**Estudo comparativo de duas estratégias
utilizadas na busca de relações funcionais
entre genes**

Juan Manuel Vidal García

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E
ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA OBTENÇÃO DO TÍTULO DE
MESTRE EM CIÊNCIAS

Programa: Mestrado em Bioinformática

Orientador: Prof. Dr. André Fujita

Agosto de 2016

ESTUDO COMPARATIVO DE DUAS
ESTRATÉGIAS UTILIZADAS NA
BUSCA DE RELAÇÕES FUNCIONAIS
ENTRE GENES

*Dedicado a
Sabrina Márquez (obviamente)*

Resumo

Um dos fundamentos teóricos e metodológicos da biologia de sistemas é a busca e interpretação das relações entre biomoléculas que têm lugar no interior da célula e que mantêm seu funcionamento. As relações podem existir entre moléculas da mesma natureza, por exemplo entre proteínas, ou entre moléculas de natureza diferente como por exemplo DNA-proteína. De todas as relações possíveis, as relações entre genes tem sido bastante estudadas utilizando dois métodos: correlacionando valores de expressão genética a partir de diferentes medidas de dependência estatística, ou utilizando técnicas biotecnológicas como as interações genéticas que identificam relações entre genes medindo o efeito fenotípico de mutações ou deleções de genes alvo.

Sem importar que sejam técnicas diferentes, o que se procura em ambos casos é identificar relações funcionais que pode existir entre um par de genes. Essa semelhança conceitual faz com que seja possível comparar o resultado das duas estratégias a fim de avaliar a proporção de relações funcionais que são identificadas de modo simultâneo pelas medidas de dependência e pelas interações genéticas. Para levar a cabo dita comparação, as medidas de dependência de Pearson e Spearman foram aqui utilizadas para obter as redes de co-expressão de três conjuntos de dados de expressão genética de *Saccharomyces cerevisiae*. As relações funcionais obtidas no passo anterior foram comparadas com aquelas relações obtidas pela técnica das interações genéticas que se encontram disponíveis nos dois principais bancos de dados.

Como resultado dessas comparações, observou-se que apesar das duas técnicas serem desenhadas com o mesmo objetivo (identificar relações funcionais entre genes), o número de relações que são comuns às duas metodologias estudadas é muito baixo. Tanto a diferença nas técnicas de obtenção das relações como a ausência de uma definição específica do que é uma relação funcional, podem ser as principais causas do baixo nível de relação entre as duas estratégias.

Conteúdo

1	Introdução	1
1.1	Contextualização	1
1.2	Motivação	4
1.3	Objetivos	6
1.3.1	Objetivos específicos	6
2	Marco teórico e estado da arte	7
2.1	Relações funcionais entre genes	7
2.2	Interações Genéticas	8
2.3	Medidas de dependência estatística	11
3	Materiais	14
3.1	Seleção e descrição dos dados utilizados	14
3.1.1	Dados de expressão em estado estacionário	14
3.1.2	Dados de expressão sob diferentes condições	15
3.1.3	Dados de expressão em serie temporal	15
3.1.4	Dados de Interações Genéticas	16
4	Métodos	18
4.1	Processamento e descrição dos dados de co-expressão	18
4.1.1	Medidas de dependência estatística utilizadas	19
4.2	Processamento e descrição dos bancos de dados de interação genética	20
4.2.1	Duplo mutante e cálculo de outras interações genéticas	21
4.3	Comparação dos coeficientes de correlação com as interações genéticas	21
5	Resultados e Discussão	23
5.1	Descrição dos dados de expressão normalizados	23
5.2	Cálculo e descrição de medidas de correlação	26
5.2.1	Dados de expressão em estado estacionário	26

5.2.2	Dados de expressão em série temporal	27
5.3	Processamento e descrição das interações genéticas	32
5.3.1	Obtenção de outras medidas de interação genética . . .	37
5.4	Comparando co-expressão com interações genéticas	39
5.4.1	Comparação da co-expressão com as IG qualitativas . .	40
5.4.2	Comparação da co-expressão com as IG quantitativas .	43
5.4.3	Comparando a co-expressão com IG mínima e com o duplo mutante	48
5.4.4	Comparação segundo processos biológicos	51
6	Conclusões	63
A 1.	Comparação dos resultados obtidos para a medida de Pearson e Spearman quando comparados com as interações genéticas do banco de dados qualitativo	67
B	Curvas ROC para todas as comparações com a medida de Pearson e Spearman segundo os dados de expressão	70
C	Resultados comparação IG mínima e duplo mutante para os dados de expressão em estado estacionário e série temporal	72
D	Resultados comparação IG por processos biológicos para to- dos os conjuntos de expressão	74
E	Gráficos de dispersão para todos os processos biológicos se- gundo cada conjunto de expressão	82
	Bibliografia	91

Lista de Figuras

1.1	Níveis de abstração biológica. Modificada de Benjamin e Sara, 2013	2
2.1	Esquema geral do experimento de interação genética. Tomado de Boone e colaboradores, 2007. a : O primeiro passo a ser feito na busca de interação genética entre um par de genes consiste em trocar cada um deles por marcadores de seleção em duas cepas haplóides diferentes ($MAT\alpha$ e $MATa$). Os marcadores de seleção conferem resistência para condições específicas de crescimento. b : Já com os marcadores (círculos preto e azul) substituindo os genes, ambas culturas são cruzadas a fim de obter células diplóides que carregem as duas mutações. c : A cultura derivada do cruzamento é induzida a esporulação para gerar células haplóides. d , e : As células haplóides resultantes são submetidas a crescimento em condições para as quais os marcadores inseridos oferecem resistência. f : Após todo o processo, só as células haplóides que levem os dois marcadores irão sobreviver e serão deixadas para crescer em meio estável. A taxa de crescimento medido representa então medida quantitativa da interação genética entre os dois genes.	10
4.1	MA-plot característico para dados de expressão normalizados (Modificado de Yang e colaboradores, 2002).	19
5.1	Comportamento dos dados de expressão obtidos em condições de crescimento controlado após o processo de normalização. . .	24
5.2	Comportamento dos dados de expressão obtidos em serie temporal já normalizados.	25
5.3	Comportamento dos dados de expressão obtidos sob múltiplas condições já normalizados.	25

5.4	Distribuição dos coeficientes de correlação para as medidas de Pearson e Spearman a partir de dados de expressão em condições de estado estacionário.	26
5.5	Distribuição dos coeficientes de correlação para as medidas de Pearson e Spearman a partir de dados de expressão em série temporal.	28
5.6	Distribuição dos coeficientes de correlação para as medidas de Pearson e Spearman a partir de dados de expressão em múltiplas condições.	30
5.7	Interações genéticas compartilhadas entre os dois bancos de dados.	33
5.8	Distribuição do número de interações genéticas por gene para cada banco de dados.	34
5.9	Comparação da distribuição dos valores de interação genética calculada a partir de quatro funções de neutralidade diferentes.	37
5.10	Distribuição dos coeficientes de correlação nos três conjuntos de expressão dos pares de genes que têm interação genética.	41
5.11	Distribuição dos coeficientes de correlação nos três conjuntos de expressão comparando os pares de genes com e sem interação genética para o banco de dados quantitativo.	44
5.12	Possíveis resultados obtidos numa curva ROC: (— — —): O teste de classificação produz respostas aleatórias. A : O teste é um classificador perfeito. B : O teste não funciona como classificador.	45
5.13	Curvas ROC para os coeficientes de correlação nos três conjuntos de dados de expressão. TPR: Taxa de verdadeiros positivos. FPR: Taxa de falsos positivos.	46
5.14	Gráfico de dispersão dos valores de interação genética contra os valores de correlação para todos os dados de expressão.	47
5.15	Comparação dos valores de correlação de Pearson com a interação genética mínima. Dados de expressão em múltiplas condições.	49
5.16	Comparação dos valores de correlação de Pearson com os valores dos duplos mutantes. Expressão sob múltiplas condições.	50
5.17	Distribuição dos coeficientes de correlação dos pares de genes com e sem interação genética para diferentes processos biológicos nos três conjuntos de expressão.	53
5.18	Distribuição dos coeficientes de correlação dos pares de genes com e sem interação genética em processos biológicos com poucos genes para os três conjuntos de expressão.	55

5.19	Gráficos de dispersão entre coeficientes de correlação e valores de interação genética para diferentes processos biológicos. . . .	57
A.1	Classificação das interações genéticas segundo diferentes intervalos de correlação. Dados de expressão em estado estacionário.	68
A.2	Classificação das interações genéticas segundo diferentes intervalos de correlação. Dados de expressão em série temporal. . .	68
A.3	Classificação das interações genéticas segundo diferentes intervalos de correlação. Dados de expressão sob múltiplas condições.	68
B.1	Curvas ROC para coeficientes de correlação de Pearson usando os dados de expressão em estado estacionário (a) e série temporal (b), e para a medida de Spearman com os dados de expressão em múltiplas condições (c). TPR: Taxa de verdadeiros positivos. FPR: Taxa de falsos positivos.	71
C.1	Resultados da comparação entre os valores de IG mínima contra a co-expressão dos dados em estado estacionário e série temporal.	72
C.2	Resultados da comparação entre os valores de duplo mutante contra a co-expressão dos dados em estado estacionário e série temporal.	73
D.1	Distribuição dos coeficientes de correlação dos pares de genes com e sem interação genética para diferentes processos biológicos nos três conjuntos de expressão (I).	74
D.2	Distribuição dos coeficientes de correlação dos pares de genes com e sem interação genética para diferentes processos biológicos nos três conjuntos de expressão (II).	75
D.3	Distribuição dos coeficientes de correlação dos pares de genes com e sem interação genética para diferentes processos biológicos nos três conjuntos de expressão (III).	76
D.4	Distribuição dos coeficientes de correlação dos pares de genes com e sem interação genética para diferentes processos biológicos nos três conjuntos de expressão (IV).	77
D.5	Distribuição dos coeficientes de correlação dos pares de genes com e sem interação genética para diferentes processos biológicos nos três conjuntos de expressão (V).	78
D.6	Distribuição dos coeficientes de correlação dos pares de genes com e sem interação genética para diferentes processos biológicos nos três conjuntos de expressão (VI).	79

D.7	Distribuição dos coeficientes de correlação dos pares de genes com e sem interação genética para diferentes processos biológicos nos três conjuntos de expressão (VII).	80
D.8	Distribuição dos coeficientes de correlação dos pares de genes com e sem interação genética para diferentes processos biológicos nos três conjuntos de expressão (VIII).	81
E.1	Gráficos de dispersão entre coeficientes de correlação e valores de interação genética para diferentes processos biológicos (I). .	82
E.2	Gráficos de dispersão entre coeficientes de correlação e valores de interação genética para diferentes processos biológicos (II). .	83
E.3	Gráficos de dispersão entre coeficientes de correlação e valores de interação genética para diferentes processos biológicos (III). .	84
E.4	Gráficos de dispersão entre coeficientes de correlação e valores de interação genética para diferentes processos biológicos (IV). .	85
E.5	Gráficos de dispersão entre coeficientes de correlação e valores de interação genética para diferentes processos biológicos (V). .	86
E.6	Gráficos de dispersão entre coeficientes de correlação e valores de interação genética para diferentes processos biológicos (VI). .	87
E.7	Gráficos de dispersão entre coeficientes de correlação e valores de interação genética para diferentes processos biológicos (VII). .	88
E.8	Gráficos de dispersão entre coeficientes de correlação e valores de interação genética para diferentes processos biológicos (VIII). .	89

Lista de Tabelas

5.1	Número de pares de genes segundo diferentes valores de corte nas medidas de Pearson e Spearman para os três conjuntos de expressão	31
5.2	Descrição das interações genéticas (IG) presentes nos dois bancos de dados utilizados.	34
5.3	Genes mais conectados para cada banco de interação genética.	35
5.4	Comparação das relações funcionais que podem ser identificadas só pelos coeficientes de correlação e pelos coeficientes de correlação junto à interação genética.	42
5.5	Coeficientes de determinação da regressão.	47
5.6	Lista de todos os processos biológicos utilizados para comparações específicas.	52
5.7	Resultados da regressão para diferentes processos biológicos I (EE: Estado estacionário, ST: Série temporal, MC: Múltiplas condições)	58
5.8	Resultados da regressão para diferentes processos biológicos II (EE: Estado estacionário, ST: Série temporal, MC: Múltiplas condições)	59
5.9	Resultados do teste de Mantel para diferentes processos biológicos	61
A.1	Relações funcionais comuns aos métodos de correlação e de interação genética.	69

Capítulo 1

Introdução

1.1 Contextualização

O sequenciamento do primeiro genoma completo em meados dos anos 90 [1] ampliou as perspectivas de estudo no campo da genética e da biologia molecular, hoje caracterizadas pela velocidade em que se produzem e são classificadas as informações bioquímicas de todos os componentes celulares (genes, RNA, proteínas, metabólitos, etc). Usando estas informações tem sido possível passar de aproximações reducionistas, onde os estudos eram focados em componentes isolados, para começar a estudar conjuntos de componentes em termos de estrutura e dinâmica, e assim, fazer uma reconstrução no nível de “sistema” do quadro das relações que existem entre esses componentes e sua contribuição ao funcionamento celular.

Embora essa perspectiva tenha permitido identificar, por exemplo, vias metabólicas importantes como as descritas em [2] para a degradação de polissacarídeos, ou complexos funcionais gerados por relações entre proteínas como assinalado em [3], a identificação dessas relações continua sendo um desafio dado o enorme espaço de possíveis relações não só entre moléculas da mesma natureza mas também entre aquelas de natureza diferente como por exemplo as relações descritas entre metabólitos e RNA [4]. Para contornar esses obstáculos, que já chegaram a ser definidos como “problemas indeterminados” [5], vêm sendo feitas sucessivas divisões abstratas das relações entre biomoléculas em níveis menores segundo a sua natureza. Uma dessas divisões proposta por Benjamin e Sara [6], identifica seis níveis de abstração de sistemas biológicos celulares como se observa na figura (1.1).

As relações ao interior de cada um desses níveis, podem ser modeladas

Figura 1.1: Níveis de abstração biológica. Modificada de Benjamin e Sara, 2013

em forma de rede. Assim, no primeiro nível encontram-se todas as relações que representam a estrutura física e a organização do genoma. Em estas redes, os nós podem representar segmentos de DNA como polimorfismos de nucleotídeo único (SNPs) ligados entre si segundo sua proximidade, sua organização dentro dos cromossomos ou sua homologia. O segundo nível representa a expressão desse genoma em componentes físicos, principalmente RNA. As ligações entre estes elementos representam relações de co-expressão.

No terceiro nível encontram-se as interações físicas entre diferentes elementos: proteína-proteína (PPI), proteína-DNA (PDA) ou proteína-RNA (PRI). O quarto nível de abstração está composto pelas interações genéticas: um conjunto de relações entre genes cuja descrição encontra-se estreitamente ligada ao conceito de epistasia, como redefinido por Fisher [7], onde o mecanismo de ação de um gene, vê-se afetado pela presença ou ausência de um segundo gene. As interações genéticas não necessariamente representam ligações físicas entre moléculas [8].

O quinto nível pode representar processos biológicos como geralmente descritos; isto é: uma serie de eventos biológicos que precisam da ação co-

ordenada de um, ou mais funções moleculares. Neste nível, as redes contêm relações entre moléculas diferentes, implicadas no mesmo processo biológico. Já no sexto nível, os autores incluem redes de fenótipos nas quais as ligações estão dadas entre diferentes tipos de elementos que são comuns entre dois ou mais fenótipos.

Ainda que artificial e por vezes difícil de classificar dentro das categorias acima descritas, a síntese das associações entre biomoléculas feita em [6], além de representar com clareza a complexidade do funcionamento celular, são uma prova da importância que tem a identificação e o estudo das relações entre componentes moleculares, naquilo que hoje se conhece como redes biológicas [9]. Dentre todas elas, duas são objeto de estudo neste trabalho: as redes formadas pelas relações de expressão do nível dois e as redes de interação genética do nível quatro.

Cada um desses níveis já se tem posicionado como um área ativa de pesquisa. No caso das redes de co-expressão, inúmeros exemplos podem ser citados, entre seus destaques vale ressaltar o trabalho de Stuart e colaboradores [10], no qual identificaram a partir de análise de mRNA, módulos de genes com níveis de expressão conservada entre vários organismos diferentes; ou o trabalho de Spellman [11] que identifica os genes envolvidos na regulação do ciclo celular. No caso das interações genéticas, o trabalho de Tong [12] considera-se um referente, dado que em ele são procuradas as interações genéticas para mais de 1000 genes de levedura. Além desse, também vale a pena destacar o trabalho de Constanzo e colaboradores onde são obtidas interações genéticas para mais de 30% de todos os possíveis pares de genes de levedura [13].

Após feitos grandes avanços dentro desses dois níveis de maneira independente, o próximo passo consiste então em analisar de forma conjunta os resultados que são obtidos em cada um, a fim de oferecer informações relevantes que permitam complementar ou depurar o quadro de relações descritas entre genes e seus produtos.

Sob essa perspectiva já vem sendo feitos trabalhos que tentam relacionar dois ou mais níveis daqueles descritos na figura 1.1. No caso particular dos níveis dois (expressão genética) e quatro (interações genéticas), existem trabalhos que analisam de forma conjunta as interações de proteínas (PPI) bem seja com a co-expressão [14], ou com as interações genéticas [15], porém, não existem ainda estudos que relacionem co-expressão com interações genéticas e escassas referências são feitas ao respeito. Assim por exemplo,

no trabalho de Constanzo e colaboradores [13], só foi feita uma breve menção de que grupos ou “clusters” de interações genéticas tendem a ter níveis de expressão altos, mas nada é dito respeito à similaridade das relações obtidas pelas interações genéticas e pela co-expressão.

1.2 Motivação

Embora classificadas em níveis diferentes, as relações obtidas por meio de expressão genética e de interação genética compartilham o mesmo objetivo: identificar relações entre genes, conhecidas como relações funcionais. Uma relação funcional entre genes pode ser entendida como um efeito ou atividade celular quantificável que possa ser atribuído à associação entre os produtos derivados de um gene. Para exemplificar uma relação funcional, considere a seguinte situação hipotética: Um gene X que sob diferentes condições sempre aparece co-ativado com um segundo gene Y . Apesar de não sabermos com certeza como interage X com Y para que suas atividades estejam sincronizadas, podemos atribuir a esse par XY uma relação funcional.

