homogeneidade
1	211	0,375
2	161	0,574

k = 4

Os 372 genes da base de dados CCSc foram agrupados em 4 grupos, conforme mostrado na Tabela A.25.

Tabela A.25: Agrupamento $k = 4$.

Grupos	Quantidade de genes	Homogeneidade
1	28	0,332
2	39	0,424
3	156	0,562
4	149	0,496

k = 5

O resultado deste agrupamento é apresentado na Tabela A.26. O resultado do primeiro, segundo e terceiro grupo são iguais aos resultados do agrupamento $k = 4$.

Tabela A.26: Agrupamento $k = 5$.

Grupos	Quantidade de genes	Homogeneidade
1	28	0,332
2	39	0,424
3	156	0,562
4	143	0,488
5	6	0,835

k = 8

Os 372 genes da base de dados CCSc foram agrupados em 8 grupos, conforme mostrado na Tabela A.27.

Tabela A.27: Agrupamento $k = 8$.

Grupos	Quantidade de genes	Homogeneidade
1	2	0,748
2	11	0,478
3	115	0,642
4	51	0,641
5	6	0,835
6	14	0,813
7	167	0,299
8	6	0,936

k = 10

Os 372 genes da base de dados CCSc foram agrupados em 10 grupos, conforme mostrado na Tabela A.28.

Tabela A.28: Agrupamento $k = 10$.

Grupos	Quantidade de genes	Homogeneidade
1	2	0,768
2	6	0,63
3	113	0,65
4	40	0,652
5	3	0,861
6	20	0,833
7	163	0,305
8	6	0,904
9	5	0,862
10	14	0,617

A.2.2 Agrupamento SOM

O algoritmo SOM foi aplicado com as dimensões da matriz definidas por 2x2, 5x1, 2x3, 2x4 e 2x5. Os resultados obtidos foram os seguintes:

SOM = 2x2

A quantidade de genes e a índice de homogeneidade de cada grupo são apresentadas na Tabela A.29.

Tabela A.29: Agrupamento SOM = 2x2.

Grupos	Quantidade de genes	Homogeneidade
1	91	0,726
2	87	0,5
3	82	0,551
4	112	0,576

SOM = 5x1

A quantidade de genes e a índice de homogeneidade de cada grupo são apresentadas na Tabela A.30.

Tabela A.30: Agrupamento SOM = 5x1.

Grupos	Quantidade de genes	Homogeneidade
1	76	0,638
2	81	0,471
3	51	0,628
4	103	0,436
5	61	0,752

SOM = 2x3

A quantidade de genes e a índice de homogeneidade de cada grupo são apresentadas na Tabela A.31.

Tabela A.31: Agrupamento SOM = 2x3.

Grupos	Quantidade de genes	Homogeneidade
1	52	0,723
2	81	0,437
3	41	0,779
4	72	0,463
5	48	0,619
6	78	0,7

SOM = 2x4

A quantidade de genes e a índice de homogeneidade de cada grupo são apresentadas na Tabela A.32.

Tabela A.32: Agrupamento SOM = 2x4.

Grupos	Quantidade de genes	Homogeneidade
1	19	0,659
2	27	0,579
3	25	0,574
4	24	0,818
5	13	0,805
6	31	0,486
7	16	0,579
8	15	0,802

SOM = 2x5

A quantidade de genes e a índice de homogeneidade de cada grupo são apresentados na Tabela A.33.

Tabela A.33: Agrupamento SOM = 2x5.

Grupos	Quantidade de genes	Homogeneidade
1	14	0,675
2	21	0,647
3	4	0,465
4	28	0,504
5	12	0,69
6	20	0,832
7	14	0,812
8	16	0,561
9	24	0,547
10	17	0,78

A.2.3 Agrupamento SAMBA

Tabela A.34: Agrupamento SAMBA.

Grupo	Escore	Condições	Genes
1	76,6124	7	13
2	84,9407	9	14
3	164,798	10	23
4	246,771	20	19
5	124,37	7	22
6	133,454	12	14
7	255,504	17	23
8	129,842	6	31
9	108,928	8	20
10	204,885	11	22
11	105,373	10	19
12	115,007	7	18
13	75,3509	6	18
14	87,4347	5	19
15	291,64	25	13
16	154,876	10	16
17	66,9991	7	12
18	94,1067	13	12
19	100,402	12	11
20	95,5788	6	18
21	75,6715	7	14
22	79,3204	6	13
23	183,909	9	28
24	92,2522	6	16
25	124,615	13	14
26	199,345	17	22
27	89,4856	7	15

Continua na próxima página

Tabela A.34 – continuação da página anterior

28	81,2041	9	11
29	165,525	11	24
30	161,175	6	29
31	144,308	15	13
32	61,6834	5	19
33	119,53	9	18
34	120,615	10	13
35	55,6148	9	9
36	102,933	16	10
37	138,983	7	25
38	99,817	8	14
39	60,2557	10	9
40	78,13	4	20
41	104,157	8	19
42	246,021	23	18
43	112,395	10	16
44	65,5143	8	11
45	116,82	8	17

A.2.4 Validação estatística k-médias

Os resultados da aplicação das técnicas de validação estatística na base de dados CCSc com a aplicação de filtros de dados foram os mesmos obtidos da base de dados sem a aplicação dos filtros, a não ser o índice Davies Bouldin que identificou o melhor agrupamento $k = 10$ e na base sem filtro $k = 9$ A partir desses resultados, a idéia é que, se realmente os grupos são formados com a mesma qualidade, é muito mais vantajoso a utilização de um conjunto reduzido de dados, pelo tempo computacional dos programas e pela facilidade de analisar uma quantidade de dados menor.

Tabela A.35: Validação estatística dos agrupamentos k-médias.

k	Homogeneidade	Separação	C	D. Bouldin	Dunn	Silhueta	Isolamento
2	0,448	**-0,119**	0,29	1,805	**1,09**	**0,165**	**0,869**
4	**0,524**	-0,075	0,291	1,927	0,886	0,043	0,514
5	0,522	-0,068	0,28	1,914	0,838	0,039	0,481
8	0,425	-0,026	0,198	1,775	0,466	0,038	0,475
10	0,431	-0,022	**0,183**	**1,677**	0,63	0,042	0,46

A.2.5 Validação estatística SOM

Da mesma forma ocorreu com os resultados do algoritmo SOM, somente o índice C e o Davies Bouldin deram resultados diferentes. O índice C indicou o agrupamento SOM = 2x5 sendo o melhor, enquanto na base de dados sem a aplicação dos filtros ele indicou o melhor agrupamento SOM = 2x4. O índice Davies Bouldin indicou como melhor agrupamento SOM = 2x3, enquanto na base de dados sem a aplicação dos filtros ele indicou o melhor quando SOM = 5x1.

Tabela A.36: Validação estatística dos agrupamentos SOM.

k	Homogeneidade	Separação	C	D. Bouldin	Dunn	Silhueta	Isolamento
2x2	0,59	**-0,055**	0,209	1,787	0,865	**0,124**	**0,733**
5x1	0,541	-0,02	0,194	1,771	0,891	0,084	0,579
2x3	0,58	-0,016	0,163	**1,69**	**0,903**	0,101	0,582
2x4	0,604	-0,001	0,165	1,804	0,782	0,078	0,534
2x5	**0,606**	0,005	**0,134**	1,766	0,82	0,09	0,511

A.2.6 Validação biológica k-médias

k = 2

A Figura A.13 ilustra o perfil médio da expressão dos genes dos 2 grupos.

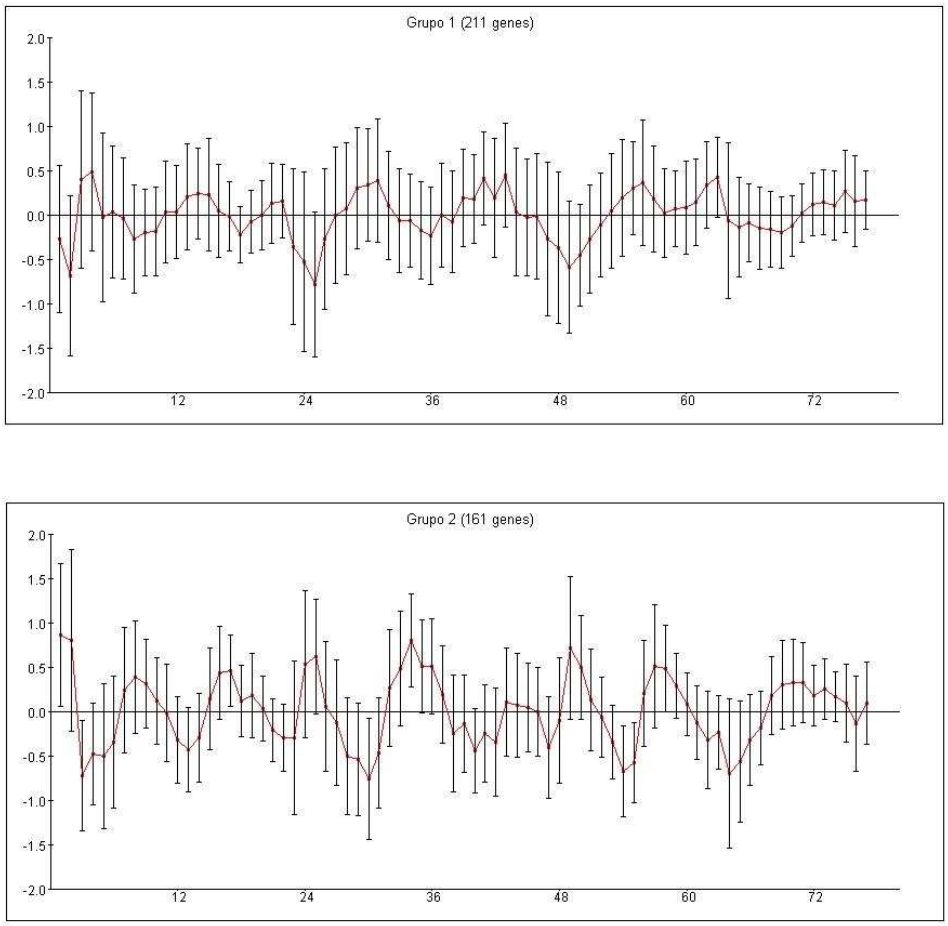


Figura A.13: Perfil médio da expressão dos genes dos 2 grupos ($k = 2$).

A distribuição do genes nos grupos foi a mesma obtida da base de dados sem a aplicação de filtros. Porém, as funções biológicas identificadas foram um pouco diferentes, influenciadas pela quantidade reduzida de genes. O grupo 1 foi formado por genes da fase G2 e M e o grupo 2 por genes das fases G1 e S. Ambos os grupos são potencialmente siginificativos se considerada a ordem do ciclo celular.

A Tabela A.37 apresenta as funções biológicas e fatores de transcrição associados a cada grupo do agrupamento $k = 2$.

A validação biológica de fatores de transcrição identificou os mesmos 2 fatores do

grupo 2 identificados no agrupamento $k = 2$ sem filtros, embora não tenha identificado o fator AZF1 no grupo 1.

Tabela A.37: Validação biológica do agrupamento $k = 2$.

Grupos	TANGO	PRIMA
1	GO:0000749 - Resposta ao feromônio	
2	GO:0044238 - Metabolismo primário	MBP1
	GO:0006259 - Metabolismo do DNA	STB1
	GO:0009719 - Resposta a estímulos endógenos	
	GO:0006996 - Organização da organela e biogênese	
	GO:0008283 - Proliferação celular	

k = 4

A Figura A.14 ilustra o perfil médio da expressão dos genes dos 4 grupos. Os grupos não correspondem aos mesmos grupos quando aplicado na base sem a aplicação dos filtros de dados. E essa diferença é, principalmente porque não foi formado o grupo somente com o gene CLN3. Neste agrupamento o gene CLN3 foi alocado no grupo 1 junto com outros 27 genes, possibilitando a melhor distribuição dos outros genes nos grupos.

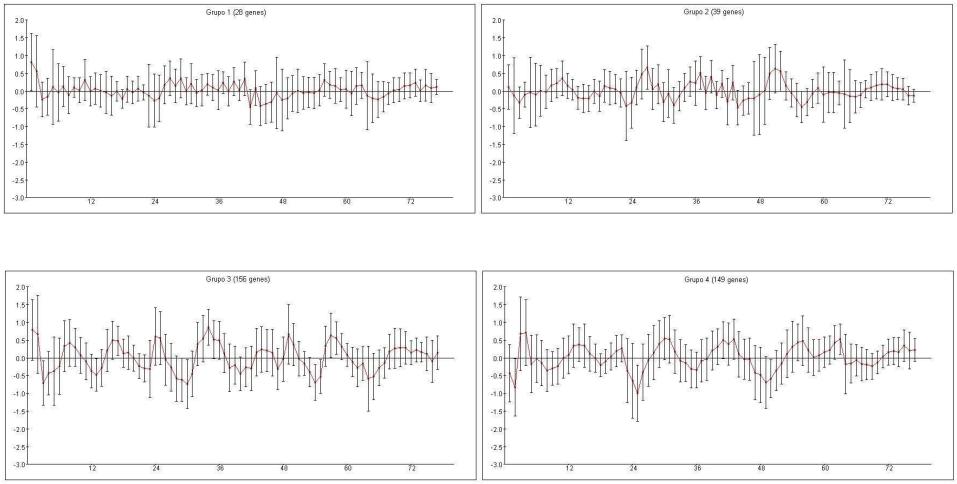


Figura A.14: Perfil médio da expressão dos genes dos 4 grupos ($k = 4$).

O grupo 1 foi formado por genes das fases S, G2 e M, o grupo 2 por genes das fases S e G2, o grupo 3 por genes da fase G1 e o grupo 4 por genes da fase M (G2/M e M/G1).

A Tabela A.38 apresenta as funções biológicas e fatores de transcrição associados a cada grupo do agrupamento $k = 4$.

Tabela A.38: Validação biológica do agrupamento $k = 4$.

Grupos	TANGO	PRIMA
3	GO:0044238 - Metabolismo primário	MBP1
	GO:0043283 - Metabolismo de biopolímeros	STB1
	GO:0006281 - Reparo do DNA	
	GO:0006139 - Metabolismo de ácido nucléico	
4	GO:0044238 - Transporte	MCM1

k = 5

Os genes foram distribuídos conforme a Figura A.15 que ilustra o perfil médio da expressão dos genes de cada um dos 5 grupos.

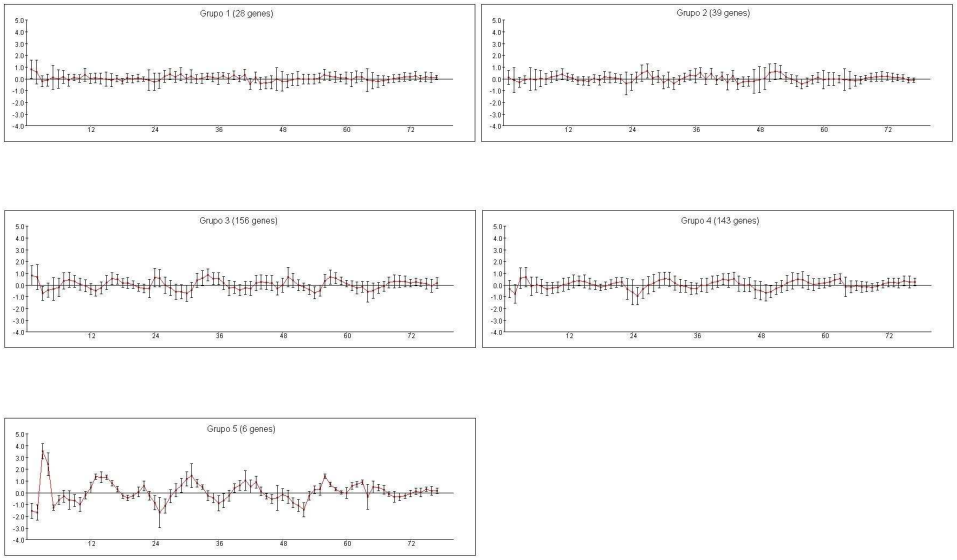


Figura A.15: Perfil médio da expressão dos genes dos 5 grupos ($k = 5$).

Este agrupamento não acrescentou nenhuma informação adicional se comparado com o agrupamento $k = 4$. O grupo 1 foi formado por genes envolvidos nas fases M (G2/M e M/G1), S e G2, o grupo 2 por genes das fases M (M/G1), G1, S e G2, o grupo 3 por genes da fase G1, o grupo 4 por genes da fase M (G2/M e M/G1) e o grupo 5 foi formado por genes da fase M (G2/M).

A Tabela A.39 apresenta as funções biológicas e fatores de transcrição associados a cada grupo do agrupamento $k = 5$.

Tabela A.39: Validação biológica do agrupamento $k = 5$.

Grupos	TANGO	PRIMA
3	GO:0044238 - Metabolismo primário	MBP1
	GO:0043283 - Metabolismo de biopolímeros	STB1
	GO:0006281 - Reparo do DNA	
	GO:0006139 - Metabolismo de ácido nucléico	
4	GO:0044238 - Transporte	MCM1

k = 8

O perfil médio da expressão dos genes de cada grupo é ilustrado na Figura A.16.

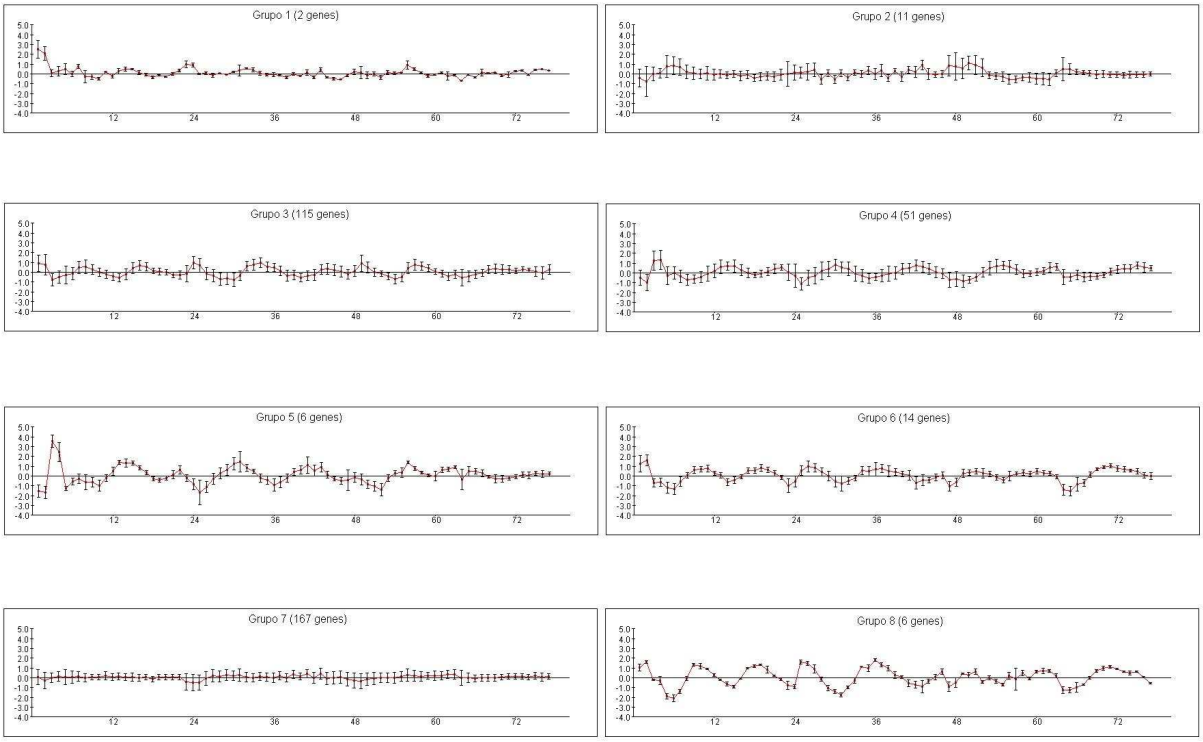


Figura A.16: Perfil da expressão dos genes dos 8 grupos ($k = 8$).

O grupo 1 foi formado por genes envolvidos na fase M (G2/M e M/G1), o grupo 2 por genes das fases S e M (M/G1), o grupo 3 por genes da fase G1, o grupo 4 por genes da fase M (G2/M e M/G1), o grupo 5 por genes da fase M (G2/M), o grupo 6 por genes das fases S e G2, o grupo 7 por genes das fases S, G2 3 M (G2/M e M/G1) e o grupo 8 foi formado por genes envolvidos na fase S do ciclo celular.

A distribuição dos genes em um número maior de grupos resultou em grupos pequenos de alta homogeneidade, por exemplo o grupo 8. Este grupo apresentou alta homogeneidade, com 6 genes caracterizados como histonas e todos eles envolvidos na fase S do ciclo, conforme o *heat map* da Figura A.17.

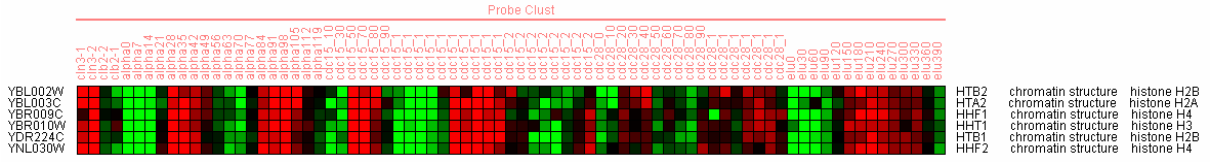


Figura A.17: *Heat map* do grupo 8 quando $k = 8$ (captura da tela do programa Expander).

A Tabela A.40 apresenta as funções biológicas e fatores de transcrição associados a cada grupo do agrupamento $k = 8$.

Tabela A.40: Validação biológica do agrupamento $k = 8$.

Grupos	TANGO	PRIMA
3	GO:0006260 - Replicação do DNA GO:0043283 - Metabolismo de biopolímeros GO:0006281 - Reparo do DNA GO:0008283 - Proliferação celular	MBP1 STB1
4	GO:0008283 - Proliferação celular	MCM1
7	GO:0006810 - Transporte	
8	GO:0006333 - União ou separação da cromatina	

k = 10

A Figura A.18 ilustra o perfil médio da expressão dos genes dos 10 grupos.

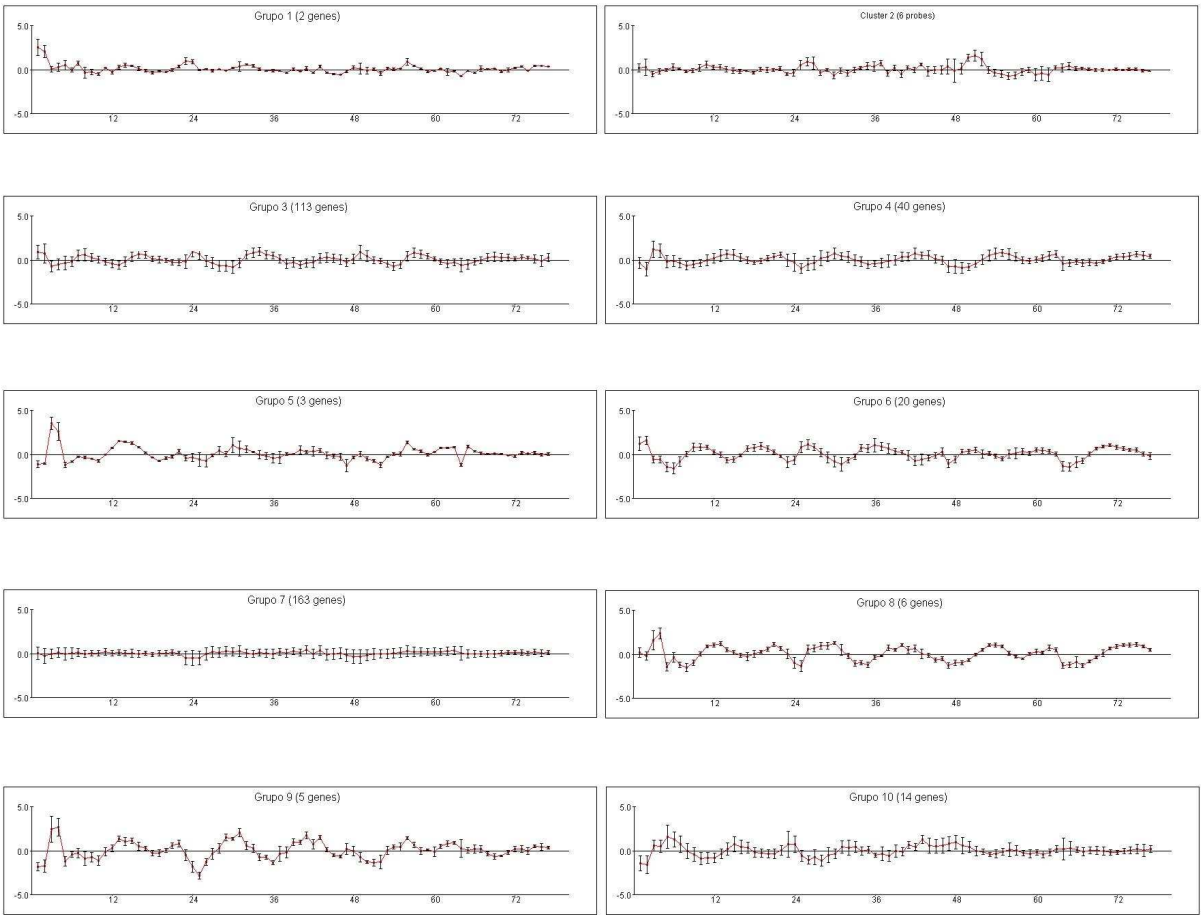


Figura A.18: Perfil médio da expressão dos genes dos 10 grupos ($k = 10$).

Neste agrupamento os genes foram mais distribuídos, não houve a formação de um grupo para representar cada fase do ciclo celular. A idéia, neste caso, é explorar o potencial de cada grupo individualmente, embora tenham sido formados grupos pouco expressivos com 2 genes apenas, como o grupo 1.

O grupo 2 foi formado por genes envolvidos na fase S do ciclo, o grupo 3 por genes da fase G1, o grupo 4 por genes da fase M (G2/M e M/G1), o grupo 5 por genes da fase M (G2/M), o grupo 6 por genes da fase S, o grupo 7 por genes das fases G1, S e M (G2/M e M/G1), o grupo 8 por genes da fase M (G2/M), o grupo 9 por genes da fase M (G2/M) e o grupo 10 foi formado por genes da fase M (M/G1).

O grupo 5 foi formado por 3 genes com a função de metabolismo de fosfato, o grupo 6 por 20 genes, todos eles histonas e o grupo 8 por genes caracterizados com a

função de ciclo celular.

A Tabela A.41 apresenta as funções biológicas e fatores de transcrição associados a cada grupo do agrupamento $k = 10$.

Tabela A.41: Validação biológica do agrupamento $k = 10$.

Grupos	TANGO	PRIMA
3	GO:0006260 - Replicação do DNA GO:0043283 - Metabolismo de biopolímeros GO:0006281 - Reparo do DNA GO:0008283 - Proliferação celular GO:0006139 - Metabolismo de ácido nucléico	MBP1 STB1
4		MCM1
5		PHO4
6	GO:0006333 - União ou separação da cromatina GO:0006139 - Metabolismo de ácido nucléico GO:0003677 - Ligação do DNA GO:0016043 - Organização celular e biogênese	HAC1
7	GO:0006810 - Transporte	
10	GO:0005199 - Constituinte estrtutural da parede celular GO:0000749 - Resposta ao feromônio durante conjugação celular	STE12

A.2.7 Validação biológica SOM

SOM = 2x2

O perfil médio da expressão dos genes dos grupos é ilustrado na Figura A.19.

O grupo 1 foi formado por genes envolvidos na fase G1 do ciclo celular. O grupo 2 por genes das fases M (M/G1) e G1, o grupo 3 por genes das fases S e G2 e o grupo 4 formado por genes das fases G2 e M (G2/M). Os grupos são os mesmos que os formados nos agrupamento sem filtro.