Para identificar relações funcionais a tarefa principal consiste na quantificação do grau de relação que possa existir entre cada par de genes que sejam objeto de estudo. Uma associação entre um par de genes, bem seja identificada a partir de medidas de dependência estatística como quando são usados valores de expressão [16], ou a partir de medidas fenotípicas como quando são usadas técnicas de interação genética por deleção de genes [8] são consideradas *per se* uma relação funcional.

No entanto, considerando que um gene pode ter interação com qualquer outro dentro do genoma, cria-se um enorme espaço de possíveis relações, pois num estudo com N genes podem existir 2^N relações. Essa característica, junto à dinâmica da célula fazem com que o significado biológico da relação entre um par de genes seja na maior parte dos casos difícil de interpretar, tanto pela presença de relações espúrias, quanto pela dificuldade para atribuir significado biológico a uma determinada relação. Para encarar esses problemas têm sido utilizadas técnicas de agrupamento de genes segundo o perfil da suas relações ou a modelagem das mesmas com estratégias derivadas da teoria dos grafos. O propósito em ambos casos é, além de oferecer uma visão mais integral, obter aquelas relações que destacam acima das outras e que explicam melhor as características biológicas sob estudo.

Por exemplo, usando medidas de dependência estatística sob valores de

expressão do mRNA há trabalhos que vale destacar pelo uso que fazem de técnicas de modelagem de redes para obter relações funcionais, tal é o caso do trabalho de De La Fuente [17] e o trabalho de Studham e colaboradores [18]. O primeiro deles mostrou que os métodos de redes podem ajudar a definir funções biológicas para genes não caracterizados graças a suas relações funcionais com outros genes de função muito bem definida, e o segundo propõe o uso de relações funcionais para inferir redes de regulação genética. Usando relações entre genes vindas de experimentos de disrupção do genoma como é o caso das interações genéticas, também têm utilizado métodos derivados da teoria dos grafos para modelar relações funcionais; dentre os trabalhos mais importantes podem ser citados os seguintes estudos [13, 19], os quais estão desenhando a arquitetura genética global das relações funcionais para *Saccharomyces cerevisiae*.

Embora as metodologias de estudo das relações funcionais baseadas em grafos tenham aportado informação para elucidar importantes relações funcionais entre genes, ainda continua sendo uma tarefa complexa filtrar de dentro do espaço de todas as relações funcionais possíveis, só aquelas que melhor caracterizam uma determinada função ou condição biológica. Prova disso são a ampla gama de métodos que têm sido desenvolvidos tanto para agrupamento como para criar redes de relações funcionais [5, 20–22].

Tendo em vista essas dificuldades, e considerando que tanto os métodos de agrupamento como os de redes baseiam-se na identificação de relações funcionais pareadas bem seja por meio de coeficientes de correlação para dados de expressão, ou por meio de valores de interação genética vindos de experimentos de deleção de genes; cabe-se perguntar o quanto das relações funcionais pareadas que são obtidas usando coeficientes de correlação são também identificadas pelas interações genéticas. Dita comparação, além de contrastar os resultados que são obtidos por duas estratégias desenhadas para identificar relações funcionais entre genes, pode ser útil como filtro do amplo espaço de possíveis relações funcionais, e como alternativa para sua interpretação biológica.

Assim, as motivações do presente trabalho estão inseridas na idéia de comparar as relações funcionais entre genes que podem ser colhidas segundo sejam utilizados coeficientes de correlação ou interações genéticas. Perguntas desse tipo já tem sido feitas para comparar relações de co-expressão com interações de proteínas [23] e abordagens semelhantes também estão sendo utilizadas para combinar a informação de interações genéticas e de proteínas [24], mas uma busca de literatura evidenciou que não existe ainda uma comparação

sistemática das relações obtidas por ambos métodos: interações genéticas e co-expressão.

Este trabalho deve ser considerado como um estudo exploratório necessário antes de fazer qualquer tipo de predição ou integração de dados usados para identificar relações funcionais.

1.3 Objetivos

Avaliar se a co-expressão e as interações genéticas identificam as mesmas relações funcionais entre os genes de um organismo modelo (*Saccharomyces cerevisiae*).

1.3.1 Objetivos específicos

- Pré-processar os dados de expressão genética de *Saccharomyces cerevisiae*.
- Pré-processar os dados de interação genética de *Saccharomyces cerevisiae*.
- Construir as redes de co-expressão da *Saccharomyces cerevisiae*
- Construir as redes de interação genética da *Saccharomyces cerevisiae*
- Comparar as redes de co-expressão com as redes de interação genética.

Capítulo 2

Marco teórico e estado da arte

2.1 Relações funcionais entre genes

Uma busca extensiva na literatura especializada permitiu observar que, embora seja frequentemente utilizado, o conceito de relação funcional entre genes encontra-se vagamente definido. Porém, essa revisão de literatura faz evidente que uma relação funcional consiste em identificar associações entre genes, assumindo que essas relações terão uma funcionalidade que deve aparecer representada biologicamente em um, ou vários dos níveis que tem sido descritos na figura 1.1. Portanto, uma relação funcional entre genes é um conceito abstrato que permite atribuir utilidade aos genes.

A principal estratégia que tem sido utilizada para identificar relações funcionais vem dada pelos estudos de co-expressão [4, 25, 26]. Neles, as relações entre genes são identificadas a partir da associação estatística calculada sob níveis de expressão genética (mRNA). Porém, com o avanço tecnológico feito na última década para adquirir dados em grande escala, uma segunda estratégia vem se posicionando como alternativa na identificação de relações funcionais. Essa estratégia consiste em identificar a relação entre pares de genes a partir da quantificação do efeito fenotípico que produz a mutação ou deleção desses genes, num experimento que se conhece como interação genética [12].

Em ambos os casos, a tentativa de validar essas relações em outros níveis é feita, mas não se garante que todas elas encontrem sua complementariedade em outros níveis. Assim por exemplo, em [27], Kasemi e colaboradores identificam relações entre valores de expressão de genes por meio do coeficiente de correlação de Pearson, e tentam validar essas relações utilizando

informação das redes de interação de proteínas. Sob essa mesma perspectiva, Bandyopadhyay e colaboradores [28] utilizam o conhecimento das redes PPI para avaliar relações entre genes, mas desta vez identificadas por meio de interações genéticas.

Mesmo que grandes esforços tem sido feitos para encontrar a representação física dessas relações abstratas, o certo é que não todas as relações entre genes conseguem ser explicadas em níveis superiores [29]. Por exemplo em [30], é reportada uma baixa correlação entre os níveis de expressão genética e os proteínas, e em [22], é reportada a dificuldade para interpretar as interações genéticas à luz das PPI.

Ainda assim, os estudos que identificam relações funcionais entre genes continuam sendo feitos, mesmo sem que essas relações sejam observadas em outros níveis. Por exemplo no trabalho de Stuart e colaboradores [10], são utilizados dados de expressão vindos de microarranjos para identificar relações entre genes compartilhadas entre *Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans* e *Saccharomyces cerevisiae*. Outro trabalho, de Ostojic e colaboradores [31] identifica as relações funcionais de três genes mitocondriais que são essenciais para o processo de respiração celular, valendo-se das suas interações genéticas.

Apesar que não sempre é possível determinar a reciprocidade das relações funcionais em outros níveis da complexidade celular, as estratégias de co-expressão e de interações genéticas ocupam um amplo espaço dos estudos em genômica funcional, como indicado em [32], e tem oferecido informação biológica valiosa para identificar genes essenciais [33], ou relações em processos de regulação transcricional [11, 18], por citar dois exemplos. A seguir, será feita uma descrição de cada uma dessas técnicas.

2.2 Interações Genéticas

As interações genéticas têm sido amplamente estudadas em organismos modelo [31, 34] como um método para identificar relações funcionais entre genes [35]. O desenho experimental mostrado na figura 2.1 descreve de modo geral os passos a serem feitos para obter uma medida quantitativa do “*fitness*” de uma cultura dupla-mutante. Sendo o “*fitness*” entendido como o sucesso reprodutivo atribuído a um genótipo, geralmente medido no tamanho (crescimento) alcançado por uma cultura que carrega uma dupla mutação [36].

Em síntese, um processo contínuo de cruzamento e seleção é feito começando com duas culturas que levam cada uma, mutações em genes de interesse até chegar a obter por cruzamento uma única cultura que carrega ambos os genes modificados. A medida do crescimento dessa cultura dupla-mutante comparado com o crescimento de uma cultura sem mutações, representa o grau de interação entre os genes de interesse.

Segundo Boone e colaboradores [8] é muito comum encontrar que o fenótipo produzido pela mutação ou deleção de dois genes ao mesmo tempo resulte em um efeito diferente daquele gerado pela mutação ou deleção dos mesmos genes de forma individual. Esse fenômeno, que define uma interação genética, pode revelar relações funcionais entre genes [6]. Assim, em contraste com as interações físicas onde o contato direto entre duas moléculas é medido, as interações genéticas representam uma relação funcional entre dois genes, que é observável e quantificável pelo seu efeito conjunto no fenótipo.

A definição de interação genética apresenta dois componentes; o primeiro deles é a medida quantitativa do fenótipo, sendo este na maior parte dos casos a taxa de crescimento em fase exponencial de uma cultura celular mutante com respeito a uma selvagem [37, 38], mas também é possível que seja medido como o crescimento relativo de cepas mutantes em relação a cepas selvagens [39], ou como o número da descendência de células mutantes em referência a células selvagens [40]. Essa medida quantitativa do fenótipo deve ser obtida não só para a cultura dupla-mutante, mas também para as culturas de células que carregam cada um dos genes mutados de forma isolada.

O segundo componente usado para definir uma interação genética é a função de neutralidade. A função de neutralidade representa o valor esperado do fenótipo duplo mutante quando a célula carrega uma dupla mutação ou deleção de genes, e se assume que eles não apresentam interação [35]. Da mesma forma como ocorre com a medida do fenótipo, a função de neutralidade também oferece várias alternativas para ser calculada, no entanto quatro delas têm-se destacado como as mais utilizadas [35].

Geralmente, o nome que recebe a função de neutralidade é também atribuído à interação genética. Assim por exemplo, a definição conhecida como “interação genética mínima” aparece da função de neutralidade mínima. Nessa função, espera-se que uma cultura dupla-mutante não apresente interação genética se o efeito no fenótipo gerado pelo duplo mutante é menor ou igual ao efeito fenotípico que produz a mutação de um só dos genes alvo.

Figura 2.1: Esquema geral do experimento de interação genética. Tomado de Boone e colaboradores, 2007. **a:** O primeiro passo a ser feito na busca de interação genética entre um par de genes consiste em trocar cada um deles por marcadores de seleção em duas cepas haploides diferentes (MAT α e MAT α). Os marcadores de seleção conferem resistência para condições específicas de crescimento. **b:** Já com os marcadores (círculos preto e azul) substituindo os genes, ambas culturas são cruzadas a fim de obter células diplóides que carregem as duas mutações. **c:** A cultura derivada do cruzamento é induzida a esporulação para gerar células haplóides. **d, e:** As células haplóides resultantes são submetidas a crescimento em condições para as quais os marcadores inseridos oferecem resistência. **f:** Após todo o processo, só as células haploides que levem os dois marcadores irão sobreviver e serão deixadas para crescer em meio estável. A taxa de crescimento medido representa então medida quantitativa da interação genética entre os dois genes.

Já a definição de interação genética multiplicativa, vem dada pela função de neutralidade multiplicativa onde assume-se que dois genes não apresentam interação genética, se o valor medido no fenótipo é menor ou igual ao produto dos efeitos fenotípicos da mutação de cada gene de forma isolada. Esta definição, que foi originalmente utilizada para medir o “fitness” em termos da frequência de alelos [41], é na atualidade, junto com a definição mínima, as de uso mais comum [35].

Modificações da definição multiplicativa tem gerado outras duas definições de interação genética conhecidas como aditiva e logarítmica [35]. Ambas são utilizadas somente quando o efeito no fenótipo é medido usando o número da descendência de células mutantes em referência a células selvagens, o que, como indicado em [35] é pouco frequente.

Independentemente da definição utilizada, é possível concluir que a interação genética pode ser entendida como o desvio do fenótipo duplo mutante com respeito ao fenótipo esperado, definido pela função de neutralidade. O resultado obtido nos experimentos de interação genética gera valores comumente no intervalo de $[-1, 1]$, pois embora não existe um intervalo estandarizado, a dificuldade de medir fenótipos muito extremos faz com que os resultados mais frequentes estejam dentro do intervalo de $[-1, 1]$ [19].

2.3 Medidas de dependência estatística

De modo geral, a dependência estatística ou a correlação no amplo sentido da palavra segundo Renyi [42], é um parâmetro usado para medir o grau de associação entre duas variáveis aleatórias [43]. Formalmente, duas variáveis aleatórias X e Y com funções de densidade de probabilidade $f_x(X)$ e $f_y(Y)$, são dependentes se:

$$f_{xy}(X, Y) \neq f_x(X)f_y(Y) \quad (2.1)$$

e são independentes se:

$$f_{xy}(X, Y) = f_x(X)f_y(Y) \quad (2.2)$$

Em se tratando de genes, o sinal dos seus níveis de expressão medidos em termos absolutos (qPCR) ou relativos (Microarranjos) são geralmente tomados como as variáveis aleatórias com as quais são calculados diferentes tipos de medidas de dependência. Boa parte dessas medidas de dependência podem ser situadas dentro de duas categorias principais: informação mútua ou coeficientes de correlação.

Embora as medidas baseadas em informação mútua possam oferecer interpretações teóricas para associações não lineares, elas estão bem definidas para variáveis discretas ou categóricas. Dessa forma, a aplicação de informação mútua não resulta trivial para variáveis quantitativas [44].

Por outra parte, as medidas baseadas em coeficientes de correlação fornecem uma relativa facilidade na interpretação, e seu cálculo também é simples. Porém, talvez o aspecto que tem potenciado o uso de coeficientes de correlação é o fato que, biologicamente é muito mais aceito que os objetos (moléculas, tecidos, órgãos, etc) possam estar biologicamente mais relacionados entanto compartilhem uma maior similaridade, o que tem sido chamado como tendência de similaridade [45] e que no caso da expressão de genes suporta a idéia de que a maior similaridade entre os valores de expressão (padrões de aumento ou queda nos valores de expressão em forma coordenada, como definido em [46]), maior será sua relação [45]. Como consequência desse raciocínio, valores de correlação fortes obtidos por coeficientes de correlação de Pearson ou por variações dele [47], têm sido amplamente adotados como indicadores de relação funcional entre genes.

O coeficiente de correlação de Pearson [48] mede o grau de associação linear entre duas variáveis, no intervalo $[-1, +1]$, onde o sinal indica se a correlação é positiva ou negativa, e o modulo indica a força dessa associação. O coeficiente de correlação de Pearson é dado pela seguinte expressão, onde cov e var indicam a covariância e variância, respectivamente:

$$p = \frac{cov(X, Y)}{\sqrt{var(X)var(Y)}} \quad (2.3)$$

Por sua vez, a correlação de Spearman [49] é uma medida não paramétrica para avaliar o nível de associação não linear entre dois níveis de expressão

gênica X e Y . O coeficiente de correlação de Spearman é dado por:

$$r = 1 - \frac{6 \sum (d_i^2)}{n(n^2 - 1)} \quad (2.4)$$

Em que n é o número dos pares dos valores (medições da expressão gênica) e d_i é a diferença entre i -ésima estatística de ordem [47]. A medida de Spearman assume valores entre -1 e 1 e sua interpretação é similar ao do coeficiente de Pearson.

Tanto a medida de correlação de Pearson quanto a de Spearman capturam relações funcionais que tem-se mostrado em muitos casos, consistentes na identificação de relações funcionais a partir de valores de expressão [10,50,51]. Além disso alguns estudos que avaliam estimadores de associação destacam o uso dessas medidas para inferência de redes de regulação; como é o caso de Olsen e colaboradores [52]. Em 2014, Kurt e colaboradores [53] também avaliaram mais de 20 medidas de associação diferentes que tem sido utilizadas na identificação de relações funcionais e na inferência de redes. Neste trabalho, sinalizam que os coeficientes de correlação de Pearson e Spearman oferecem um grande desempenho tanto pelo seu tempo de execução como pelos resultados que oferecem dentro dos algoritmos de inferência.

Capítulo 3

Materiais

A levedura *Saccharomyces cerevisiae*, pelas suas características de rápido crescimento e fácil manipulação converteu-se num organismo modelo amplamente descrito. Assim, a quantidade de informação biológica obtida a partir da levedura propicia as tentativas de comparar ou integrar informações celulares vindas de níveis e/ou técnicas diferentes [54]. Munidos dessas características, a presente pesquisa utiliza três conjuntos de co-expressão de levedura, identifica as relações entre seus genes usando medidas de dependência estatística e compara esses resultados com as relações obtidas por meio de interações genéticas.

3.1 Seleção e descrição dos dados utilizados

3.1.1 Dados de expressão em estado estacionário

O primeiro conjunto de dados foi escolhido para obter informação da expressão genética a partir de culturas em estado de crescimento estacionário. O objetivo é identificar relações entre genes que não sejam influenciadas por fatores externos induzidos propositalmente, como por exemplo, deficiência de aminoácidos, mutação de genes, presença de antibióticos, ou em geral qualquer alteração intencional da cultura de crescimento que possa afetar drasticamente a expressão de um ou vários genes.

A série identificada com o numero de acesso GSE42215 da base de dados “Gene Expression Omnibus” (-GEO- <http://www.ncbi.nlm.nih.gov/geo/>) cumpre com essas condições e portanto foi selecionada. O grupo esta composto de 100 microarranjos de DNA, os quais foram desenhados com 15.552 son-

das que mapeiam todos os genes caracterizados para levedura [54], além de regiões intergênicas e várias sondas de controle. O microarranjo é identificado com o nome A-UMCU-Y16k-1-3, e foi criado na Universidade de Utrecht.

As cepas de levedura utilizadas provêm de diferentes culturas celulares. Os 100 microarranjos foram hibridizados em um desenho de dois canais, utilizando o fluorocromo Cy3 para marcar o cDNA que vem de células cultivadas sob condições padrão de crescimento e Cy5 para marcar um “pool” de cDNA de referência obtido a partir da mistura de todas as culturas [55]. Neste trabalho, o valor estatístico da mediana de expressão do canal Cy3 foi escolhido como referência para identificar relações funcionais. Os valores de expressão presentes no canal Cy5 não foram aqui considerados.

3.1.2 Dados de expressão sob diferentes condições

Este conjunto de dados contém informação da expressão dos genes da levedura sob diferentes condições. O intuito agora é identificar aquelas relações que possam aparecer como resultado de algum tipo de alteração externa em condições biológicas específicas, como por exemplo, algum tipo de stress.

Em 2008, Faith e colaboradores [56] criaram um compêndio de expressão de genes sob diversas condições utilizando microarranjos com desenho de um canal e provenientes de uma plataforma única da Affimetrix tanto para *E. coli* quanto para *S. cerevisiae*. Os dados encontram-se disponíveis já normalizados pelo método RMA [57] e continuam sendo atualizados. Quando obtidos na presente pesquisa, o banco de dados contava com 530 microarranjos e 8.119 sondas que mapeiam todos os genes para *Saccharomyces cerevisiae*, além de regiões intergênicas e de controle.

Da mesma forma como foi feito com os valores de expressão do conjunto de dados anterior, os valores de expressão aqui obtidos foram utilizados para identificar relações funcionais usando medidas de dependência estatística.

3.1.3 Dados de expressão em serie temporal

O terceiro conjunto de dados é formado por microarranjos de DNA que medem a expressão dos genes em diferentes momentos do ciclo celular. Dado que muitos genes atuam em processos que ocorrem só uma vez durante todo o ciclo celular (como por exemplo, a síntese de DNA ou a citocinese) [58], a escolha destes dados tenta identificar justamente aquelas possíveis relações entre genes que possam aparecer especificamente em alguma das diferentes

etapas do ciclo.

Os dados, originalmente publicados no trabalho de Spellman [11] foram obtidos do site “<http://genome-www.stanford.edu/cellcycle>” e estão divididos em três subconjuntos: *Fator Alpha*, *Elutriação* e *CDC15*, cada um deles representando um método diferente de sincronização do ciclo celular. Nos dois primeiros casos as células de levedura são sincronizadas no período inicial da fase G1 e contam com 16 e 14 micro-arranjos respectivamente, quanto que no terceiro subconjunto a sincronização é feita na fase tardia da mitose, perto do início da fase G1 e contam 25 microarranjos. Para cada método de sincronização, as amostras foram tomadas em intervalos diferentes de tempo. Assim, para os dados sincronizados com o *Fator Alpha* as amostras foram coletadas a cada sete minutos durante 140 minutos; para os dados sincronizados com *Elutriação* a coleta se fez a cada 30 minutos durante sete horas; e finalmente, para os dados sincronizados por *CDC15*, foram feitas coletas a cada dez minutos durante um período de cinco horas.

O total de 55 microarranjos foram hibridizados num desenho de dois canais no qual o fluorocromo Cy3 foi utilizados para marcar amostras de controle não sincronizadas e o fluorocromo Cy5 foi utilizado para marcar as amostras sincronizadas. Os valores da mediana de expressão do canal marcado com Cy5 foram utilizados para identificar as possíveis relações entre genes utilizando medidas de dependência estatística sob o conjunto total dos 55 microarranjos. As medidas de controle presentes no canal Cy3 foram desconsideradas para análise neste trabalho.

3.1.4 Dados de Interações Genéticas

Os dados de interações genéticas que foram utilizados para comparar com as relações obtidas por correlação, encontram-se divididos em três conjuntos. No primeiro deles, as interações genéticas que vêm de diversos estudos e que foram depositadas no site do SGD (*Saccharomyces Genome Database*) foram obtidas. No total, encontram-se disponíveis 340.355 interações, as quais foram filtradas para evitar relações redundantes ou computacionalmente preditas. Após o filtro, a lista resultou em 280.755 relações entre pares de genes biologicamente testadas. Neste trabalho, chamaremos essa lista como “banco qualitativo de interações genéticas”.

Um segundo conjunto de interações genéticas vindos do trabalho de Costanzo e colaboradores [13], também foi utilizado no presente estudo. A dife-

rença do anterior, esse banco de dados inclui não só os nomes dos pares de genes que compartilha uma interação, mas também os valores medidos para cada uma dessas interações genéticas, junto com outras medidas. No total, esse banco de dados contém 763.854 interações genéticas, e para diferenciá-las do conjunto anterior, foram chamadas de “banco de interações genéticas quantitativas”.

O terceiro, e último conjunto utilizado abrange as interações genéticas pertencentes a diferentes processos biológicos. Para criar esse banco de dados, dois passos básicos foram feitos: Um primeiro passo foi obter do *site* Gene Ontology (geneontology.org) as listas de genes que participam de todos os processos biológicos para *Saccharomyces cerevisiae*. No total, foram recuperadas listas de genes para 63 processos biológicos específicos de levedura.

Após isso, cada uma das listas obtidas no passo anterior foi utilizada como base para filtrar dos bancos de dados “qualitativo” e “quantitativo” só aquelas interações que pertencem a cada processo. Assim por exemplo, o termo “ATPase Activity” tem associados quinze genes e cada um deles por sua vez tem um número determinado de interações com outros genes, que bem podem estar dentro do processo “ATPase Activity” ou fora dele. Tanto as interações genéticas restritas aos genes dentro do processo biológico, como aquelas que ficam fora dele foram obtidas e armazenados em arquivos independentes para ser utilizadas em análises posteriores. Ao final do processo, foram obtidas interações genéticas para 55 termos ontológicos diferentes.

Capítulo 4

Métodos

Nesta seção são descritos os procedimentos que foram desenhados para avaliar se as técnicas de co-expressão e as de interação genética conseguem identificar as mesmas relações funcionais entre genes. Os procedimentos iniciais foram desenhados com o intuito de descrever o modo padrão em que os valores de expressão medidos no mRNA, e os valores do “fitness” medidos pelos experimentos de interação genética são utilizados para identificar relações funcionais entre genes. Após isso, os métodos utilizados para comparar ambas estratégias serão descritos. O procedimento metodológico está dividido em três seções principais, para detalhar cada um dos passos feitos.

4.1 Processamento e descrição dos dados de co-expressão

Dos três conjuntos de expressão escolhidos, somente os dados de expressão de células em estado estacionário não se encontram normalizados; portanto, eles foram submetidos a um processo de normalização. O primeiro procedimento efetuado foi a subtração do ruído produzido pela intensidade no fundo de cada sonda. Para isso, uma correção local foi efetuada subtraindo o valor medido no fundo de cada “spot” da mediana da intensidade. Tal procedimento é descrito por Draghici [59] como o método geralmente utilizado na correção do ruído do fundo ou “background”.

Em seguida à correção do *background*, um processo de normalização para os microarranjos foi feita utilizando a técnica conhecida como “*Faster Cyclic Loess*” descrita em [60]. Essa técnica faz uma normalização dos dados utilizando uma modificação do algoritmo concebido para normalização tipo

“*Loess*” [61], fazendo com que seu desempenho seja melhor sem perda de qualidade. Aqui, o método de normalização foi aplicado sob o valor da mediana de expressão contido da sonda Cy3.

Uma ferramenta fundamental para o análise de microarranjos é o gráfico conhecido como *MA-plot*. Nesse gráfico, a M , geralmente no eixo vertical, representa a diferença nos valores de expressão com transformação logarítmica, entanto que o A , no eixo horizontal, representa a media dos valores de expressão após transformação logarítmica [62]. Num gráfico *MA-plot* feito para um par de microarranjos normalizados, espera-se então uma nuvem de pontos em torno de zero no eixo vertical, como mostrado na figura 4.1, modificada de [63]. O método “*Faster Cyclic Loess*” normaliza dois microarranjos ao mesmo tempo, aplicando um fator de correção que é obtido a partir do gráfico *MA-plot* feito para os dois microarranjos que estão sendo normalizados [60]. Esse procedimento é seguido para todos os pares possíveis de microarranjos que façam parte do estudo.

Figura 4.1: MA-plot característico para dados de expressão normalizados (Modificado de Yang e colaboradores, 2002).

Para visualizar os resultados do processo, além dos gráficos *MA-plots*, gráficos de caixa ou *Box-plots* foram utilizados.

4.1.1 Medidas de dependência estatística utilizadas

Para obter os valores de co-expressão que tem sido utilizados na comparação com as interações genéticas, as medidas de dependência de Pearson [48] e

Spearman [49] foram utilizadas. Como já foi indicado (Secção 2.3), o coeficiente de correlação de Pearson varia no intervalo de $+1 - 1$ sendo que o sinal positivo indica que quando o valor medido da expressão de um gene aumenta, o do outro aumenta de forma linear; já no caso do sinal negativo, quanto maior o nível de expressão de um gene, menor será o do outro gene, e quando o coeficiente de correlação é zero, indica que não foi identificada uma relação linear entre os valores de expressão desse par de genes. No caso do coeficiente de correlação de Spearman, a interpretação é similar à do coeficiente de correlação de Pearson, a diferença está no fato que a relação pode ser também não-linear e monotônica.

Essas medidas foram aplicadas nos três grupos de expressão (Estado estacionário, Série temporal e Múltiplas condições), gerando como resultado matrizes simétricas que contém os coeficientes de correlação entre cada par de genes. Esses resultados são descritos, fazendo ênfase no modo como eles são comumente utilizados para definir relações funcionais a partir de valores de corte. O objetivo neste caso não é caracterizar biologicamente as relações funcionais que podem ser obtidas; antes disso, o que se pretende é avaliar a capacidade de diminuição do espaço de possíveis relações funcionais, além de obter resultados que sirvam como referência na comparação com as relações funcionais obtidas por interação genética.

4.2 Processamento e descrição dos bancos de dados de interação genética

Com o intuito de aprimorar as informações presentes nos bancos de dados qualitativo e quantitativo, dois procedimentos foram feitos. O primeiro deles consistiu em verificar se existem interações genéticas presentes em ambos bancos de dados. Caso existam, elas serão mantidas somente em um banco de dados.

Já o segundo procedimento, consistiu em criar um grafo não dirigido para cada banco de dados, onde os vértices representam cada um dos genes, e as arestas representam as interações genéticas entre eles. A partir desses grafos, matrizes de adjacência foram obtidas. Ditas matrizes foram utilizadas na comparação com as matrizes que vêm do cálculo da correlação nos dados de expressão.

Após processadas, as interações genéticas presentes em cada banco de dados, foram descritas considerando suas principais características tais como o número total de interações, o número médio de interações, e os genes com maior quantidade de interações.

4.2.1 Duplo mutante e cálculo de outras interações genéticas

Lembrando que o banco de dados quantitativo, além de conter os valores de interação genética (os quais foram calculados usando a definição de interação genética multiplicativa), contém as medidas do “fitness” tanto para os mutantes simples quanto para os duplos mutantes, esses valores foram aqui aproveitados de duas formas: Primeiro, os valores medidos dos “fitness” para os duplos mutantes foram utilizados para comparar contra os valores de correlação. Neste caso o que se pretende é verificar se existe relação entre os valores de co-expressão e o resultado fenotípico de uma dupla mutação antes que nela seja aplicada uma função de neutralidade que defina uma interação genética.

Além disso, os valores dos mutantes simples também foram utilizados para calcular as interações genéticas definidas pelas funções de neutralidade mínima, aditiva e logarítmica (Ver secção 2.2). Os resultados obtidos foram comparados entre si, para avaliar que tanto variam os valores que definem uma interação genética segundo a função de neutralidade que seja utilizada.

4.3 Comparação dos coeficientes de correlação com as interações genéticas

Com o objetivo de verificar se os valores de correlação e as interações genéticas conseguem identificar o mesmo conjunto de relações funcionais, foram feitas várias comparações. Inicialmente, para as relações entre genes presentes no banco de dados qualitativo avaliou-se a distribuição dos coeficientes de correlação, para os mesmos pares de genes, obtidos nos três conjuntos de expressão. Isto com o intuito de verificar se as relações funcionais identificadas por meio de interação genética, também apresentam valores fortes de correlação.

Já para as interações genéticas presentes no banco de dados qualitativo, foram comparados os coeficientes de correlação de Pearson e Spearman, para aqueles pares de genes que têm interação genética e para aqueles que não têm. O intuito é descobrir, se existe alguma faixa nos valores de correlação a partir da qual exista um maior número relações funcionais identificadas pela técnica das interações genéticas. Utilizando esses resultados, foram desenhadas curvas ROC a fim de avaliar o poder de predição que pode ser obtido para identificar relações funcionais com dados de co-expressão, utilizando como controle as relações funcionais obtidas nos experimentos de interação genética.

Além das anteriores comparações, regressões lineares simples e testes de “Mantel” foram calculados entre os valores absolutos da interação genética presentes no banco de dados quantitativo e os valores absolutos da correlação. Com esses procedimentos, espera-se quantificar o grau de relação que pode existir entre essas duas estratégias, as quais são utilizadas para identificar relações funcionais entre genes. Esse procedimento, também foi feito em forma separada para o banco de dados de interação genética quantitativa contra todas as matrizes de correlação.

Capítulo 5

Resultados e Discussão

Nesta seção são apresentados os principais resultados obtidos, ao tempo que se faz uma discussão dos mesmos. A ordem em que são apresentados corresponde à ordem em que foi exposta a metodologia (Seção 4). Assim, nas primeiras três seções são descritos e discutidos os resultados do processamento e da obtenção dos pares de genes alvo de relação funcional.

Já na seção 5.4, são mostrados os resultados da comparação das interações genéticas contra os coeficientes de correlação. Nessas seções, a discussão estará orientada a explicar, baseado na informação biológica existente, os resultados obtidos.

5.1 Descrição dos dados de expressão normalizados

Como já tem sido indicado em vários estudos [59,63], os valores de expressão obtidos pela técnica dos microarranjos, se encontram expostos a uma ampla fonte de variação não biológica. Entre as condições que podem introduzir esse tipo de variação, podem ser citados o processo de manufatura da lâmina de microarranjo, a preparação da amostra biológica, a hibridização das amostras e a quantificação da intensidade do sinal em cada *spot*. Portanto, com o objetivo de minimizar os efeitos dessa variação sistemática, os dados de expressão foram aqui normalizados.

O resultado da normalização dos dados de expressão de células sob estado estacionário é exibido na figura (5.1). Nela mostram-se quatro diagramas de dispersão tipo *MA-plots* em que são comparadas oito amostras escolhidas de

forma aleatória (Figura 5.1a). Pode se observar que grande parte dos pontos encontram-se distribuídos perto de zero no eixo vertical. No caso dos gráficos de caixa, as vinte amostras aleatoriamente escolhidas mostram que existe uma distribuição de valores de expressão semelhante entre os microarranjos (Figura 5.1b).

(a) MA-plots microarranjos já normalizados.

(b) Gráficos de caixa dos valores de expressão normalizados.

Figura 5.1: Comportamento dos dados de expressão obtidos em condições de crescimento controlado após o processo de normalização.

Com respeito aos dados de expressão em série temporal e aos dados de expressão sob diferentes condições, que foram obtidos já normalizados, também foram feitas o mesmo tipo de figuras de diagnóstico. Assim como nos casos anteriores, a figura (5.2) e a figura (5.3), expõem um comportamento de dados que é consoante com dados de expressão normalizados. Neste tipo de comportamento, assume-se que a grande maioria de genes apresentam poucas mudanças nos seus níveis de expressão entre as diferentes amostras. Esse comportamento, como descrito em [60], resulta biologicamente plausível.

5.1. DESCRIÇÃO DOS DADOS DE EXPRESSÃO NORMALIZADOS 25

(a) MA-plots microarranjos já normalizados.

(b) Gráficos de caixa dos valores de expressão normalizados.

Figura 5.2: Comportamento dos dados de expressão obtidos em serie temporal já normalizados.

(a) MA-plots microarranjos já normalizados.

(b) Gráficos de caixa dos valores de expressão normalizados.

Figura 5.3: Comportamento dos dados de expressão obtidos sob múltiplas condições já normalizados.

Já normalizados, os dados de expressão foram utilizados para o cálculo de diferentes medidas de correlação como descrito na seção 5.2

5.2 Cálculo e descrição de medidas de correlação

Após o processamento e a normalização, medidas de dependência estatística foram utilizadas para calcular a correlação entre os níveis de expressão dos genes que compõem cada conjunto de dados. Baseado na ampla utilização [32] e na qualidade dos resultados que podem oferecer [16] (como também foi descrito na seção 2), as medidas de Pearson [48] e Spearman [49] foram aqui utilizadas. O esquema geral de resultados para cada medida está representado por matrizes que contêm os valores de correlação que identificam a força de associação para cada par de genes. Além de ser descritos, os resultados do cálculo da correlação foram utilizados para discutir brevemente como eles podem ser utilizados para definir relações funcionais.

5.2.1 Dados de expressão em estado estacionário

Como produto do cálculo da correlação de Pearson e Spearman para os dados de expressão sob condições de estado estacionário, matrizes quadradas de 6.170 genes foram obtidas. A figura (5.4) mostra um resumo descritivo dos resultados nos quais são apresentados dois histogramas.

Figura 5.4: Distribuição dos coeficientes de correlação para as medidas de Pearson e Spearman a partir de dados de expressão em condições de estado estacionário.

As figuras 5.4a e 5.4b, descrevem a distribuição de todos os valores de correlação obtidos. O que pode ser destacado nesses gráficos é a forma mostrada pela distribuição dos valores de correlação, os quais se apresentam semelhantes a uma distribuição normal com média de 0.016 e desvio padrão de 0.26 no caso da medida de Pearson e de média 0.024 e desvio padrão 0.28 para a medida de Spearman. Essa característica indica que existem poucos pares de genes relacionados por coeficientes de correlação fortes, os quais são um tipo de critério aceito para definir se um par de genes possui (ou não) uma relação funcional. Esse critério, que é conhecido como “força mínima aceitável” (*minimum acceptable strength -MAS-*) [64], é comumente utilizado para definir relações funcionais entre genes a partir de valores de expressão [65–67].

Desse modo, a distribuição da figura 5.4a mostra que 99% dos valores de correlação encontram-se no intervalo de $[-0.7, +0.7]$ e que a pequena percentagem de valores que em termos absolutos encontram-se acima desse intervalo, e que poderiam, a partir do critério do *MAS* ser considerados como fortes, constitui um grupo de pouco mais de 130 mil possíveis relações funcionais. Claramente essas percentagens bem poderiam ser maiores ou menores dependendo da escolha do valor de corte, de fato, se o valor de corte fosse diminuído para $|0.5|$, o número de pares de genes que poderiam ser considerados alvo de relação funcional aumenta a mais de um milhão. Porém sem uma definição específica sob qual deve ser esse limiar, que além disso, é muitas vezes escolhido de forma subjetiva [68,69], o valor de $> |0.7|$ é aqui utilizado só como um valor de referência para descrição dos resultados e não como um corte definitivo para identificar relações funcionais a partir dos dados de expressão que são utilizados neste trabalho.