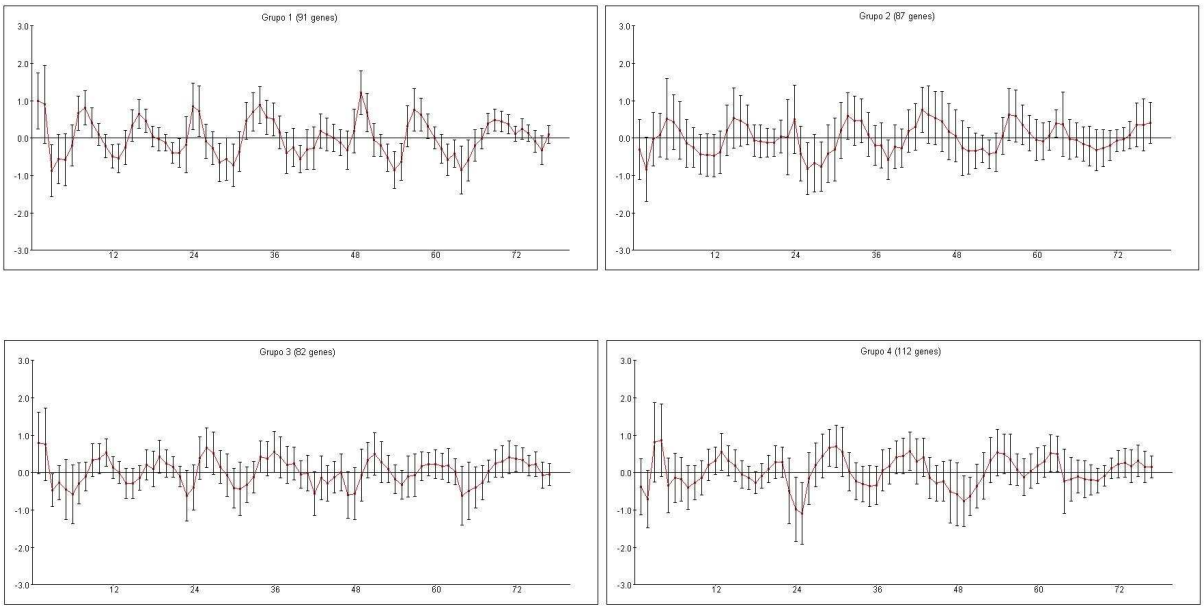


Figura A.19: Perfil médio da expressão dos genes dos 4 grupos (SOM = 2x2).

A Tabela A.42 apresenta as funções biológicas e fatores de transcrição associados a cada grupo do agrupamento SOM = 2x2.

Tabela A.42: Validação biológica do agrupamento SOM = 2x2.

Grupos	TANGO	PRIMA
1	GO:0006260 - Replicação do DNA GO:0006259 - Metabolismo do DNA GO:0009719 - Resposta a estímulos endógenos GO:0008283 - Proliferação celular	MBP1 STB1
2	GO:0000749 - Resposta do feromônio durante conjugação com fusão celular	
3	GO:0006333 - União ou separação da cromatina GO:0016043 - Organização celular e biogênese GO:0003677 - Metabolismo de enxofre	

SOM = 5x1

O perfil médio da expressão dos genes dos grupos é ilustrado na Figura A.20.

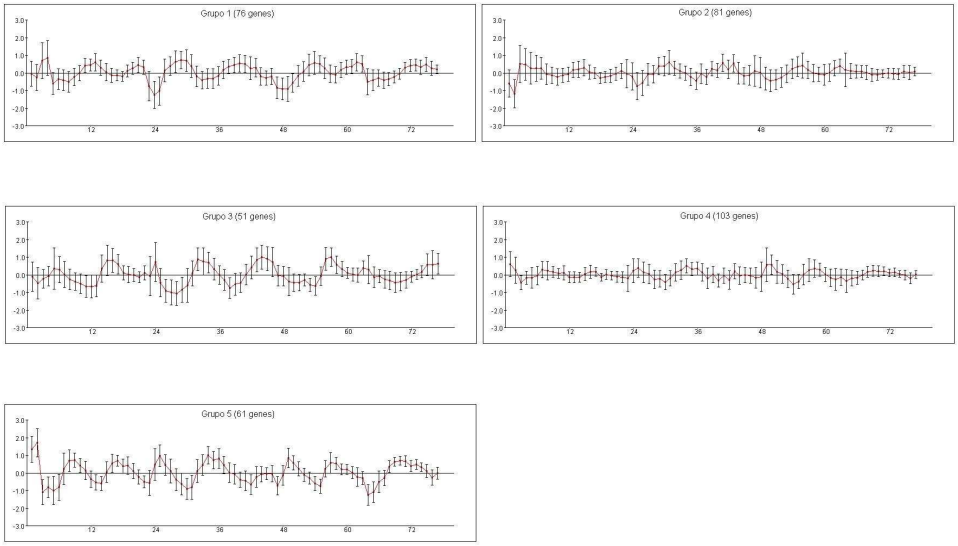


Figura A.20: Perfil médio da expressão dos genes dos 4 grupos (SOM = 5x1).

O grupo 1 foi formado por genes envolvidos nas fases G2 e M (G2/M), o grupo 2 por genes da fase M (G2/M e M/G1), o grupo 3 por genes das fases M (M/G1) e G1, o grupo 4 por genes das fases G1, S e G2 e o grupo 5 por genes da fase G1.

Somente o grupo 5 foi enriquecido com funções biológicas. A Tabela A.43 apresenta as funções biológicas e fatores de transcrição associados a cada grupo do agrupamento SOM = 5x1.

Tabela A.43: Validação biológica do agrupamento SOM = 5x1.

Grupos	TANGO	PRIMA
3		SWI5
5	GO:0044238 - Metabolismo primário GO:0006259 - Metabolismo do DNA GO:0006333 - União ou separação da cromatina	STB1

SOM = 2x3

O perfil médio da expressão dos genes dos 6 grupos é ilustrado na Figura A.21.

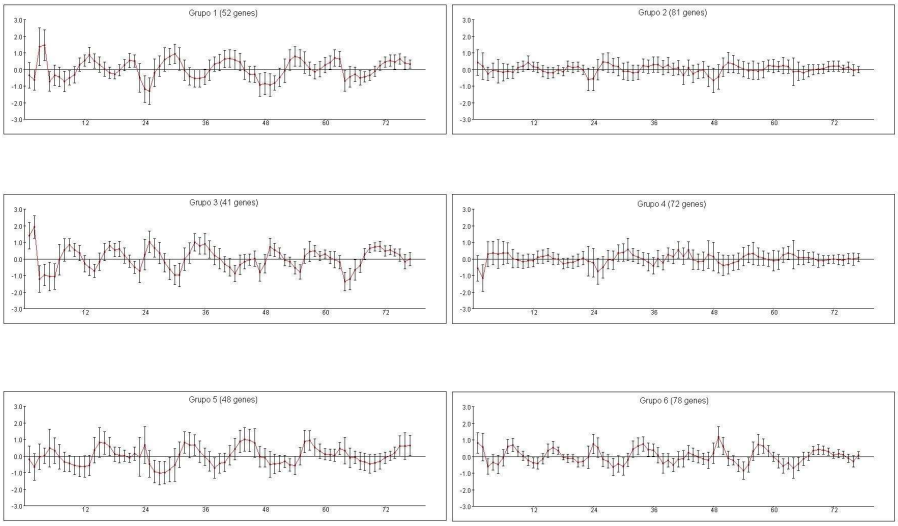


Figura A.21: Perfil médio da expressão dos genes dos 6 grupos (SOM = 2x3).

O grupo 1 foi formado por genes envolvidos na fase M (G2/M) do ciclo celular. O grupo 2 por genes das fases S e G2, o grupo 3 por genes das fase G1, o grupo 4 por genes da fase M (M/G1) e G1 e o grupo 5 foi formado por genes envolvidos na fase G1 do ciclo.

A Tabela A.44 apresenta as funções biológicas e fatores de transcrição associados a cada grupo do agrupamento SOM = 2x3.

Tabela A.44: Validação biológica do agrupamento SOM = 2x3.

Grupos	TANGO	PRIMA
1		MCM1
2	GO:0009308 - Metabolismo de amido	
3	GO:0003677 - Ligação do DNA GO:0006333 - União ou separação da cromatina	STB1
5		SWI5
6	GO:0008283 - Proliferação celular GO:0006260 - Replicação do DNA GO:0006259 - Metabolismo do DNA GO:0009719 - Resposta à estímulos endógenos	MBP1

SOM = 2x4

O perfil médio da expressão dos genes dos 8 grupos é ilustrado na Figura A.22.

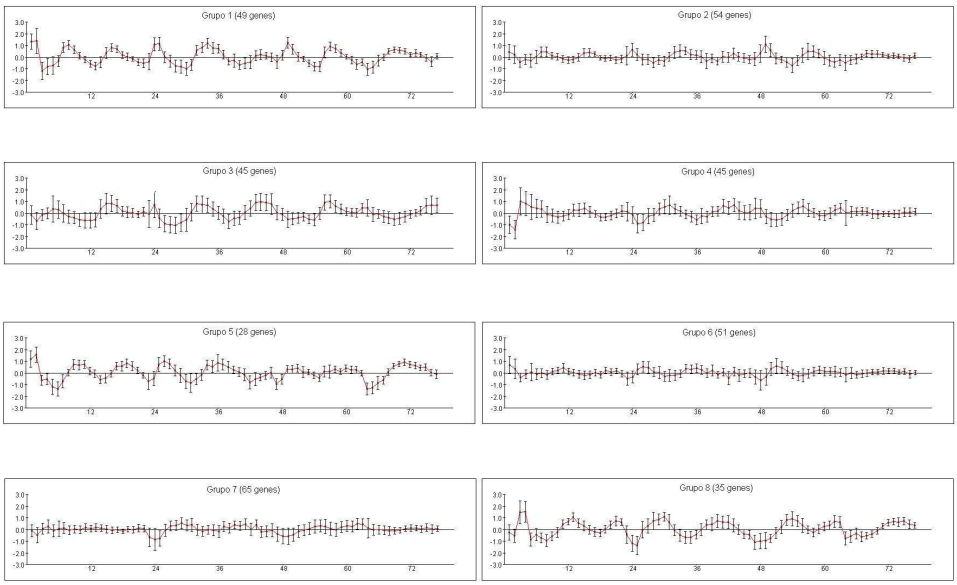


Figura A.22: Perfil médio da expressão dos genes dos 8 grupos (SOM = 2x4).

Os grupos 1 e 3 foram formados por genes envolvidos na fase G1 do ciclo celular. O grupo 3 por genes da fase M (M/G1) e G1, o grupo 4 por genes da fase M (G2/M e M/G1), o grupo 5 por genes das fases S e G2, o grupo 6 por genes das fases G1, S e G2, o grupo 7 por genes das fases M(G2/M) e G2 e o grupo 8 por genes da fase M (G2/M).

No grupo 1, formado por 49 genes, foram identificadas as funções de replicação do DNA e proliferação celular, no grupo 4, com 45 genes, a função de resposta ao feromônio, no grupo 5, com 28 genes, as funções de organização celular e biogênese e união ou separação da cromatina, no grupo 6, com 51 genes, a função de metabolismo de enxofre e no grupo 7, formado por 65 genes, foi identificada a função de transporte.

A Tabela A.45 apresenta as funções biológicas e fatores de transcrição associados a cada grupo do agrupamento SOM = 2x4.

Tabela A.45: Validação biológica do agrupamento SOM = 2x4.

Grupos	TANGO	PRIMA
1	GO:0008283 - Proliferação celular GO:0006260 - Replicação do DNA	MBP1 STB1
3		SWI5
4	GO:0000749 - Resposta do feromônio durante conjugação com fusão celular	MCM1
5	GO:0006333 - União ou separação da cromatina GO:0016043 - Organização celular e biogênese	
6	GO:0006790 - Metabolismo de enxofre	MET31
7	GO:0006810 - Transporte	MCM1

SOM = 2x5

O perfil médio da expressão dos genes de cada um dos 10 grupos é ilustradona Figura A.23.

O grupo 1 foi formado por genes envolvidos na fase M (G2/M), o grupo 2 por genes da fase M (M/G1), o grupo 3 por genes das fases M (M/G1) e G1, os grupos 4 e 5 por genes da fase G1, os grupos 6 e 7 por genes da fase M (G2/M), o grupo 8 por genes da fase M (G2/M e M/G1) e os grupos 9 e 10 foram formados por genes envolvidos nas fases S e G2.

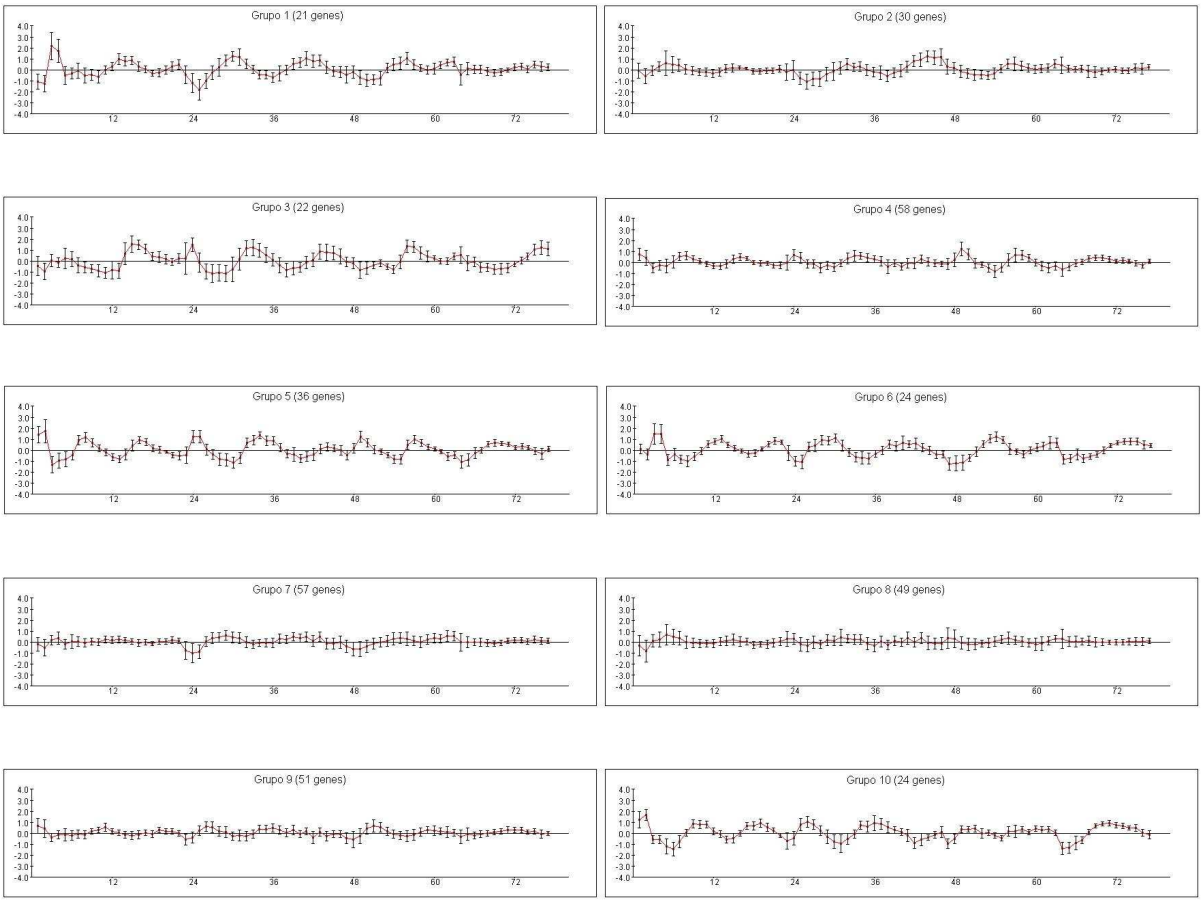


Figura A.23: Perfil médio da expressão dos genes dos 10 grupos (SOM = 2x5).

A Tabela A.46 apresenta as funções biológicas e fatores de transcrição associados a cada grupo do agrupamento SOM = 2x5.

Tabela A.46: Validação biológica do agrupamento SOM = 2x5.

Grupos	TANGO	PRIMA
1	GO:0042578 - Atividade de hidrolase éster fosfórico	
2	GO:0000746 - Conjugação	
3	GO:0006259 - Metabolismo do DNA GO:0009719 - Resposta a estímulos endógenos	ACE2 SWI5
4	GO:0000746 - Conjugação	MBP1
5	GO:0008283 - Proliferação celular GO:0006260 - Replicação do DNA	MBP1 STB1
6		MCM1
7	GO:0006810 - Transporte	
9	GO:0006790 - Metabolismo de enxofre	MET31
10	GO:0006333 - União ou separação da cromatina GO:0016043 - Organização celular e biogênese	

A.2.8 Validação biológica SAMBA

Os resultados da aplicação do algoritmo bidimensional na base de dados filtrada não apresentou diferença significativa se comparado com a aplicação na base de dados sem filtros de dados. Os mesmos grupos de funções mais evidentes, como a função de estrutura de cromatina e replicação, por exemplo, estão presentes em ambos os agrupamentos, embora neste os grupos sejam menores.

no grupo 26 da fase G1 a função de proliferação celular e no grupo 44 também representativo da fase G1 foram identificadas as funções de atividade de hidrolase e cito-quinase, funções não identificadas em nenhum dos agrupamentos anteriores.

A Tabela A.47 apresenta as funções biológicas e fatores de transcrição associados a cada grupo do agrupamento SAMBA.

Tabela A.47: Validação biológica do agrupamento SAMBA.

Grupos	TANGO	PRIMA
2	GO:0019236 - Resposta do feromônio durante conjugação celular	
4		ACE2 SWI5
6	GO:0016043 - Organização celular e biogênese	
7		MCM1
8		SWI5
11	GO:0000746 - Conjugação	
15	GO:0006333 - União ou separação da cromatina GO:0003677 - Ligação do DNA	
16	GO:0006333 - União ou separação da cromatina GO:0003677 - Ligação do DNA	
23	GO:0006260 - Replicação do DNA	MBP1
26	GO:0008283 - Proliferação celular	
29		SWI5
37		MBP1 STB1
39		MCM1
43		MSN2 MSN4 ADR1
44	GO:0008283 - Atividade de hidrolase GO:0000910 - Citoquinase	ACE2
45		ACE2

A.3 Conclusão dos resultados da base de dados CCSc

A utilização da base de dados CCSc trouxe vantagens como modelo experimental. O conhecimento *a priori* dos dados permitiu avaliar o potencial das técnicas de validação estatística e biológica e o comportamento das diferentes abordagens de agrupamento de dados.

k-médias

A maioria dos índices estatísticos indicaram o $k = 2$ como a melhor solução de agrupamento, para os experimentos realizados com e sem a aplicação de filtros de dados. Do ponto de vista biológico (não considerando o resultado da validação biológica) este agrupamento demonstrou potencial significância, uma vez que os grupos foram representativos das 2 fases do ciclo celular, obedecendo a ordem que ocorrem: as fases G2 e M no grupo 1 e as fases G1 e S no grupo 2.

Já de acordo com as técnicas de validação biológica, o melhor agrupamento foi $k = 10$, onde foi associado maior número de funções biológicas e fatores de transcrição, 17 e 5 respectivamente. Este agrupamento, no entanto, foi indicado com a melhor solução somente pelo índice estatístico C.

Esta análise levantou a questão de que o índice C foi o único corroborado pelos resultados da validação biológica.

Todos os agrupamentos do k-médias (sem o uso de filtros) foram prejudicados pela formação de um grupo contendo somente um gene. A formação deste grupo ocorreu em decorrência da implementação do algoritmo k-médias de definir os centróides como os primeiros elementos da base de dados.

SOM

Todos os agrupamentos do SOM foram indicados como melhor solução por algum dos índices estatísticos. No entanto, o agrupamento SOM = 2x2 foi indicado por 3 índices diferentes, para os experimentos realizados com e sem a aplicação de filtros de dados.

Os agrupamentos SOM = 2x2 demonstraram bastante significância biológica. Os grupos não foram representativos das 4 fases do ciclo celular como era esperado, mas foram agrupados de acordo com a transição das fases.

Os resultados do SOM não foram prejudicados pela presença dos genes pouco representativos, como ocorreu com o k-médias e ainda assim os grupos apresentaram alta homogeneidade.

Os algoritmos k-médias e SOM são popularmente utilizados para a análise de dados de expressão gênica devido às suas características de fácil utilização e apresentarem bom desempenho. Estas características foram confirmadas neste trabalho, embora o algoritmo SOM tenha apresentado resultados melhores.

Aplicação de filtros de dados

Um comparativo dos resultados do algoritmo k-médias com e sem o uso de filtro

de dados é apresentado na Tabela A.48.

De acordo com a maioria dos índices estatísticos, os melhores agrupamentos do k-médias foram obtidos com o uso de filtros de dados.

Tabela A.48: Comparação dos agrupamentos k-médias com e sem aplicação de filtros.

k	Homogeneidade	Separação	C	D-Bouldin	Dunn	Silhueta	Isolamento
2	0.431	-0.094	0.341	**1.761**	1.088	0.143	0.863
2 Filtro	**0.448**	**-0.119**	**0.29**	1.805	**1.09**	**0.165**	**0.869**

k	Homogeneidade	Separação	C	D-Bouldin	Dunn	Silhueta	Isolamento
4	0.491	-0.049	**0.287**	**1.765**	0.702	**0.087**	**0.526**
4 Filtro	**0.524**	**-0.075**	0.291	1.927	**0.886**	0.043	0.514

k	Homogeneidade	Separação	C	D-Bouldin	Dunn	Silhueta	Isolamento
5	0.364	-0.017	0.257	1.749	0.672	**0.101**	0.392
5 Filtro	**0.522**	**-0.068**	**0.28**	**1.914**	**0.838**	0.039	**0.481**

k	Homogeneidade	Separação	C	D-Bouldin	Dunn	Silhueta	Isolamento
8	0.38	0.006	0.236	**1.72**	**0.605**	**0.058**	0.365
8 Filtro	**0.425**	**-0.026**	**0.198**	1.775	0.466	0.038	**0.475**

k	Homogeneidade	Separação	C	D-Bouldin	Dunn	Silhueta	Isolamento
10	0.372	0.012	0.206	1.768	0.579	**0.06**	0.352
10 Filtro	**0.431**	**-0.022**	**0.183**	**1.677**	**0.63**	0.042	**0.46**

Um comparativo dos resultados do algoritmo SOM com e sem o uso de filtro de dados é apresentado na Tabela A.49.

Tabela A.49: Comparação dos agrupamentos SOM com e sem aplicação de filtros.

SOM	Homogeneidade	Separação	C	D-Bouldin	Dunn	Silhueta	Isolamento
2x2	0.576	-0.045	0.231	1.787	0.865	0.101	0.672
2x2 Filtro	**0.59**	**-0.055**	**0.209**	1.787	0.865	**0.124**	**0.733**

SOM	Homogeneidade	Separação	C	D-Bouldin	Dunn	Silhueta	Isolamento
5x1	**0.554**	**-0.031**	**0.188**	**1.663**	0.828	**0.094**	0.579
5x1 Filtro	0.541	-0.02	0.194	1.771	**0.891**	0.084	0.579

Continua na próxima página

Tabela A.49 – continuação da página anterior

SOM	Homogeneidade	Separação	C	D-Bouldin	Dunn	Silhueta	Isolamento
2x4	0.557	**-0.004**	**0.142**	**1.696**	0.755	**0.083**	0.499
2x4 Filtro	**0.604**	-0.001	0.165	1.804	**0.782**	0.078	**0.534**

SOM	Homogeneidade	Separação	C	D-Bouldin	Dunn	Silhueta	Isolamento
2x5	0.588	0.008	0.144	1.814	0.733	0.065	0.461
2x5 Filtro	**0.606**	**0.005**	**0.134**	**1.766**	**0.82**	**0.09**	**0.511**

Para o algoritmo SOM a vantagem da aplicação dos filtros de dados não ficou tão evidente assim como no k-médias. Os índices estatísticos indicaram como melhores soluções de agrupamento o SOM = 2x2 e 2x5 com filtro, o SOM = 5x1 e 2x4 sem filtro, conforme a Tabela A.49. Análises biológicas, inclusive, demonstraram que o algoritmo SOM não é sensível à ruídos como o k-médias.

A Tabela A.50 ilustra a comparação dos agrupamentos dos algoritmos k-médias e SOM, de acordo com os índices estatísticos.

Tabela A.50: Comparação dos agrupamentos k-médias e SOM com e sem aplicação de filtros.

Grupo	Homogeneidade	Separação	C	D-Bouldin	Dunn	Silhueta	Isolamento
4	0.491	-0.049	0.287	**1.765**	0.702	0.087	0.526
4 Filtro	0.524	**-0.075**	0.291	1.927	**0.886**	0.043	0.514
2x2	0.576	-0.045	0.231	1.787	0.865	0.101	0.672
2x2 Filtro	**0.59**	-0.055	**0.209**	1.787	0.865	**0.124**	**0.733**

Grupo	Homogeneidade	Separação	C	D-Bouldin	Dunn	Silhueta	Isolamento
5	0.364	-0.017	0.257	1.749	0.672	**0.101**	0.392
5 Filtro	0.522	**-0.068**	0.28	1.914	0.838	0.039	0.481
5x1	**0.554**	-0.031	**0.188**	**1.663**	0.828	0.094	**0.579**
5x1 Filtro	0.541	-0.02	0.194	1.771	**0.891**	0.084	**0.579**

Grupo	Homogeneidade	Separação	C	D-Bouldin	Dunn	Silhueta	Isolamento
8	0.38	0.006	0.236	1.72	0.605	0.058	0.365
8 Filtro	0.425	**-0.026**	0.198	1.775	0.466	0.038	0.475
2x4	0.557	-0.004	**0.142**	**1.696**	0.755	**0.083**	0.499
2x4 Filtro	0.604	-0.001	0.165	1.804	**0.782**	0.078	**0.534**

Continua na próxima página

Tabela A.50 – continuação da página anterior

Grupo	Homogeneidade	Separação	C	D-Bouldin	Dunn	Silhueta	Isolamento
10	0.372	0.012	0.206	1.768	0.579	0.06	0.352
10 Filtro	0.431	**-0.022**	0.183	**1.677**	0.63	0.042	0.46
2x5	0.588	0.008	0.144	1.814	0.733	0.065	0.461
2x5Filtro	**0.606**	0.005	**0.134**	1.766	**0.82**	**0.09**	**0.511**

Para o agrupamento com 4 grupos, os índices de homogeneidade, C, Silhueta e Isolamento indicaram a solução do SOM = 2x2 com o uso de filtros como a melhor opção.

Para o agrupamento com 5 grupos, os índices de homogeneidade, C, D-Bouldin e Isolamento indicaram a solução do SOM = 5x1 sem o uso de filtros como a melhor opção.

Para o agrupamento com 8 grupos, os índices de homogeneidade , Dunn e Isolamento indicaram a solução do SOM = 2x4 com o uso de filtros como a melhor opção junto com o SOM = 2x4 sem o uso de filtros, indicado pelos índices C, D-Bouldin e Silhueta.

Para o agrupamento com 10 grupos, os índices de homogeneidade, C, Dunn, Silhueta e Isolamento indicaram a solução SOM = 2x5 com o uso de filtros como a melhor opção.

Para o índice de separação, as melhores soluções de agrupamento foram todas resultantes do algoritmo k-médias. No entanto, de acordo com a significância biológica obtida das análises de validação biológica, as indicações do índice de separação não se apresentam como uma boa opção.

Baseado nas indicações dos índices estatísticos, o SOM demonstrou ser a melhor opção de algoritmo de agrupamento se comparado ao k-médias.

A Tabela A.51 ilustra a comparação dos agrupamentos dos algoritmos k-médias e SOM, de acordo com as ferramentas de validação biológica TANGO e PRIMA. A indicação das melhores soluções de agrupamento é indicada na tabela com valores em negrito.

Tabela A.51: Comparação biológica dos agrupamentos k-médias e SOM.

Grupo	TANGO	PRIMA
4	12	3
2x2	**15**	**5**

Grupo	TANGO	PRIMA
5	8	3
5x1	**16**	**5**

Continua na próxima página

Tabela A.51 – continuação da página anterior

Grupo	TANGO	PRIMA
8	10	**5**
2x4	**14**	**5**

Grupo	TANGO	PRIMA
10	17	5
2x5	**18**	**6**

Em todos os casos, as ferramentas de validação biológica indicaram os agrupamentos do algoritmo SOM com a melhor solução de agrupamento, se comparado com os resultados do algoritmo k-médias.

Portanto, das análises dos índices estatísticos combinadas com as análises das técnicas de validação biológica, é possível concluir que os índices de homogeneidade e o índice C são mais apropriados para o problema de análise de dados de expressão gênica. Esta comparação é baseada somente nos agrupamentos sem o uso de filtro de dados, já que é intrínseca à quantidade de funcões biológicas e fatores de transcrição associadas a cada agrupamento.

Além da indicação dos índices estatísticos mais apropriados, também é possível identificar as soluções do algoritmo SOM como melhores que do algoritmo k-médias.

SAMBA

Das análises obtidas do algoritmo SAMBA de agrupamento bidimensional é possível concluir suas vantagens de custo computacional e sua habilidade na identificação de subgrupos de genes e condições, comparados com a performance dos algoritmos k-médias e SOM de agrupamento unidimensional.

Do ponto de vista biológico, a vantagem da utilização da abordagem bidimensional é evidente, embora não faça sentido afirmar que esta aplicação seja melhor que de um algoritmo unidimensional como o SOM, por exemplo. A definição de qual técnica utilizar depende do objetivo do problema. A técnica bidimensional é indicada para quando o objetivo é identificar grupos mais específicos, de genes expressos somente em determinadas condições. Porém, não é indicada quando se deseja encontrar um número pré-definido de grupos, como neste caso, com a utilização da base de dados CCSc. Também não é uma técnica prática quando o objetivo é ter uma visão geral do comportamento dos genes da base de dados.

Uma aplicação interessante da abordagem bidimensional é utilizá-la em conjunto com algoritmos unidimensionais, pois ela é capaz de revelar uma variedade de estruturas que podem estar presentes no conjunto de dados.

Esta técnica é mais indicada para bases de dados que contenham uma grande quantidade de condições, envolvidas em diferentes processos biológicos, o que não era o caso da base de dados CCSc, cujas as condições são todas relacionadas ao processo de ciclo celular. Já para base de dados como a GSc, com condições de ciclo celular e estresse nutricional, as vantagens dessa abordagem foram mais destacadas.

O algoritmo bidimensional também demonstrou não ser sensível a ruídos, pois mesmo sem a aplicação dos filtros, todos os genes foram agrupados em grupos significativos.

Funções biológicas e fatores de transcrição do agrupamento SAMBA

As técnicas de validação biológica de enriquecimento funcional e identificação de fatores de transcrição foram bastante úteis para enriquecer a significância dos resultados dos agrupamentos das abordagens unidimensional e bidimensional. Estes resultados ainda permitiram a associação de funções biológicas e fatores de transcrição a cada fase do ciclo celular, conforme apresentados na Tabela A.52.

Tabela A.52: Validação biológica dos agrupamentos.

Fase do ciclo celular	Função biológica	TF
G1	Metabolismo do DNA	MBP1
	Replicação do DNA	STB1
	Reparo do DNA	ACE1
	Recombinação do DNA	ADR1
	Alongamento da fita de DNA	
	Condensação do cromossomo	
	Ciclo celular	
	Metabolismo primário	
	Atividade de endonuclease	
	União ou separação da cromatina	
	Manutenção dos telômeros	
	Transição da fase G1/S do ciclo celular mitótico	
	Recombinação mitótica	
	Resposta à estímulos de danos ao DNA	
	Resposta á estímulos endógenos	
	Metabolismo de biopolímero	
Continua na próxima página		

Tabela A.52 – continuação da página anterior

	Metabolismo de ácido nucléico Atividade de hidrolase Citoquinase	
S e G2	Metabolismo de enxofre Metabolismo de aminoácido Metabolismo de amido União ou separação da cromatina Organização celular e biogênese Org. da estrutura de encapsulamento externo e biogênese	HAC1
G2/M	Transporte de cálcio Localização Atividade de transporte de cálcio inorgânico Atividade da helicase dependente de ATP Atividade de transporte de açúcar	ABF1 GCN4 BAS1 MET31
M/G1	Resposta ao feromônio Comunicação celular Conjugação Atividade de transporte de açúcar — — —	AZF1 PHO4 MSN2 MSN4 SWI5 AZF1 ADR1

Aplicação de filtros de dados

Finalmente, os resultados dos agrupamentos quando aplicados filtros na base de dados foram semelhantes aos resultados obtidos sem a aplicação dos filtros, embora tenha a qualidade maximizada nos resultados do algoritmo k-médias.

Os processos de validação estatística e biológica indicaram os melhores agrupamentos como sendo os mesmos indicados quando não aplicados os filtros. As funções biológicas e os fatores de transcrição identificados foram em menor quantidade porque são intrínsecos à quantidade de genes nos grupos.

A utilização de filtros de dados serve como alternativa para minimizar a quantidade de dados redundantes, ruidosos ou irrelevantes, que podem influenciar na tarefa de agrupamento. Além disso, oferece a vantagem da diminuição do tempo computacional.

No agrupamento bidimensional a aplicação dos filtros não apresentou vantagem. Foram formados 45 grupos, comparados aos 50 grupos da base de dados sem filtros, o tempo de execução de ambos os agrupamentos foi praticamente o mesmo e os grupos

formados foram menores e menos expressivos.

Teoricamente a utilização de filtros com a abordagem bidimensional não faz sentido. Nesta abordagem, genes pouco representativos ou ruidosos não são agrupados, genes que apresentam ruídos em somente algumas condições são agrupados somente considerando as condições significativas.

A aplicação dos filtros depende, portanto, do objetivo do problema. Se a intenção é identificar genes participantes de um mesmo processo biológico, os filtros podem ser prejudiciais. Mas essa decisão depende, entre outros fatores, da qualidade dos dados e das informações contidas na base de dados. Dependendo do problema, uma opção interessante seria a eliminação somente da redundância e não da irrelevância. Este processo é mais comum para a tarefa de classificação, mas serve como sugestão de trabalho futuro. Uma outra opção seria a ponderação das condições ao invés da eliminação. Assim as informações não seriam descartadas, apenas seria atribuído peso maior às condições mais relevantes. O programa *Cluster* implementa esta opção [MBEB98].

A.4 Base de dados GSc

A.4.1 Agrupamento k-médias

Ao contrário da base de dados CCSc em que havia conhecimento prévio dos dados, na base GSc não havia nenhum conhecimento a priori. Os 6621 correspondem ao genoma completo do organismo *Sacharomyces cerevisiae*.

O algoritmo k-médias foi aplicado com o valor de $k = 5, 10, 20, 30$ *e* 50. Estes valores foram escolhidos de maneira a facilitar a análise dos resultados dos agrupamentos e por serem valores que pudessem indicar padrões dos genes agrupados de acordo com às 80 condições experimentais do microarranjo. Os resultados obtidos foram os seguintes:

k = 5

Os 6621 genes foram agrupados em 5 grupos, conforme mostrado na Tabela A.53.

Tabela A.53: Agrupamento $k = 5$.

Grupos	Quantidade de genes	Homogeneidade
1	278	0,564
2	687	0,655
Continua na próxima página		

Tabela A.53 – continuação da página anterior

3	104	0,878
4	2732	0,359
5	2420	0,427

k = 10

Os 6621 genes foram agrupados em 10 grupos, conforme mostrado na Tabela A.54.

Tabela A.54: Agrupamento $k = 10$.

Grupos	Quantidade de genes	Homogeneidade
1	201	0,63
2	509	0,7
3	102	0,88
4	1148	0,482
5	731	0,524
6	95	0,831
7	1747	0,374
8	298	0,513
9	2	0,99
10	1388	0,489

k = 20

Os 6621 genes foram agrupados em 20 grupos, conforme mostrado na Tabela A.55.

Tabela A.55: Agrupamento $k = 20$.

Grupos	Quantidade de genes	Homogeneidade
1	121	0,677
2	386	0,676
3	72	0,911

Continua na próxima página

Tabela A.55 – continuação da página anterior

4	855	0,51
5	432	0,487
6	95	0,831
7	1400	0,39
8	169	0,544
9	1	1,0
10	854	0,537
11	339	0,552
12	34	0,672
13	50	0,632
14	25	0,612
15	484	0,449
16	33	0,743
17	126	0,571
18	579	0,615
19	105	0,888
20	61	0,663

k = 30

Os 6621 genes foram agrupados em 30 grupos, conforme mostrado na Tabela A.56.

Tabela A.56: Agrupamento $k = 30$.

Grupos	Qtd de genes	Homogeneidade
1	101	0,678
2	254	0,71
3	72	0,913
4	718	0,502
5	394	0,503
6	91	0,844
7	1101	0,398

Continua na próxima página

Tabela A.56 – continuação da página anterior

8	156	0,552
9	1	1,0
10	648	0,558
11	292	0,566
12	16	0,688
13	43	0,675
14	16	0,63
15	369	0,456
16	24	0,75
17	118	0,586
18	428	0,638
19	102	0,893
20	18	0,697
21	609	0,561
22	16	0,594
23	46	0,606
24	83	0,478
25	40	0,683
26	27	0,812
27	33	0,52
28	263	0,621
29	18	0,58
30	124	0,652

k = 50

Os 6621 genes foram agrupados em 50 grupos, conforme mostrado na Tabela A.57.

Tabela A.57: Agrupamento $k = 50$.

Grupos	Qtd de genes	Homogeneidade
1	70	0,704
		Continua na próxima página

Tabela A.57 – continuação da página anterior

2	202	0,702
3	45	0,937
4	946	0,481
5	443	0,462
6	81	0,848
7	944	0,505
8	170	0,496
9	1	1,0
10	370	0,533
11	31	0,696
12	43	0,645
13	12	0,718
14	17	0,805
15	338	0,595
16	55	0,885
17	26	0,697
18	18	0,576
19	41	0,601
20	95	0,488
21	45	0,682
22	24	0,817
23	41	0,505
24	24	0,654
25	154	0,613
26	199	0,66
27	23	0,715
28	2	0,993
29	44	0,748
30	1	1,0
31	29	0,928
32	244	0,556
33	278	0,405
34	50	0,779

Continua na próxima página

Tabela A.57 – continuação da página anterior

35	1	1,0
36	30	0,663
37	8	0,673
38	35	0,717
39	17	0,698
40	84	0,89
41	13	0,853
42	219	0,758
43	202	0,592
44	18	0,609
45	12	0,932
46	156	0,689
47	139	0,731
48	1	1,0
49	170	0,719
50	10	0,633

A.4.2 Agrupamento SOM

O algoritmo SOM foi aplicado com as dimensões da matriz definidas por 5x1, 2x5, 5x5 e 5x10. Os resultados obtidos foram os seguintes:

SOM = 5x1

A quantidade de genes e o índice de homogeneidade de cada grupo são apresentadas na Tabela A.58.

Tabela A.58: Agrupamento SOM = 5x1.

Grupos	Quantidade de genes	Homogeneidade
1	617	0,714
2	2670	0,332
3	1035	0,521
4	1590	0,489

Continua na próxima página

Tabela A.58 – continuação da página anterior

5	309	0,801

SOM = 2x5

A quantidade de genes e o índice de homogeneidade de cada grupo são apresentadas na Tabela A.59.

Tabela A.59: Agrupamento SOM = 2x5.

Grupos	Quantidade de genes	Homogeneidade
1	213	0,836
2	656	0,56
3	1025	0,573
4	614	0,51
5	320	0,802
6	654	0,628
7	702	0,447
8	1234	0,333
9	386	0,594
10	417	0,686

SOM = 5x5

A quantidade de genes e o índice de homogeneidade de cada grupo são apresentadas na Tabela A.60.

Tabela A.60: Agrupamento SOM = 5x5.

Grupos	Qtd de genes	Homogeneidade
1	116	0,595
2	402	0,575
3	149	0,71
		Continua na próxima página

Tabela A.60 – continuação da página anterior

4	144	0,817
5	104	0,895
6	440	0,527
7	164	0,513
8	268	0,542
9	340	0,659
10	220	0,747
11	243	0,513
12	668	0,443
13	196	0,619
14	370	0,62
15	213	0,548
16	199	0,801
17	282	0,626
18	301	0,653
19	230	0,583
20	454	0,464
21	143	0,878
22	103	0,774
23	216	0,697
24	109	0,744
25	147	0,646

SOM = 5x10

A quantidade de genes e o índice de homogeneidade de cada grupo são apresentadas na Tabela A.61.

Tabela A.61: Agrupamento SOM = 5x10.

Grupos	Qtd de genes	Homogeneidade
1	107	0,672
2	123	0,605
		Continua na próxima página

Tabela A.61 – continuação da página anterior

3	157	0,657
4	166	0,691
5	101	0,674
6	120	0,582
7	188	0,505
8	121	0,588
9	91	0,623
10	74	0,751
11	98	0,629
12	160	0,665
13	125	0,644
14	148	0,656
15	130	0,607
16	204	0,571
17	184	0,445
18	134	0,604
19	143	0,677
20	80	0,765
21	91	0,804
22	101	0,739
23	81	0,753
24	87	0,579
25	96	0,546
26	177	0,566
27	157	0,559
28	138	0,585
29	164	0,688
30	81	0,743
31	51	0,89
32	116	0,771
33	145	0,623
34	158	0,412
35	147	0,618

Continua na próxima página

Tabela A.61 – continuação da página anterior

36	152	0,525
37	39	0,596
38	163	0,584
39	141	0,705
40	61	0,797
41	82	0,908
42	136	0,791
43	132	0,635
44	147	0,526
45	130	0,644
46	58	0,695
47	164	0,608
48	140	0,745
49	108	0,838
50	124	0,894

A.4.3 Agrupamento SAMBA

Tabela A.62: Agrupamento bidimensional da base de dados GSc.

Grupos	Escores	Condições	Genes
1	493,856	5	134
2	800,811	7	156
3	730,992	8	120
4	689,543	7	154
5	557,317	5	137
6	785,764	9	110
7	639,561	8	137
8	897,656	6	239
9	657,4	11	121
10	942,877	5	208
			Continua na próxima página

Tabela A.62 – continuação da página anterior

11	950,175	4	250
12	398,213	9	95
13	649,743	6	111
14	1388,92	7	236
15	570,272	6	125
16	799,493	5	201
17	815,766	5	216
18	827,276	6	205
19	918,171	7	173
20	733,46	7	142
21	1628,57	7	232
22	1334,62	5	248
23	1838,6	5	308
24	521,401	7	126
25	2147,72	6	280
26	364,853	5	93
27	783,672	5	218
28	211,754	6	33
29	421,574	9	39
30	222,639	7	23
31	338,014	9	31
32	226,227	4	41
33	347,283	8	40
34	335,992	6	64
35	226,678	4	56
36	305,061	8	56
37	357,873	6	61
38	193,286	6	52
39	333,497	8	35
40	276,789	6	63
41	241,694	10	24
42	418,324	10	40
43	408,664	8	50

Continua na próxima página

Tabela A.62 – continuação da página anterior

44	283,887	12	33
45	604,798	12	69
46	96,8258	7	17
47	512,468	11	40
48	200,699	6	37
49	655,26	8	104
50	1056,94	16	97
51	460,65	10	66
52	1061,62	10	159
53	1167,61	13	78
54	594,752	11	47
55	715,281	8	69
56	1119,25	16	94
57	1532,37	14	168
58	1989,58	12	186
59	1939,75	19	118
60	464,669	22	33
61	871,159	12	143
62	310,277	15	42
63	2693,92	8	291
64	1900,34	13	189
65	235,221	7	53
66	928,101	10	73
67	1073,0	12	114

A.4.4 Validação estatística k-médias

Os melhores agrupamentos da base de dados GSc são indicados nas Tabelas A.63 e A.64. Dos resultados obtidos do algoritmo k-médias, o melhor agrupamento foi $k = 50$, de acordo com os índices de homogeneidade e C, para o índice Davies Bouldin quando $k = 10$ e para os índices de separação, Dunn, Silhueta e Isolamento quando $k = 5$.

Tabela A.63: Validação estatística dos agrupamentos k-médias.

k	Homogeneidade	Separação	C	D. Bouldin	Dunn	Silhueta	Isolamento
5	0,399	**-0,005**	0,198	1,85	**0,765**	**0,069**	**0,489**
10	0,452	0,019	0,213	**1,77**	0,75	0,034	0,396
20	0,476	0,031	0,168	1,856	0,498	0,019	0,263
30	0,496	0,037	0,162	1,846	0,385	0,008	0,221
50	**0,519**	0,039	**0,146**	1,779	0,457	0,018	0,218

A.4.5 Validação estatística SOM

Para os resultados do algoritmo SOM, o melhor agrupamento foi SOM = 5x10, de acordo com o índice de homogeneidade. O índice C indicou o melhor agrupamento SOM = 5x1, os índices Davies Bouldin e Dunn quando SOM = 5x1, e os demais índices quando SOM = 2x2.

Tabela A.64: Validação estatística dos agrupamentos SOM.

Matriz	Homogeneidade	Separação	C	D. Bouldin	Dunn	Silhueta	Isolamento
5x1	0,402	0,007	0,193	**1,777**	**0,927**	**0,078**	**0,521**
2x5	0,494	0,026	0,165	1,845	0,685	0,04	0,397
5x5	0,563	0,041	0,148	1,848	0,445	0,022	0,308
5X10	**0,627**	0,048	0,125	1,854	0,354	0,007	0,258

A.4.6 Validação biológica k-médias

k = 5

A Tabela A.65 apresenta as funções biológicas e os fatores de transcrição identificados nos grupos.

Tabela A.65: Validação biológica do agrupamento $k = 5$.

Grupos	PRIMA	TANGO
Grupo 1	GO: 0016491 - Atividade de oxidoredutase	HAP4
	GO: 0005975 - Metabolismo de carboidrato	STRE
	GO: 0006100 - Metabolismo intermediário do ciclo de ácido tricarboxílico	MSN2
	GO: 0006119 - Fosforilação oxidativa	MSN4
	GO: 0006091 - Geração de metabólitos precursores e energia	PUT3
	GO: 0006536 - Metabolismo de glutamato	ADR1
	GO: 0006099 - Ciclo do ácido tricarboxílico	MIG1
	GO: 0005489 - Atividade de transporte de elétron	SUT1
	GO: 0051186 - Metabolismo de cofator	HAP 2/3/4
	GO: 0015078 - Atividade de transporte de íon	
	GO: 0005386 - Atividade de carregamento	
	GO: 0006118 - Transporte de elétron	
Grupo 2	GO: 0009112 - Metabolismo de nucleotídeo	ABF1
	GO: 0003724 - Atividade de RNA helicase	SFP1
	GO: 0003735 - Constituinte estrutural do ribossomo	AZF1
	GO: 0042273 - Biogênese da subunidade ribossomal maior	RAP1
	GO: 0003743 - Atividade do fator de tradução	
	GO: 0003899 - Atividade da RNA polimerase	
	GO: 0009451 - Modificação do RNA	
	GO: 0003723 - Ligação do RNA	
	GO: 0009059 - Biosíntese de macromolécula	
	GO: 0006163 - Metabolismo de nucleotídeo (purina)	
	GO: 0006412 - Biosíntese de proteína	
	GO: 0016886 - Formação de éster fosfato / atividade ligase	

Continua na próxima página

Tabela A.65 – continuação da página anterior

	GO: 0042255 - União do ribossomo GO: 0043037 - Tradução GO: 0042257 - União da subunidade ribossomal GO: 0030515 - Ligação de snRNAs GO: 0006520 - Metabolismo de aminoácido GO: 0007028 - Organização do citoplasma e biogênese GO: 0016070 - Metabolismo do RNA GO: 0019320 - Catabolismo da hexose GO: 0030490 - Processamento do pré-RNA GO: 0044249 - Biosíntese celular GO: 0044238 - Metabolismo primário	
Grupo 3	GO: 0007126 - Meiose GO: 0042244 - Formação da membrana celular de um esporo GO: 0030435 - Esporulação GO: 0000279 - Fase M	SUM1
Grupo 4	GO: 0051641 - Localização celular GO: 0030029 - Processo baseado no filamento de actina GO: 0005515 - Ligação de proteína GO: 0051179 - Localização GO: 0019725 - Homeostase celular GO: 0016192 - Vesículo mediado por transporte GO: 0006366 - Transcrição da RNA polimerase II	ADR1
Grupo 5		FKH1

k = 10

A Tabela A.66 apresenta as funções biológicas e os fatores de transcrição identificados nos grupos.

Tabela A.66: Validação biológica do agrupamento $k = 10$.

Grupos	TANGO	PRIMA
Grupo 1	GO: 0016491 - Atividade de oxiredutase	HAP 4
	GO: 0005975 - Metabolismo de Carboidrato	STRE
	GO: 0006099 - Ciclo do ácido tricarboxílico	MSN2
	GO: 0006119 - Fosforilação oxidativa	MSN4
	GO: 0006091 - Geração de metabólitos precursores e energia	PUT3
	GO: 0006536 - Metabolismo de glutamato	ADR1
	GO: 0005489 - Atividade de transporte de elétron	MIG1
	GO: 0051186 - Metabolismo de cofator	SUT1
	GO: 0015078 - Atividade de transporte de íon	HAP 2/3/4
	GO: 0006118 - Transporte de elétron	
Grupo 2	GO: 0009112 - Metabolismo de nucleotídeo	SFP1
	GO: 0003735 - Constituinte estrutural do ribossomo	AZF1
	GO: 0003899 - Atividade da RNA polimerase	RAP1
	GO: 0006399 - Metabolismo do tRNA	
	GO: 0009451 - Modificação do RNA	
	GO: 0044237 - Metabolismo celular	
	GO: 0009059 - Biosíntese de macromolécula	
	GO: 0006163 - Metabolismo de nucleotídeo (purina)	
	GO: 0006412 - Biosíntese de proteína	
	GO: 0016886 - Formação de éster fosfato / atividade ligase	
	GO: 0006519 - Metabolismo de aminoácido	
	GO: 0043037 - Tradução	
	GO: 0042257 - União da subunidade ribossomal	
	GO: 0007028 - Organização do citoplasma e biogênese	
	GO: 0019320 - Catabolismo da hexose	
	GO: 0044249 - Biosíntese celular	
	GO: 0044238 - Metabolismo primário	

Continua na próxima página

Tabela A.66 – continuação da página anterior

Grupo 3	GO: 0007126 - Meiose GO: 0042244 - Formação da membrana celular de um esporo GO: 0030435 - Esporulação GO: 0000279 - Fase M	SUM1
Grupo 4	GO: 0030036 - Organização do citoesqueleto de actina e biogênese GO: 0005515 - Ligação de proteína GO: 0006366 - Transcrição do promotor da Pol II GO: 0000866 - Organização do citoesqueleto de actina cortical e biogênese	ADR1
Grupo 5	GO: 0007017 - Processo dependente de microtúbulo GO: 0000819 - Segregação da cromátide irmã GO: 0016458 - Silenciamento de genes GO: 0007091 - Transição das fases mitóticas metáfase/anáfase	MBP1
Grupo 6	GO: 0003724 - Atividade da RNA helicase GO: 0003723 - Ligação do RNA GO: 0000154 - Modificação do rRNA GO: 0007046 - Biogênese do ribossomo GO: 0030515 - Ligação do snRNA GO: 0007028 - Organização do citoplasma e biogênese GO: 0030490 - Processamento do pré-rRNA	
Grupo 7		MBP1
Grupo 8	GO: 0006526 - Biosíntese de arginina GO: 0009309 - Biosíntese de amido GO: 0006082 - Metabolismo de ácido orgânico GO: 0006807 - Metabolismo de nitrogênio	MBP1 GCN4

k = 20

A Tabela A.67 apresenta as funções biológicas e os fatores de transcrição identificados nos grupos.

Tabela A.67: Validação biológica do agrupamento $k = 20$.

Grupos	TANGO	PRIMA
Grupo 1	GO: 0016491 - Atividade de oxidoredutase	HAP4
	GO: 0005975 - Metabolismo de carboidrato	MSN4
	GO: 0006119 - Fosforilação oxidativa	HAP2/3/4
	GO: 0006091 - Geração de metabólitos precursores e energia	
	GO: 0005489 - Atividade de transporte de elétron	
	GO: 0051186 - Metabolismo de cofator	
	GO: 0009060 - Respiração aeróbica	
	GO: 0015078 - Atividade de transporte de íon	
	GO: 0005386 - Atividade de transporte	
	GO: 0006118 - Transporte de elétron	
	GO: 0009109 - Catabolismo de coenzima	
Grupo 2	GO: 0009112 - Metabolismo de nucleotídeo	SFP1
	GO: 0003735 - Constituinte estrutural do ribossomo	AZF1
	GO: 0042257 - União da subunidade ribossomal	
	GO: 0003899 - Atividade da RNA polimerase	
	GO: 0006399 - Metabolismo do tRNA	
	GO: 0009059 - Biosíntese de macromolécula	
	GO: 0003723 - Ligação do RNA	
	GO: 0006163 - Metabolismo de nucleotídeo (purina)	
	GO: 0006412 - Biosíntese de proteína	
	GO: 0016886 - Formação de éster fosfato / atividade ligase	
	GO: 0006519 - Metabolismo de aminoácido	
	GO: 0043037 - Tradução	
	GO: 0007028 - Organização do citoplasma e biogênese	
	GO: 0009259 - Metabolismo de ribonucleotídeo	

Continua na próxima página

Tabela A.67 – continuação da página anterior

	GO: 0009451 - Modificação do RNA GO: 0044237 - Metabolismo celular GO: 0019320 - Catabolismo da hexose GO: 0044249 - Biosíntese celular GO: 0044238 - Metabolismo primário	
Grupo 3	GO: 0030476 - Formação da membrana celular de um esporo (Fungo) GO: 0030435 - Esporulação	SUM1
Grupo 4	GO: 0006366 - Transcrição da RNA polimerase II	
Grupo 6	GO: 0003724 - Atividade da RNA helicase GO: 0009451 - Modificação do RNA GO: 0003723 - Ligação do RNA GO: 0000154 - Modificação do rRNA GO: 0007046 - Biogênese do ribossomo GO: 0030515 - Ligação do snRNA GO: 0007028 - Organização do citoplasma e biogênese GO: 0016070 - Metabolismo do RNA GO: 0030490 - Processamento do pré-rRNA	
Grupo 8	GO: 0009309 - Biosíntese de amido	
Grupo 11	GO: 0005975 - Metabolismo de carboidrato GO: 0006091 - Geração de metabólitos precursores e energia	
Grupo 14		DAL80 DAL82
Grupo 17	GO: 0006974 - Resposta à estímulos de danos ao DNA	
Grupo 18	GO: 0000278 - Ciclo celular mitótico GO: 0007017 - Processo dependente de microtúbulo GO: 0051726 - Regulação do ciclo celular GO: 0007091 - Transição mitótica metáfase/anáfase GO: 0030476 - Formação da membrana celular de um esporo (Fungo) GO: 0007052 - Organização e biogênese do eixo mitótico GO: 0000279 - Fase M	SUM1
Grupo 19	GO: 0003735 - Constituinte estrutural do citoesqueleto	SFP1

Continua na próxima página

Tabela A.67 – continuação da página anterior

	GO: 0006090 - Metabolismo de piruvato GO: 0009059 - Alongamento traducional GO: 0019320 - Catabolismo da hexose GO: 0042273 - Biogênese da subunidade ribossomal maior	RAP1
Grupo 20	GO: 0007127 - Meiose I	CAR1[1] MSN2

k = 30

A Tabela A.68 apresenta as funções biológicas e os fatores de transcrição identificados nos grupos.

Tabela A.68: Validação biológica do agrupamento $k = 30$.