Por sua parte, a figura 5.4b, que descreve a distribuição dos valores de correlação para a medida de Spearman, também mostra que ao redor de 1% do total de pares de genes apresenta coeficientes de correlação $> |0.7|$. Essa percentagem, que representa quase 150 mil possíveis relações, indica um pequeno aumento em relação aos pares de genes que são identificados pela medida de Pearson acima desse mesmo limiar.

5.2.2 Dados de expressão em série temporal

O cálculo dos coeficientes de correlação de Pearson e Spearman para os dados de expressão ao longo do tempo (Seção 3.1.3), deu como resultado matrizes quadradas de 6102 genes. Os histogramas da figura 5.5, descrevem a distribuição dos coeficientes de correlação obtidos.

(a) Coeficientes de correlação de Pearson para os dados em série temporal.

(b) Coeficientes de correlação de Spearman para os dados em série temporal.

Figura 5.5: Distribuição dos coeficientes de correlação para as medidas de Pearson e Spearman a partir de dados de expressão em série temporal.

Assim como descrito para os dados de expressão em estado estacionário (seção 5.2.1), a distribuição dos coeficientes de correlação dos dados de expressão em série temporal tanto para Pearson quanto para Spearman, apresenta-se próxima de uma distribuição normal. No caso da medida de Pearson, as estatísticas de resumo mostram uma média de 0.0031 e um desvio padrão de 0.38, enquanto que para a medida de Spearman a média é 0.007 e o desvio padrão 0.4.

Para estes dados de expressão, se fosse assumido um valor de corte para a correlação de $|0.7|$, poderiam ser obtidas algo mais que um milhão duzentos mil pares de genes no caso da medida de Pearson, e um milhão quatrocentos mil no caso de Spearman. Esses valores representam 6,9% e 7,8% do total de pares de genes. Já se esse valor de corte fosse diminuído até $|0.5|$, essas percentagens aumentam para 22,3% e 26,4% no caso de Pearson e Spearman respectivamente.

Aqui, é importante destacar o aumento na quantidade de possíveis relações entre pares de genes que são obtidas com coeficientes de correlação $> |0.7|$, quando comparados com os resultados para o mesmo valor de corte nos dados de expressão em estado estacionário. Essa diferença pode ser explicada pela própria natureza dos dados; pois para condições estáveis, como no caso de uma cultura celular em estado estacionário, espera-se que poucos genes

se encontrem diferencialmente expressos. Esta característica, que tem sido descrita como nível de expressão basal [70], permite o funcionamento celular sem grandes alterações nos níveis de expressão.

Já no caso dos dados em série temporal, o que pode ser esperado não é um tipo de expressão basal, mas pelo contrario, fortes oscilações coordenadas nos níveis de expressão de muitos genes como resposta às diferentes funções e transformações que se sucedem a traves do ciclo celular [11, 71]. Essas diferenças na natureza dos dados, pode explicar as diferenças no número de relações com coeficientes $> |0.7|$.

Dados de expressão sob múltiplas condições

Finalmente, os resultados do cálculo da correlação de Pearson e Spearman para os dados de expressão obtidos para diferentes condições são mostrados na figura 5.6. No total, uma matriz quadrada de 6119 genes foi obtida. Dessa figura é possível observar que se mantêm a tendência descrita com respeito à simetria das distribuições que foram descritas nos dois anteriores conjuntos de dados (Figuras 5.4 e 5.5). Assim, no histograma da esquerda (Figura 5.6a) que representa a medida de Pearson, a média dos coeficientes de correlação é de 0.0026 e o desvio padrão é 0.26. Já no caso da medida de Spearman (Figura 5.6b), a média da correlação é 0,003 e o desvio padrão é 0.25; valores próximos aos obtidos para os outros dois conjuntos de dados.

(a) Coeficientes de correlação de Pearson para os dados de múltiplas condições.

(b) Distribuição dos p-valores obtidos para os dados de múltiplas condições.

Figura 5.6: Distribuição dos coeficientes de correlação para as medidas de Pearson e Spearman a partir de dados de expressão em múltiplas condições.

Com respeito à quantidade de pares de genes que apresentam coeficientes de correlação acima de $|0.7|$, foi observado que 1,05% do total de possíveis relações encontram-se acima desse valor de corte na medida de Pearson, e 0,6% no caso da medida de Spearman. Caso esse valor fosse menos restritivo, por exemplo $|0.5|$, as percentagens de pares de genes acima desse coeficiente de correlação aumentam até 6,7% e 5% para a medida de Pearson e Spearman respectivamente.

Em geral, os resultados descritos até agora, mostram que dos três conjuntos de expressão utilizados, os que foram obtidos em série temporal apresentam a maior quantidade de pares de genes com coeficientes de correlação acima de $|0.5|$ e $|0.7|$, quando comparados com os dados de expressão vindos de células em estado estacionário e múltiplas condições, para os quais percentagens menores de relações acima desses mesmos valores de corte foram obtidos.

Se comparados, os resultados obtidos pela medida de Pearson e de Spearman para os três conjuntos de dados, é possível perceber que eles são muito semelhantes. Esta característica coincide com o descrito no trabalho de Kurt e colaboradores [53], onde se afirma que, ainda que a medida de Spearman seja útil para identificar um espectro maior de relações entre variáveis, os resultados que podem ser obtidos usando dados de expressão são muito se-

melhantes.

Entre as diferenças, destaca o fato de que a medida de Pearson consegue identificar uma maior quantidade de pares de genes do que a medida de Spearman, se fossem assumidos valores de corte de $|0, 5|$ ou de $|0, 7|$ nos dados de expressão em múltiplas condições (Ver tabela 5.1). Já nos outros dois conjuntos de dados, a medida de Spearman apresenta um melhor desempenho.

Esses resultados exemplificam um dos métodos mais comuns de seleção de possíveis relações funcionais a partir da definição de valores de corte para os coeficientes de correlação. De fato, este tipo de estratégia como método de identificação e construção de redes funcionais foi um dos primeiros a serem amplamente utilizados desde que foi introduzido por Butte e colaboradores [72] para criar redes de co-expressão. Além dele, outros autores tem utilizado essa mesma estratégia para definir relações funcionais [17, 23, 68, 69], mostrando resultados semelhantes aos aqui discutidos.

No entanto, além do valor de corte na correlação, muitas outras estratégias tem sido utilizadas para definir relações funcionais, como por exemplo estratégias de validação cruzada [73], o uso de correlações parciais [74], técnicas baseadas em posicionamento [75], ou técnicas mais complexas de aleatorização [76]. Sem dúvida, o continuo desenvolvimento de estratégias mostra que a definição de uma relação funcional continua sendo um problema a ser resolvido, onde talvez a comparação com relações funcionais vindas de técnicas de interação genética abra uma janela para explorar alternativas diferentes aos métodos baseados em modelos estatísticos.

Tabela 5.1: Número de pares de genes segundo diferentes valores de corte nas medidas de Pearson e Spearman para os três conjuntos de expressão

	Estado estacionário		Série temporal		Múltiplas condições	
	$> 0, 5 $	$> 0, 7 $	$> 0, 5 $	$> 0, 7 $	$> 0, 5 $	$> 0, 7 $
PEARSON	1.145.851	130.485	4.151.926	1.275.122	2.223.834	346.644
SPEARMAN	1.467.948	149.133	4.919.309	1.460.210	1.634.860	168.769

Finalmente, vendo que os resultados obtidos pelas medidas de Pearson e Spearman são altamente semelhantes para os três dados de expressão que foram aqui utilizados (Tabela 5.1), as análises comparativas serão feitas dando prioridade a aquela medida que identificou um maior número de relações funcionais segundo cada conjunto de dados de expressão. Assim, para os dados

de estado estacionário e série temporal será utilizada a medida de Spearman, enquanto que para os dados sob múltiplas condições será utilizada a medida de Pearson.

5.3 Processamento e descrição das interações genéticas

Considerando que dos dois bancos de dados escolhidos, um provém do site SGD (Saccharomyces Genome Database) e o outro do trabalho de Constanzo [13], que o primeiro é apresentado em forma qualitativa enquanto que o segundo é quantitativo, e que existe diferença no número total de interações que cada um possui; um primeiro passo antes de fazer as comparações entre elas e os coeficientes de correlação foi identificar o quanto do total de interações que cada banco possui são diferentes. Esse procedimento permitiu reconhecer que 46.8% das interações que compõem o banco de dados qualitativo também estão presentes no banco de dados quantitativo (Figura 5.7), fazendo assim desnecessário manter ditas interações redundante em ambos bancos de dados. Portanto, essas interações redundantes foram retiradas do banco qualitativo.

Deste modo, as interações genéticas restantes no banco de dados qualitativo foram utilizadas no processo de comparação com os coeficientes de correlação. Para tal, um grafo não dirigido foi gerado com ajuda do pacote *iGraph* [77] incluído no software R (<http://www.r-project.org>) onde os nós representam os genes e as arestas as interações genéticas existentes entre cada par de genes. A partir desse grafo foi criada uma matriz de adjacências $N \times N$ com os genes tanto nas colunas quanto nas linhas. Essa matriz é utilizada como representação das interações genéticas. Assim, uma posição i,j da matriz contém valor zero (0) quando não existe interação genética entre dois genes e um (1) quando a relação existe dentro do banco de dados.

Uma descrição dos dois bancos de dados é apresentada na tabela 5.2 e na figura 5.8. A partir delas é possível destacar as seguintes informações:

Comparado com o descrito em [78], onde os autores assumem que em média cada gene compartilha relações funcionais com outros dez genes, os resultados mostrados na tabela 5.2 indicam um número médio maior de relações por gene, o que se faz muito mais extremo nos resultados obtidos

Figura 5.7: Interações genéticas compartilhadas entre os dois bancos de dados.

para o banco de dados quantitativo, no qual o número médio de interações por gene é de 456,8.

Como pode ser observado tanto na tabela 5.2 quanto na figura 5.8, as estatísticas resumo para cada banco de dados resultaram muito diferentes. Portanto, para verificar se a remoção das interações genéticas compartilhadas, como explicado no começo desta seção, é a responsável pelas diferenças encontradas entre os dois bancos de dados, as mesmas medidas estatísticas foram calculadas no banco de dados qualitativo prévio ao processo de filtragem, obtendo como resultado valores semelhantes (média: 35 e mediana: 16); indicando assim que o fato de ter filtrado o banco qualitativo não influenciou de forma drástica os resultados descritos na tabela 5.2.

Não só as diferentes interações genéticas encontradas entre os dois bancos de dados como também a maior quantidade média de relações, podem ser o reflexo da forma como foram obtidas as interações genéticas de cada banco, pois no caso do quantitativo, todas as interações genéticas foram calculadas usando a definição “multiplicativa” de interação genética (Ver seção 2.2), entanto que no caso do banco qualitativo existem interações genéticas calculadas não só a partir da definição multiplicativa, mas também a partir de diferentes definições de interação genética.

Considerando isso, e vendo que o trabalho de Mani [35], mostrou que algumas dessas definições (como a multiplicativa) conseguem uma melhor quantificação do efeito fenotípico de uma dupla mutação, as diferenças entre os bancos de dados podem encontrar nessa condição uma possível explicação.

Tabela 5.2: Descrição das interações genéticas (IG) presentes nos dois bancos de dados utilizados.

	IG totais	IG media por gene	IG mediana por gene	Maior número de IG para um gene	Menor número de IG para um gene
Qualitativo	287.755	28	11	585	1
Quantitativo	763.854	456.8	430.5	3738	27

(a) BD. Qualitativo

(b) BD. Quantitativo

Figura 5.8: Distribuição do número de interações genéticas por gene para cada banco de dados.

Além das diferenças na definição de interação genética utilizada, a preferência na busca de interações genéticas de uns genes em relação a outros também pode explicar as diferenças encontradas entre ambos os bancos de dados. Estas preferências podem ser o reflexo da dificuldade para gerar experimentos de interação genética para alguns genes, pois segundo Many [35] se um gene tem um efeito fenotípico considerado forte, a identificação de interações genéticas para esse gene é mais difícil dado que muitas vezes a introdução de uma segunda mutação inviabiliza o crescimento da célula.

Assim, considerando que no trabalho de Constanzo [13] (de onde vem o banco de dados quantitativo) foi dada preferência à busca de interações genéticas pra genes chamados “nulos” pelo seu baixo efeito no fenótipo, a possibilidade de achar uma maior quantidade de interações genéticas é mais alto do que no banco de dados qualitativo no qual não há uma distinção segundo o tipo de efeito fenotípico.

Finalmente, da tabela 5.2 e dos histogramas da figura 5.8, também é importante destacar o fato que poucos genes apresentam uma grande quantidade de interações genéticas. Apesar que essa característica pode ser causada pelo interesse científico que de maneira particular recebem uns genes com respeito de outros, o certo é que se procuradas as funções biológicas onde participam os genes como maior número de interações genéticas é possível observar que eles fazem parte de importantes processos biológicos (Tabela 5.3), o que sem dúvida aumenta a possibilidade de cada um deles manter grande número de interações com outros genes.

Tabela 5.3: Genes mais conectados para cada banco de interação genética.

	Genes	Processo Biológico	No. Interações
BD. Qualitativo	YLR039C	Transporte intracelular de proteínas	585
	YDR432W	Síntese de RBP's	517
	YER083C	Subunidade do complexo GET	434
	YDL220C	Gene essencial. Regulação da telomerasa	430
	YMR186W	Chaperona da família Hsp82p	414
	YKR001C	Atividade hidrolítica (GTPase)	408
	YOL012C	Proteína histona	395
	YLR262C	Atividade hidrolítica (GTPase)	383
	YGR078C	Proteína “chaperona”	370
	YNL298W	Proteína kinase	356
BD. Quantitativo	YFL039C	Polarização celular	3738
	YBR060C	Replicação do DNA	2456
	YKL203C	Proteína kinase	2145
	YLR298C	Subunidade do splicing	1453
	YJL166W	Subunidade complexo COX	723
	YOL126C	Enzima catalítica	689
	YAL041W	Estabelecimento de polaridade	543
	YDR168W	Chaperona da família Hsp90	497
	YLR268W	Proteína R-Snare	406
	YEL002C	Subunidade complexo glico-proteico	321

Assim por exemplo, os dez genes com maior número de interações genéticas presentes no banco de dados qualitativo, fazem parte de processos como a regulação da telomerase (fundamental para a estabilidade dos cromossomos) e a síntese de proteínas de ligação ao RNA (RBP's que agem em diversos processos de regulação). Já no caso do banco de dados quantitativo, os dez genes que apresentam o maior número de interações genéticas também encontram-se envolvidos em processos importantes, tais como a replicação do DNA, a respiração celular e o complexo COX.

Até agora têm sido descritos os dois principais bancos de interação genética que são utilizados nos estudos de relações funcionais [22]. Apesar que os resultados aqui descritos mostram que os genes com maior número de interações genéticas podem estar envolvidos em importantes processos biológicos (Tabela 5.3), a interpretação biológica para módulos de interações genéticas é uma tarefa muito mais complexa do que quando utilizadas relações funcionais obtidas por co-expressão, pois apesar de que com as relações de co-expressão os problemas passam pela escolha de um valor de corte que permita definir uma relação funcional (como mostrado na seção 5.2), a modularidade das relações de co-expressão pode ser em muitos casos, biologicamente explicada. Este é o caso dos módulos de co-expressão identificados para processo como o ciclo celular [11], transporte de lipídios ou transporte de proteínas [79], só para citar alguns exemplos.

No entanto, esse não é o caso das relações funcionais obtidas por meio de interação genética, onde apesar de que a relação entre dois genes pode ser observada no fenótipo (no caso, o crescimento da célula), a formação de módulos de genes não tem uma interpretação biológica simples. Esse problema, que foi inicialmente descrito em [78], tem fomentado o desenvolvimento de estratégias de modelagem de redes que procuram dar uma interpretação biológica para uma rede de relações funcionais obtida por meio dos experimentos de interação genética [22].

Essa dificuldade, que se faz mais evidente na medida em que mais genes são conectados até formar pequenos módulos, como descrito em [78], poderia ser entendida por meio da comparação das relações funcionais que podem ser identificadas ao interior de processos biológicos tanto pelas interações genéticas como pela co-expressão.

5.3.1 Obtenção de outras medidas de interação genética

Como explicado no marco teórico (Seção 2.2), a interação genética pode ser calculada de diversas maneiras, dependendo da função de neutralidade que seja utilizada [35]. Portanto, para verificar se existem diferenças nas relações funcionais que podem ser obtidas quando usadas funções de neutralidade diferentes, foram calculadas as interações genéticas definidas pelas funções de neutralidade mínima, aditiva e logarítmica.

A figura 5.9, contém uma comparação das distribuições dos valores de interação genética do banco de dados quantitativo, as quais foram obtidas usando a definição multiplicativa [13], junto aos valores calculados para as definições mínima, aditiva e logarítmica. Dessa figura, é possível observar que a distribuição dos valores de interação genética que são obtidos usando as funções de neutralidade multiplicativa, aditiva e logarítmica, são muito semelhantes.

Figura 5.9: Comparação da distribuição dos valores de interação genética calculada a partir de quatro funções de neutralidade diferentes.

Resultados semelhantes foram obtidos no trabalho de Mani [35]. Nesse trabalho, o autor afirma que: “as definições multiplicativa e logarítmica oferecem resultados equivalentes quando usadas para calcular interação genética

para genes que, mutados de modo individual não apresentam efeitos fenotípicos graves. De fato, para a criação das interações genéticas do banco de dados quantitativo, foi dada preferência para genes definidos como “nulos”, dado os leves efeitos fenotípicos que apresentam na célula [13].

Já com respeito aos valores obtidos pela função de neutralidade mínima, é possível perceber valores com uma tendência positiva. Assim, o valor da mediana das interações genéticas “mínima” é de 0.44, enquanto que para as outras três definições esse valor encontra-se em torno de zero (0). Esses resultados, continuam sendo concordantes com o descrito por Mani [35], pois o autor afirma que “de entre todas as medidas, aquela definida pela função de neutralidade mínima é a que apresenta uma maior variação com respeito às outras medidas já que atribui valores positivos as interações genéticas”. Esses resultados, que foram obtidos para bancos de dados com aproximadamente 26.000 interações genéticas, são aqui confirmadas com um conjunto com mais de 600.000 interações genéticas.