Grupos	TANGO	PRIMA
Grupo 1	GO: 0016491 - Atividade de oxiredutase GO: 0005975 - Metabolismo de carboidrato GO: 0015002 - Atividade da oxidase terminal GO: 0006119 - Fosforilação Oxidativa GO: 0006091 - Geração de metabólitos precursores e energia GO: 0051186 - Metabolismo de cofator GO: 0006118 - Transporte de elétron GO: 0009060 - Respiração aeróbica GO: 0015078 - Atividade de transporte de íon GO: 0005386 - Atividade de carregamento GO: 0009109 - Catabolismo de coenzima	HAP4 MSN4 PHO1 HAP 2/3/4
Grupo 2	GO: 0009112 - Metabolismo de nucleotídeo GO: 0003735 - Constituinte estrutural do ribossomo GO: 0000028 - Manutenção e união da subunidade ribossomal menor GO: 0006412 - Biosíntese de proteína	SFP1 AZF1
Continua na próxima página		

[1]Repressor da expressão de CAR1

Tabela A.68 – continuação da página anterior

	GO: 0016072 - Metabolismo de rRNA GO: 0006164 - Biosíntese de nucleotídeo (purina) GO: 0043037 - Metabolismo de aminoácido GO: 0007028 - Organização e biogênese do citoplasma GO: 0044249 - Biosíntese celular GO: 0044238 - Metabolismo primário	
Grupo 3	GO: 0030476 - Formação da membrana celular de um esporo (Fungo) GO: 0030435 - Esporulação	SUM1
Grupo 4		INO4
Grupo 6	GO: 0003724 - Atividade da RNA helicase GO: 0009451 - Modificação do RNA GO: 0003723 - Ligação do RNA GO: 0007046 - Biogênese do ribossomo GO: 0030515 - Ligação do snRNA GO: 0007028 - Organização do citoplasma e biogênese GO: 0016070 - Metabolismo do RNA	
Grupo 7	GO: 0007010 - Organização e biogênese do citoesqueleto GO: 0000902 - Morfogênese celular GO: 0006259 - Metabolismo do DNA GO: 0006312 - Recombinação mitótica GO: 0043283 - Metabolismo de biopolímero	
Grupo 8	GO: 0009309 - Biosíntese de amido	
Grupo 11	GO: 0005975 - Metabolismo de carboidrato GO: 0006091 - Geração de metabólitos precursores e energia	STRE MSN2 MSN4 ADR1
Grupo 14	GO: 0030036 - Organização do citoesqueleto de actina e biogênese GO: 0048308 - Classes de organela	DAL80 DAL82
Grupo 15		ADR1
Grupo 16		ADR1

Continua na próxima página

Tabela A.68 – continuação da página anterior

Grupo 18	GO: 0048610 - Processo fisiológico de reprodução celular GO: 0000226 - Organização e biogênese do Citoesqueleto do microtúbulo GO: 0051726 - Regulação do ciclo celular GO: 0007067 - Mitose GO: 0007091 - Transição mitótica metáfase/anáfase GO: 0007052 - Organização e biogênese do eixo mitótico GO: 0030476 - Formação da membrana celular de um esporo (Fungo)	
Grupo 19	GO: 0003735 - Constituinte estrutural do ribossomo GO: 0006090 - Metabolismo de piruvato GO: 0009059 - Biosíntese de macromolécula GO: 0006414 - Alongamento traducional GO: 0000027 - Manutenção e união da subunidade ribossomal maior GO: 0019320 - Catabolismo de hexose	SFP1 RAP1
Grupo 20		CAR1 MSN2 ZAP1 PHO2 ACE2 STP1 SUT1
Grupo 21	GO: 0051327 - Fase M do ciclo celular meiótico GO: 0007049 - Ciclo celular GO: 0007127 - Meiose I GO: 0000279 - Fase M GO: 0006310 - Recombinação do DNA	CAR1 REB1
Grupo 22	GO: 0006555 - Metabolismo da metionina	CBF1
Grupo 23	GO: 0016072 - Metabolismo de rRNA GO: 0007046 - Biogênese do ribossomo	
Grupo 24	GO: 0004553 - Atividade da hidrolase	ACE2
Grupo 26	GO: 0006091 - Geração de metabólitos precursores e energia	STRE

Continua na próxima página

Tabela A.68 – continuação da página anterior

	GO: 0016051 - Biosíntese de carboidrato GO: 0044262 - Metabolismo de carboidrato GO: 0006112 - Metabolismo de reserva de enegia	MSN2 MSN4 ADR1
Grupo 29		GAL4
Grupo 30	GO: 0009451 - Modificação do RNA GO: 0016072 - Metabolismo de rRNA GO: 0007028 - Organização e biogênese do citoplasma GO: 0016070 - Metabolismo do RNA	

k = 50

A Tabela A.69 apresenta as funções biológicas e fatores de transcrição identificados nos grupos.

Tabela A.69: Validação biológica do agrupamento $k = 50$.

Grupo 1	GO: 0016491 - Atividade de oxidoredutase GO: 0015002 - Atividade de oxidase GO: 0006119 - Fosforilação oxidativa GO: 0006091 - Geração de metabólitos precursores e energia GO: 0051186 - Metabolismo de cofator GO: 0009060 - Respiração aeróbica GO: 0015078 - Atividade de transporte de íon GO: 0005386 - Atividade de carregamento GO: 0006118 - Transporte de elétron GO: 0009109 - Catabolismo de coenzima	HAP2/3/4
Grupo 2	GO: 0009127 - Biosíntese de nucleotídeo (purina) GO: 0009112 - Metabolismo de nucleotídeo GO: 0016875 - Atividade da ligase GO: 0003743 - Atividade do fator de iniciação da tradução GO: 0006399 - Metabolismo de tRNA GO: 0016072 - Metabolismo de rRNA	
Continua na próxima página		

Tabela A.69 – continuação da página anterior

	GO: 0006164 - Biosíntese de nucleotídeo (purina) GO: 0009058 - Biosíntese GO: 0043037 - Tradução GO: 0009066 - Metabolismo de aminoácido GO: 0007028 - Organização e biogênese do citoplasma GO: 0016070 - Metabolismo de RNA GO: 0044238 - Metabolismo primário	
Grupo 3	GO: 0042244 - Formação da membrana celular de um esporo GO: 0030435 - Esporulação	SUM1
Grupo 4	GO: 0051641 - Localização celular GO: 0030036 - Organização do citoesqueleto de actina e biogênese GO: 0016192 - Transporte mediado por vesículo GO: 0006366 - Transcrição do promotor da Pol II	
Grupo 5	GO: 0006351 - Transcrição, dependente do DNA	
Grupo 6	GO: 0003724 - Atividade da RNA helicase GO: 0009451 - Modificação do RNA GO: 0003723 - Ligação do RNA GO: 0007046 - Biogênese do ribossomo GO: 0030515 - Ligação do snRNA GO: 0007028 - Organização do citoplasma e biogênese GO: 0016070 - Metabolismo do RNA	
Grupo 8	GO: 0009309 - Biosíntese de amido GO: 0006520 - Metabolismo de aminoácido	
Grupo 10		ADR1
Grupo 13		DAL80
Grupo 16	GO: 0003735 - Constituinte estrutural do ribossomo GO: 0006090 - Metabolismo de piruvato GO: 0009059 - Biosíntese de macromolécula GO: 0006414 - Alongamento traducional GO: 0019320 - Catabolismo da hexose	RAP1
Grupo 17		ACE2
Grupo 18	GO: 0006555 - Metabolismo da metionina	BAS1

Continua na próxima página

Tabela A.69 – continuação da página anterior

	GO: 0006807 - Metabolismo do nitrogênio	
Grupo 19	GO: 0016072 - Metabolismo do rRNA GO: 0007046 - Biogênese do ribossomo	
Grupo 22	GO: 0006091 - Geração de metabólitos precursores e energia GO: 0016051 - Biosíntese de carboidrato GO: 0044262 - Metabolismo de carboidrato GO: 006112 - Metabolismo de reserva de energia	ADR1
Grupo 25	GO: 0009451 - Modificação do RNA GO: 0016072 - Metabolismo do rRNA GO: 0007046 - Biogênese do ribossomo GO: 0016070 - Metabolismo do RNA	
Grupo 27	GO: 0046943 - Atividade de transporte de ácido carboxílico	DAL80 DAL82
Grupo 29	GO: 0006396 - Processamento do RNA GO: 0007028 - Organização e biogênese do citoplasma	
Grupo 31	GO: 0051327 - Fase M do ciclo celular meiótico GO: 0043566 - Estrutura específica de ligação do DNA GO: 0007127 - Meiose I	
Grupo 32	GO: 0003735 - Constituinte estrutural do ribossomo GO: 0009059 - Biosíntese de macromolécula GO: 0045333 - Respiração celular GO: 0019538 - Metabolismo de proteína	
Grupo 33	GO: 0003678 - Atividade da DNA helicase GO: 0000722 - Manutenção de telômero independente da telomerase	
Grupo 40	GO: 0003735 - Constituinte estrutural do ribossomo GO: 0000027 - Manutenção e união da subunidade ribossomal maior	RAP1
Grupo 42	GO: 0000226 - Organização e biogênese do citoesqueleto do microtúbulo GO: 0007059 - Segregação do cromossomo GO: 0000279 - Fase M	
Grupo 43	GO: 0051603 - Catabolismo de proteína durante proteólise	

Continua na próxima página

Tabela A.69 – continuação da página anterior

	GO: 0004175 - Atividade de endopeptidase	
Grupo 44		SWI5
Grupo 45	GO: 0048622 - Esporulação reprodutiva GO: 0030476 - Formação da membrana celular de um esporo (Fungo)	SUM1
Grupo 46	GO: 0006974 - Resposta à estímulos de danos ao DNA GO: 0007049 - Ciclo celular GO: 0006259 - Metabolismo do DNA GO: 0006260 - Replicação do DNA GO: 0006271 - Alongamento da fita de DNA GO: 0006310 - Recombinação do DNA	
Grupo 47	GO: 0000278 - Ciclo celular mitótico	FKH1

A.4.7 Validação biológica SOM

SOM = 5x1

A Tabela A.70 apresenta as funções biológicas e fatores de transcrição identificados nos grupos.

Tabela A.70: Validação biológica do agrupamento SOM = 5x1.

Grupos	TANGO	PRIMA
Grupo 1	GO: 0008173 - Atividade da RNA metiltransferase GO: 0030490 - Processamento do pré-rRNA GO: 0044237 - Metabolismo celular GO: 0006414 - Alongamento da tradução GO: 0006412 - Biosíntese de proteína GO: 0006164 - Biosíntese de nucleotídeo (purina) GO: 0006360 - Transcrição do promotor Pol I GO: 0003723 - Ligação do RNA GO: 0003724 - Atividade da RNA helicase	SUM1 AZF1 RAP1
Continua na próxima página		

Tabela A.70 – continuação da página anterior

	GO: 0030515 - Ligação do snRNA GO: 0019329 - Catabolismo de hexose GO: 0003899 - Atividade da RNA polimerase GO: 0009126 - Metabolismo de monofostato GO: 0003735 - Constituinte estrutural do ribossomo GO: 0009451 - Modificação do RNA GO: 0019843 - Ligação do rRNA GO: 0044238 - Metabolismo primário GO: 0044249 - Biosíntese celular GO: 0043170 - Metabolismo de macromolécula GO: 0007028 - Organização do citoplasma e biogênese GO: 0009112 - Metabolismo de nucleotídeo GO: 0042257 - União da subunidade ribossomal	
Grupo 2	GO: 0008173 - Atividade da RNA metiltransferase GO: 0006139 - Metabolismo de ácido nucléico GO: 0006999 - Biogênese e organização de núcleo GO: 0016192 - Transporte mediado por vesículo GO: 0006888 - Transporte mediado por vesículo de Golgi para RE GO: 0006403 - Localiação do RNA GO: 0016043 - Organização celular e biogênese GO: 0016070 - Metabolismo do RNA GO: 0051179 - Importação nuclear GO: 0016071 - Metabolismo mRNA GO: 0046903 - Secreção GO: 0051169 - Transporte nuclear GO: 0031323 - Metabolismo celular de regulação GO: 0008380 - Splicing do RNA GO: 0043283 - Metabolismo de biopolímero GO: 0006366 - Transcrição da RNA polimerase II GO: 0006351 - Transcrição, dependente do DNA GO: 0045941 - Transcrição da regulação positiva GO: 0046907 - Transporte intracelular	
Grupo 3	GO: 0005386 - Atividade de carregamento	HAP4

Continua na próxima página

Tabela A.70 – continuação da página anterior

	GO: 0015075 - Atividade de transporte de íon	STRE
	GO: 0015002 - Atividade da oxidase terminal	MBP1
	GO: 0006119 - Fosforilação oxidativa	MSN4
	GO: 0005975 - Metabolismo de carboidrato	PUT3
	GO: 0006118 - Transporte de elétron	ADR1
	GO: 0015078 - Atividade de transporte de íon	MIG1
	GO: 0051186 - Metabolismo de cofator	SUT1
	GO: 0006811 - Transporte de íon	HAP 2/3/4
	GO: 0045333 - Respiração celular	
	GO: 0003735 - Constituinte estrutural do ribossomo	
	GO: 0015672 - Transporte de cálcio inorgânico	
	GO: 0051187 - Metabolismo de cofator	
	GO: 0009310 - Catabolismo de amido	
	GO: 0016491 - Atividade de oxidoredutase	
	GO: 0006091 - Geração de metabólitos precursores e energia	
Grupo 4	GO: 0000723 - Manutenção de telômeros	MBP1
	GO: 0048519 - Processo biológico de regulação negativa	STB1
	GO: 0004175 - Atividade da endopeptidase	FKH1
	GO: 0000279 - Fase M	
	GO: 0009719 - Resposta à estímulos endógenos	
	GO: 0006260 - Replicação do DNA	
	GO: 0007049 - Ciclo celular	
	GO: 0006271 - Alongamento da fita de DNA	
	GO: 0051603 - Catabolismo de proteína durante proteólise	
	GO: 0006259 - Metabolismo do DNA	
	GO: 0006950 - Resposta à estresse	
	GO: 0043283 - Metabolismo de biopolmero	
	GO: 0003677 - Ligação do DNA	
	GO: 0051276 - Organização e biogênese do cromossomo	
	GO: 0007059 - Segregação cromossomal	
Grupo 5	GO: 0030435 - Esporulação	CAR1
	GO: 0007127 - Meiose I	SUM1
	GO: 0000279 - Fase M	
	GO: 0000226 - Organização e biogênese do Citoesqueleto do	

Continua na próxima página

Tabela A.70 – continuação da página anterior

	microtúbulo GO: 0051726 - Regulação do ciclo celular GO: 0007091 - Transição mitótica metáfase/anáfase GO: 0051327 - Fase M do ciclo celular meiótico GO: 0030476 - Formação da membrana celular de um esporo (Fungo) GO: 0005200 - Constituinte estrutural do citoesqueleto GO: 0048610 - Processo fisiológico de reprodução celular GO: 0007052 - Organização e biogênese do eixo mitótico GO: 0050790 - Regulação da atividade enzimática	

SOM = 2x5

A Tabela A.71 apresenta as funções biológicas e os fatores de transcrição identificados nos grupos.

Tabela A.71: Validação biológica do agrupamento SOM = 2x5.

Grupos	TANGO	PRIMA
Grupo 1	GO: 0030435 - Esporulação GO: 0000279 - Fase M GO: 0000226 - Organização e biogênese do citoesqueleto do microtúbulo GO: 0051726 - Regulação do ciclo celular GO: 0007091 - Transição das fases mitóticas metáfase/anáfase GO: 0051327 - Fase M do ciclo celular meiótico GO: 0030476 - Formação da membrana celular de um esporo (Fungo) GO: 0050790 - Regulação da atividade enzimática GO: 0007059 - Segregação cromossomal	CAR1 SUM1
Grupo 4	GO: 0006139 - Metabolismo de ácido nucléico GO: 0008168 - Atividade da metiltransferase	

Continua na próxima página

Tabela A.71 – continuação da página anterior

	GO: 0043037 - Tradução GO: 0008135 - Ligação de ácido nucléido/ Ativação de fator de tradução GO: 0003899 - Atividade da RNA polimerase GO: 0016070 - Metabolismo do RNA GO: 0009451 - Modificação do RNA GO: 0016072 - Metabolismo de rRNA GO: 0051169 - Transporte nuclear GO: 0006413 - Iniciação da tradução GO: 0006399 - Metabolismo do tRNA GO: 0006365 - Processamento do transcrito primário - 35S GO: 0007028 - Organização celular e biogênese	
Grupo 5	GO: 0030490 - Processamento do pré-rRNA GO: 0044237 - Metabolismo celular GO: 0006414 - Alongamento da tradução GO: 0006412 - Biosíntese de proteína GO: 0003723 - Ligação do RNA GO: 0030515 - Ligação do snRNA GO: 0003735 - Constituinte estrutural do ribossomo GO: 0044238 - Metabolismo primário GO: 0043170 - Metabolismo de macromolécula GO: 0007028 - Organização do citoplasma e biogênese	SFP1 AZF1 RAP1
Grupo 6	GO: 0000723 - Manutenção dos telômeros GO: 0006310 - Recombinação do DNA GO: 0007127 - Meiose I GO: 0000279 - Fase M GO: 0006260 - Replicação do DNA GO: 0007049 - Ciclo celular GO: 0007017 - Processo dependente de microtúbulo GO: 0009719 - Resposta à estímulos endógenos GO: 0006301 - Reparo pós-replicação GO: 0007010 - Organização e biogênese do citoesqueleto GO: 0007064 - Coesão mitótica da cromátide irmã GO: 0006271 - Alongamento da fita do DNA	CAR1 MBP1 RPN4 STB1 FKH1

Continua na próxima página

Tabela A.71 – continuação da página anterior

	GO: 0005200 - Constituinte estrutural do citoesqueleto	
	GO: 0006259 - Metabolismo do DNA	
	GO: 0006298 - Reparo de erro	
	GO: 0006996 - Organização celular e biogênese	
	GO: 0043283 - Metabolismo de biopolímero	
	GO: 0003677 - Ligação do DNA	
	GO: 0051276 - Organização do cromossomo e biogênese	
	GO: 0000075 - Checkpoint do ciclo celular	
	GO: 0007059 - Segregação cromossomal	
	GO: 0000724 - Interrupção para reparo da fita de DNA	
Grupo 7	GO: 0051641 - Localização celular	
	GO: 0005515 - Ligação de proteína	
	GO: 0016043 - Organização celular e biogênese	
	GO: 0016568 - Modificação da cromatina	
	GO: 0006351 - Transcrição, dependente do DNA	
	GO: 0045045 - Via metabólica de secreção	
Grupo 9	GO: 0005386 - Atividade de carregamento	HAP4
	GO: 0015075 - Atividade de transporte de íon	STRE
	GO: 0015002 - Atividade da oxidase terminal	MSN2
	GO: 0006119 - Fosforilação oxidativa	MSN4
	GO: 0005975 - Metabolismo de carboidrato	PUT3
	GO: 0006118 - Transporte de elétron	SKN7
	GO: 0015078 - Atividade de transporte de íon	ADR1
	GO: 0051186 - Metabolismo de cofator	SUT1
	GO: 0006811 - Transporte de íon	HAP2/3/4
	GO: 0045333 - Respiração celular	
	GO: 0003735 - Constituinte estrutural do ribossomo	
	GO: 0009109 - Catabolismo da coenzima	
	GO: 0006091 - Geração de metabólitos precursores e energia	
	GO: 0016491 - Atividade de oxidoredutase	
	GO: 0006100 - Metabolismo intermediário do ciclo de ácido tricarboxílico	
Grupo 10	GO: 0006066 - Metabolismo de álcool	MSN4
	GO: 0009165 - Biosíntese de nucleotídeo	ADR1

Continua na próxima página

Tabela A.71 – continuação da página anterior

	GO: 0006006 - Metabolismo de glicose GO: 0046365 - Catabolismo de monossacarídeo	

SOM = 5x5

A Tabela A.72 apresenta as funções biológicas e os fatores de transcrição identificados nos grupos.

Tabela A.72: Validação biológica do agrupamento SOM = 5x5.

Grupos	TANGO	PRIMA
Grupo 1	GO: 0051082 - Ligação de proteína desestruturada GO: 0005515 - Ligação de proteína	HSF MAT1 GCR1 STB5
Grupo 2	GO: 0004175 - Atividade de endopeptidase GO: 0019941 - Catabolismo de proteína dependente-modificação	RPN4
Grupo 3	GO: 0006284 - Reparo da base de excisão GO: 0006310 - Recombinação do DNA GO: 0006260 - Replicação do DNA GO: 0007064 - Coesão mitótica da cromátide irmã GO: 0006974 - Resposta à estímulos de danos ao DNA GO: 0003887 - Atividade da DNA polimerase GO: 0006271 - Alongamento da fita de DNA GO: 0006259 - Metabolismo do DNA GO: 0006298 - Reparo de erro GO: 0006289 - Reparo da excisão de nucleotídeo GO: 0007059 - Segregação cromossomal	MBP1 STB1
Grupo 4	GO: 0007127 - Meiose I GO: 0000279 - Fase M GO: 0007017 - Processo dependente de microtúbulo	CAR1 ADR1

Continua na próxima página

Tabela A.72 – continuação da página anterior

	GO: 0051327 - Fase M do ciclo celular meiótico	
Grupo 5	GO: 0030435 - Esporulação GO: 0030476 - Formação da membrana celular de um esporo GO: 0005200 - Constituinte estrutural do citoesqueleto	SUM1
Grupo 6	GO: 0051641 - Localização celular GO: 0046903 - Secreção	
Grupo 7	GO: 0006486 - Glicosilação de proteína GO: 0006259 - Metabolismo do DNA GO: 0007001 - Organização cromossomal e biogênese (Eucariota)	STB1
Grupo 10	GO: 0016458 - Silenciamento de gene GO: 0007091 - Transição das fases mitóticas metáfase/anáfase GO: 0048610 - Processo fisiológico de reprodução celular	SUM1
Grupo 11	GO: 0003743 - Atividade do fator de iniciação da tradução GO: 0008452 - Atividade da RNA ligase GO: 0006402 - Catabolismo do mRNA GO: 0006139 - Metabolismo de ácido nucléico GO: 0043037 - Tradução GO: 0003899 - Atividade da RNA polimerase GO: 0016070 - Metabolismo do RNA GO: 0008408 - Atividade da exonuclease 3–5- GO: 0006413 - Iniciação da tradução GO: 0044238 - Metabolismo primário GO: 0006520 - Metabolismo de aminoácido GO: 0006399 - Metabolismo de tRNA GO: 0006365 - Processamento do transcrito primário - 35S GO: 0007046 - Biogênese do ribossomo	
Grupo 12	GO: 0031323 - Metabolismo celular da regulação GO: 0006366 - Transcrição do promotor da Pol II	
Grupo 16	GO: 0030490 - Processamento do pré-rRNA GO: 0003723 - Ligação do RNA GO: 0030515 - Ligação do snRNA GO: 0004004 - Atividade da RNA helicase dependente-ATP	

Continua na próxima página

Tabela A.72 – continuação da página anterior

	GO: 0016070 - Metabolismo do RNA GO: 0009451 - Modificação do RNA GO: 0019843 - Ligação do rRNA GO: 0016072 - Metabolismo do rRNA GO: 0007028 - Organização do citoplasma e biogênese GO: 0007046 - Biogênese do ribossomo	
Grupo 17		AZF1
Grupo 19	GO: 0005975 - Metabolismo de carboidrato GO: 0006092 - Metabolismo de carboidrato de vias principais GO: 0006081 - Metabolismo de aldeído GO: 0006091 - Geração de metabólitos precursores e energia	STRE MSN2 MSN4 ADR1 STP1
Grupo 20	GO: 0003735 - Consituinte estrutural do ribossomo GO: 0044260 - Metabolismo celular de macromolécula	
Grupo 21	GO: 0006414 - Alongamento traducional GO: 0006412 - Biosíntese de proteína GO: 0003735 - Constituinte estrtutural do ribossomo GO: 0042257 - União da subunidade ribossomal	SFP1 RAP1
Grupo 22	GO: 0006094 - Gliconeogênese GO: 0019320 - Catabolismo da hexose GO: 0006082 - Metabolismo de ácido orgânico GO: 0006144 - Metabolismo de base (purina)	RCS1 AFT2
Grupo 24	GO: 0006754 - Biosíntese de ATP GO: 0015002 - Atividade da oxidase terminal GO: 0006119 - Fosforilação oxidativa GO: 0005975 - Metabolismo de carboidrato GO: 0015078 - Atividade de transporte de íon GO: 0035251 - Atividade de glicosiltransferase - UDP GO: 0006732 - Metabolismo da coenzima GO: 0006091 - Geração de metabólitos precursores e energia GO: 0016491 - Atividade de oxidoredutase	HAP4 STRE MSN2 MSN4 ADR1 HAP2/3/4
Grupo 25	GO: 0006119 - Fosforilação oxidativa GO: 0009060 - Respiração aeróbica	

Continua na próxima página

Tabela A.72 – continuação da página anterior

	GO: 0006118 - Transporte de elétron GO: 0015749 - Transporte de monossacarídeo GO: 0006091 - Geração de metabólitos precursores e energia	

SOM = 5x10

A Tabela A.73 apresenta as funções biológicas e os fatores de transcrição identificados nos grupos.

Tabela A.73: Validação biológica do agrupamento SOM = 5x10.

Grupos	TANGO	PRIMA
Grupo1	GO: 0005975 - Metabolismo de carboidrato GO: 0006092 - Metabolismo de carboidrato de vias principais GO: 0015980 - Derivação de energia por oxidação de compostos orgânicos	
Grupo 9	GO: 0006119 - Fosforilação oxidativa GO: 0006118 - Transporte de elétron GO: 0006084 - Metabolismo de acetil-CoA GO: 0051186 - Respiração celular GO: 0006091 - Geração de metabólitos precursores e energia GO: 0016491 - Atividade de oxidoredutase	
Grupo 10	GO: 0015002 - Atividade da oxidase terminal GO: 0006119 - Fosforilação oxidativa GO: 0006118 - Transporte de elétron GO: 0015078 - Atividade de transporte de íon GO: 0005353 - Atividade de transporte de frutose GO: 0006091 - Geração de metabólitos precursores e energia GO: 0009142 - Biosíntese de nucleosídeo trifosfato GO: 0016491 - Atividade de oxidoredutase	
Grupo 13	GO: 0006260 - Replicação do DNA	
Grupo 18	GO: 0003735 - Constituinte estrutural do ribossomo	

Continua na próxima página

Tabela A.73 – continuação da página anterior

Grupo 19	GO: 0009058 - Biosíntese	
Grupo 21	GO: 0006310 - Recombinação do DNA GO: 0000279 - Fase M	
Grupo 22	GO: 0031109 - Polimerização ou despolimerização de microtúbulo GO: 0000279 - Fase M GO: 0006260 - Replicação do DNA GO: 0007017 - Processo dependente de microtúbulo GO: 0007062 - Coesão da cromátide irmã GO: 0006271 - Alongamento da fita de DNA GO: 0005200 - Constituinte estrutural do citoesqueleto GO: 0006259 - Metabolismo do DNA GO: 0000075 - Checkpoint do ciclo celular GO: 0006281 - Reparo do DNA GO: 0030472 - Organização mitótica e biogênese no núcleo GO: 0007059 - Segregação cromossomal	
Grupo 23	GO: 0006260 - Replicação do DNA GO: 0003678 - Atividade da DNA helicase GO: 0000722 - Manutenção de telômero independente da telomerase GO: 0006271 - Alongamento da fita de DNA GO: 0006259 - Metabolismo do DNA GO: 0006298 - Reparo de erro	
Grupo 24	GO: 0008610 - Biosíntese de lipídio GO: 0006486 - Glicosilação da proteína GO: 0009058 - Biosíntese	
Grupo 25	GO: 0030473 - Migração nuclear / mediado por microtúbulo	
Grupo 28	GO: 0030036 - Organização do citoesqueleto de actina e biogênese	
Grupo 31	GO: 0007127 - Meiose I GO: 0000279 - Fase M GO: 0051327 - Fase M do ciclo celular meiótico	
Grupo 32	GO: 0007017 - Processo dependente de microtúbulo GO: 0007067 - Mitose	

Continua na próxima página

Tabela A.73 – continuação da página anterior

	GO: 0040029 - Regulação da expressão gênica	
Grupo 37	GO: 0004553 - Atividade da hidrolase GO: 0000910 - Citoquinase	
Grupo 38	GO: 0016070 - Metabolismo do RNA	
Grupo 39	GO: 0006412 - Biosíntese de proteína GO: 0043037 - Tradução GO: 0003899 - Atividade da RNA polimerase GO: 0007028 - Organização do citoplasma e biogênese GO: 0009058 - Biosíntese	
Grupo 40	GO: 0006066 - Metabolismo de álcool GO: 0006090 - Metabolismo de piruvato GO: 0016829 - Atividade de liase GO: 0019320 - Catabolismo da hexose GO: 0006082 - Metabolismo de ácido orgânico GO: 0006826 - Transporte de íon ferro GO: 0006144 - Metabolismo de base purina	
Grupo 41	GO: 0030435 - Esporulação GO: 0030476 - Formação da membrana celular de um esporo (Fungo)	
Grupo 42	GO: 0007091 - Transição mitótica metáfase/anáfase GO: 0030476 - Formação da membrana celular de um esporo (Fungo) GO: 0030154 - Diferenciação celular GO: 0048610 - Processo fisiológico celular reprodutivo	
Grupo 45	GO: 0006508 - Proteólise GO: 0004175 - Atividade Endopeptidase GO: 0019941 - Catabolismo de proteína dependente-modificação GO: 0044260 - Metabolismo de macromolécula celular	
Grupo 48	GO: 0008173 - Atividade RNA metiltransferase GO: 0003724 - Atividade da RNA helicase GO: 0016070 - Metabolismo do RNA GO: 0009451 - Modificação do RNA	

Continua na próxima página

Tabela A.73 – continuação da página anterior

	GO: 0016072 - Metabolismo do rRNA GO: 0007028 - Organização do citoplasma e biogênese GO: 0007046 - Biogênse do ribossomo	
Grupo 49	GO: 0030490 - Processamento do pré-RNA GO: 0003723 - Ligação do RNA GO: 0030515 - Ligação do snRNA GO: 0007028 - Organização do citoplasma e biogênese GO: 0007046 - Biogênese do ribossomo	
Grupo 50	GO: 0006414 - Alongamento traducional GO: 0006412 - Biosíntese de proteína GO: 0003735 - Constituinte estrutural do ribossomo GO: 0042257 - União da subunidade ribossomal	

A.4.8 Validação biológica SAMBA

O agrupamento bidimensional resultou em 67 grupos, 78 funções biológicas e 27 fatores de transcrição. Alguns desses grupos representativos somente da condição de ciclo celular ou somente da condição de esporulação ou somente da condição de mudança diáuxica. Também foram formados grupos de diferentes condições, que podem até mesmo revelar associações entre elas. A Tabela A.74 mostra a distribuição dos 67 grupos de acordo com as condições do microarranjo.