A importância de uma medida gerar valores positivos para as interações genéticas em maior proporção do que as outras medidas, não se reflete na identificação de relações funcionais diferentes e sim, na interpretação biológica dessas relações funcionais. Isto deve-se ao fato das interações genéticas ter sido classificadas em duas categorias diferentes. Por um lado, estão as interações genéticas negativas, as quais têm sido chamadas como “agravantes”, o que significa que seu efeito no fenótipo é muito pior do que é esperado a partir da função de neutralidade. A hipótese biológica utilizada para explicar ditas interações assinala que as interações genéticas “agravantes” se apresentam quando os genes mutados ou deletados fazem parte de vias ou de processos diferentes, fazendo com que sua não funcionalidade seja muito mais grave para o fenótipo. Já no caso das interações genéticas positivas, que têm sido classificadas como “aliviantes” [22], dariam-se quando os genes mutados fazem parte da mesma via ou processo biológico, fazendo com que seu efeito fenotípico seja melhor que aquele predito pela função de neutralidade.

Se fossem comparadas as interpretações biológicas que são dadas para as relações de co-expressão, com aquelas que são dadas para as interações genéticas, pode ser desenhado um paralelo entre elas. Por exemplo, vendo que as interações genéticas positivas, aparecem com maior frequência dentro de processos biológicos e que no caso da co-expressão positiva ela é entendida como a co-ativação de dois genes que comumente fazem parte do mesmo processo biológico, como sucede por exemplo no caso de processos de pro-

liferação celular [80], ou de ligação de proteínas [81], é razoável pensar que possa existir uma relação na interpretação do signo para a co-expressão e a interação genética.

No caso dos valores de interação genética negativa, sua interpretação biológica também pode encontrar uma relação com a interpretação dada para valores de correlação negativa. O que acontece neste último caso é que a atividade de um gene aumenta enquanto a do outro decresce indicando assim a supressão de alguma função. Este tipo de comportamento já foi observado na transição de estados durante regulações de processos metabólicos [4], onde processos diferentes precisam ser ativados e inibidos de forma sincronizada. Essas características de regulação entre processos também poderiam ser o reflexo de uma interação genética negativa.

Porém, estas não são uma generalidade, nem no caso da co-expressão nem no caso das interações genéticas, o que faz da interpretação de uma relação funcional um processo difícil de generalizar.

5.4 Comparando co-expressão com interações genéticas

O uso de valores de corte aplicados aos coeficientes de correlação como estratégia para definir relações funcionais que foi discutido na seção 5.2, é comumente conhecido como identificação por limites rígidos [32] (em contraposição aos limites chamados “fracos”, nos quais o principal expoente é o método de construção de redes WGCNA [82]), os quais são basicamente critérios de seleção de base estatística. No entanto, esses métodos nem sempre refletem completamente a conectividade dos dados e muitas vezes são carentes de sentido biológico [18]. Além disso, a eleição de valores de corte mais ou menos restritivos para escolher pares de genes é feita em alguns casos de forma arbitrária [68].

Por outro lado, as relações funcionais identificadas por meio de interações genéticas apresentam dificuldades para serem interpretadas biologicamente quando são utilizados em análises sistêmicas. Em [12] essa dificuldade é explicada assumido que, a diferença das relações de co-expressão, muitas das interações genéticas são produto da interação de genes que fazem parte de vias biológicas diferentes. Mas em contraste, Costanzo e colaboradores [13],

têm descrito interações genéticas para genes essenciais que fazem parte das mesmas vias ou processos biológicos.

Portanto, a comparação dos coeficientes de correlação obtidos por meio de medidas de dependência estatística com os resultados dos experimentos de interações genéticas pode ser útil não só para comparar duas técnicas que compartilham o objetivo de identificar relações funcionais, mas também porque as informações obtidas como produto dessa comparação, podem oferecer critérios diferentes aos comumente utilizados para auxiliar na escolha de uma faixa ou de um valor limite que defina relações funcionais a partir de co-expressão. Além disso, as relações de co-expressão também podem ser utilizadas como uma ferramenta de auxílio na interpretação biológica de grupos de relações funcionais que são obtidas por meio de interação genética.

Lembrando que a eleição de um valor de corte para definir relações funcionais usando coeficientes de correlação não é uma metodologia definida e que os valores de $|0.5|$ e $|0.7|$ foram aqui utilizados simplesmente como exemplo do processo de determinação de relações funcionais, as comparações que serão descritas a continuação utilizam todo o intervalo de possíveis resultados para os coeficientes de correlação $[-1, 1]$ e não se limitam em aqueles que podem estar acima de um determinado limiar.

Para facilitar a leitura do documento e lembrando que, como foi descrito no final da seção 5.2.2, os resultados obtidos por ambas medidas de dependência são semelhantes, o ênfase das análise de comparação será dada para a medida de Spearman. Porém, os resultados completos estarão disponíveis num capítulo de anexos (Seção 6).

5.4.1 Comparação da co-expressão com as IG qualitativas

Para verificar se as relações que podem ser obtidas através da co-expressão podem também ser obtidas pelas interações genéticas, foi avaliada a força dos coeficientes de correlação daqueles pares de genes que dentro da matriz de adjacência (ver seção 5.3) têm valor de um (1). Isto é, para aqueles pares de genes que foram identificados por manter uma relação funcional a partir dos experimentos de interação genética. Assim, pretende-se definir se os pares de genes que apresentam interação genética, também apresentam coeficientes de correlação fortes.

Antes do processo, as matrizes de correlação e adjacência foram filtradas e ordenadas para deixar unicamente os genes que pertencem a ambos conjuntos de dados, dando como resultado matrizes quadradas de tamanho 5362. A figura 5.10, descreve as distribuições dos coeficientes de correlação daqueles pares de genes que apresentam interação genética.

Observando os três gráficos de caixa da figura 5.10, é possível perceber que em média, os valores de correlação dos pares de genes que têm sido identificados por apresentar interação genética é próxima de zero. Assim, para os dados de expressão em estado estacionário a média é 0.04, no caso dos dados em série temporal é 0.025 e no caso dos dados em múltiplas condições é 0.039.

Figura 5.10: Distribuição dos coeficientes de correlação nos três conjuntos de expressão dos pares de genes que têm interação genética.

Esse resultado mostra que não só, se forem usados valores fortes de correlação é possível identificar um par de genes com interação genética, mas que em geral para todo o intervalo de coeficientes de correlação não existe nenhum tipo de tendência que possa sugerir que um par de genes com relação funcional identificada por co-expressão, possa também ser identificada por interação genética. De fato, se fosse assumido um valor de corte para a correlação de $[0, 7]$, na tabela 5.4, é possível observar que a quantidade de pares de genes com interação genética e coeficiente de correlação $> |0.7|$ é menor que a dos pares de genes que usando esse mesmo valor de correlação, não apresentam interação genética.

Tabela 5.4: Comparação das relações funcionais que podem ser identificadas só pelos coeficientes de correlação e pelos coeficientes de correlação junto à interação genética.

	$ \rho > 0.7$ e IG presente	$ \rho > 0.7$ e IG ausente
Estado estacionário	831	87.548
Série temporal	11.241	1.078.643
Múltiplas condições	1.823	127.686

Quando comparados entre se os resultados dos três conjuntos de expressão, é possível notar que para aqueles obtidos em série temporal, do total de pares de genes com coeficiente de correlação acima de $|0.7|$, 7,4% deles também apresenta interação genética, entanto que para os outros dois conjuntos essas percentagens não superam o 1,5% (Tabela 5.4). Essas diferenças entre os dados de expressão em série temporal e os outros dois conjuntos, não necessariamente indicam uma maior relação entre as interações genéticas e a co-expressão de genes em série temporal uma vez que como já foi mostrado na seção 5.2.2, a partir de dados de expressão em série temporal são identificados um maior número de pares de genes com correlação acima de $|0,7|$ dos que são identificados pelos outros dois conjuntos de dados, aumentado assim a chance desses pares de genes possuírem também interação genética. De fato, a quantidade de pares de genes que para os dados em série temporal não apresentam interação genética, mas que têm coeficientes de correlação $> |0,7|$, é muito maior que nos outros dois conjuntos de dados. Resultados completos para as medidas de Pearson e Spearman são mostrados no anexo (A).

Os resultados descritos até agora, revelam uma baixa conexão entre as relações funcionais que podem ser identificadas pela co-expressão e pelas interações genéticas. Isso significa que ainda quando a co-expressão e as interações genéticas têm como objetivo identificar relações funcionais entre genes, o número dessas relações que podem ser identificadas de forma conjunta é muito baixa. De fato, a validação das relações de co-expressão é uma dificuldade que ainda não tem uma resposta definitiva, pois embora seja comum considerar valores fracos de correlação como sem significância biológica [83], a escolha do valor de corte que permita fazer essa afirmação não esta ainda estabelecido.

Entre as estratégias utilizadas para validar relações de co-expressão, a comparação com redes de PPI tem sido amplamente utilizada [84]; e embora os resultados não sejam muito bons com respeito à quantidade de relações funcionais que consegue validar (como também é mostrado aqui para interações genéticas), esse método tem descoberto algumas relações funcionais comuns, principalmente entre genes que fazem parte dos mesmos complexos proteicos como no caso de algumas proteínas em complexos ribossomais [85]. Levando em conta isso, os resultados mostrados na tabela 5.4, mostram que, embora sejam poucos, existe um grupo de relações funcionais que podem ser alvo de validação biológica em estudos posteriores.

Até aqui, tem sido avaliados os coeficientes de correlação utilizados nas interações genéticas presentes no banco de dados qualitativo. Porém, no decorrer da pesquisa, foram obtidos os valores medidos do “fitness” que foram utilizados para definir quais genes apresentam interação genética no estudo de Constanzo e colaboradores [13]. Esse conjunto de informações, que tem sido aqui chamadas com banco de dados quantitativo foram aproveitadas para avaliar se existe alguma tendência presente entre os valores de correlação que podem definir uma relação funcional e os valores de interação genética, utilizados com o mesmo objetivo. Essa avaliação pode ajudar a estabelecer se existe alguma faixa de coeficientes dentro do intervalo de correlação (não necessariamente valores fortes) que tenha uma maior relação com os valores do “fitness” que definem uma interação genética. A descrição desses resultados são apresentados a seguir.

5.4.2 Comparação da co-expressão com as IG quantitativas

As comparações feitas na seção anterior, tinham o intuito de identificar se as relações funcionais identificadas por meio de interação genética também possuíam coeficientes de correlação forte para algum dos três conjunto de expressão utilizados. Já nesta seção, aproveitando que o banco de dados quantitativo oferece informação não só ao respeito de quais pares de genes possuem interação, mas também de quais não possuem, as comparações que são aqui apresentadas, mostram a distribuição dos coeficientes de correlação para esses dois grupos. O objetivo agora não é simplesmente identificar se valores fortes de correlação apresentam interação genética e sim tentar identificar algum tipo de tendência diferente entre aqueles pares de genes com

interação genética em referência daqueles que não têm.

Como resultado dessa comparação, os histogramas mostrados na figura 5.11, indicam que não existe diferença na distribuição dos coeficientes de correlação no grupo de genes com interação genética com respeito ao grupo que não têm, confirmando assim os resultados obtidos a partir do banco de quantitativo de interações genéticas.

Embora a partir dos dados aqui utilizados a escassa relação seja evidente, vale a pena quantificar o poder de predição que pode-se obter se forem utilizadas as interações genéticas como critério de classificação dos valores de co-expressão. Em outras palavras, o que se pretende verificar é se na medida em que aumenta o coeficiente de correlação é possível aumentar o número de interações genéticas. Com esse objetivo foram criadas as curvas ROC (“*Receiver Operating Characteristic curve*”) mostradas na figura 5.13.

Figura 5.11: Distribuição dos coeficientes de correlação nos três conjuntos de expressão comparando os pares de genes com e sem interação genética para o banco de dados quantitativo.

De maneira breve, as curvas ROC são uma técnica gráfica amplamente utilizada para visualizar, organizar e selecionar classificadores binários baseados em seu desempenho [86]. A curva ROC, é desenhada como resultado da comparação da taxa de verdadeiros positivos (TPR) contra a taxa de falsos positivos (FPR) que são obtidos após aplicação de um teste binário. No caso particular da presente pesquisa, o teste binário foi definido como a presença ou ausência de interações genéticas que são identificadas na medida em que é aumentado o valor de corte nos coeficientes absolutos da correlação, começando em zero (0) e aumentando de forma gradual em intervalos de 0.1

até chegar num valor de correlação um (1).

O teste pode oferecer quatro resultados possíveis: Verdadeiro positivo, Falso positivo, Verdadeiro negativo e Falso negativo. Por exemplo se o teste esta sendo aplicado para um valor de correlação de 0.2 um resultado é considerado verdadeiro positivo (VP), se o par de genes que apresenta esse valor de correlação também apresenta interação genética. Um resultado falso positivo (FP), refere-se ao par de genes que com esse valor de correlação não apresenta interação genética. Já um resultado verdadeiro negativo (TN), aparece quando o par de genes não tem esse valor de correlação nem interação genética, e finalmente um resultado falso negativo (FN) aparece quando o par de genes não apresenta esse valor de correlação, mas apresenta interação genética.

O que se espera de um bom classificador é que a taxa de verdadeiros positivos (TPR) aumente, e que a taxa de falsos positivos diminua (FPR), conforme vão sendo avaliados coeficientes de correlação mais fortes até chegar num valor de correlação de um (1). Desse modo, ao serem representados num gráfico, a taxa de verdadeiros positivos contra a de falsos positivos espera-se que exibam algum tipo de comportamento similar aos descritos na figura de exemplo 5.12.

Figura 5.12: Possíveis resultados obtidos numa curva ROC: (— — —): O teste de classificação produz respostas aleatórias. **A**: O teste é um classificador perfeito. **B**: O teste não funciona como classificador.

Considerando o descrito anteriormente e observando as curvas ROC mostradas na figura 5.13, é razoável concluir que a comparação dos dados de expressão aqui utilizados com as interações genéticas do banco de dados quantitativo, não é possível saber se existe algum valor de correlação que possa ser associado a relações funcionais identificadas por meio de interação

genética. Isso confirma que não só, poucas relações funcionais podem ser identificadas de modo conjunto entre as medidas de dependência e as interações genéticas, mas também, que a identificação de uma relação funcional por meio de interação genética não garante que esse mesmo par de genes tenha um coeficiente de correlação forte. No anexo B, são mostradas as curvas ROC para todas as comparações feitas com a medida de Pearson e Spearman, os resultados ali mostrados são semelhantes aos aqui descritos.

Além das comparações anteriores, aproveitando os valores medidos das interações genéticas do banco quantitativo, foi quantificada a relação entre os coeficientes de correlação e os valores das interações genéticas, usando regressões simples que foram calculadas a partir dos valores absolutos das interações genéticas e dos coeficientes de correlação. Os gráficos de dispersão mostrados nas figuras 5.14a, 5.14b e 5.14c para os dados de expressão em estado estacionário, série temporal e múltiplas condições respectivamente, mostram que a relação entre os valores que definem uma interação genética e os coeficientes de correlação que definem uma co-expressão é muito baixa. Os resultados indicam que as relações de co-expressão que podem ser explicadas pelas interações genéticas são praticamente nulas.

A tabela 5.5, mostra um resumo dos resultados do cálculo da regressão. Para todos os casos, tanto os coeficientes de determinação r^2 , quanto os valores da pendente β são muito baixos, confirmando assim que a variação de ambos conjuntos de dados é altamente independente uma da outra.

Figura 5.13: Curvas ROC para os coeficientes de correlação nos três conjuntos de dados de expressão. TPR: Taxa de verdadeiros positivos. FPR: Taxa de falsos positivos.

Figura 5.14: Gráfico de dispersão dos valores de interação genética contra os valores de correlação para todos os dados de expressão.

Tabela 5.5: Coeficientes de determinação da regressão.

Dados de expressão	Correlação		
	r^2	β	p-valor
Estado estacionário	$1.7 * 10^{-5}$	0.011	0.0003
Série temporal	$3.2 * 10^{-5}$	-0.021	$5.1 * 10^{-7}$
Múltiplas condições	0.00078	0.083	$5.6 * 10^{-132}$

Embora a importância de utilizar os níveis de expressão do mRNA como medida da atividade dos genes tem sido amplamente aceita [17], a relação que a quantidade de mRNA medido possa chegar a ter com outros níveis de abstração como os descritos na figura 1.1, é ainda uma tarefa difícil de interpretar. Os resultados até agora descritos dão conta dessas dificuldades, mas não são o único exemplo, já que em [29] e [30] também são descritas relações globais fracas entre os níveis de expressão e os de proteínas em levedura, e resultados semelhantes também foram descritos para a bactéria *Desulfovibrio vulgaris* em [87], mostrando que o baixo nível de relação não se encontra restrito unicamente para *Saccharomyces cerevisiae*.

No entanto, quando não se trata das relações entre os níveis absolutos de mRNA (ou relativos dependendo de como sejam medidos) e os de proteínas, e sim da conexão entre as relações funcionais identificadas por co-expressão e por interações de proteínas ou interações genéticas, as referências embora escassas, também dão conta de uma baixa correspondência. Assim, no trabalho de Grigoriev [84] mostram-se variações menores do que $|0.2|$ nos coeficientes

de correlação de Pearson entre os pares de genes que têm uma interação de proteínas contra aqueles que não têm, em forma semelhante aos resultados aqui descritos na figura 5.11. Comparações desse tipo não têm sido feitas ainda entre co-expressão e interações genéticas e a referência mais próxima pode ser o trabalho de Colak e colaboradores [78] no qual, coeficientes de correlação fracos são utilizados junto a interações genéticas e de proteínas para identificar módulos de genes densamente conectados.

5.4.3 Comparando a co-expressão com IG mínima e com o duplo mutante

Apesar que até agora tem-se mostrado difícil que a co-expressão e as interações genéticas identifiquem as mesmas relações funcionais, mais duas estratégias foram utilizadas para confirmar os resultados até agora obtidos. A primeira dessas estratégias consistiu em comparar os valores de co-expressão com as interações genéticas definidas pela função de neutralidade mínima, as quais tinham sido previamente calculadas (Ver seção 5.3.1). A escolha particular desse tipo de interação genética, deve-se ao fato dela apresentar a maior variação nos valores das interações genéticas (Figura 5.9) dentre todas as medidas que foram calculadas: aditiva, logarítmica, multiplicativa e mínima.