Lembrando que a condição de mudança diáuxica refere-se à mudanças de concentrações de glicose e o processo de esporulação refere-se à formação de novos organismos.

Tabela A.74: Grupos bidimensionais separados por condições.

Condição	Grupo
Ciclo Celular	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15, 16,17,18,19,20,34,35,36,37,40,46 e 48
Esporulação	21,22,23 e 25
Mudança diáuxica	26
Ciclo celular + Esporulação	24,28,29,30,31,32,33,38,39,41,42,
Continua na próxima página	

Tabela A.74 – continuação da página anterior

	43,47,51,53,54,55,61,62,63 e 65
Ciclo celular + Mudança diáuxica	27 e 52
Esporulação + Mudança diaúxica	66
Ciclo celular + Esporulação + Mudança diáuxica	44,45,49,50,56, 57 58,59,60,64 e 67

De acordo com a Tabela A.74, a maioria dos grupos foi formada por genes expressos nas condições do ciclo celular, seguido do grupo de genes expressos nas condições de ciclo celular e esporulação, grupos de genes expressos nas três condições, grupos de genes somente na condição de esporulação, dois grupos de genes nas condições de ciclo celular e mudança diáuxica, um grupo de genes somente na condição de mudança diáuxica e um grupo de genes nas condições de esporulação e mudança diáuxica.

A maior parte dos grupos foi formada pelas condições de ciclo celular e esporulação. Esta relação faz sentido, uma vez que ambas as condições são referentes ao processo de reprodução de *S. cerevisiae*. A última linha da tabela apresenta os grupos que foram formados pelas condições de ciclo, esporulação e mudança diáuxica. Esta relação das 3 condições é porque nos experimentos que resultaram na base de dados GCSc o ciclo celular e a esporulação foram induzidos por limitação nutricional.

Nesta seção são apresentadas as análises de alguns grupos através da interpretação dos *heat maps*. Os resultados do processo de validação biológica de enriquecimento funcional e identificação de fatores de transcrição são discutidos junto com cada um desses grupos analisados. Infomações de todo o agrupamento e comparações com os resultados dos agrupamentos unidimensionais são apresentadas na seção X de conclusão da base de dados GSc.

A Figura A.24 ilustra o *heat map* do grupo 29. O grupo foi formado por 39 genes e 9 condições, sendo 2 delas referentes ao ciclo celular e as demais referentes ao processo de esporulação. Alguns genes estão caracterizados com a função de meiose, mitose, esporulação ou ciclo celular. De acordo com a imagem do *heat map*, enquanto os genes são reprimidos nas 2 condições de ciclo celular, eles tem a expressão induzida em quase todas as condições de esporulação, sugerindo que não são nessas condições do ciclo que ocorre o processo de esporulação.

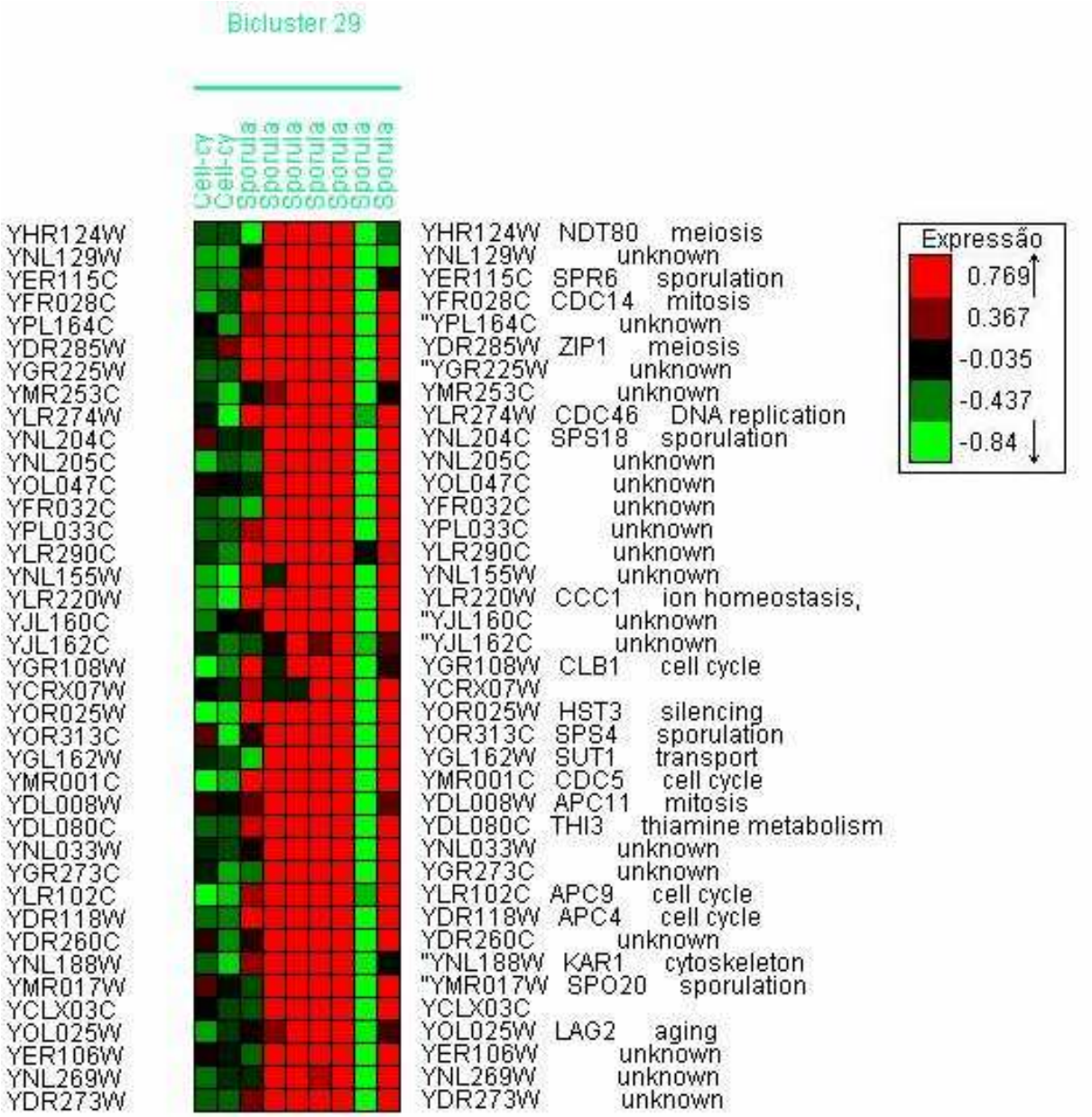


Figura A.24: *Heat map* de alguns genes do grupo biodimensional 29 (captura da tela do programa Expander).

Nesse grupo foram identificadas as funções de esporulação (GO: 0030435), fase M (GO: 0000279 e transição das fases mitóticas de metáfase/anáfase (GO: 0007091).

O fator de transcrição identificado neste grupo foi o SUM1. Este fator também foi identificado em mais 13 grupos bidimensionais, todos eles contendo a condição de esporulação, sugerindo que é um fator característico dos genes envolvidos nesse processo.

A Figura A.25 ilustra um trecho do *heat map* do grupo 66. Este grupo foi formado por 73 genes e 10 condições, sendo 9 delas de esporulação e uma de mudança diáuxica a 19.0g/L. Este resultado mostra que os genes se expressam nestas condições de esporulação são todos reprimidos quando submetidos à alta concentração de glicose.

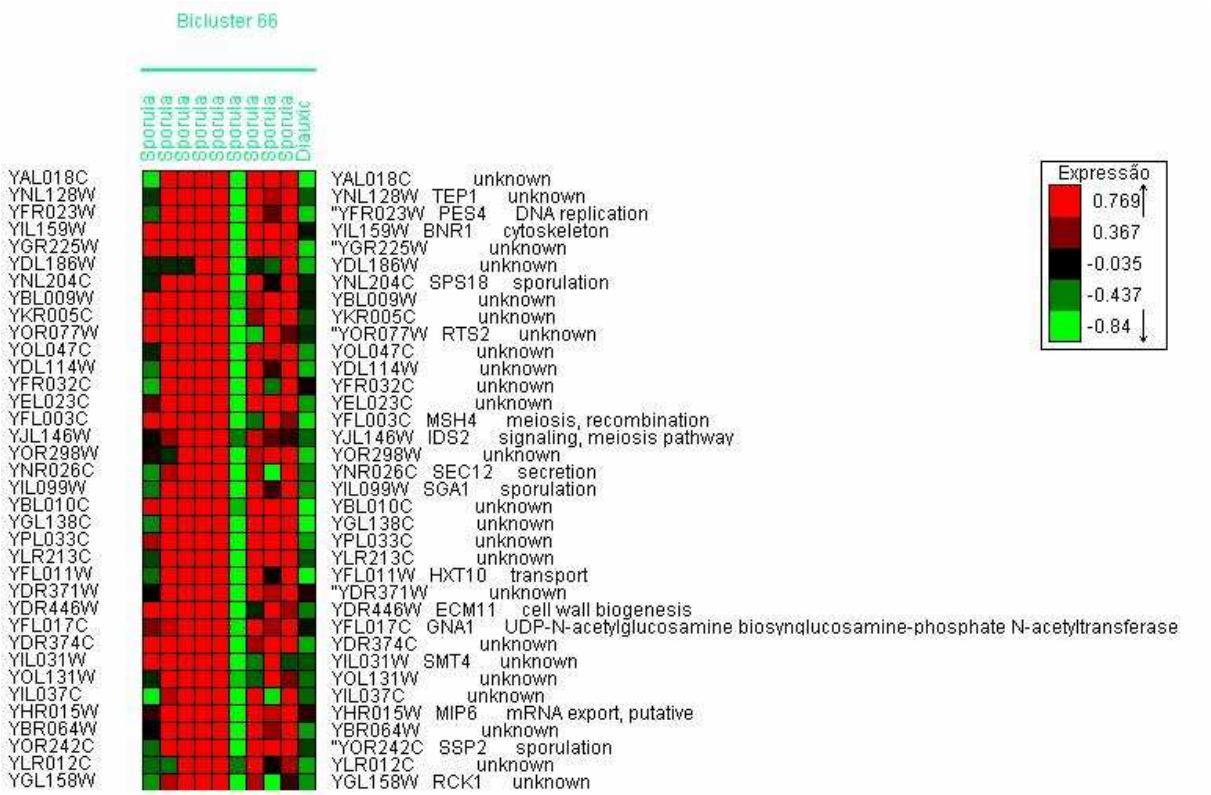


Figura A.25: *Heat map* de alguns genes do grupo biodimensional 66 (captura da tela do programa Expander).

Nesse grupo foram identificadas as funções de esporulação (GO: 0030435), meiose (GO: 0007126) e *spore wall assembly* (GO: 00042244) e o fator de transcrição SUM1.

A Figura A.26 ilustra o *heat map* do grupo 56. São 94 genes agrupados nas 3 diferentes condições: ciclo celular, esporulação e mudança diáuxica. A maioria dos genes desempenham a função de síntese de proteínas, conforme a última coluna da imagem correspondente à anotação funcional dos genes. As condições de mudança diáuxica deste

grupo corresponde à baixa concentração de glicose. Sendo assim, os 94 genes se expressam na maioria das condições de ciclo celular e em uma das condições de esporulação e são reprimidos quando submetidos à baixa concentração de glicose.

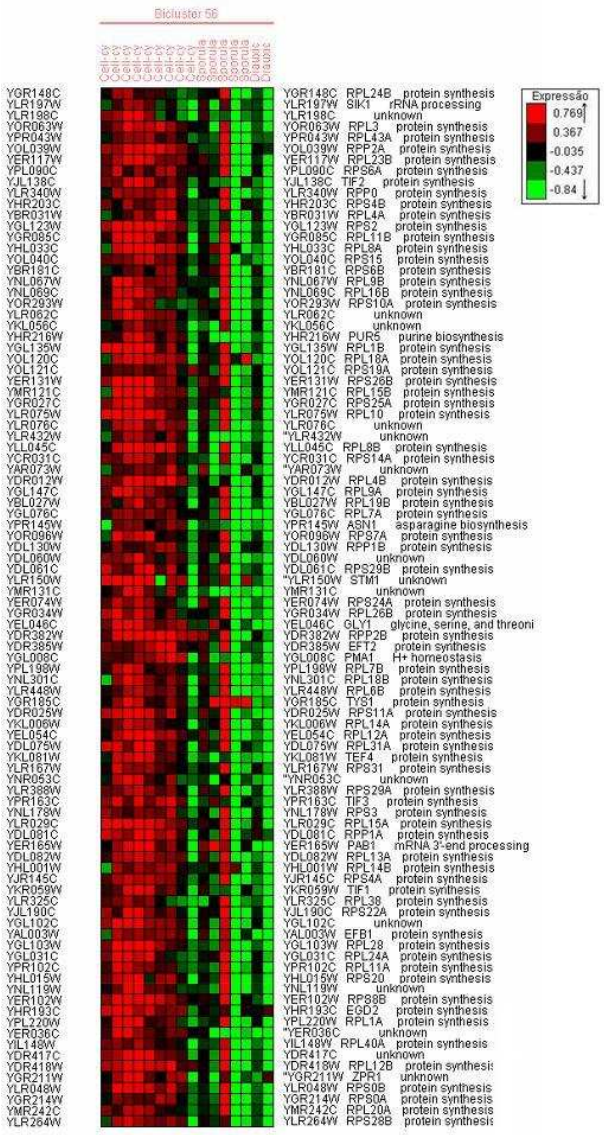


Figura A.26: *Heat map* do grupo biodimensional 56 (captura da tela do programa Expander).

Neste grupo foram identificadas as funções de alogamento da tradução do mRNA (GO: 0006414), biosíntese de proteína (GO: 0006412) , união com o ribossomo (GO: 0042255), constituinte estrutural do ribossomo (GO: 0003735), biosíntese celular (GO: 0044249) e organização do citoplasma e biogênese (GO: 0007028). Foram identificados os fatores de transcrição SFP1 e RAP1.

A.5 Base de dados GSc - com a aplicação de filtros de dados

A.5.1 Agrupamento k-médias

O algoritmo k-médias também foi aplicado com o valor de $k = 5, 10, 20, 30$ *e* 50. Os resultados obtidos foram os seguintes:

k = 5

Os 3456 genes foram agrupados em 5 grupos, conforme mostrado na Tabela A.75.

Tabela A.75: Agrupamento $k = 5$.

Grupos	Qtd de genes	Homogeneidade
1	996	0,38
2	2134	0,304
3	196	0,838
4	115	0,733
5	15	0,643

k = 10

Os 3456 genes foram agrupados em 10 grupos, conforme mostrado na Tabela A.76.

Tabela A.76: Agrupamento $k = 10$.

Grupos	Qtd de genes	Homogeneidade
1	241	0,629
2	936	0,601
3	98	0,889
4	1	1,0
5	258	0,555
6	118	0,493
7	1120	0,389

Continua na próxima página

Tabela A.76 – continuação da página anterior

8	53	0,535
9	583	0,485
10	48	0,757

k = 20

Os 3456 genes foram agrupados em 20 grupos, conforme mostrado na Tabela A.77.

Tabela A.77: Agrupamento $k = 20$.

Grupos	Qtd de genes	Homogeneidade
1	137	0,631
2	453	0,633
3	64	0,917
4	99	0,824
5	1	1,0
6	167	0,611
7	79	0,609
8	545	0,437
9	12	0,639
10	158	0,428
11	23	0,748
12	169	0,495
13	104	0,889
14	33	0,615
15	27	0,527
16	30	0,801
17	764	0,639
18	29	0,608
19	2	0,993
20	560	0,548

k = 30

Os 3456 genes foram agrupados em 30 grupos, conforme mostrado na Tabela A.78.

Tabela A.78: Agrupamento $k = 30$.

Grupos	Qtd de genes	Homogeneidade
1	136	0,658
2	316	0,679
3	63	0,929
4	93	0,843
5	1	1,0
6	102	0,634
7	62	0,641
8	523	0,472
9	13	0,677
10	154	0,481
11	21	0,794
12	199	0,501
13	514	0,599
14	102	0,893
15	44	0,647
16	25	0,535
17	66	0,501
18	31	0,817
19	85	0,469
20	506	0,571
21	88	0,6
22	28	0,623
23	2	0,993
24	2	0,864
25	57	0,895
26	54	0,723
27	1	1,0
28	78	0,551

Continua na próxima página

Tabela A.78 – continuação da página anterior

29	14	0,66
30	76	0,678

k = 50

Os 3456 genes foram agrupados em 50 grupos, conforme mostrado na Tabela A.79.

Tabela A.79: Agrupamento $k = 50$.

Grupos	Qtd de genes	Homogeneidade
1	72	0,688
2	415	0,656
3	40	0,943
4	1	1,0
5	199	0,619
6	74	0,633
7	9	0,717
8	211	0,418
9	16	0,815
10	120	0,509
11	101	0,894
12	26	0,652
13	16	0,566
14	23	0,824
15	24	0,666
16	2	0,993
17	384	0,549
18	1	1,0
19	21	0,932
20	34	0,769
21	1	1,0
22	73	0,52

Continua na próxima página

Tabela A.79 – continuação da página anterior

23	8	0,66
24	32	0,663
25	270	0,736
26	8	0,851
27	293	0,734
28	193	0,53
29	36	0,536
30	12	0,932
31	1	1,0
32	42	0,621
33	9	0,796
34	3	0,801
35	23	0,873
36	15	0,662
37	35	0,881
38	13	0,826
39	45	0,66
40	15	0,734
41	36	0,737
42	63	0,663
43	81	0,77
44	12	0,906
45	14	0,632
46	171	0,568
47	29	0,577
48	73	0,656
49	54	0,651
50	7	0,808

A.5.2 Agrupamento SOM

O algoritmo SOM foi aplicado com as dimensões da matriz definidas por 5x1, 2x5, 5x5 e 5x10. Os resultados obtidos foram os seguintes:

SOM = 5x1

A quantidade de genes e o índice de homogeneidade de cada grupo são apresentadas na Tabela A.80.

Tabela A.80: Agrupamento SOM = 5x1.

Grupos	Quantidade de genes	Homogeneidade
1	234	0,825
2	892	0,54
3	1192	0,351
4	530	0,583
5	608	0,71

SOM = 2x5

A quantidade de genes e o índice de homogeneidade de cada grupo são apresentadas na Tabela A.81

Tabela A.81: Agrupamento SOM = 2x5.

Grupos	Quantidade de genes	Homogeneidade
1	245	0,827
2	374	0,607
3	355	0,419
4	440	0,673
5	134	0,861
6	370	0,68
7	239	0,635
8	725	0,489
Continua na próxima página		

Tabela A.81 – continuação da página anterior

9	237	0,545
10	337	0,688

SOM = 5x5

A quantidade de genes e o índice de homogeneidade de cada grupo são apresentadas na Tabela A.82.

Tabela A.82: Agrupamento SOM = 5x5.

Grupos	Qtd de genes	Homogeneidade
1	67	0,616
2	177	0,607
3	182	0,741
4	60	0,883
5	74	0,913
6	153	0,47
7	94	0,583
8	183	0,629
9	175	0,702
10	118	0,811
11	151	0,62
12	114	0,498
13	168	0,623
14	286	0,448
15	116	0,646
16	126	0,837
17	144	0,711
18	221	0,655
19	128	0,622
20	96	0,591
21	130	0,894
22	73	0,791

Continua na próxima página

Tabela A.82 – continuação da página anterior

23	202	0,692
24	102	0,762
25	116	0,683

SOM = 5x10

A quantidade de genes e o índice de homogeneidade de cada grupo são apresentadas na Tabela A.83.

Tabela A.83: Agrupamento SOM = 5x10.

Grupos	Qtd de genes	Homogeneidade
1	84	0,915
2	39	0,829
3	33	0,774
4	61	0,746
5	79	0,588
6	53	0,636
7	56	0,682
8	71	0,74
9	77	0,841
10	58	0,931
11	55	0,84
12	65	0,723
13	77	0,785
14	60	0,784
15	64	0,662
16	62	0,695
17	74	0,533
18	94	0,697
19	78	0,798
20	36	0,904
21	57	0,783
		Continua na próxima página

Tabela A.83 – continuação da página anterior

22	81	0,697
23	37	0,685
24	74	0,712
25	75	0,691
26	84	0,558
27	80	0,625
28	90	0,728
29	64	0,75
30	34	0,873
31	60	0,86
32	81	0,722
33	48	0,577
34	98	0,538
35	61	0,66
36	112	0,691
37	43	0,592
38	78	0,566
39	96	0,669
40	97	0,777
41	73	0,848
42	99	0,632
43	95	0,515
44	44	0,688
45	48	0,679
46	104	0,534
47	73	0,505
48	84	0,449
49	46	0,664
50	64	0,764

A.5.3 Agrupamento SAMBA

Tabela A.84: Agrupamento bidimensional da base de dados GSc com filtro de dados.

Grupos	Escores	Condições	Genes
1	383,959	6	88
2	389,388	5	93
3	522,524	8	87
4	590,422	10	67
5	550,474	8	88
6	746,559	12	72
7	583,018	11	103
8	476,992	6	124
9	384,96	7	80
10	510,339	5	116
11	336,319	10	58
12	284,043	7	73
13	272,952	5	55
14	632,904	6	135
15	641,855	8	112
16	635,837	13	94
17	265,049	12	16
18	399,489	5	96
19	316,315	8	61
20	456,639	9	50
21	811,815	10	91
22	480,88	6	133
23	375,3	7	80
24	359,794	10	42
25	907,175	5	157
26	1332,66	5	191
27	245,055	6	30
28	1528,68	7	164

Continua na próxima página

Tabela A.84 – continuação da página anterior

29	226,006	5	63
30	498,797	5	128
31	166,987	10	12
32	232,2	20	15
33	181,917	9	20
34	197,109	7	20
35	233,735	12	40
36	250,994	10	27
37	251,694	12	33
38	174,68	17	12
39	130,952	5	31
40	95,3225	12	8
41	186,362	9	28
42	272,352	12	19
43	159,325	10	16
44	346,41	9	25
45	75,7631	5	23
46	132,446	7	22
47	239,27	11	26
48	351,167	22	24
49	479,514	17	41
50	318,968	6	38
51	247,46	6	61
52	131,224	8	17
53	217,903	12	30
54	186,806	15	26
55	773,586	9	110
56	459,348	8	37
57	410,098	9	55
58	748,205	13	106
59	602,544	13	48
60	1170,01	16	95
61	392,889	11	75

Continua na próxima página

Tabela A.84 – continuação da página anterior

62	1774,91	8	180
63	430,063	13	55
64	515,681	4	105
65	467,221	13	68
66	114,744	7	26
67	380,983	9	34
68	598,724	12	74
69	530,222	15	57
70	225,527	9	31
71	1224,89	17	99

A.5.4 Validação estatística k-médias

De acordo com os índices de validação estatística de homogeneidade, C e Davies Bouldin, a melhor solução de agrupamento do algoritmo k-médias foi $k = 50$. O índice de separação indicou $k = 20$ como a melhor solução. O índice Dunn e Isolamento indicaram $k = 5$ como a melhor solução e o índice Silhueta $k = 10$.

Tabela A.85: Validação estatística dos agrupamentos k-médias.

k	Homogeneidade	Separação	C	D. Bouldin	Dunn	Silhueta	Isolamento
5	0,322	-0,012	0,286	1,87	0,806	0,056	0,46
10	0,485	-0,012	0,203	1,764	0,526	0,062	0,334
20	0,577	6,40E-04	0,177	1,804	0,454	0,033	0,319
30	0,569	0,016	0,149	1,804	0,454	0,026	0,274
50	0,602	0,022	0,12	1,737	0,398	0,025	0,251

A.5.5 Validação estatística SOM

Para os resultados do algoritmo SOM, o melhor agrupamento foi SOM = 5x10, de acordo com o índice de homogeneidade. O índice C indicou o melhor agrupamento SOM = 5x1, os índices Davies Bouldin e Dunn quando SOM = 5x1, e os demais índices quando SOM = 2x2.

Tabela A.86: Validação estatística dos agrupamentos SOM.

Matriz	Homogeneidade	Separação	C	D. Bouldin	Dunn	Silhueta	Isolamento
5x1	0,479	-0,021	0,182	1,917	0,617	0,097	0,606
2x5	0,58	0,003	0,148	1,782	0,642	0,082	0,468
5x5	0,702	0,028	0,116	1,815	0,5	0,047	0,367
5x10	0,682	0,035	0,109	1,822	0,406	0,028	0,309

A.5.6 Validação biológica k-médias

k = 5

A Tabela A.87 apresenta as funções biológicas e os fatores de transcrição identificados nos grupos.

Tabela A.87: Validação biológica do agrupamento $k = 5$.