Já a segunda estratégia consistiu em comparar os coeficientes de correlação não com os valores da interação genética, e sim com os valores do duplo mutante os quais representam o valor medido do “*fitness*” (crescimento) para quando dois genes são mutados simultaneamente numa cultura celular. Com isso, espera-se verificar se os valores medidos do duplo mutante, antes de serem usados para calcular as interações genéticas, têm uma maior relação com as relações de co-expressão. Além disso, o aumento na quantidade de pares de genes que podem ser comparados com os valores de correlação (pois os duplo mutante representam todas as mutações testadas e não só aquelas que foram consideradas interações genéticas), pode oferecer resultados diferentes aos que têm sido descritos até agora.

A figura 5.15, mostra o resultado do cálculo da regressão, e a comparação das distribuições da correlação dos pares de genes quando têm interação genética mínima e quando não têm, para os dados de expressão de múltiplas condições. Por sua vez, a figura 5.16 mostra o mesmo tipo de resultado mas desta vez usando os valores do duplo mutante. Resultados semelhantes

foram obtidos para os dados de expressão em série temporal e em estado estacionário com ambas medidas de dependência; esses resultados são mostrados no anexo C.

Assim como acontece com as interações genéticas definidas por meio da função de neutralidade multiplicativa, e considerando também que não existem diferenças importantes entre esta e as outras três funções de neutralidade aqui calculadas, é possível afirmar que os coeficientes de correlação não apresentam variação quando os pares de genes têm interação genética e quando não têm. Resultado que se repete de maneira invariável ao tipo de interação genética calculada.

Figura 5.15: Comparação dos valores de correlação de Pearson com a interação genética mínima. Dados de expressão em múltiplas condições.

Como assinalado na seção anterior, o banco de dados quantitativo de interações genéticas também inclui os valores medidos para todos os duplos mutantes que foram testados (quase seis milhões de pares de genes). Esse valor de “*fitness*” para o duplo-mutante foi comparado com os coeficientes de correlação vindos a partir dos dados de expressão. Aqui, vale a pena lembrar que as interações genéticas são calculadas como a diferença entre o “*fitness*” do duplo-mutante e a função de neutralidade, independentemente de qual seja escolhida (Ver secção 2.2). Isso significa que a comparação não será feita com o valor excedente, mas com o efeito medido da dupla mutação sem que seja assumida nenhuma função que defina se tal efeito deve ou não, ser considerado como uma interação genética. A modo de analogia, é como se para os coeficientes de correlação obtidos a partir de medidas de dependência, não

fosse aplicado nenhum valor de corte.

Figura 5.16: Comparação dos valores de correlação de Pearson com os valores dos duplos mutantes. Expressão sob múltiplas condições.

Na figura 5.16 mostra-se o resultado dessa comparação para os coeficientes de correlação de Pearson obtidos a partir dos dados de expressão sob múltiplas condições. Resultados semelhantes foram obtidos para os outros dois conjuntos de dados de expressão, incluindo também aqueles onde a medida de dependência de Spearman é utilizada. O resultado, coincidente com o que até agora tem sido mostrado, indica que o coeficiente de correlação não apresenta diferença entre os pares de genes que têm valor medido de duplo-mutante e aqueles que não o apresentam.

Considerando-se que aqui não é calculada nenhum tipo de interação genética e que, portanto aumenta a quantidade de pares de genes para os que podem ser verificados seus respectivos coeficientes de correlação, o resultado mostrado na figura 5.16 indica que o efeito fenotípico que produz a mutação de dois genes ao mesmo tempo não pode ser relacionado com os coeficientes de correlação que podem definir uma co-expressão no mesmo par de genes.

Com os resultados até agora descritos, se faz evidente que quando comparadas em forma geral, as interações genéticas e as relações de co-expressão não conseguem identificar as mesmas relações funcionais. Para avaliar se esse comportamento se mantém ao nível de processos biológicos específicos, os mesmos procedimentos de comparação até aqui utilizados, foram também aplicados sob diferentes processos biológicos.

5.4.4 Comparação segundo processos biológicos

Até agora, as informações apresentadas indicam que de modo global não existe relação entre a co-expressão e as interações genéticas, porém é possível que exista algum processo biológico no qual a relação entre co-expressão e interação genética seja maior à mostrada até agora, mas que não possa ser observado durante a análise global. Portanto, após comparar de modo geral todos os coeficientes de correlação usando suas respectivas interações genéticas, uma comparação para diferentes processos biológicos foi feita. A seleção dos processos que foram utilizados durante a comparação pode ser descrita nos seguintes passos:

1. Aproveitando o fato que o banco de dados do *site* Gene Ontology (geneontology.org) oferece listados de termos específicos para organismos modelo, uma busca de todos os termos inclusos na categoria de “Processos Biológicos” anotados de maneira exclusiva para *Saccharomyces cerevisiae* foi feita. Como resultado, 63 termos foram obtidos.
2. Para cada termo obtido no passo anterior, foram identificados todos os genes a ele associados a partir das informações contidas na base de dados “Saccharomyces Genome Database” (yeastgenome.org). Como resultado, para cada processo foi criada uma lista que contém os nomes dos genes que pertencem a um determinado processo.
3. Finalmente, para cada uma dessas listas, foram procuradas suas interações genéticas usando os bancos de dados qualitativo e quantitativo (Seção 5.3). Essa busca foi feita de duas maneiras, sendo que no primeiro caso só foram procuradas as interações genéticas dos genes que fazem parte do próprio processo biológico. Esse grupo de interações foi chamado aqui como “*intra* processo”. Já a segunda busca consistiu em identificar as interações genéticas não só dos genes dentro de cada processo, mas de todas as interações que um gene possa ter, dentro ou fora do processo ao qual ele pertence. Esse grupo foi chamado aqui de interações genéticas “*entre* processos”. Dos 63 termos inicialmente obtidos, foram encontradas interações genéticas para 55 deles. A tabela 5.6 contém o nome desses 55 processos biológicos.

Tabela 5.6: Lista de todos os processos biológicos utilizados para comparações específicas.

1. Transporte de aminoácidos	12.Organização do citoesqueleto	23. Transporte de Ions	34. Glicosilação de proteínas	45. Biogênese do ribossomo
2. Transporte de carboidratos	13. Recombinação do DNA	24. Fusão de membrana	35. Maturação de proteínas	46. Exportação dos ribossomos
3. Gemulação	14. Reparo do DNA	25. Tradução na mitocôndria	36. Fosforilação de proteína	47. Montagem do ribossomo
4.Morfogênese celular	15. Replicação do DNA	26.Organização da mitocôndria	37.Proteólise	48. Catabolismo do RNA
5. Homeostase Ions celulares	16. Elongação da transcrição	27. Ciclo celular mitótico	38.Crescimento pseudo-hifas	49. Splicing RNA
6.Respiração celular	17. Terminação da transcrição	28. Processamento do mRNA	39. Regulação do ciclo celular	50. Processamento do rRNA
7. Resposta a danos	18. Endocitose	29. Organização nuclear	40. Regulação da tradução	51.Transcrição RNAPol I
8. Organização da cromatina	19. Transporte no endossoma	30. Organização do peroxissoma	41. Resposta ao calor	52.Transcrição RNAPol II
9. Segregação dos cromossomos	20. Exocitose	31. Defosforilação de proteína	42. Stress osmótico	53.Transcrição RNAPol III
10. Tradução no citoplasma	21. Limitação de glucose	32. Enovelamento de proteína	43. Biogênese do ribossomo	54.Elongação da tradução
11.Início da tradução	22. Transporte através membrana	33. Processamento tRNA	44. Organização da vacuola	55. Organização da vesícula

A comparação para os processos biológicos segue a mesma metodologia que foi utilizada nas anteriores comparações, é dizer, métodos gráficos que mostram a distribuição dos coeficientes de correlação dos pares de genes que têm e que não têm interação genética. Além disso, regressões simples calculadas para cada processo utilizando os valores absolutos das interações genéticas e dos coeficientes de correlação. A seguir, são descritos esses resultados.

IG e processos biológicos I

Os resultados mostrados nos gráficos de caixa da figura 5.17, descrevem o comportamento geral dos coeficientes de correlação para todas as sub-matrizes que representam as relações “*intra* processos”. Nesse gráfico, mostram-se várias características importantes. A primeira dessas características indica que não existe diferença na distribuição dos coeficientes de correlação quando comparados os pares de genes que dentro do processo possuem interação genética contra aqueles que não possuem. Este resultado confirma aquilo que foi descrito na seção 5.4.2, onde foram comparados de modo global as interações genéticas e os coeficientes de correlação, mostrando assim, que ainda para diferentes processos biológicos específicos, não existe relação entre a co-expressão e as interações genéticas.

(a) Segregação dos cromossomos. Expressão em estado estacionário

(b) Tradução no citoplasma. Expressão em estado estacionário

(c) Reparação do DNA. Expressão em estado estacionário

(d) Endocitosis. Expressão em série temporal

(e) Exocitosis. Expressão em série temporal

(f) Organização do mitocôndrio. Expressão em série temporal

(g) Defosforilação de proteínas. Expressão em múltiplas condições

(h) Síntese de proteínas. Expressão em múltiplas condições

(i) Fosforilação de proteínas. Expressão em múltiplas condições

Figura 5.17: Distribuição dos coeficientes de correlação dos pares de genes com e sem interação genética para diferentes processos biológicos nos três conjuntos de expressão.

Uma segunda característica que pode ser destacada é o fato de que dentro dos processos biológicos, são identificadas poucas interações genéticas,

como se indica na lenda de cada gráfico de caixa. Este resultado pode oferecer uma explicação ao problema discutido na seção 5.3 e referenciado nos trabalhos de Colak e colaboradores [78] e de Bellay e colaboradores [22] no que diz respeito à dificuldade de interpretar em termos de modularidade as relações funcionais obtidas por meio de interação genética. Possivelmente esse fato indique que os experimentos de interação genética não são uma boa técnica para identificar relações funcionais dentro de processos biológicos, na forma como eles têm sido tradicionalmente definidos. Ao respeito, Bellay [22] afirma que “cada interação genética oferece uma fotografia de um estado celular que dificilmente se repete dentro das condições que geram uma outra interação genética”. Assim, a diferença das relações funcionais obtidas por co-expressão onde todas essas relações representam um instante de tempo, nas interações genéticas cada relação funcional é um instante diferente.

Como descrito anteriormente, os resultados mostrados na figura 5.17, exemplificam o comportamento geral para todas as comparações. Porém, alguns processos mostram resultados que embora não sejam diferentes aos já descritos, se apresentam algumas alterações que vale a pena comentar. Esses processos são mostrados na figura 5.18.

O que se pretende destacar nesses gráficos, é o fato de todos esses processos apresentar poucos genes e portanto poucas relações funcionais possíveis. Assim, apesar de que sejam observadas diferenças entre os pares de genes com interação genética e os que não têm para alguns processos como a respiração celular ou a organização da cromatina, a pouca quantidade de pares de genes em ambos conjuntos impede que possa ser feita qualquer hipótese biológica ao respeito. Quando comparados entre se, os resultados obtidos para os três conjuntos de expressão, não foram encontradas diferenças significativas entre os diferentes processos. Todos os resultados são mostrados no anexo D.

(a) Morfogénese. Expressão em estado estacionário

(b) Respiração celular. Expressão em estado estacionário

(c) Organização da cromatina. Expressão em estado estacionário

(d) Tradução na mitocôndria. Expressão em série temporal

(e) Ciclo celular mitótico. Expressão em série temporal

(f) Organização do peroxisoma. Expressão em série temporal

(g) Regulação ciclo celular. Expressão em múltiplas condições

(h) Iniciação da tradução. Expressão em múltiplas condições

(i) Organização de vesículas. Expressão em múltiplas condições

Figura 5.18: Distribuição dos coeficientes de correlação dos pares de genes com e sem interação genética em processos biológicos com poucos genes para os três conjuntos de expressão.

IG e processos biológicos II

Na seção anterior foram descritos os resultados das comparações para as interações genéticas obtidas só para os genes de cada processo (“*intra* processo”). Agora, são descritos os resultados obtidos para os grupos de interações genéticas que foram chamadas como “*entre* processos”. Os gráficos de dispersão mostrados na figura 5.19 e nas tabelas 5.7 e 5.8, mostram uma baixa relação entre os valores de interação genética e os coeficientes de correlação.

Segundo Colak e colaboradores [78], as redes de co-expressão conseguem refletir uma modularidade funcional quando usados diferentes termos do GO (Gene Ontology), no entanto elas não conseguem identificar relações funcionais que dificilmente se manifestam no nível do mRNA. Por sua parte, as interações genéticas conseguem uma maior cobertura de relações funcionais independentes de tecidos ou condições de crescimento, porém, não refletem modularidade em termos do GO dado que existe um alto grau de incerteza como para afirmar que duas interações genéticas estarão ativas ao mesmo tempo. Portanto uma aproximação combinada como a que foi aqui feita, no sentido de utilizar os termos de modularidade definidos pelo GO, mas procurando ampliar o espectro de possíveis relações identificando não só as interações genéticas que fazem parte do termo específico e sim todas aquelas descritas para os genes do modulo, pode melhorar o balanço entre cobertura e modularidade.

No entanto, os resultados descritos na figura 5.19 e nas tabelas 5.7 e 5.8 mostram que se mantém a tendência dos resultados obtidos para as comparações globais (seção 5.4.2) no sentido que, tanto os coeficientes de determinação r^2 quanto os β para todas as regressões são muito baixos, confirmando assim que para os dados de expressão aqui utilizados não é possível definir uma relação funcional a partir daquelas que são definidas por interação genética. Os gráficos de dispersão para todos os processos biológicos encontram-se no anexo E.

(a) Resposta a danos no DNA. Expressão em estado estacionário

(b) Recombinação do DNA. Expressão em estado estacionário

(c) Replicação do DNA. Expressão em estado estacionário

(d) Maturação de proteínas. Expressão em série temporal

(e) Resposta ao calor. Expressão em série temporal

(f) Resposta stress osmótico. Expressão em série temporal

(g) Morfogênese celular. Expressão em Múltiplas condições

(h) Processamento do rRNA. Expressão em múltiplas condições

(i) Elongação na tradução. Expressão em múltiplas condições

(j) Morfogênese celular. Expressão em Estado estacionário

(k) Processamento do rRNA. Expressão em série temporal

(l) Iniciação da tradução. Expressão em múltiplas condições

Figura 5.19: Gráficos de dispersão entre coeficientes de correlação e valores de interação genética para diferentes processos biológicos.

Tabela 5.7: Resultados da regressão para diferentes processos biológicos I (EE: Estado estacionário, ST: Série temporal, MC: Múltiplas condições)

PROCESSO	EE			ST			MC		
	r^2	β	p	r^2	β	p	r^2	β	p
Transporte de aminoácidos	$2 * 10^{-5}$	-0.014	0.81	0.003	0.21	0.003	0.0001	-0.027	0.57
Organização do citoesqueleto	0.004	0.15	$7 * 10^{-15}$	0.0093	0.32	$7 * 10^{-7}$	0.0003	-0.04	0.34
Transporte de Ions	0.005	0.34	0.4	0.001	-0.26	0.61	0.0004	-0.15	0.80
Glicosilação de proteínas	0.0009	-0.05	0.02	0.007	0.25	$1.7 * 10^{-7}$	0.004	0.139	$2 * 10^{-11}$
Biogenesis do ribossomo	0.0002	-0.06	0.12	0.0008	0.14	0.007	0.01	0.47	$1 * 10^{-10}$
Recombinação do DNA	$5 * 10^{-7}$	0.001	0.95	$2 * 10^{-6}$	-0.004	0.88	0.002	0.12	$2 * 10^{-5}$
Fusão de membrana	0.0004	0.04	0.43	0.0001	0.06	0.6	0.0014	0.16	0.14
Maturação de proteínas	0.001	0.089	0.05	0.003	-0.19	0.005	0.01	0.24	$4 * 10^{-7}$
Exportação dos ribossomos	$9 * 10^{-6}$	-0.009	0.86	0.0001	-0.05	0.5	0.01	0.47	$1 * 10^{-15}$
Gemulação	0.0002	0.03	0.39	0.0003	0.06	0.24	0.004	-0.17	0.0005
Reparo do DNA	$4 * 10^{-5}$	-0.015	0.32	$9 * 10^{-5}$	-0.03	0.17	0.001	0.001	$6 * 10^{-10}$
Tradução na mitocôndria	0.0002	0.04	0.16	0.02	-0.56	$1 * 10^{-15}$	0.0003	0.05	0.13
Fosforilação de proteína	0.0003	0.04	0.0003	0.0004	0.07	$4 * 10^{-5}$	$1 * 10^{-7}$	-0.01	0.4
Montagem do ribossomo	0.003	0.17	0.12	0.0001	0.03	0.76	0.004	0.24	0.07
Morfogênese celular	$5 * 10^{-5}$	0.02	0.71	0.0007	-0.09	0.18	0.0046	-0.19	0.0011
Replicação do DNA	0.0004	-0.045	0.06	0.002	-0.13	$3 * 10^{-5}$	$1 * 10^{-5}$	-0.007	0.76
Organização da mitocôndria	0.0002	0.038	0.09	$6 * 10^{-5}$	-0.02	0.37	0.0002	-0.04	0.12
Proteólise	0.015	0.35	0.0005	$3 * 10^{-5}$	-0.02	0.87	0.0013	0.13	0.3
Catabolismo do RNA	$2 * 10^{-5}$	0.019	0.81	0.007	0.42	0.0001	0.001	-0.13	0.13
Homeostase	0.001	0.11	0.006	0.004	0.25	$4 * 10^{-6}$	$2 * 10^{-7}$	0.001	0.97
Processamento do rRNA	0.0001	-0.02	0.28	0.0009	-0.12	0.001	0.0002	-0.05	0.08
Gemulação	0.0016	0.0015	8.10^{-11}	0.0003	0.06	0.24	0.004	-0.17	$5 * 10^{-5}$
Organização nuclear	0.003	0.08	0.13	0.001	-0.07	0.33	0.03	0.41	$9 * 10^{-7}$
Regulação da tradução	0.0015	0.12	0.13	0.007	0.49	0.0006	$6 * 10^{-6}$	0.007	0.91
Transcrição RNAPol I	0.00042	-0.05	0.27	0.002	0.17	0.009	0.0001	-0.03	0.54
Organização da cromatina	0.0007	0.06	0.08	0.14	1.03	$9 * 10^{-15}$	0.0006	0.075	0.1
Transporte no endossoma	0.001	-0.13	0.2	0.0019	0.17	0.16	0.0003	0.08	0.5
Organização do peroxissoma	$7.5 * 10^{-5}$	-0.026	0.55	0.004	-0.25	$8 * 10^{-6}$	0.0007	0.09	0.05
Resposta ao calor	0.002	-0.11	$5 * 10^{-5}$	0.001	0.11	0.005	0.0007	-0.07	0.02
Transcrição RNAPol II	0.0006	0.06	0.007	0.0008	0.10	0.001	0.0006	-0.06	0.005