Grupos	PRIMA	TANGO
Grupo 1	GO: 0004601 - Atividade de peroxidase	HAP4
	GO: 0015077 - Transporte de cálcio inorgânico	CAR1
	GO: 0051186 - Metabolismo de cofator	STRE
	GO: 0006091 - Geração de metabólitos precursores e energia	MSN2
	GO: 0004175 - Atividade de endopeptidase	MSN4
	GO: 0006119 - Fosforilação oxidativa	RPN4
	GO: 0005975 - Metabolismo de carboidrato	ADR1
	GO: 0016491 - Atividade de oxidoredutase	SUT1
	GO: 0043285 - Catabolismo de biopolímero	
	GO: 0006099 - Ciclo do ácido tricarboxílico	
	GO: 0006118 - Transporte de elétron	
	GO: 0045333 - Respiração celular	
	GO: 0009056 - Catabolismo	
Grupo 2	GO: 0050875 - Processo fisiológico celular	SFP1

Continua na próxima página

Tabela A.87 – continuação da página anterior

	GO: 0006139 - Metabolismo de ácido nucléico	
	GO: 0006997 - Organização nuclelar e biogênese	
	GO: 0006351 - Transcrição, dependente do DNA	
	GO: 0044238 - Metabolismo primário	
	GO: 0006412 - Biosíntese de proteína	
	GO: 0044249 - Biosíntese celular	
Grupo 3	GO: 0008283 - Proliferação celular	SUM1
	GO: 0005200 - Constituinte estrutural do citoesqueleto	
	GO: 0007017 - Processo dependente de microtúbulo	
	GO: 0000279 - Fase M	
	GO: 0042244 - Formação da membrana celular de um esporo	
	GO: 0000819 - Segregação da cromátide irmã	
	GO: 0030154 - Diferenciação celular	
	GO: 0007091 - Transição das fases mitóticas metáfase/anáfase	
Grupo 4	GO: 0007028 - Organização celular e biogênese	
	GO: 0003723 - Ligação do RNA	
	GO: 0007046 - Biogênese do ribossomo	
	GO: 0003724 - Atividade de DNA helicase	
	GO: 0016070 - Metabolismo do RNA	

k = 10

A Tabela A.88 apresenta as funções biológicas e os fatores de transcrição identificados nos grupos.

Tabela A.88: Validação biológica do agrupamento $k = 10$.

Grupos	PRIMA	TANGO
Grupo 1	GO: 0016491 - Atividade de oxidoredutase	HAP4
	GO: 0005975 - Metabolismo de carboidrato	CAR1
	GO: 0015002 - Atividade da oxidase terminal	STRE
	GO: 0006119 - Fosforilação oxidativa	MSN2

Continua na próxima página

Tabela A.88 – continuação da página anterior

	GO: 0006091 - Geração de metabólitos precursores e energia	MSN4
	GO: 0006536 - Metabolismo do glutamato	RPN4
	GO: 0006099 - Ciclo do ácido tricarboxílico	ADR1
	GO: 0051186 - Metabolismo de cofator	SUT1
	GO: 0015077 - Transporte de cálcio inorgânico	HAP2/3/4
	GO: 0045333 - Respiração celular	
	GO: 0006118 - Transporte de elétron	
Grupo 2	GO: 0009112 - Metabolismo de nucleotídeo	SFP1
	GO: 0003735 - Constituinte estrutural do ribossomo	
	GO: 0003743 - Atividade do fator de iniciação da tradução	
	GO: 0003899 - Atividade da RNA polimerase	
	GO: 0009451 - Modificação do RNA	
	GO: 0003723 - Ligação do RNA	
	GO: 0006360 - Transcrição do promotor Pol I	
	GO: 0006412 - Biosíntese de proteína	
	GO: 0043037 - Tradução	
	GO: 0030515 - Ligação do snRNA	
	GO: 0007028 - Organização do citoplasma e biogênese	
	GO: 0016070 - Metabolismo do RNA	
	GO: 0044249 - Biosíntese celular	
	GO: 0044238 - Metabolismo primário	
Grupo 3	GO: 0042244 - Formação da membrana celular de um esporo	SUM1
	GO: 0030435 - Esporulação	
Grupo 6	GO: 0006613 - Co-tradução de proteína de membrana	
	GO: 0006457 - Dobramento de proteína	
	GO: 0004553 - Atividade da hidrolase	
	GO: 0051082 - Ligação de proteína desdobrada	
Grupo 7	GO: 0008283 - Proliferação celular	
	GO: 0019219 - Regulação de nucleotídeo e metabolismo de ácido nucléico	
	GO: 0003723 - Ligação do DNA	
	GO: 0006259 - Metabolismo do DNA	

Continua na próxima página

Tabela A.88 – continuação da página anterior

	GO: 0000723 - Manutenção de telômeros GO: 0007127 - Meiose I GO: 0006310 - Recombinação do DNA	
Grupo 9	GO: 0007017 - Processo dependente de microtúbulo GO: 0008134 - Fator de ligação da transcrição GO: 0000090 - Anáfase (Mitose) GO: 0000087 - Fase M do ciclo celular mitótico GO: 0007049 - Ciclo celular GO: 0007059 - Segregação do cromossomo GO: 0006259 - Metabolismo do DNA GO: 0043283 - Metabolismo de biopolímero GO: 0000279 - Fase M GO: 0019941 - Catabolismo de proteína dependente-modificação GO: 0000067 - Ciclo cromossomal e Replicação do DNA	
Grupo 10	GO: 0006094 - Gliconeogênese GO: 0016620 - Atividade de oxidoredutase GO: 0019320 - Catabolismo da hexose	

k = 20

A Tabela A.89 apresenta as funções biológicas e os fatores de transcrição identificados nos grupos.

Tabela A.89: Validação biológica do agrupamento $k = 20$.

Grupos	PRIMA	TANGO
Grupo 1		HAP4 MSN2 MSN4 ADR1 HAP2/3/4

Continua na próxima página

Tabela A.89 – continuação da página anterior

Grupo 2	GO: 0008135 - Atividade de fator de tradução	SFP1
	GO: 0009112 - Metabolismo de nucleotídeo	AZF1
	GO: 0016874 - Atividade da ligase	
	GO: 0003743 - Atividade do fator de iniciação da tradução	
	GO: 0006399 - Metabolismo de tRNA	
	GO: 0003723 - Ligação do RNA	
	GO: 0006412 - Biosíntese de proteína	
	GO: 0016072 - Metabolismo rRNA	
	GO: 0008452 - Atividade da RNA ligase	
	GO: 0006519 - Metabolismo de aminoácido	
	GO: 0043037 - Tradução	
	GO: 0009066 - Metabolismo de aminoácido	
	GO: 0009126 - Metabolismo de monofostato	
	GO: 0003676 - Ligação de ácido nucléico	
	GO: 0007028 - Organização do citoplasma e biogênese	
	GO: 0016070 - Metabolismo do RNA	
	GO: 0044249 - Biosíntese celular	
	GO: 0044238 - Metabolismo primário	
Grupo 3	GO: 0030435 - Esporulação	SIP4
	GO: 0030476 - Formação da membrana celular de um esporo (Fungo)	
Grupo 4	GO: 0003724 - Atividade da RNA helicase	
	GO: 0030515 - Ligação do snRNA	
	GO: 0009451 - Modificação do RNA	
	GO: 0007046 - Biogênese do ribossomo	
	GO: 0007028 - Organização do citoplasma e biogênese	
	GO: 0016070 - Metabolismo do RNA	
Grupo 9		SIP4
Grupo 13	GO: 0003735 - Constituinte estrutural do ribossomo	SFP1
	GO: 0006090 - Metabolismo de piruvato	RAP1
	GO: 0009059 - Biosíntese de macromolécula	
	GO: 0006414 - Alongamento da tradução	
	GO: 0019320 - Catabolismo da hexose	

Continua na próxima página

Tabela A.89 – continuação da página anterior

Grupo 14	GO: 0004175 - Atividade da endopeptidase	CAR1 RPN4
Grupo 15	GO: 0006555 - Metabolismo de metionina	
Grupo 16	GO: 0006091 - Geração de metabólitos precursores e energia GO: 0044262 - Metabolismo celular de carboidrato GO: 0006112 - Metabolismo de reserva de energia	STRE MSN2 MSN4 ADR1
Grupo 17	GO: 0007017 - Processo dependente de microtúbulo GO: 0051603 - Catabolismo de proteína durante proteólise GO: 0003677 - Ligação do DNA GO: 0005200 - Constituinte estrutural do citoesqueleto GO: 0045934 - Metabolismo de ácido nucléico / regulação negativa GO: 0007049 - Ciclo celular GO: 0007059 - Segregação cromossomal GO: 0006259 - Metabolismo do DNA GO: 0007091 - Transição das fases mitóticas metáfase/anáfase GO: 0051244 - Regulação do processo fisiológico celular GO: 0051276 - Organização cromossomal e biogênese GO: 0007127 - Meiose I GO: 0043283 - Metabolismo de biopolímero GO: 0000279 - Fase M GO: 0006310 - Recombinação do DNA	CAR1 FKH1
Grupo 18	GO: 0046943 - Atividade de transporte de ácido carboxílico	DAL80 DAL82
Grupo 20	GO: 0030036 - Organização do citoesqueleto de actina e biogênese	ADR1

k = 30

Tabela A.90: Validação biológica do agrupamento $k = 30$.

Grupos	PRIMA	TANGO
Grupo 1	GO: 0016491 - Atividade de oxidoredutase	HAP4
	GO: 0005975 - Metabolismo de carboidrato	MSN2
	GO: 0051187 - Catabolismo de cofator	MSN4
	GO: 0006119 - Fosforilação oxidativa	ADR1
	GO: 0006091 - Geração de metabólitos precursores e energia	HAP2/3/4
	GO: 0051186 - Metabolismo de cofator	
	GO: 0009060 - Respiração aeróbica	
	GO: 0005386 - Atividade de transporte	
	GO: 0015077 - Transporte de cálcio inorgânico	
	GO: 0006118 - Transporte de elétron	
Grupo 2	GO: 0008135 - Ligação de ácido nucléido/ Ativação de fator de tradução	SFP1
	GO: 0009112 - Metabolismo de nucleotídeo	AZF1
	GO: 0003723 - Ligação do RNA	
	GO: 0044237 - Metabolismo celular	
	GO: 0006163 - Metabolismo de nucleotídeo (purina)	
	GO: 0006412 - Biosíntese de proteína	
	GO: 0043037 - Tradução	
	GO: 0006520 - Metabolismo de aminoácido	
	GO: 0009126 - Metabolismo de monofostato	
	GO: 0007028 - Organização do citoplasma e biogênese	
	GO: 0016070 - Metabolismo do RNA	
	GO: 0044249 - Biosíntese celular	
	GO: 0044238 - Metabolismo primário	
Grupo 3	GO: 0030435 - Esporulação	SUM1
	GO: 0030476 - Formação da membrana celular de um esporo (Fungo)	
Grupo 4	GO: 0003724 - Atividade da RNA helicase	
	GO: 0009451 - Modificação do RNA	

Continua na próxima página

Tabela A.90 – continuação da página anterior

	GO: 0003723 - Ligação do RNA GO: 0007046 - Biogênese do ribossomo GO: 0007028 - Organização do citoplasma e biogênese GO: 0016070 - Metabolismo do RNA	
Grupo 10	GO: 0003712 - Atividade de cofator na transcrição	SPT23
Grupo 11		RGT1
Grupo 12	GO: 0019941 - Catabolismo de proteína dependente-modificação	
Grupo 13	GO: 0051726 - Regulação do ciclo celular GO: 0007091 - Transição das fases mitóticas metáfase/anáfase GO: 0000279 - Fase M GO: 0007052 - Organização e biogênese do eixo mitótico	
Grupo 14	GO: 0003735 - Constituinte estrutural do ribossomo GO: 0006090 - Metabolismo de piruvato GO: 0009059 - Biosíntese de macromolécula GO: 0006414 - Alongamento traducional GO: 0019320 - Catabolismo da hexose	SFP1 RAP1
Grupo 15	GO: 0004175 - Atividade da endopeptidase	CAR1 RPN4 STP1
Grupo 16	GO: 0006555 - Metabolismo de metionina	
Grupo 18	GO: 0044262 - Metabolismo celular de carboidrato GO: 0006112 - Metabolismo de reserva de energia GO: 0006950 - Resposta a estresse	STRE MSN2 MSN4 ADR1
Grupo 19	GO: 0007005 - Organização e biogênese mitocondrial	
Grupo 22	GO: 0046943 - Atividade de transporte de ácido carboxílico	DAL80 DAL82
Grupo 25	GO: 0051327 - Fase M do ciclo celular meiótico GO: 0007127 - Meiose I GO: 0000279 - Fase M	CAR1
Grupo 26	GO: 0006974 - Resposta à estímulos de danos ao DNA GO: 0007049 - Ciclo celular	MBP1 STB1

Continua na próxima página

Tabela A.90 – continuação da página anterior

	GO: 0006259 - Metabolismo do DNA GO: 0006260 - Replicação do DNA	
Grupo 28		MSN4 GCR1 STB5
Grupo 29	GO: 0000910 - Citoquinase GO: 0004553 - Atividade da hidrolase	ACE2
Grupo 30		MSN4

k = 50

A Tabela A.91 apresenta as funções biológicas e os fatores de transcrição identificados nos grupos.

Tabela A.91: Validação biológica do agrupamento $k = 50$.

Grupos	PRIMA	TANGO
Grupo 1	GO: 0016491 - Atividade de oxidoredutase GO: 0051187 - Metabolismo de cofator GO: 0006119 - Fosforilação oxidativa GO: 0006732 - Metabolismo da coenzima GO: 0006091 - Geração de metabólitos precursores e energia GO: 0009060 - Respiração aeróbica GO: 0005386 - Atividade de carregamento GO: 0015077 - Transporte de cálcio inorgânico GO: 0006118 - Transporte de elétron	HAP4 MSN4 HAP2/3/4
Grupo 2	GO: 0008135 - Atividade de fator de tradução GO: 0003743 - Atividade do fator de iniciação da tradução GO: 0003899 - Atividade da RNA polimerase GO: 0006399 - Metabolismo do tRNA GO: 0003723 - Ligação do RNA GO: 0006412 - Biosíntese de proteína	SFP1
Continua na próxima página		

Tabela A.91 – continuação da página anterior

	GO: 0016072 - Metabolismo de RNAr GO: 0006519 - Metabolismo de aminoácido GO: 0043037 - Tradução GO: 0008175 - Atividade da metiltransferase tRNA GO: 0003676 - Ligação de ácido nucléico GO: 0007028 - Organização do citoplasma e biogênese GO: 0016070 - Metabolismo do RNA GO: 0044249 - Biosíntese celular GO: 0044238 - Metabolismo primário	
Grupo 3	GO: 0030435 - Esporulação	SUM1
Grupo 8	GO: 0003712 - Atividade de cofator na transcrição	
Grupo 11	GO: 0003735 - Constituinte estrtutural do ribossomo GO: 0006090 - Metabolismo de piruvato GO: 0009059 - Biosíntese de macromolécula GO: 0006414 - Alongamento traducional GO: 0019320 - Catabolismo da hexose	SFP1 RAP1
Grupo 13	GO: 0006555 - Metabolismo de metionina	MET31
Grupo 14	GO: 0006112 - Metabolismo de reserva de energia	STRE MSN2 MSN4 ADR1
Grupo 15	GO: 0006807 - Metabolismo de composto de nitrogênio	DAL80 DAL82
Grupo 17	GO: 0030036 - Organização do citoesqueleto de actina e biogênese GO: 0003779 - Ligação de actina	
Grupo 19	GO: 0007127 - Meiose I	CAR1
Grupo 20		MBP1 STB1
Grupo 22		STB5
Grupo 23		ACE2
Grupo 25	GO: 0048610 - Processo fisiológico de reprodução celular GO: 0016567 - Proteína de ubiquitinação	SUM1

Continua na próxima página

Tabela A.91 – continuação da página anterior

	GO: 0007091 - Transição das fases mitóticas metáfase/anáfase GO: 0030476 - Formação da membrana celular de um esporo (Fungo) GO: 0007052 - Organização e biogênese do eixo mitótico	
Grupo 27	GO: 0007017 - Processo dependente de microtúbulo GO: 0045132 - Segregação cromossomal (meiose) GO: 0009719 - Resposta à estímulos endógenos GO: 0007020 - Microtúblulo GO: 0000087 - Fase M do ciclo celular mitótico GO: 0007059 - Segregação cromossomal GO: 0006259 - Metabolismo do DNA GO: 0006260 - Replicação do DNA GO: 0000075 - Checkpoint do ciclo celular GO: 0043283 - Metabolismo de biopolímero GO: 0000279 - Fase M	MBP1
Grupo 28	GO: 0006508 - Proteólise GO: 0004175 - Atividade da endopeptidase GO: 0043285 - Catabolismo de biopolímero GO: 0044248 - Catabolismo celular GO: 0043283 - Metabolismo de biopolímero GO: 0016787 - Atividade da hidrolase	RPN4
Grupo 29		MCM1
Grupo 30	GO: 0048622 - Esporulação reprodutiva	SUM1
Grupo 35	GO: 0006396 - Processamento do RNA GO: 0007028 - Organização citoplasmática e biogênese GO: 0004004 - Atividade da RNA helicase dependente ATP GO: 0006365 - Processamento do transcrito primário - 35S	
Grupo 37	GO: 0003723 - Ligação do RNA GO: 0007046 - Biogênese do ribossomo	
Grupo 38	GO: 0006082 - Metabolismo de ácido orgânico	CAT8 CAR1
Grupo 39		MSN4
Grupo 40	GO: 0006766 - Metabolismo de vitamina	

Continua na próxima página

Tabela A.91 – continuação da página anterior

Grupo 41		INO4 XBP1
Grupo 44	GO: 0000279 - Fase M	CAR1
Grupo 45	GO: 0008652 - Biosíntese de aminoácido GO: 0009082 - Biosíntese de aminoácido II	LEU3 BAS1
Grupo 48	GO: 0006091 - Geração de metabólitos precursores e energia	ADR1
Grupo 49	GO: 0003677 - Ligação do DNA GO: 0006333 - União ou separação da cromatina GO: 0006325 - Manutenção e/ou estabilização da estrutura da cromatina	FKH1
Grupo 50	GO: 0006732 - Metabolismo da coenzima	

A.5.7 Validação biológica SOM

SOM = 5x1

A Tabela A.92 apresenta as funções biológicas e os fatores de transcrição identificados nos grupos.

Tabela A.92: Validação biológica do agrupamento SOM = 5x1.

Grupos	PRIMA	TANGO
Grupo 1	GO: 0007017 - Processo dependente de microtúbulo GO: 0008283 - Proliferação celular GO: 0000819 - Segregação da cromátide irmã GO: 0030154 - Diferenciação celular GO: 0042244 - Formação da membrana celular de um esporo GO: 0007091 - Transição das fases mitóticas metáfase/anáfase GO: 0000279 - Fase M	HAP4 CAR1 STRE MSN2 MSN4 RPN4 ADR1 SUT1 HAP2/3/4
Grupo 2	GO: 0007017 - Processo dependente de microtúbulo	SFP1

Continua na próxima página

Tabela A.92 – continuação da página anterior

	GO: 0009719 - Resposta à estímulos endógenos GO: 0003677 - Ligação do DNA GO: 0004175 - Atividade de endopeptidase GO: 0009056 - Catabolismo GO: 0007049 - Ciclo celular GO: 0006259 - Metabolismo do DNA GO: 0051276 - Organização do cromossomo e biogênese GO: 0000723 - Manutenção dos telômeros GO: 0045005 - Manutenção da fidelidade durante a replicação do DNA dependente do DNA GO: 0007127 - Meiose I GO: 0006271 - Alongamento da fita do DNA GO: 0043283 - Metabolismo de biopolímero GO: 0006310 - Recombinação do DNA GO: 0019941 - Catabolismo de proteína dependente-modificação GO: 0000067 - Replicação do DNA e ciclo cromossomal	
Grupo 3	GO: 0051179 - Localização GO: 0030695 - Atividade reguladora de GTPase GO: 0006366 - Transcrição do promotor da Pol II GO: 0045045 - Via de secreção GO: 0000910 - Citoquinase GO: 0016192 - Transporte mediado por vesículo GO: 0008610 - Biosíntese de lipídio GO: 0046907 - Transporte intracelular	SUM1
Grupo 4	GO: 0016491 - Atividade de oxidoredutase GO: 0006818 - Transporte de hidrogênio GO: 0005975 - Metabolismo de carboidrato GO: 0006119 - Fosforilação oxidativa GO: 0006732 - Metabolismo da coenzima GO: 0006091 - Geração de metabólitos precursores e energia GO: 0009117 - Metabolismo de nucleotídeo GO: 0006099 - Ciclo do ácido tricarboxílico GO: 0009060 - Respiração aeróbica	

Continua na próxima página

Tabela A.92 – continuação da página anterior

	GO: 0015077 - Transporte de cálcio inorgânico	
Grupo 5	GO: 0009112 - Metabolismo de nucleotídeo GO: 0003735 - Constituinte estrutural do ribossomo GO: 0003899 - Atividade da RNA polimerase GO: 0009451 - Modificação do RNA GO: 0003723 - Ligação do RNA GO: 0044237 - Metabolismo celular GO: 0006360 - Transcrição do promotor da Pol I GO: 0006412 - Biosíntese de proteína GO: 0007028 - Organização do citoplasma e biogênese GO: 0044249 - Biosíntese celular GO: 0044238 - Metabolismo primário	

SOM = 2x5

A Tabela A.93 apresenta as funções biológicas e os fatores de transcrição identificados nos grupos.

Tabela A.93: Validação biológica do agrupamento SOM = 2x5.

Grupos	PRIMA	TANGO
Grupo 1	GO: 0003735 - Constituinte estrutural do ribossomo GO: 0003723 - Ligação do RNA GO: 0044237 - Metabolismo celular GO: 0009059 - Biosíntese de macromolécula GO: 0006412 - Biosíntese de proteína GO: 0042255 - União com o ribossomo GO: 0030515 - Ligação do snRNA GO: 0007028 - Organização e biogênese do citoplasma GO: 0006414 - Alongamento traducional GO: 0030490 - Processamento do pré-RNA GO: 0043170 - Metabolismo de macromolécula	SFP1 RAP1

Continua na próxima página

Tabela A.93 – continuação da página anterior

Grupo 2	GO: 0008757 - Atividade da metiltransferase	
	GO: 0006139 - Metabolismo de ácido nucléico	
	GO: 0003724 - Atividade da RNA helicase	
	GO: 0003743 - Atividade do fator de iniciação da tradução	
	GO: 0003899 - Atividade da RNA polimerase	
	GO: 0009451 - Modificação do RNA	
	GO: 0003723 - Ligação RNA	
	GO: 0006383 - Transcrição da RNA polimerase III	
	GO: 0016072 - Metabolismo rRNA	
	GO: 0007046 - Biogênese do ribossomo	
	GO: 0043037 - Tradução	
	GO: 0007028 - Organização e biogênese do citoplasma	
	GO: 0016070 - Metabolismo RNA	
	GO: 0006365 - Processamento do transcrito primário	
	GO: 0044238 - Metabolismo primário	
Grupo 3	GO: 0051641 - Localização celular	
	GO: 0045045 - Via de secreção	
	GO: 0006457 - Estruturação de proteína	
	GO: 0019941 - Catabolismo de proteína dependente-modificação	
Grupo 4	GO: 0007017 - Processo baseado em microtúbulo	CAR1
	GO: 0006974 - Resposta à estímulos de danos ao DNA	MBP1
	GO: 0003677 - Ligação DNA	RPN4
	GO: 0006302 - Reparo da dupla fita	STB1
	GO: 0000087 - Fase M do ciclo celular mitótico	
	GO: 0005200 - Constituinte estrutural do citoesqueleto	
	GO: 0051656 - Estabelecimento da localização de organela	
	GO: 0007059 - Segregação cromossomal	
	GO: 0006259 - Metabolismo do DNA	
	GO: 0050791 - Regulação do processo fisiológico	
	GO: 0045002 - Reparo da dupla fita	
	GO: 0030472 - Organização mitótica e biogênese no núcleo	
	GO: 0051276 - Organização cromossomal e biogênese	
	GO: 0006260 - Replicação do DNA	

Continua na próxima página

Tabela A.93 – continuação da página anterior

	GO: 0000075 - Checkpoint do ciclo celular GO: 0007127 - Meiose I GO: 0006271 - Alongamento da fita de DNA GO: 0043283 - Metabolismo de biopolímero GO: 0000279 - Fase M GO: 0006310 - Recombinação do DNA GO: 0007064 - Coesão mitótica da cromátide irmã	
Grupo 5	GO: 0051327 - Fase M do ciclo celular meiótico GO: 0030435 - Esporulação GO: 0030476 - Formação da membrana celular de um esporo (Fungo) GO: 0005200 - Constituinte estrutural do citoesqueleto GO: 0000279 - Fase M	SUM1
Grupo 6	GO: 0006006 - Metabolismo de glicose GO: 0006066 - Metabolismo de álcool GO: 0046365 - Catabolismo de monossacarídeo	MSN4
Grupo 7	GO: 0016491 - Atividade de oxidoredutase GO: 0006119 - Fosforilação oxidativa GO: 0006091 - Geração de metabólitos precursores e energia GO: 0051186 - Metabolismo de cofator GO: 0009060 - Respiração aeróbica GO: 0015077 - Transporte de cálcio inorgânico GO: 0006118 - Transporte de elétron	HAP4 MSN2 MSN4 SUT1 HAP2/3/4
Grupo 9	GO: 0016491 - Atividade de oxidoredutase GO: 0005975 - Metabolismo de carboidrato GO: 0051187 - Catabolismo de cofator GO: 0006732 - Metabolismo da coenzima GO: 0006091 - Geração de metabólitos precursores e energia GO: 0009056 - Catabolismo GO: 0006092 - Metabolismo de carboidrato de vias principais GO: 0006081 - Metabolismo de aldeído GO: 0006118 - Transporte de elétron	STRE MSN2 MSN4 ADR1 SUT1
Grupo 10	GO: 0048610 - Processo fisiológico de reprodução celular	

Continua na próxima página

Tabela A.93 – continuação da página anterior

	GO: 0006629 - Metabolismo de lipídio GO: 0007091 - Transição das fases mitóticas metáfase/anáfase	

SOM = 5x5

A Tabela A.94 apresenta as funções biológicas identificadas e os fatores de transcrição identificados nos grupos.

Tabela A.94: Validação biológica do agrupamento SOM = 5x5.