Tabela 5.8: Resultados da regressão para diferentes processos biológicos II (EE: Estado estacionário, ST: Série temporal, MC: Múltiplas condições)

PROCESSO	EE			ST			MC		
	r^2	β	p	r^2	β	p	r^2	β	p
Segregação dos cromossomos	$1.4 * 10^{-6}$	0.002	0.9	0.001	-0.08	0.0009	0.0009	-0.06	0.001
Exocitose	0.0002	-0.03	0.06	0.002	0.21	$2.9 * 10^{-11}$	0.001	0.11	$1 * 10^{-6}$
Defosforilação de proteína	0.0009	-0.08	0.0009	0.005	-0.27	$2 * 10^{-15}$	0.0001	-0.03	0.16
Transcrição RNAPol III	0.001	0.07	0.37	0.0018	0.11	0.24	0.003	0.15	0.09
Tradução no citoplasma	0.0004	-0.06	$6 * 10^{-6}$	0.0011	0.17	$1 * 10^{-11}$	0.005	0.36	$1 * 10^{-5}$
Limitação de glucose	$2 * 10^{-5}$	0.01	0.59	0.001	0.13	$8 * 10^{-5}$	0.001	0.11	$5 * 10^{-6}$
Enovelamento de proteína	$7 * 10^{-5}$	0.021	0.26	0.0007	-0.09	0.0004	0.002	0.12	$2 * 10^{-10}$
Biogênese do ribossomo	0.0002	-0.06	0.12	0.0008	0.14	0.007	0.01	0.47	$1 * 10^{-15}$
Elongação da tradução	$2 * 10^{-6}$	-0.003	0.92	0.005	0.28	$1 * 10^{-6}$	0.001	0.14	0.03
Início da tradução	0.003	0.16	0.0001	0.0004	-0.10	0.15	0.004	0.25	$1 * 10^{-5}$
Elongação da transcrição	$2 * 10^{-6}$	-0.003	0.92	0.005	0.28	$1 * 10^{-6}$	0.001	0.14	0.03
Ciclo celular mitótico	0.003	0.09	0.003	0.0005	0.06	0.23	0.0008	0.06	0.12
Crescimento pseudo-hifas	$1.4 * 10^{-6}$	-0.01	0.58	0.001	0.12	$6 * 10^{-6}$	0.0006	0.06	0.0004
Splicing RNA	0.02	0.35	$9 * 10^{-9}$	0.02	-0.33	$8 * 10^{-10}$	0.015	-0.3	$7 * 10^{-7}$
Respiração celular	0.0002	0.03	0.38	0.004	-0.21	$7.6 * 10^{-5}$	0.001	-0.12	0.03
Processamento do mRNA	0.008	-0.24	0.021	0.02	-0.62	$2.2 * 10^{-5}$	0.002	-0.12	0.25
Regulação do ciclo celular	0.0024	0.10	$2.6 * 10^{-5}$	0.00018	0.03	0.25	$9 * 10^{-5}$	-0.02	0.4
Transporte através membrana	0.00015	0.04	0.12	0.0007	0.12	0.0008	0.0001	-0.04	0.14
Processamento tRNA	0.01	-0.24	$1.7 * 10^{-7}$	0.019	-0.51	$6.5 * 10^{-7}$	0.001	0.1	0.07
Organização da vacúola	$5.6 * 10^{-6}$	-0.006	0.84	0.009	0.34	$4.7 * 10^{-15}$	0.0001	-0.02	0.33
Organização da vesícula	0.0006	-0.04	0.062	0.006	0.24	$9.2 * 10^{-10}$	0.006	-0.15	$6 * 10^{-9}$
Regulação do ciclo celular	0.0024	0.10	$2.6 * 10^{-5}$	0.0001	0.037	0.25	$9 * 10^{-5}$	-0.02	0.4

IG e processos biológico III

Finalmente, o teste de Mantel foi utilizado para quantificar em termos estatísticos a relação entre as interações genéticas e os coeficientes de correlação. O teste de Mantel é um teste estatístico utilizado para medir o grau de relação que pode existir entre duas matrizes, neste caso as matrizes de correlação e as de interação genética obtidas para os diferentes proces-

tos biológicos mostrados na tabela 5.6. O teste de Mantel, que foi descrito em [88] tem sido amplamente utilizado em pesquisa ecológica [89], para relacionar dados de distanciamento geográfico entre sub-populações de espécies. A hipótese que deve ser assumida no teste é que o grau de dissimilaridade de uma das matrizes corresponde ao grau de dissimilaridade da outra matriz sempre que ambos conjuntos de dados tenham sido independentemente obtidos. De modo que para aplicar o teste sob os dados de correlação e de interação genética, ambos os valores foram dispostos e transformados em matrizes de dissimilaridade.

O teste retorna um valor no intervalo $[0, 1]$, sendo 1 o máximo grau de correlação. Sob essas bases, a aplicação do teste de Mantel tanto ao nível de processo quanto de todo o conjunto de dados, devolve valores de correlação que embora para alguns processos seja forte, em geral todos os resultados não apresentam p-valores significativos. Quando aplicado sob todo o conjunto de dados, o coeficiente de Mantel é -0.0045 com um p-valor de 0.987 para os dados de expressão em estado estacionário, de 0.001 e p-valor de 0.587 para os dados em série temporal e de 0.0104 e p-valor 0.001 para os dados de expressão em múltiplas condições. Já no caso dos processos biológicos, a tabela 5.9 mostra o resultado para 27 dos 55 processos biológicos que apresentaram interação genética. De todos os processos, somente o de morfogênese apresenta valores fortes de correlação para os três conjuntos de expressão, mas em nenhum desses casos o p-valor é significativo.

Além desse processo, existem outros como os de replicação, terminação da transcrição, glicolização de proteínas ou biogênese do ribossomo que no caso dos dados em estado estacionário, também apresentam valores de correlação acima de $|0.5|$, porém sem p-valores significativos. Dentre todos os processos, só o de organização da vacuola nos dados de série temporal e de organização da mitocôndria nos dados em múltiplas condições apresentam valores de correlação acima de $|0.5|$ e p-valores abaixo de 0.05. Apesar que se corrigidos para múltiplos testes pelo método de FDR [90], esses p-valores perdem a significância estatística, esses processos de organização de estruturas se constituem em alvos de avaliação biológica.

Para os 28 processos biológicos restantes, não foi possível calcular o teste de Mantel, pois apesar que eles também foram testados pelo experimento de interação genética, os resultados desses experimentos não identificaram relações funcionais, corroborando assim que esse tipo de abordagem não oferecem bons resultados em se tratando de identificar relações funcionais segundo processos biológicos.

Tabela 5.9: Resultados do teste de Mantel para diferentes processos biológicos

Processo	Estado estacionário		Série temporal		Múltiplas condições	
	Mantel	p-valor	Mantel	p-valor	Mantel	p-valor
Morfogênese celular	0.8	0.66	0.71	0.65	-0.78	0.666
Respiração celular	-0.07	0.865	-0.58	0.141	-0.33	0.615
Resposta a danos no DNA	-0.1	0.358	0.0015	0.992	-0.13	0.291
Organização da cromatina	-0.38	0.37	0.076	0.97	-0.076	0.78
Segregação dos cromossomos	0.07	0.8	-0.27	0.297	-0.18	0.445
Tradução no citoplasma	0.03	0.523	0.11	0.075	0.082	0.258
Recombinação do DNA	-0.03	0.917	-0.31	0.149	-0.015	0.957
Reparo do DNA	-0.08	0.437	-0.08	0.631	0.23	0.054
Replicação	-0.73	0.172	-0.15	0.83	0.71	0.162
Terminação da transcrição	0.9	0.349	-0.89	0.341	-0.41	1
Endocitose	-0.15	0.125	0.17	0.167	0.074	0.489
Limitação de glicose	0.04	0.752	-0.2	0.073	0.068	0.601
Tradução na mitocôndria	0.12	1			0.99	0.328

Organização da mitocôndria	-0.02	0.937	-0.19	0.384	0.51	0.017
Defosforilação de proteínas	-0.09	0.503	-0.25	0.107	-0.078	0.583
Montagem de proteínas	-0.07	0.457	-0.13	0.286	0.19	0.043
Glicosilação de proteínas	-0.65	0.114	-0.16	0.605	0.17	0.654
Exocitose	-0.08	0.588	0.11	0.548	-0.012	0.943
Fosforilação de proteínas	0.05	0.158	0.03	0.482	-0.017	0.699
Crescimento de pseudo-hifas	0.1	0.263	-0.001	0.984	-0.09	0.35
Regulação ciclo celular	0.45	0.674	0.52	0.478	0.56	0.328
Resposta ao aquecimento	0.16	0.457	-0.11	0.572	0.016	0.958
Stress osmótico	-0.16	0.407	0.3	0.038	0.47	0.015
Biogênese do ribossomo	-0.94	0.345	-0.45	1	0.88	0.334
Transporte através da membrana	0.02	0.803	0.06	0.5	0.013	0.906
Organização da vacuola	-0.27	0.488	-0.71	0.024	-0.17	0.644
Organização da vesícula	-0.12	0.625	0.049	0.835	-0.021	0.932

Capítulo 6

Conclusões

Como produto da revisão de literatura e do análise dos dados que têm sido apresentados neste documento, é possível apresentar as seguintes conclusões.

As relações funcionais entre genes, são entendidas como o reflexo da atividade conjunta de dois ou mais genes. Esse princípio é assumido tanto pelas interações genéticas quanto pela co-expressão. Por exemplo, em se tratando de interações genéticas, vários autores utilizam quase que de modo literal expressões como: “...uma interação genética só implica que o par de genes compartilha uma relação funcional...” [6, 31, 34]. No caso dos estudos de co-expressão, não é difícil encontrar a mesma premissa [16, 32, 76]. O fato das duas estratégias estarem baseadas na mesma hipótese junto ao fato de que os níveis de expressão de mais de 25% dos mRNA em levedura possam ser explicados quase linearmente pela taxa de crescimento como medida do fenótipo [91, 92], suportam a tentativa de procurar o quanto as duas estratégias conseguem identificar as mesmas relações funcionais.

Portanto, o presente trabalho deve ser entendido como a busca de associação entre dois recursos de informação existente, na tentativa de identificar algum tipo de tendência entre eles. Este é um estudo exploratório prévio a qualquer tipo de integração ou predição, no qual o conhecimento implícito nos estudos de co-expressão e de interação genética ao respeito do que é uma relação funcional entre genes é avaliado de forma conjunta.

Apesar de que o tratamento das relações funcionais entre genes em forma de redes tem oferecido valiosa informação ao respeito da dinâmica e da estrutura da célula, definir quando dois genes apresentam uma relação funcional não é uma tarefa simples. Neste documento têm sido estudadas e comparadas as duas principais estratégias desenhadas para identificar relações funcio-

nais: co-expressão, a partir de medidas de dependência estatística e interação genética por meio da medida do “*fitness*” de uma cultura dupla-mutante. Para tal, foram utilizados métodos gráficos e estatísticos os quais mostraram que tanto num nível global quanto para diferentes processos biológicos as relações funcionais que podem ser identificadas por ambas estratégias é quase nulo. Além disso, verificou-se que os valores que definem uma dupla mutação ou uma interação genética também não conseguem explicar os valores de correlação que definem uma relação funcional por co-expressão. Varias características tanto técnicas como biológicas podem explicar a baixa conexão que existe entre ambas as estratégias. A continuação serão comentadas algumas dessas possíveis causas.

Diferença das técnicas de captura de dados

Talvez a diferença mais notória entre a co-expressão e as interações genéticas é a forma que cada estratégia utiliza para obter os dados a partir dos quais são descritas as relações funcionais. Assim por exemplo, as interações genéticas são experimentos nos quais a dupla mutação que se faz para identificar uma relação funcional deve ser feito em culturas diferentes para cada par de genes que deseja-se testar; e apesar de que para todos se mantém as mesmas condições de cultura, não é possível garantir que duas interações estejam presentes ao mesmo tempo. Contudo, autores como Colak e colaboradores [78], e Costanzo e colaboradores [13] afirmam que as interações genéticas oferecem uma visão mais abrangente da dinâmica celular dado que não se encontram limitadas a condições específicas.

Já no caso da co-expressão, os experimentos de microarranjos permitem obter os valores de expressão de muitos genes em forma simultânea e portanto identificar as relações funcionais por medidas de dependência para todos eles. O problema neste caso está no fato dessas relações funcionais estarem restritas às condições nas quais foram obtidos os dados de expressão, de modo que é difícil garantir que as mesmas relações funcionais estejam presentes em condições diferentes.

Esta diferença nas técnicas pode ser um dos motivos que explique o baixo nível de relação que foi encontrado entre a co-expressão e as interações genéticas.

Problemas de conceitualização

Visto que as técnicas são diferentes, é lógico pensar que isso tenha uma influência nos resultados da comparação. No entanto, sabendo que ambas estratégias pretendem identificar o mesmo tipo de relações, também é válido explorar a possibilidade de que ambas estratégias estejam considerando tipos de relações diferentes dentro de uma mesma classe chamada relação funcional. Assim, considerando que os diferentes níveis de abstração em que podem ser divididos os sistemas biológicos (uma dessas divisões é apresentada no começo deste documento. Figura 1.1), serve não só para facilitar o estudo mas também como reflexo da complexidade biológica, na qual os níveis superiores representam graus maiores de complexidade [6], dificilmente uma relação do nível dois possa também estar presente em níveis superiores.

Porém, em caso que seja possível identificar o mesmo tipo de relações em dois níveis diferentes, uma definição precisa ao respeito do que é, e do que não é uma relação funcional, ou em caso que existam vários tipos, uma categorização delas pode ajudar a definir o tipo de relação que pode existir entre as duas estratégias. Portanto, só depois de definir com precisão o que é uma relação funcional e quais tipos de relações funcionais são identificadas pela co-expressão e pelas interações genéticas, poderão ser aplicadas os métodos de integração de dados como as que tem sido amplamente utilizadas entre as interações de co-expressão e as de proteínas, ou entre estas últimas e as interações genéticas.

Identificação de “não-relações funcionais”

Como assinalado nas linhas anteriores, a ausência de relação observada entre co-expressão e interação genética pode ser o reflexo da presença de diferentes tipos de relações funcionais. Assim por exemplo, alguns processos biológicos já muito bem definidos, têm sido considerados como permanentes, bem seja porque as atividades que eles realizam são necessárias para o bom funcionamento celular durante todo o ciclo de vida, ou porque a função que eles desenvolvem é levada a cabo por complexos proteicos onde todas as subunidades encontram-se altamente co-reguladas. Essa condição de atividade permanente tem sido amplamente validada com estudos de co-expressão onde os genes envolvidos em cada processo, ou complexo proteico encontram-se fortemente co-regulados. No entanto, propriedades diferentes têm sido descritas para as interações genéticas, nas quais a medida em que uma maior quantidade de interações genéticas são coletadas, menor é sua correspondência

para processos biológicos menores.

Além disso, a co-expressão visto de uma forma simples, se refere ao aumento ou diminuição da função que desenvolvem um par de genes, enquanto que a interação genética se refere à ausência total dessa função. Portanto poderíamos afirmar que no primeiro caso o que se identifica é uma relação funcional enquanto que no segundo o que se identifica é uma relação de não-função, sendo essa uma diferença que pode criar a lacuna entre as duas estratégias aqui comparadas.

Perspectivas

Mesmo que os resultados das comparações aqui realizadas mostrem a dificuldade de que as duas estratégias aqui estudadas obtenham as mesmas relações funcionais, alguns procedimentos que podem ser feitos para aprimorar os resultados aqui obtidos e oferecer uma perspectiva mais clara da relação entre co-expressão e interação genética são:

- Desenhar uma série de experimentos que obtenham medidas de expressão de culturas celulares que carregam uma dupla-mutação a fim de avaliar o efeito na expressão gerado pela ausência de dois genes em forma simultânea.
- Desenvolver uma metodologia que permita “enriquecer” relações pareadas do mesmo modo como é feito, por exemplo, no método conhecido como GSEA [93]. Pois apesar de já existir algumas estratégias que fazem algo parecido [94,95] elas se baseiam na comparação dos níveis de expressão absolutos e não no resultado de medidas de correlação que identificam relações funcionais.
- Comparar os resultados que podem oferecer as técnicas de construção de redes como WGCNA [82] tanto para relações de co-expressão como para interações genéticas. Isto, considerando que a partir desse tipo técnicas as relações funcionais são entendidas principalmente como o efeito conjunto de todos os genes e não só daqueles que por meio de um valor de corte são definidas como funcionalmente relacionadas.

Apêndice A

1. Comparação dos resultados obtidos para a medida de Pearson e Spearman quando comparados com as interações genéticas do banco de dados qualitativo

Os gráficos de barras das figuras A.1, A.2 e A.3, mostram uma classificação das interações genéticas segundo diferentes intervalos de correlação para as medidas de Pearson e Spearman. Na tabela A.1, é mostrado um resumo comparativo do número de relações obtidas de modo conjunto pelas interações genéticas e as medidas de Pearson e Spearman.

Figura A.1: Classificação das interações genéticas segundo diferentes intervalos de correlação. Dados de expressão em estado estacionário.

Figura A.2: Classificação das interações genéticas segundo diferentes intervalos de correlação. Dados de expressão em serie temporal.

Figura A.3: Classificação das interações genéticas segundo diferentes intervalos de correlação. Dados de expressão sob múltiplas condições.

Tabela A.1: Relações funcionais comuns aos métodos de correlação e de interação genética.

	Pearson e interações genéticas	Spearman e interações genéticas	Comum em Pearson e Spearman
Estado estacionário	840	831	525
Série temporal	9.647	11.241	7.408
Múltiplas condições	1.823	830	696

Apêndice B

Curvas ROC para todas as comparações com a medida de Pearson e Spearman segundo os dados de expressão

A figura B.1, mostra todas as curvas ROC obtidas para a medida de Pearson, Spearman segundo cada um dos dados de expressão.

(a) Curva ROC $|\rho|$. Expressão em Estado estacionário

(b) Curva ROC $|\rho|$. Expressão em Série temporal

(c) Curva ROC $|\rho|$. Expressão sob múltiplas condições

Figura B.1: Curvas ROC para coeficientes de correlação de Pearson usando os dados de expressão em estado estacionário (a) e série temporal (b), e para a medida de Spearman com os dados de expressão em múltiplas condições (c). TPR: Taxa de verdadeiros positivos. FPR: Taxa de falsos positivos.