Grupos	PRIMA	TANGO
Grupo 1	GO: 0005515 - Ligação de proteína GO: 0006457 - Estruturação de proteína GO: 0051082 - Ligação de proteína desestruturada	HSF GCR1 STB5
Grupo 2	GO: 0006974 - Resposta à estímulos de danos ao DNA GO: 0051603 - Catabolismo de proteína durante proteólise GO: 0004175 - Atividade da endopeptidase GO: 0006259 - Metabolismo do DNA GO: 0006950 - Resposta à estresse GO: 0043283 - Metabolismo de biopolímero	RPN4
Grupo 3	GO: 0007017 - Processo dependente de microtúbulo GO: 0045132 - Segregação cromossomal (meiose) GO: 0006974 - Resposta à estímulos de danos ao DNA GO: 0006302 - Reparo da dupla fita GO: 0005200 - Constituinte estrutural do citoesqueleto GO: 0007059 - Segregação cromossomal GO: 0006259 - Metabolismo do DNA GO: 0006260 - Replicação do DNA GO: 0000279 - Fase M GO: 0006310 - Recombinação do DNA GO: 0007064 - Coesão mitótica da cromátide irmã	CAR1 MBP1
Grupo 4	GO: 0051327 - Fase M do ciclo celular meiótico	CAR1

Continua na próxima página

Tabela A.94 – continuação da página anterior

	GO: 0007127 - Meiose I GO: 0000279 - Fase M	
Grupo 5	GO: 0030435 - Esporulação GO: 0030476 - Formação da membrana celular de um esporo (Fungo)	SUM1
Grupo 6	GO: 0045045 - Via de secreção	
Grupo 7	GO: 0003677 - Ligação do DNA GO: 0051276 - Organização cromossomal e biogênese	MBP1 STB1
Grupo 10	GO: 0006512 - Ubiquitina GO: 0030154 - Diferenciação celular GO: 0007091 - Transição das fases mitóticas metáfase/anáfase GO: 0030476 - Formação da membrana celular de um esporo (Fungo)	SUM1
Grupo 11	GO: 0003743 - Atividade do fator de iniciação da tradução GO: 0043037 - Tradução GO: 0006520 - Metabolismo de aminoácido GO: 0007028 - Organização do citoplasma e biogênese GO: 0016070 - Metabolismo do RNA GO: 0044238 - Metabolismo primário	
Grupo 12		MCM1
Grupo 16	GO: 0009451 - Modificação do RNA GO: 0003723 - Ligação do RNA GO: 0016072 - Metabolismo rRNA GO: 0007046 - Biogênese do ribossomo GO: 0030515 - Ligação do snRNA GO: 0007028 - Organização do citoplasma e biogênese GO: 0004004 - Atividade da RNA helicase dependente ATP	
Grupo 17	GO: 0006139 - Metabolismo de ácido nucléico GO: 0003899 - Atividade da RNA polimerase GO: 0016070 - Metabolismo do RNA GO: 0007028 - Organização do citoplasma e biogênese	
Grupo 19	GO: 0005975 - Metabolismo de carboidrato GO: 0015980 - Derivação de energia por oxidação de compostos	MSN4 ADR1

Continua na próxima página

Tabela A.94 – continuação da página anterior

	orgânicos GO: 0006092 - Metabolismo de carboidrato de vias principais	
Grupo 20	GO: 0006793 - Metabolismo de fósforo GO: 0051187 - Metabolismo de cofator GO: 0006091 - Geração de metabólitos precursores e energia GO: 0045333 - Respiração celular GO: 0006118 - Transporte de elétron	MSN2 MSN4 ADR1
Grupo 21	GO: 0003735 - Constituinte estrutural do ribossomo GO: 0006412 - Biosíntese de proteína	SFP1 RAP1
Grupo 22	GO: 0006090 - Metabolismo de piruvato GO: 0006066 - Metabolismo de álcool GO: 0006082 - Metabolismo de ácido orgânico GO: 0019320 - Catabolismo da hexose	
Grupo 24	GO: 0005975 - Metabolismo de carboidrato GO: 0016614 - Atividade de oxidoredutase GO: 0006091 - Geração de metabólitos precursores e energia GO: 0006112 - Metabolismo de reserva de energia GO: 0006950 - Resposta à estresse	STRE MSN2 MSN4 ADR1
Grupo 25	GO: 0006091 - Geração de metabólitos precursores e energia GO: 0009060 - Respiração aeróbica GO: 0015077 - Transporte de cálcio inorgânico	PUT3

SOM = 5x10

A Tabela A.95 apresenta as funções biológicas e os fatores de transcrição identificados nos grupos.

Tabela A.95: Validação biológica do agrupamento SOM = 5x10.

Grupos	PRIMA	TANGO
Grupo 1	GO: 0003735 - Constituinte estrutural do citoesqueleto GO: 0000027 - Manutenção e união da subunidade maior do	STRE1 MSN2

Continua na próxima página

Tabela A.95 – continuação da página anterior

	ribossomo	MSN4 ADR1
Grupo 2	GO: 0006826 - Transporte de íon ferro GO: 0006090 - Metabolismo de piruvato GO: 0006066 - Metabolismo de álcool GO: 0019320 - Catabolismo da hexose	
Grupo 4	GO: 0015672 - Transporte de cálcio inorgânico GO: 0006091 - Geração de metabólitos precursores e energia GO: 0009060 - Respiração aeróbica	
Grupo 5	GO: 0016491 - Atividade da oxidoredutase GO: 0005975 - Metabolismo de carboidrato GO: 0006793 - Metabolismo de fósforo GO: 0006091 - Geração de metabólitos precursores e energia GO: 0009060 - Respiração aeróbica GO: 0006118 - Transporte de elétron	
Grupo 9	GO: 00300154 - Diferenciação celular GO: 0007091 - Transição mitótica metáfase/anáfase GO: 0030476 - Formação da membrana celular de um esporo (Fungo)	HAP4
Grupo 10	GO: 0030435 - Esporulação GO: 0030476 - Formação da membrana celular de um esporo (Fungo)	MSN2 MSN4 ADR1 MIG1
Grupo 11	GGO: 0003735 - Constituinte estrutural do ribossomo GO: 0006412 - Biosíntese de proteína	CAR1
Grupo 14	GO: 0006950 - Resposta à estresse	
Grupo 20	GO: 0007127 - Meiose I GO: 0000279 - Fase M	MSN4 ADR1
Grupo 21	GO: 0003723 - Ligação do RNA GO: 0007046 - Biogênese do ribossomo GO: 0016070 - Metabolismo do RNA	CAR1

Continua na próxima página

Tabela A.95 – continuação da página anterior

Grupo 22		MBP1 STB1
Grupo 23		MBP1 MCM1 STB1 UME1
Grupo 24		STB1
Grupo 25		MCM1
Grupo 30	GO: 0005137 - Fase M do ciclo celular meiótico GO: 0007127 - Meiose I GO: 0000279 - Fase M	
Grupo 31	GO: 0003723 - Ligação do RNA GO: 0007046 - Biogênese do ribossomo GO: 0030515 - Ligação do snRNA GO: 0007028 - Organização e biogênese do citoplasma GO: 0016070 - Metabolismo do RNA	CAR1
Grupo 32	GO: 0009152 - Biosíntese de ribonucleotídeo (purina) GO: 0009127 - Biosíntese de nucleosídeo monofosfato (purina) GO: 0006399 - Metabolismo do tRNA GO: 0043037 - Tradução	
Grupo 37		ACE2 SWI5
Grupo 39	GO: 0004175 - Atividade da endopeptidase GO: 0019941 - Catabolismo de proteína dependente-modificação	
Grupo 40	GO: 0007017 - Processo dependente de microtúbulo GO: 0005200 - Constituinte estrutural do citoesqueleto GO: 0007059 - Segregação cromossomal GO: 0000087 - Fase M do ciclo celular mitótico GO: 0031109 - Polimerização ou despolimerização de microtúbulo	RCS1
Grupo 41	GO: 0003724 - Atividade da RNA helicase GO: 0009451 - Modificação do RNA	SUM1

Continua na próxima página

Tabela A.95 – continuação da página anterior

	GO: 0007028 - Organização do citoplasma e biogênese GO: 0007046 - Biogênese do ribossomo GO: 0003723 - Ligação do RNA	
Grupo 42	GO: 0007028 - Organização do citoplasma e biogênese GO: 0016070 - Metabolismo do RNA	SUM1
Grupo 44	GO: 0005515 - Ligação de proteína GO: 0006457 - Estruturação de proteína GO: 0051082 - Ligação de proteína desestruturada	
Grupo 45		REB1 RPN4
Grupo 46		HSF GCR1 STB5
Grupo 48	GO: 0006807 - Metabolismo de composto de nitrogênio	
Grupo 49	GO: 0003677 - Ligação do DNA GO: 0006333 - União ou separação da cromatina GO: 0006259 - Metabolismo do DNA GO: 0051276 - Organização e biogênese do cromossomo	
Grupo 50	GO: 0006974 - Resposta à estímulos de danos ao DNA GO: 0006271 - Alongamento da fita de DNA GO: 0006259 - Metabolismo do DNA GO: 0006298 - Reparo de erro GO: 0006260 - Replicação do DNA GO: 0007064 - Coesão mitótica da cromátide irmã	RAP1

A.5.8 Validação biológica SAMBA

As funções biológicas e os fatores de transcrição identificados nos grupos do algoritmo SAMBA foram as mesmas nos experimentos com e sem a aplicação de filtros de dados e, por esta razão são todos citados a seguir.

Funções biológicas e fatores de transcrição que enriqueceram o resultado dos agrupamentos bidimensionais:

Tabela A.96: Validação biológica do agrupamento SAMBA

PRIMA	**TANGO**
GO: 0006826 - Transporte de íon ferro	MCM1
GO: 0016491 - Atividade de oxidoredutase	CAR1
GO: 0005515 - Ligação de proteína	HAP4
GO: 0006139 - Metabolismo de ácido nucléico	MSN2
GO: 0009084 - Biosíntese de aminoácido e glutamina	MSN4
GO: 0051327 - Fase M do ciclo celular meiótico	ADR1
GO: 0003724 - Atividade da RNA helicase	SUM1
GO: 0003735 - Constituinte estrutural do ribossomo	ACE2
GO: 0042273 - Biogênese da subunidade maior do ribossomo	MIG1
GO: 0006090 - Metabolismo de piruvato	MBP1
GO: 0005975 - Metabolismo de carboidrato	RAP1
GO: 0006974 - Resposta à estímulos de danos ao DNA	HSF
GO: 0051234 - Estabelecimento da localização	REB1
GO: 0006066 - Metabolismo de álcool	RCS1
GO: 0003723 - Ligação do RNA	RPN4
GO: 0009059 - Biosíntese de macromolécula	STRE
GO: 0003677 - Ligação do DNA	STB1
GO: 0000278 - Ciclo celular mitótico	STB5
GO: 0006512 - Ubiquitina	UME1
GO: 0006100 - Metabolismo intermediário do ciclo de ácido tricarboxílico	SFP1
GO: 0000226 - Organização e biogênese do Citoesqueleto do microtúbulo	SWI5
GO: 0005515 - Ligação de proteína	
GO: 0006457 - Dobramento de proteína	
GO: 0006333 - União ou separação da cromatina	
GO: 0006412 - Biosíntese de proteína	
GO: 0051726 - Regulação do ciclo celular	
GO: 0042255 - União com o ribossomo	
GO: 0006091 - Geração de metabólitos precursores e energia	
GO: 0000041 - Transporte de íon metal	
GO: 0004842 - Atividade ligase da proteína ubiquitina	

Continua na próxima página

Tabela A.96 – continuação da página anterior

GO: 0007046 - Biogênese do ribossomo	
GO: 0009058 - Biosíntese	
GO: 0051186 - Metabolismo de cofator	
GO: 0015031 - Proteína de transporte	
GO: 0000027 - Manutenção e união da subunidade maior do ribossomo	
GO: 0005355 - Atividade de transporte de glicose	
GO: 0007059 - Segregação do cromossomo	
GO: 0006259 - Metabolismo do DNA	
GO: 0030515 - Ligação do snRNA	
GO: 0030154 - Diferenciação celular	
GO: 0009060 - Respiração aeróbica	
GO: 0006092 - Metabolismo de carboidrato de vias principais	
GO: 0044262 - Metabolismo celular de carboidrato	
GO: 0005386 - Atividade de transporte	
GO: 0006084 - Metabolismo de acetil-CoA	
GO: 0007091 - Transição das fases mitóticas metáfase/anáfase	
GO: 0045333 - Respiração celular	
GO: 0007028 - Organização do citoplasma e biogênese	
GO: 0005353 - Atividade de transporte de frutose	
GO: 0006414 - Alongamento traducional	
GO: 0019843 - Ligação do rRNA	
GO: 0000079 - Regulação da proteína ciclina dependente da atividade quinase	
GO: 0030435 - Esporulação	
GO: 0006260 - Replicação do DNA	
GO: 0000027 - Manutenção e união da subunidade maior do ribossomo	
GO: 0007127 - Meiose I	
GO: 0030476 - Formação da membrana celular de um esporo (Fungo)	
GO: 0051082 - Ligação de proteína desdobrada	
GO: 0016070 - Metabolismo do RNA	
GO: 0007052 - Organização e biogênese do eixo mitótico	
GO: 0000279 - Fase M	
GO: 0019320 - Catabolismo da hexose	

Continua na próxima página

Tabela A.96 – continuação da página anterior

GO: 0006118 - Transporte de elétron	
GO: 0006310 - Recombinação do DNA	
GO: 0007062 - Coesão da cromátide irmã	
GO: 0044249 - Biosíntese celular	
GO: 0044238 - Metabolismo primário	

A.5.9 Conclusão dos resultados da base de dados GSc

Diferente da base de dados anterior, nesta base não havia nenhum conhecimento prévio dos dados, portanto, as análises foram conduzidas baseadas na comclusão das análises da base de dados CCSc.

Os índices estatísticos indicaram os agrupamentos $k = 5$ e $k = 50$ como as melhores soluções de agrupamento. No entanto, de acordo com as técnicas de validação biológica, o agrupamento $k = 50$ é a melhor solução, pois a ele foi atribuída maior quantidade de funções biológicas e fatores de transcrição.

Esta conclusão reafirma a conclusão da base CCSc de que o índice C é o potencialmente significativo do ponto de vista biológico.

A Tabela A.97 apresenta os resultados dos índices aplicados nos agrupamentos do algoritmo k-médias com e sem a aplicação de filtros de dados.

Tabela A.97: Comparação dos agrupamentos k-médias com e sem aplicação de filtros.

k	Homogeneidade	Separação	C	D-Bouldin	Dunn	Silhueta	Isolamento
5	**0,399**	-0,005	0,198	**1,85**	0,765	**0,069**	**0,489**
5 Filtro	0,322	**-0,012**	**0,286**	1,87	**0,806**	0,056	0,046

k	Homogeneidade	Separação	C	D-Bouldin	Dunn	Silhueta	Isolamento
10	0,452	0,019	0,213	1,77	**0,75**	0,034	**0,396**
10 Filtro	**0,485**	**-0,012**	**0,203**	**1,764**	0,526	**0,062**	0,334

k	Homogeneidade	Separação	C	D-Bouldin	Dunn	Silhueta	Isolamento
20	0,476	0,031	**0,168**	1,856	**0,498**	0,019	0,263
20 Filtro	**0,577**	**6,40E-04**	0,177	**1,804**	0,454	**0,033**	**0,319**

k	Homogeneidade	Separação	C	D-Bouldin	Dunn	Silhueta	Isolamento
30	0,496	0,037	0,162	1,846	0,385	0,008	0,221
30 Filtro	**0,569**	**0,016**	**0,149**	**1,804**	**0,454**	**0,026**	**0,274**

k	Homogeneidade	Separação	C	D-Bouldin	Dunn	Silhueta	Isolamento
50	0,519	0,039	0,146	1,779	**0,457**	0,018	0,218
50 Filtro	**0,602**	**0,022**	**0,12**	**1,737**	0,398	**0,025**	**0,251**

A maioria dos índices estatísticos indicou o melhor agrupamento do k-médias com a aplicação de filtros de dados, da mesma forma que nos experimentos da base de dados CCSc.

A Tabela A.98 apresenta os resultados dos índices aplicados nos agrupamentos do algoritmo SOM com e sem a aplicação de filtros de dados. As melhores soluções foram indicadas quando utilizado filtros de dados.

Tabela A.98: Comparação dos agrupamentos SOM com e sem aplicação de filtros.

SOM	Homogeneidade	Separação	C	D-Bouldin	Dunn	Silhueta	Isolamento
5x1	0,402	0,007	0,193	**1,777**	**0,927**	0,078	0,521
5x1 Filtro	**0,479**	**-0,021**	**0,182**	1,917	0,617	**0,097**	**0,606**

SOM	Homogeneidade	Separação	C	D-Bouldin	Dunn	Silhueta	Isolamento
2x5	0,494	0,026	0,165	1,845	**0,685**	0,04	0,397
2x5 Filtro	**0,58**	**0,003**	**0,148**	**1,782**	0,642	**0,082**	**0,468**

SOM	Homogeneidade	Separação	C	D-Bouldin	Dunn	Silhueta	Isolamento
5x5	0,563	0,041	0,148	**1,848**	0,445	0,022	0,308
5x5 Filtro	**0,702**	**0,028**	**0,116**	1,815	**0,5**	**0,047**	**0,367**

SOM	Homogeneidade	Separação	C	D-Bouldin	Dunn	Silhueta	Isolamento
5x10	0,627	0,048	0,125	**1,854**	0,354	0,007	0,258
5x10 Filtro	**0,682**	**0,035**	**0,109**	1,822	**0,406**	**0,028**	**0,309**

Comparando os índices de ambos os algoritmos é possível concluir que o algoritmo SOM é indicado como a melhor opção de agrupamento na maioria dos casos, confirmando os resultados obtidos da base de dados CCSc, conforme a Tabela A.99.

Tabela A.99: Comparação dos agrupamentos k-médias e SOM com e sem o uso de filtros.

Grupo	Homogeneidade	Separação	C	D-Bouldin	Dunn	Silhueta	Isolamento
5	0,399	-0,005	0,198	1,85	0,765	0,069	0,489
5 Filtro	0,322	-0,012	0,286	1,87	0,806	0,056	0,46
5x1	0,402	0,007	0,193	**1,777**	**0,927**	0,078	0,521
5x1 Filtro	**0,479**	**-0,021**	**0,182**	1,917	0,617	**0,097**	**0,606**

Grupo	Homogeneidade	Separação	C	D-Bouldin	Dunn	Silhueta	Isolamento
10	0,452	0,019	0,213	1,77	**0,75**	0,034	0,396
10 Filtro	0,485	**-0,012**	0,203	**1,764**	0,526	0,062	0,334
2x5	0,494	0,026	0,165	1,845	0,685	0,04	0,397
2x5 Filtro	**0,58**	0,003	**0,148**	1,782	0,642	**0,082**	**0,468**

Grupo	Homogeneidade	Separação	C	D-Bouldin	Dunn	Silhueta	Isolamento
50	0,519	0,039	0,146	1,779	**0,457**	0,018	0,218
50 Filtro	0,602	**0,022**	0,12	**1,737**	0,398	0,025	0,251
5x10	0,627	0,048	0,125	1,854	0,354	0,007	0,258
5x10 Filtro	**0,682**	0,035	**0,109**	1,822	0,406	**0,028**	**0,309**

Dos resultados apresentados na Tabela acima é possível concluir o melhor desempenho do algoritmo SOM combinado com a opção de filtros de dados. Estes resultados ainda foram corroborados com os resultados das técnicas de validação biológica.

A Tabela A.100 ilustra a comparação dos agrupamentos (sem o uso de filtros de dados) dos algoritmos k-médias e SOM, de acordo com as ferramentas de validação biológica TANGO e PRIMA. A indicação das melhores soluções de agrupamento é indicada na tabela com valores em negrito.

Tabela A.100: Comparação biológica dos agrupamentos k-médias e SOM.

Grupo	TANGO	PRIMA
5	46	16
5x1	**80**	**17**

Grupo	TANGO	PRIMA
10	50	17
2x5	**76**	**21**

Grupo	TANGO	PRIMA
50	**72**	14
5x10	**72**	**28**

Em todos os casos, as ferramentas de validação biológica indicaram os agrupamentos do algoritmo SOM com a melhor solução de agrupamento, se comparado com os resultados do algoritmo k-médias.

Portanto, das análises dos índices estatísticos combinadas com as análises das técnicas de validação biológica, é possível concluir que os índices de homogeneidade e o índice C são mais apropriados para o problema de análise de dados de expressão gênica. Esta comparação é baseada somente nos agrupamentos sem o uso de filtro de dados, já que é intrínseca à quantidade de funcões biológicas e fatores de transcrição associadas a cada agrupamento.

Nesta base de dados novamente os resultados do algoritmo bidimensional apresentaram vantagens, identificando estruturas nos dados, que não foram percebidas pelos algoritmos de agrupamento unidimensional.

Apêndice B

Conceitos básicos de biologia

Neste material são apresentados os conceitos básicos de biologia molecular e um pouco de biologia celular úteis para o entendimento da etapa de validação biológica e das bases de dados utilizadas neste trabalho. A última sessão deste material apresenta os conceitos mais relevantes do organismo *Saccharomyces cerevisiae*, que serviu como referência para as análises deste trabalho.

B.1 Biologia molecular

A Biologia Molecular é a área da biologia que estuda a composição química celular e os processos biológicos que ocorrem no interior da célula.

B.1.1 Da célula a um organismo

Em um organismo pluricelular[1] todas as células são provenientes de uma única célula. O ser humano, por exemplo, é originado de uma única célula, resultado da interação do óvulo com o espermatozóide. Esta única célula passa por um processo de divisão, resultando em duas células, em seguida quatro, oito e assim por diante. No início são idênticas, mas no decorrer das etapas de divisão passam por um processo de diferenciação desenvolvendo características específicas. A medida que as células vão se especializando, elas se agrupam seguindo certos padrões formando diferentes tecidos e órgãos. Este agrupamento ocorre através do reconhecimento de "sinais de identidade" entre as células. Células do sangue, por exemplo, possuem peculiaridades que fazem com que células novas que se identifiquem, passem a exercer a mesma função no organismo.

[1] Organismo formado por mais de uma única célula

B.1.2 Célula

Todos os seres vivos são compostos por células, desde as mais simples estruturas unicelulares, as bactérias e os protozoários, até os mais complexos, como o ser humano e as plantas. Há dois tipos de células: procariotas e eucariotas. As células procariotas tem uma estrutura mais simplificada, não possuindo organelas, nem núcleo organizado. Nesse grupo inclui todo tipo de bactérias, inclusive as arqueobactérias e as cianobactérias. As células eucariotas são as células que compõem a maioria dos seres vivos, como plantas, animais e fungos, mas também protozoários e alguns organismos unicelulares como leveduras e algas verdes [Zah03]. A organização das células procariotas e eucariotas e seus componentes celulares (membrana, citoplasma, núcleo, mitocôndria, ribossomos, cloroplastos, etc) são apresentados em detalhes em [Alb04].

Todas a células que compõem um organismo possuem o mesmo material genético, ou genoma. Durante o processo de divisão este material é fielmente copiado para as células-filhas e assim sucessivamente durante toda a vida.

B.1.3 Genoma

O genoma de qualquer organismo contém toda a informação genética necessária para a realização de todas as funções vitais. Em geral, ele é composto por cromossomos que são essencialmente moléculas de DNA extremamente longas compactadas com a ajuda de certas proteínas, conforme ilustrado na Figura B.1.

O tamanho do genoma é variável de espécie a espécie e não reflete a complexidade do organismo. Por exemplo, o genoma humano contém aproximadamente 3 bilhões de pb (pares de base) e cerca de 25.000 genes (é possível que, após uma análise mais detalhada o número de genes seja ligeiramente maior), enquanto o genoma do rato possui cerca de 20.000 genes, do nematódeo *Caenorhabditis elegans* possui cerca de 19.000 genes e da levedura *Saccharomyces cerevisiae* possui aproximadamente 6.000 genes [Gen06b].

B.1.4 DNA (ácido desoxirribonucléico)

O DNA é uma macromolécula constituída de nucleotídeos e organizado em uma cadeia ou fita dupla cuja função é armazenar e transmitir a informação genética. Um nucleotídeo é composto por uma base nitrogenada, por uma pentose e um fosfato. A pentose do DNA é a desoxirribose, e as bases nitrogenadas são classificadas em purinas e pirimidinas. As purinas são a adenina (A) e a guanina (G), enquanto as pirimidinas são a citosina (C) e a timina (T).

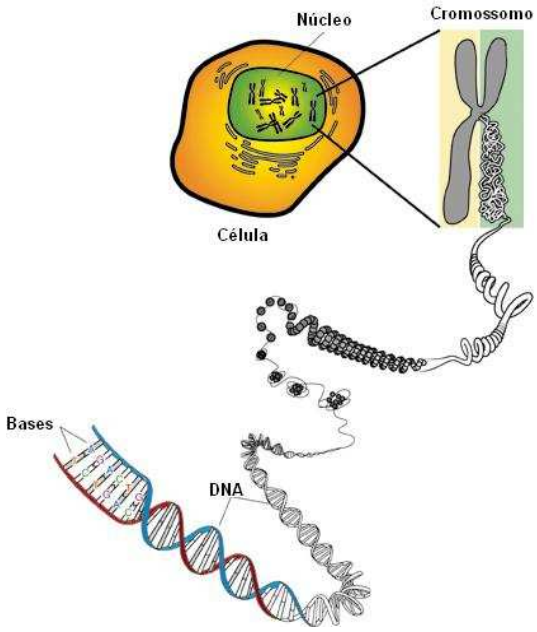


Figura B.1: Localização do DNA em uma célula eucariota. National Human Genome Research Institute (NHGRI).

Em 1953, Watson e Crick postularam um modelo tridimensional para a estrutura do DNA. Este modelo não só explicava muitas das observações sobre as propriedades físicas e químicas do DNA, como também sugeria um mecanismo pelo qual a informação genética poderia ser replicada. O modelo mostrou que o DNA é uma hélice dupla e que as duas fitas de DNA se enrolam em torno do eixo da hélice, conforme ilustrado na Figura B.2. Cada fita tem duas extremidades livres, chamadas de 3' e 5', numa alusão aos átomos de carbono que ficam livres no açúcar que compõem cada nucleotídeo. Neste contexto, duas observações são importantes. A primeira, é que a extremidade 3' de uma fita corresponde à extremidade 5' da outra e, por isso, costuma-se dizer que as fitas são antiparalelas.

```
5' ---- AATGCCGT ---- 3'
3' ---- TTACGGCA ---- 5'
```

Figura B.2: Um exemplo de trecho de DNA de fita dupla.

A segunda observação é que um A em uma fita corresponde a um T na fita oposta, e um C sempre corresponde a um G. Com isto, a seqüência de nucleotídeos de uma das fitas determina completamente a molécula de DNA. É justamente esta propriedade que

permite a auto-duplicação do DNA [MS94].

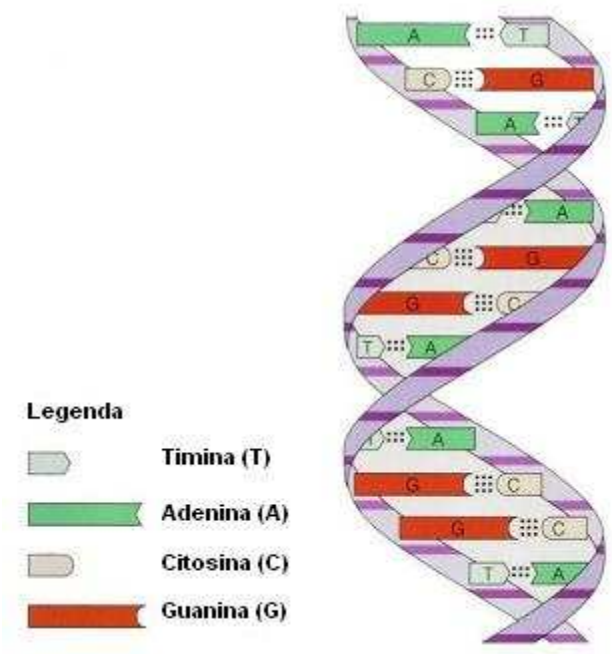


Figura B.3: Estrutura do DNA [Gen06a]

B.1.5 RNA (ácido ribonucléico)

Muito semelhante ao DNA, o RNA é constituído de nucleotídeos, mas organizado em uma única cadeia. A pentose do RNA é a ribose e as bases nitrogenadas são a adenina (A), guanina (G), citosina (C) e a uracila (U). A Tabela B.1 ilustra as duas principais diferenças entre o DNA e o RNA.

Tabela B.1: Principais diferenças entre DNA e RNA.

	Ácido desoxirribonucléico	**Ácido ribonucléico**
Localização	Primariamente no núcleo, também nas mitocôndrias e cloroplastos	No citoplasma, nucléolo e cromossomos
Bases pirimidínicas	Citosina Timina	Citosina Uracila
Bases purínicas	Adenina Guanina	Adnina Guanina

Existem outras diferenças entre DNA e RNA. Enquanto o DNA exerce essencialmente uma função (codificar informação), existem diferentes tipos de RNA na célula

que exercem diferentes funções: o RNA mensageiro (mRNA), o RNA de transferência ou transportador (tRNA) e o RNA ribossômico (rRNA).

O mRNA transfere a informação genética do DNA aos ribossomos, onde ocorre a síntese das proteínas. O tRNA identifica e transporta as moléculas de aminoácidos até o ribossomo para a síntese das proteínas, e o rRNA é o RNA encontrado em maior quantidade na célula e um dos componentes estruturais dos ribossomos, organelas que fornecem um suporte molecular para as reações químicas da montagem de uma cadeia de aminoácidos. Além desses, as células eucatiotas contém ainda outros tipos de RNA: hnRNA (RNAs heterogêneos nucleares), snRNA (pequenos RNAs nucleares) e os sRNA pequenos [Zah03].

B.1.6 Gene

Cada molécula de DNA contém muitos genes. Um gene é uma seqüência específica de nucleotídeos que carrega a informação necessária para a síntese de proteínas e é responsável pela determinação dos traços hereditários dos organismos vivos.

Um gene é constituído, basicamente, por uma região codificadora e por uma região reguladora, conforme ilustrado na Figura B.4. Por convenção, o gene é representado de 5' para 3'. A região codificadora fica posicionada depois (a jusante) da região reguladora. A região reguladora é constituída por um promotor, que sinaliza exatamente onde a síntese do RNA deve ser iniciada, e por outras seqüências reguladoras, como o operador e a UAS.

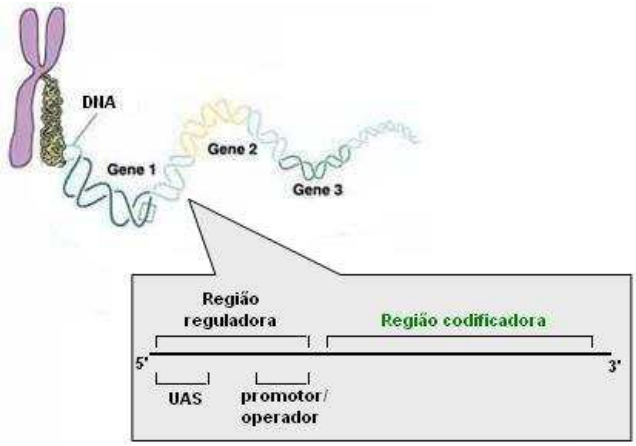


Figura B.4: Estrutura de um gene.

Operadores são sítios de ligação de proteínas repressoras da transcrição e as UAS (do inglês, *upstream activator sequences* = seqüências ativadoras a montante) são sítios de ligação para proteínas ativadoras da transcrição. O operador ocupa uma posição adjacente (anterior ou posterior) à do promotor, podendo, eventualmente haver sobreposição parcial desses dois elementos. Uma UAS típica pode ficar a dezenas ou centenas de pares de bases a montante do promotor [Zah03].

Em um gene, cada três bases específicas (códons) codifica um aminoácido específico a ser sintetizado pela maquinaria celular. A Figura B.5 ilustra um gene transcrito em mRNA. Durante o processo de tradução, a maquinaria celular substituirá cada códon pelo aminoácido correspondente.

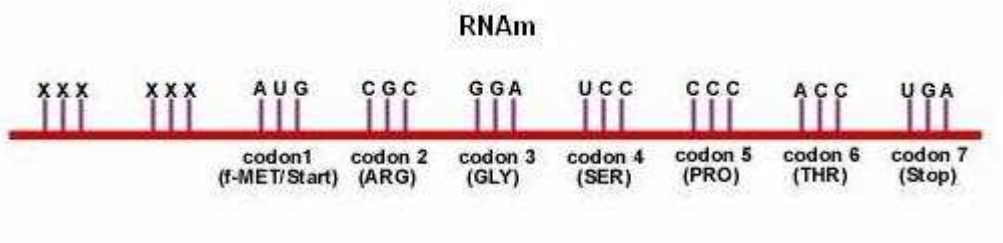


Figura B.5: Códons e seus aminoácidos correspondentes.

Na natureza existem 20 aminoácidos, conforme Tabela B.2. Vários trios de nucleotídeos podem codificar para o mesmo aminoácido; isto é, alguns trios são sinônimos. A prolina, por exemplo, é codificada por CCU, CCA, CCG e CCC. Na maioria dos casos, os códons que são sinônimos diferem somente na base que ocupa a terceira posição no trio e que as duas primeiras bases são mais inflexíveis na codificação. Em conseqüência, as mutações que atingem a terceira base freqüentemente passam desapercebidas (mutações silenciosas) pois elas podem não alterar a composição de aminoácidos da proteína.

Tabela B.2: Os vinte tipos de aminoácidos.

Nome	**Código de 3 letras**	**Código de 1 letra**
Alanina	Ala	A
Cisteína	Cys	C
Ácido aspártico	Asp	D
Ácido glutâmico	Glu	E
Fenilalanina	Phe	F
Glicina	Gly	G
Histidina	His	H
Isoleucina	Ile	I
Lisina	Lys	K
Leucina	Leu	L
Metionina	Met	M
Asparagina	Asn	N
Prolina	Pro	P
Glutamina	Gln	Q
Arginina	Arg	R
Serina	Ser	S
Treonina	Thr	T
Valina	Val	V
Triptofano	Trp	W
Tirosina	Tyr	Y

O sinal de iniciação para a síntese protéica é o códon AUG (Metionina). O sinal de terminação é fornecido pelos códons UAG, UAA, UGA.

Uma cadeia de aminoácidos denomina-se de peptideo, estas podem possuir 2 aminoácidos (dipeptídeos), 3 aminoácidos (tripeptídeos), 4 aminoácidos (tetrapeptídeos), ou muitos aminoácidos (polipeptídeos). O termo proteína é dado quando na composição do polipeptídeo entram centenas, milhares ou milhões de aminoácidos [Zah03].

B.1.7 Proteína

As proteínas são produtos dos genes, derivadas do processo de tradução do mRNA, conforme estabelece o dogma central da biologia molecular (apresentado com mais detalhes a seguir). São as macromoléculas orgânicas mais abundantes nas células, conhecidas como moléculas que realizam o trabalho celular. As unidades constituintes das proteínas são os aminoácidos. Existem muitas espécies diferentes de proteínas, cada uma especializada em uma função biológica. Algumas proteínas são estruturais, usadas para confeccionar paredes celulares, cabelo e diversos tecidos. Outras agem como catalisadores de reações específicas, chamadas de enzimas.

A seqüência de aminoácidos que compõem uma proteína é chamada de estrutura primária desta molécula. A estrutura secundária é dada por interações entre os aminoácidos, que podem formar hélices ou folhas-beta em certos trechos da molécula. A conformação tridimensional completa, incluindo pontes de hidrogênio e ligações fracas

Figura B.6: Código Genético.

entre resíduos, notadamente pontes dissulfídicas, é a estrutura terciária da proteína e a estrutura quaternária descreve a forma com que as diferentes subunidades se agrupam e se ajustam para formar a estrutura total da proteína. As proteínas constituídas por duas ou mais proteínas têm estrutura quaternária, conforme ilustrado na Figura B.7 [MS94].

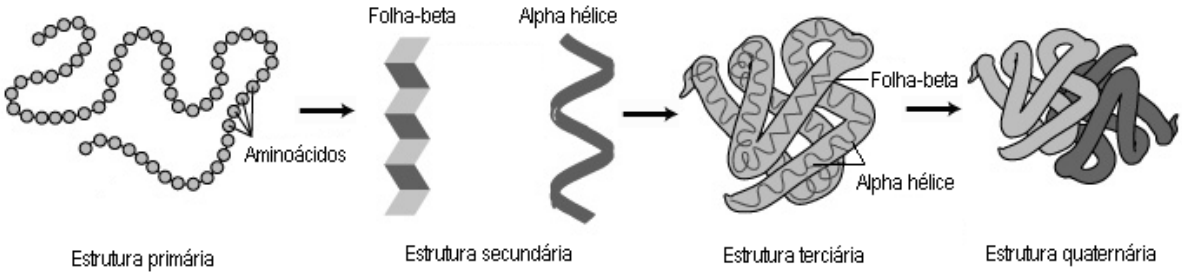


Figura B.7: Estruturas de uma proteína. National Human Genome Research Institute (NHGRI).

B.1.8 O Dogma central da biologia molecular

As instruções para a codificação das proteínas pelos genes são transmitidas através do mRNA. Para que a informação contida no gene seja expressa, uma fita complementar

de RNA é produzida (processo chamado de transcrição) de uma cópia do molde de DNA do núcleo. Este mRNA se move do núcleo até o citoplasma, onde serve de molde para síntese de proteínas. A maquinaria celular de síntese de proteínas converte os códons numa fila de aminoácidos que irão constituir uma proteína (processo chamado de tradução). Em laboratório, a molécula de mRNA pode ser isolada e usada como molde para sintetizar uma fita de DNA complementar (cDNA), que pode ser usado para localizar o gene correspondente em um mapa cromossômico.

Genes que codificam rRNA, tRNA ou outras classes menores de RNAs são transcritos mas não traduzidos, o que só ocorre com o mRNA.

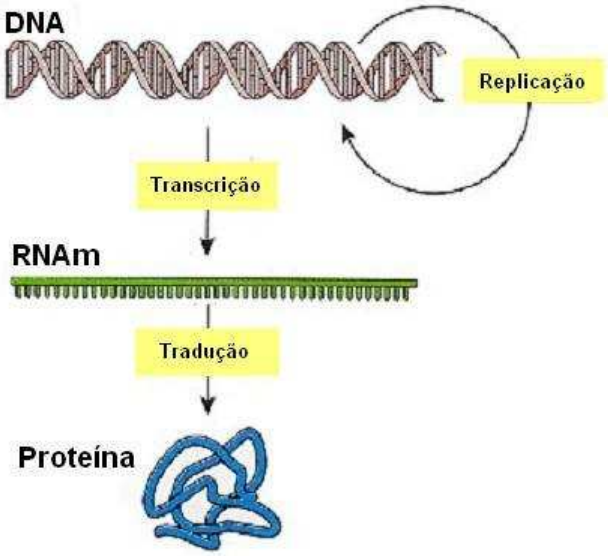


Figura B.8: Dogma central da biologia molecular. [ol06].

O termo transcrição é empregado como sinônimo de síntese do RNA, e tradução como sinônimo de síntese protéica.

O processo de tradução da informação genética pode ser melhor entendido observando a Figura B.9, onde o mRNA é ligado ao ribossomo. Os aminoácidos são levados ao ribossomo pelas moléculas de tRNA. Cada molécula de tRNA adiciona o aminoácido correspondente ao códon no mRNA. A adição vai se processando até o término da cadeia protéica [Zah03].

A expressão de um gene é o processo que inclui a sua transcrição e a eventual tradução do RNA correspondente numa seqüência de aminoácidos. Pesquisas recentes apóiam a hipótese segundo a qual as principais diferenças entre organismos próximos como, por exemplo, humanos e chimpanzés se devem à expressão de seus genes, controlados por grupos de seqüências específicas chamadas fatores de transcrição.

Fatores de transcrição (TFs) são proteínas capazes de regular a transcrição gênica.

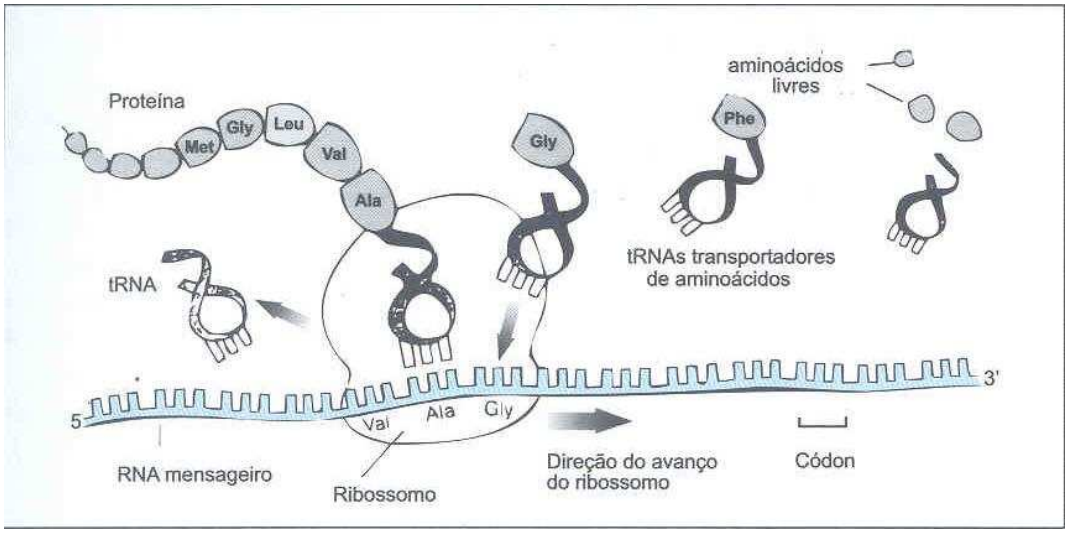


Figura B.9: Processo de tradução da informação genética [Zah03].

Em geral, estas proteínas regulatórias são dotadas de três atividades: ligação a uma seqüência específica do DNA, ativação ou repressão da transcrição, e resposta para sinais regulatórios.

Essas proteínas controlam, quando, onde e como os genes serão transcritos, sendo a base para o controle da expressão gênica. Sem os fatores de transcrição, a maioria dos genes não seria capaz de ser expresso nas células e não haveria, portanto, a chance de as células eucariotas se diferenciarem.

Os organismos eucariotos possuem uma grande variedade de fatores de transcrição que reconhecem um elemento promotor (seqüência no DNA que sinaliza o início da transcrição). A grande afinidade dos fatores de transcrição para sítios específicos determina sua especificidade de ligação. Portanto, fatores de transcrição possuem dois domínios funcionais distintos, um de ligação ao DNA e outro de ligação com a RNA polimerase [Zah03].

Na internet tem disponível alguns bancos de dados de fatores de transcrição, como o TRANSFAC (Transcription Factor Database) [TRA06].

B.1.9 Genômica funcional

O genoma é a informação completa, em termos de seqüências de DNA, de regiões que codificam para genes, assim como, de regiões não codificadoras. Os estudos genômicos compreendem basicamente três fases, das quais o seqüenciamento dos genes é a primeira; numa segunda fase, após a identificação dos genes, programas de computação analisam a similaridade dos genes com seqüências já determinadas de outros organismos e os genes

são agrupados e a eles é atribuída uma função, processo chamado de anotação funcional. A terceira fase dos estudos de genômica é conhecida como genômica funcional.

Genômica funcional é o estudo da funcionalidade dos genes e suas relações com doenças, suas associações com proteínas e sua participação em processos biológicos. Ela busca confirmar as funções atribuídas aos genes pelas análises computacionais, agregando informações de regulação gênica e níveis de expressão aos dados já conhecidos baseando principalmente em experimentos biológicos. Os experimentos executados nos estudos de genômica funcional compreendem estudos de proteômica, onde se determinam as proteínas produzidas pelos organismos frente a determinadas situações e a sua caracterização bioquímica e ainda estudos de expressão gênica. Diversas técnicas têm sido propostas para obtenção da expressão dos genes: MPSS (Massively Parallel Signature Sequence technology), SAGE, Real-time RT-PCR e Microarranjo de DNA [BJB+00, VEVW95, WMFV99, CAHR00]. Muitas dessas técnicas podem ser utilizadas em estudos de genomas inteiros, da expressão de genes ativos, no ordenamento e seqüenciamento dos genes, na determinação de variantes genéticas, em diagnósticos de doenças e várias outras aplicações [Slo02].

B.1.10 Tecnologia de microarranjo de DNA

Desenvolvida em 1990 pela Universidade de Stanford, a técnica de microarranjo de DNA é especialmente apropriada para estudos de comparação da expressão gênica em diferentes tecidos ou diferentes condições que uma população de células possa estar submetida. A preparação do microarranjo é feita a partir de uma lâmina de vidro de dimensões mínimas. Esta lâmina apresenta uma coleção de pequenos poços, ou *spots*, onde uma amostra específica de DNA representando um gene distinto é depositada. Uma vez depositado por dispositivos robotizados, o DNA é fixado através da aplicação de raios U.V.. Como resultado, conforme ilustrado na etapa de Preparação da Sonda na Figura B.10, o microarranjo conterá em cada *spot* uma amostra representativa de um gene, geralmente denominada sonda. O conjunto de sondas fornecerá uma amostragem da totalidade, ou grande parte, dos genes de um organismo qualquer. Os tipos de DNA depositados variam de cDNAs a oligonucleotídeos[2].

O experimento de expressão diferencial está baseado numa das características mais essenciais da molécula de DNA que é a complementaridade entre as duas fitas. Se uma molécula de DNA for submetida a tratamento de calor ou ação de agentes alcalinos, as

[2]Pequenas seqüências de bases de DNA, normalmente variando de 18 a 30, que servem como molde inicial para a extensão do fragmento de DNA.

ligações entre as duas fitas serão rompidas resultando na formação de duas moléculas de DNA simples fita. Tal processo é chamado desnaturação. Neste estado é possível que um DNA simples fita se pareie com uma das fitas de DNA previamente desnaturado, formando uma nova molécula de DNA dupla fita. Isso só ocorre se houver um alto grau de complementaridade entre as duas fitas, fenômeno chamado de hibridização.

O ensaio do microarranjo é justamente um experimento de hibridização. Inicia com a extração do mRNA de dois ou mais tecidos, a serem estudados comparativamente, que representem condições diferenciadas de expressão gênica. A Figura B.10 exemplifica um experimento com tecido sadio como condição controle, sendo comparado com um tecido tumoral, condição variante. Outro exemplo é um experimento com um tecido em um momento zero de um experimento, condição controle, com outro tecido recolhido uma hora depois, condição variante. Os mRNAs das duas fontes são transformados em cDNAs, através da ação de uma enzima que promove a síntese de DNA a partir de um molde de RNA. Cada grupo de cDNAs é marcado com grupamentos químicos distintos de natureza fluorescente, capazes de absorção de energia eletromagnética e subseqüente emissão em faixas de onda também distintas. As marcações mais comuns são a cianina 3 (Cy3) para cDNAs controle e a cianina 5 (Cy5) para cDNAs variantes.

O microarranjo é então submetido a um scanner e irradiado com laser. Os marcadores fluorescentes absorvem a radiação e emitem energia em um outro comprimento de onda, prontamente captado por um leitor óptico adaptado. Cada marcador emite radiação em um comprimento de onda diferente, o que permite avaliar a quantidade de cDNA hibridizado em qualquer *spot* do arranjo, relativo a cada tecido fonte diferente. Como se sabe qual gene está fixado em cada um dos milhares de *spots*, pode-se saber a expressão gênica relativa de um tecido controle em relação ao variante. Cores são computacionalmente atribuídas as faixas de emissão de cada um dos marcadores. A cianina 3, relativa a amostra controle, geralmente recebe a cor verde. A cianina 5, relativa a situação variante, recebe a cor vermelha, conforme ilustradas na Figura B.10 na etapa de preparação do alvo. Sendo assim, um *spot* que tenha cor verde representa um predomínio relativo da expressão gênica na amostra controle, o que implica em uma sub-expressão na condição variante. Ao contrário, um *spot* vermelho indica maior expressão gênica na condição variante, dizendo-se que o gene está relativamente superexpresso nesta condição. O amarelo denota uma situação intermediária de igual expressão gênica nas duas condições. A imagem de um microarranjo com as cores de cada *spot* é também ilustrada na Figura B.10.

As aplicações desta técnica são amplas, variando desde a análise de patologias como o câncer, estudo do efeito de fármacos no organismo humano, análise de tecidos submeti-

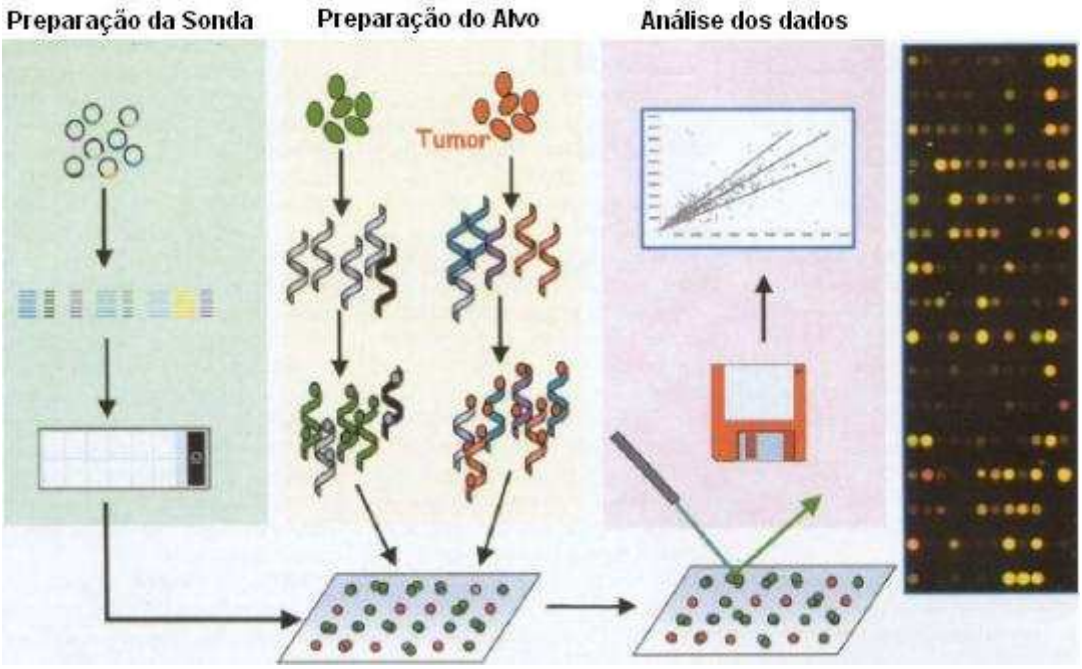


Figura B.10: Esquema de um experimento de análise de expressão gênica utilizando microarranjo de DNA [KBR06].

dos a uma determinada condição de estresse, chegando até a genotipagem e abordagens proteômicas.

A última etapa do experimento de microarranjo consiste na análise dos resultados, que exige o auxílio de técnicas computacionais por envolver a complexidade de uma grande quantidade de dados.

B.2 Biologia Celular

A Biologia Celular é a área da biologia que estuda as células, sua estrutura, funções e sua importância na complexidade dos seres vivos.

B.2.1 Ciclo celular

A capacidade de crescer e se reproduzir é um atributo fundamental de todas as células. No caso das células eucariotas, o processo que origina novas células obedece a um padrão cíclico que começa com o crescimento celular e termina com a partição de seu núcleo e citoplasma em duas células-filhas. As células originadas repetem o ciclo e o número de células aumenta exponencialmente. Este processo é chamado de ciclo celular e serve tanto para manter a vida, em organismos pluricelulares, como para gerar a vida,

no caso de organismos eucariontes unicelulares.

O ciclo celular compreende os processos que ocorrem desde a formação de uma célula até sua própria divisão em duas células-filhas, todas iguais entre si. O ciclo celular é divido em quatro fases: G1 (G0), S, G2 e M. A Figura B.11 ilustra os pontos de controle do ciclo celular. A célula diferenciada se encontra em G0, onde ela atingiu sua diferenciação terminal e está quiescente. Se a célula está destinada a proliferar, ela entra em G1, período em que aumenta de tamanho e prepara as proteínas de que necessita para a síntese de DNA. Durante essa fase, a célula é sensível às condições ambientais. Se elas não forem favoráveis, a divisão celular pára em G1. No entanto, se ultrapassar o ponto R (ponto de restrição), a divisão celular ocorrerá independente de condições ambientais. Na fase S sintetiza-se o DNA que será replicado durante a fase G2. No início de G2 existe outro ponto de controle importante, onde se verificará a qualidade do DNA replicado. Finalmente, na fase mitótica (M), o DNA duplicado será eqüitativamente dividido entre as duas células filhas. A mitose será impedida se, na checagem da mitose, forem constatadas anormalidades na divisão dos cromossomos.

Esta seqüência de fases, com seus respectivos pontos de controle, permite que a célula complete seu ciclo normal, replicando-se sem dar origem a células anormais.

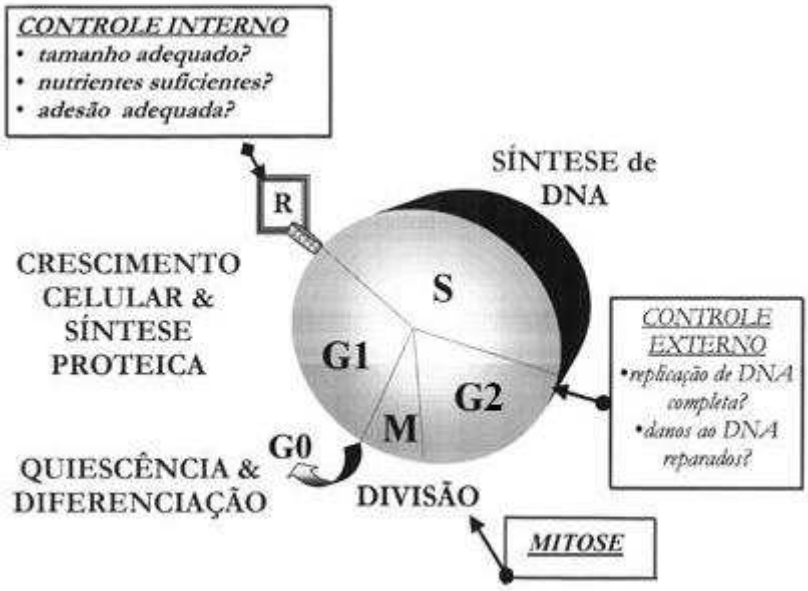


Figura B.11: As fases do ciclo celular de uma célula eucariota [ABE06].

B.2.2 Regulação do ciclo celular

Um ciclo celular bem sucedido é fundamental para a transferência de material genético sem danos de uma geração de células para outra. Se o ciclo não for controlado, as células dividirão continuamente podendo originar células cancerosas, por exemplo.

O ciclo celular pára em determinados pontos e só avança se determinadas condições se verificarem, tais como a presença de uma quantidade adequada de nutrientes ou quando a célula atinge determinadas dimensões. A regulação do ciclo celular é realizada por dois tipos de complexos de proteínas:

- CDK (quinases dependentes de ciclinas)
- Ciclinas

As quinases formam complexos com ciclinas. As ciclinas são proteínas fase-específicas do ciclo celular. Recebem esse nome porque controlam a ativação e desativação dos complexos formados por ciclinas e quinases dependentes de ciclinas (complexos ciclina-CDK) de maneira cíclica nos diferentes estágios do ciclo. Assim, existem complexos ciclina-CdKs de fase M, ciclina-CdKs de fase S, etc, cada qual responsável pela progressão controlada em cada fase do ciclo. As ciclinas não podem ser detectadas nas células que estão iniciando G1, no entanto sua síntese oscila durante o ciclo celular chegando aos níveis mais elevados durante a transição de G1 / S e G2 / M.

Se o gene relacionado à síntese de ciclinas de fase S sofrer uma alteração e se tornar hiperativo, haverá alta concentração de ciclinas de fase S, o que representa um estímulo para a atividade dos complexos ciclina-CdK de fase S e, portanto, para a replicação de DNA de maneira não controlada [Alb04].

B.2.3 Saccharomyces cerevisiae

A levedura *Saccharomyces cerevisiae* é um microrganismo eucarioto não-patogênico que apresenta vantagens como sistema experimental - trata-se de um pequeno organismo unicelular e de fácil proliferação em meios de cultura definidos e condições ambientais controladas. Há inúmeros registros da utilização de leveduras pelo homem, sendo os mais comuns: padarias, cervejarias, farmacologia, enologia e alimentação animal.

Os estudos do mecanismo de ciclo celular de organismos eucariotos têm sido realizados com leveduras como a *Saccharomyces cerevisiae*. As leveduras apresentam grandes vantagens como modelo de estudo. São adequadas para estudos de ciclo celular porque tem tempos de geração curtos (90 minutos comparado às 24 horas de duração do ciclo das

células eucariotas de animais) e seu genoma é aproximadamente cem vezes menos complexo que de uma célula de mamífero, embora mantenha o mesmo tipo de organização do ciclo celular [PTSea98]. Além disso, é um organismo fácil de ser manipulado geneticamente, sendo possível simular mutações em genes que controlam processos celulares básicos, clonar os genes identificados por estas mutações e realizar deleções de genes específicos.

Vários trabalhos sobre a *Saccharomyces cerevisiae* estão disponíveis na internet, incluindo artigos, sites e base de dados. A base de dados SGD, disponível no endereço [Uni06a], disponibiliza o acesso ao genoma completo de *Saccharomyces cerevisiae*, seus genes e produtos e a literatura dessas informações. Spellman e colaboradores disponibilizam o site do projeto de análise do ciclo celular de *Saccharomyces cerevisiae* no endereço [Uni06b], referente ao trabalho de identificação de genes regulados durante o ciclo celular de *Saccharomyces cerevisiae* através do experimento de microarranjo de DNA [PTSea98].