Apêndice C

Resultados comparação IG mínima e duplo mutante para os dados de expressão em estado estacionário e série temporal

Os resultados mostrados na figura C.1 descrevem os resultados da comparação entre os valores de interação genética mínima e os dados de expressão em estado estacionário e série temporal. Já a figura C.2 mostra os resultados da comparação entre os valores do duplo mutante e os dados de expressão em estado estacionário e série temporal.

Figura C.1: Resultados da comparação entre os valores de IG mínima contra a co-expressão dos dados em estado estacionário e série temporal.

Figura C.2: Resultados da comparação entre os valores de duplo mutante contra a co-expressão dos dados em estado estacionário e série temporal.

Apêndice D

Resultados comparação IG por processos biológicos para todos os conjuntos de expressão

Figura D.1: Distribuição dos coeficientes de correlação dos pares de genes com e sem interação genética para diferentes processos biológicos nos três conjuntos de expressão (I).

(a) Respiração celular. Múltiplas condições

(b) Resposta a danos no DNA. Estado estacionário

(c) Resposta a danos no DNA. Série temporal

(d) Resposta a danos no DNA. Múltiplas condições

(e) Organização da cromatina. Estado estacionário

(f) Organização da cromatina. Série temporal

(g) Organização da cromatina. Múltiplas condições

(h) Segregação dos cromossomos. Estado estacionário

(i) Segregação dos cromossomos. Série temporal

(j) Segregação dos cromossomos. Múltiplas condições

(k) Tradução no citoplasma. Estado estacionário

(l) Tradução no citoplasma. Série temporal

(m) Tradução no citoplasma. Múltiplas condições

(n) Recombinação do DNA. Estado estacionário

(o) Recombinação do DNA. Série temporal

(p) Recombinação do DNA. Múltiplas condições

Figura D.2: Distribuição dos coeficientes de correlação dos pares de genes com e sem interação genética para diferentes processos biológicos nos três conjuntos de expressão (II).

Figura D.3: Distribuição dos coeficientes de correlação dos pares de genes com e sem interação genética para diferentes processos biológicos nos três conjuntos de expressão (III).

(a) Limitação de glicose. Série temporal

(b) Limitação de glicose. Múltiplas condições

(c) Tradução na mitocondria. Estado estacionário

(d) Tradução na mitocondria. Série temporal

(e) Tradução na mitocondria. Múltiplas condições

(f) Organização da mitocondria. Estado estacionário

(g) Organização da mitocondria. Série temporal

(h) Organização da mitocondria. Múltiplas condições

(i) Ciclo celular mitótico. Estado estacionário

(j) Ciclo celular mitótico. Série temporal

(k) Ciclo celular mitótico. Múltiplas condições

(l) Organização do peroxisoma. Estado estacionário

(m) Organização do peroxisoma. Série temporal

(n) Organização do peroxisoma. Múltiplas condições

(o) Defosforilação de proteínas. Estado estacionário

(p) Defosforilação de proteínas. Série temporal

Figura D.4: Distribuição dos coeficientes de correlação dos pares de genes com e sem interação genética para diferentes processos biológicos nos três conjuntos de expressão (IV).

Figura D.5: Distribuição dos coeficientes de correlação dos pares de genes com e sem interação genética para diferentes processos biológicos nos três conjuntos de expressão (V).

Figura D.6: Distribuição dos coeficientes de correlação dos pares de genes com e sem interação genética para diferentes processos biológicos nos três conjuntos de expressão (VI).

Figura D.7: Distribuição dos coeficientes de correlação dos pares de genes com e sem interação genética para diferentes processos biológicos nos três conjuntos de expressão (VII).

(a) Organização da vacuola. Múltiplas condições (b) Organização da vesícula. Estado estacionário (c) Organização da vesícula. Série temporal (d) Organização da vesícula. Múltiplas condições

Figura D.8: Distribuição dos coeficientes de correlação dos pares de genes com e sem interação genética para diferentes processos biológicos nos três conjuntos de expressão (VIII).

Apêndice E

Gráficos de dispersão para todos os processos biológicos segundo cada conjunto de expressão

Figura E.1: Gráficos de dispersão entre coeficientes de correlação e valores de interação genética para diferentes processos biológicos (I).

Figura E.2: Gráficos de dispersão entre coeficientes de correlação e valores de interação genética para diferentes processos biológicos (II).

Figura E.3: Gráficos de dispersão entre coeficientes de correlação e valores de interação genética para diferentes processos biológicos (III).

Figura E.4: Gráficos de dispersão entre coeficientes de correlação e valores de interação genética para diferentes processos biológicos (IV).

Figura E.5: Gráficos de dispersão entre coeficientes de correlação e valores de interação genética para diferentes processos biológicos (V).

Figura E.6: Gráficos de dispersão entre coeficientes de correlação e valores de interação genética para diferentes processos biológicos (VI).

Figura E.7: Gráficos de dispersão entre coeficientes de correlação e valores de interação genética para diferentes processos biológicos (VII).

(a) Organização da vacuola. Múltiplas condições

(b) Organização da vesícula. Estado estacionário

(c) Organização da vesícula. Série temporal

(d) Organização da vesícula. Múltiplas condições

Figura E.8: Gráficos de dispersão entre coeficientes de correlação e valores de interação genética para diferentes processos biológicos (VIII).

Bibliografia

- [1] Kitano Hiroaki. Systems biology: a brief overview. *Science*, 295(5560):1662–1664, 2002.
- [2] Ondrej Polansky, Zuzana Sekelova, Marcela Faldynova, Alena Sebkova, Frantisek Sisak, and Ivan Rychlik. Characterisation of the most important metabolic pathways and biological processes expressed in chicken caecal microbiota. *Applied and Environmental Microbiology*, pages AEM-03473, 2015.
- [3] Bolin Chen and Fang-Xiang Wu. Identifying protein complexes based on multiple topological structures in ppi networks. *NanoBioscience, IEEE Transactions on*, 12(3):165–172, 2013.
- [4] Aleksej Zelezniak, Steven Sheridan, and Kiran Raosaheb Patil. Contribution of network connectivity in determining the relationship between gene expression and metabolite concentration changes. *PLoS Comput Biol*, 10(4):e1003572, 2014.
- [5] Riet De Smet and Kathleen Marchal. Advantages and limitations of current network inference methods. *Nature Reviews Microbiology*, 8(10):717–729, 2010.
- [6] Benjamin Boucher and Sarah Jenna. Genetic interaction networks: better understand to better predict. *Frontiers in genetics*, 4, 2013.
- [7] Ronald A Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the royal society of Edinburgh*, 52(02):399–433, 1919.
- [8] Charles Boone, Howard Bussey, and Brenda J Andrews. Exploring genetic interactions and networks with yeast. *Nature Reviews Genetics*, 8(6):437–449, 2007.

- [9] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101–113, 2004.
- [10] Joshua M Stuart, Eran Segal, Daphne Koller, and Stuart K Kim. A gene-coexpression network for global discovery of conserved genetic modules. *science*, 302(5643):249–255, 2003.
- [11] Paul T Spellman, Gavin Sherlock, Michael Q Zhang, Vishwanath R Iyer, Kirk Anders, Michael B Eisen, Patrick O Brown, David Botstein, and Bruce Futcher. Comprehensive identification of cell cycle regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular biology of the cell*, 9(12):3273–3297, 1998.
- [12] Amy Hin Yan Tong, Guillaume Lesage, Gary D Bader, Huiming Ding, Hong Xu, Xiaofeng Xin, James Young, Gabriel F Berriz, Renee L Brost, Michael Chang, et al. Global mapping of the yeast genetic interaction network. *science*, 303(5659):808–813, 2004.
- [13] Michael Costanzo, Anastasia Baryshnikova, Jeremy Bellay, Yungil Kim, Eric D Spear, Carolyn S Sevier, Huiming Ding, Judice LY Koh, Kiana Toufighi, Sara Mostafavi, et al. The genetic landscape of a cell. *science*, 327(5964):425–431, 2010.
- [14] Christoph Jüschke, Ilse Dohnal, Peter Pichler, Heike Harzer, Remco Swart, Gustav Ammerer, Karl Mechtler, and Juergen A Knoblich. Transcriptome and proteome quantification of a tumor model provides novel insights into post-transcriptional gene regulation. *Genome Biol*, 14(11):r133, 2013.
- [15] Ryan Kelley and Trey Ideker. Systematic interpretation of genetic interactions using protein networks. *Nature biotechnology*, 23(5):561–566, 2005.
- [16] Sapna Kumari, Jeff Nie, Huann-Sheng Chen, Hao Ma, Ron Stewart, Xiang Li, Meng-Zhu Lu, William M Taylor, and Hairong Wei. Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. 2012.
- [17] Alberto De La Fuente, Nan Bing, Ina Hoeschele, and Pedro Mendes. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574, 2004.

- [18] Matthew E Studham, Andreas Tjärnberg, Torbjörn EM Nordling, Sven Nelander, and Erik LL Sonnhhammer. Functional association networks as priors for gene regulatory network inference. *Bioinformatics*, 30(12):i130–i138, 2014.
- [19] Scott J Dixon, Michael Costanzo, Anastasia Baryshnikova, Brenda Andrews, and Charles Boone. Systematic mapping of genetic interaction networks. *Annual review of genetics*, 43:601–625, 2009.
- [20] Adam A Margolin and Andrea Califano. Theory and limitations of genetic network inference from microarray data. *Annals of the New York Academy of Sciences*, 1115(1):51–72, 2007.
- [21] Naoki Nariai, Yoshinori Tamada, Seiya Imoto, and Satoru Miyano. Estimating gene regulatory networks and protein–protein interactions of *saccharomyces cerevisiae* from multiple genome-wide data. *Bioinformatics*, 21(suppl 2):ii206–ii212, 2005.
- [22] Jeremy Bellay, Gowtham Atluri, Tina L Sing, Kiana Toufighi, Michael Costanzo, Philippe Souza Moraes Ribeiro, Gaurav Pandey, Joshua Baller, Benjamin VanderSluis, Magali Michaut, et al. Putting genetic interactions in context through a global modular decomposition. *Genome research*, 21(8):1375–1387, 2011.
- [23] Nitin Bhardwaj and Hui Lu. Correlation between gene expression profiles and protein–protein interactions within and across genomes. *Bioinformatics*, 21(11):2730–2738, 2005.
- [24] Andreas Beyer, Sourav Bandyopadhyay, and Trey Ideker. Integrating physical and genetic maps: from genomes to interaction networks. *Nature Reviews Genetics*, 8(9):699–710, 2007.
- [25] Patrik Dhaeseleer, Shoudan Liang, and Roland Somogyi. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16(8):707–726, 2000.
- [26] Christopher A Penfold and David L Wild. How to infer gene networks from expression profiles, revisited. *Interface Focus*, 1(6):857–870, 2011.
- [27] Ali Kazemi-Pour, Bahram Goliaei, and Hamid Pezeshk. Protein complex discovery by interaction filtering from protein interaction networks using mutual rank coexpression and sequence similarity. *BioMed research international*, 2015, 2015.

- [28] Sourav Bandyopadhyay, Ryan Kelley, Nevan J Krogan, and Trey Ideker. Functional maps of protein complexes from quantitative genetic interaction data. *PLoS Comput Biol*, 4(4):e1000065, 2008.
- [29] Tobias Maier, Marc Güell, and Luis Serrano. Correlation of mrna and protein in complex biological samples. *FEBS letters*, 583(24):3966–3973, 2009.
- [30] Steven P Gygi, Yvan Rochon, B Robert Franza, and Ruedi Aebersold. Correlation between protein and mrna abundance in yeast. *Molecular and cellular biology*, 19(3):1720–1730, 1999.
- [31] Jelena Ostojić, Annie Glatigny, Christopher J Herbert, Geneviève Du-jardin, and Nathalie Bonnefoy. Does the study of genetic interactions help predict the function of mitochondrial proteins in *saccharomyces cerevisiae*? *Biochimie*, 100:27–37, 2014.
- [32] Andrea Gobbi and Giuseppe Jurman. A null model for pearson coexpression networks. *PloS one*, 10(6):e0128115, 2015.
- [33] Elizabeth A Winzeler, Daniel D Shoemaker, Anna Astromoff, Hong Liang, Keith Anderson, Bruno Andre, Rhonda Bangham, Rocio Benito, Jef D Boeke, Howard Bussey, et al. Functional characterization of the *s. cerevisiae* genome by gene deletion and parallel analysis. *science*, 285(5429):901–906, 1999.
- [34] Benjamin Boucher, Anna Y Lee, Michael Hallett, and Sarah Jenna. Structural and functional characterization of a *caenorhabditis elegans* genetic interaction network within pathways. *PLoS Comput Biol*, 12(2):e1004738, 2016.
- [35] Ramamurthy Mani, Robert P St Onge, John L Hartman, Guri Giaever, and Frederick P Roth. Defining genetic interaction. *Proceedings of the National Academy of Sciences*, 105(9):3461–3466, 2008.
- [36] Erol Akçay and Jeremy Van Cleve. There is no fitness but fitness, and the lineage is its bearer. *Phil. Trans. R. Soc. B*, 371(1687):20150085, 2016.
- [37] Daniel Segre, Alexander DeLuna, George M Church, and Roy Kishony. Modular epistasis in yeast metabolism. *Nature genetics*, 37(1):77–83, 2005.

- [38] Robert P St Onge, Ramamurthy Mani, Julia Oh, Michael Proctor, Eula Fung, Ronald W Davis, Corey Nislow, Frederick P Roth, and Guri Giaever. Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nature genetics*, 39(2):199–206, 2007.
- [39] Lukasz Jasnos and Ryszard Korona. Epistatic buffering of fitness loss in yeast double deletion strains. *Nature genetics*, 39(4):550–554, 2007.
- [40] Rafael Sanjuán and Santiago F Elena. Epistasis correlates to genomic complexity. *Proceedings of the National Academy of Sciences*, 103(39):14402–14405, 2006.
- [41] Amit Puniyani, Uri Liberman, and Marcus W Feldman. On the meaning of non-epistatic selection. *Theoretical population biology*, 66(4):317–321, 2004.
- [42] Alfréd Rényi. On measures of dependence. *Acta mathematica hungarica*, 10(3-4):441–451, 1959.
- [43] Marcos Nascimento Magalhães and Antonio Carlos Pedroso de Lima. *Noções de probabilidade e estatística*. IME-USP, 2000.
- [44] Lin Song, Peter Langfelder, and Steve Horvath. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC bioinformatics*, 13(1):328, 2012.
- [45] Pablo A Jaskowiak, Ricardo JGB Campello, and Ivan G Costa. Evaluating correlation coefficients for clustering gene expression profiles of cancer. In *Brazilian Symposium on Bioinformatics*, pages 120–131. Springer, 2012.
- [46] Laurie J Heyer, Semyon Kruglyak, and Shibu Yooseph. Exploring expression data: identification and analysis of coexpressed genes. *Genome research*, 9(11):1106–1115, 1999.
- [47] Suzana de Siqueira Santos, Daniel Yasumasa Takahashi, Asuka Nakata, and André Fujita. A comparative study of statistical methods used to identify dependencies between gene expression signals. *Briefings in bioinformatics*, page bbt051, 2013.
- [48] Karl Pearson. Notes on the history of correlation. *Biometrika*, 13(1):25–45, 1920.
- [49] Charles Spearman. "general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904.

- [50] Curtis Huttenhower, K Tsheko Mutungu, Natasha Indik, Woongcheol Yang, Mark Schroeder, Joshua J Forman, Olga G Troyanskaya, and Hilary A Collier. Detailing regulatory networks through large scale data integration. *Bioinformatics*, 25(24):3267–3274, 2009.
- [51] Johanna Hardin, Aya Mitani, Leanne Hicks, and Brian VanKoten. A robust measure of correlation between two genes on a microarray. *BMC bioinformatics*, 8(1):220, 2007.
- [52] Catharina Olsen, Patrick E Meyer, and Gianluca Bontempi. On the impact of entropy estimation on transcriptional regulatory network inference based on mutual information. *EURASIP Journal on Bioinformatics and Systems Biology*, 2009(1):1, 2008.
- [53] Zeyneb Kurt, Nizamettin Aydin, and Gökmen Altay. A comprehensive comparison of association estimators for gene network inference algorithms. *Bioinformatics*, page btu182, 2014.
- [54] David Botstein and Gerald R Fink. Yeast: an experimental organism for 21st century biology. *Genetics*, 189(3):695–704, 2011.
- [55] Patrick Kemmeren, Katrin Sameith, Loes AL van de Pasch, Joris J Benschop, Tineke L Lenstra, Thanasis Margaritis, Eoghan O’Duibhir, Eva Apweiler, Sake van Wageningen, Cheuk W Ko, et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, 157(3):740–752, 2014.
- [56] Jeremiah J Faith, Michael E Driscoll, Vincent A Fusaro, Elissa J Cosgrove, Boris Hayete, Frank S Juhn, Stephen J Schneider, and Timothy S Gardner. Many microbe microarrays database: uniformly normalized affymetrix compendia with structured experimental metadata. *Nucleic acids research*, 36(suppl 1):D866–D870, 2008.
- [57] Rafael A Irizarry, Benjamin M Bolstad, Francois Collin, Leslie M Cope, Bridget Hobbs, and Terence P Speed. Summaries of affymetrix genechip probe level data. *Nucleic acids research*, 31(4):e15–e15, 2003.
- [58] Joachim J Li. Dna replication: Once, and only once. *Current Biology*, 5(5):472–475, 1995.
- [59] Sorin Draghici. *Statistics and data analysis for microarrays using R and bioconductor*. CRC Press, 2011.

- [60] Karla V Ballman, Diane E Grill, Ann L Oberg, and Terry M Therneau. Faster cyclic loess: normalizing rna arrays via linear models. *Bioinformatics*, 20(16):2778–2786, 2004.
- [61] William S Cleveland and Susan J Devlin. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical association*, 83(403):596–610, 1988.
- [62] Sandrine Dudoit, Yee Hwa Yang, Matthew J Callow, and Terence P Speed. Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica sinica*, pages 111–139, 2002.
- [63] Yee Hwa Yang, Sandrine Dudoit, Percy Luu, David M Lin, Vivian Peng, John Ngai, and Terence P Speed. Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic acids research*, 30(4):e15–e15, 2002.
- [64] Dongxiao Zhu, Alfred O Hero, Zhaohui S Qin, and Anand Swaroop. High throughput screening of co-expressed gene pairs with controlled false discovery rate (fdr) and minimum acceptable strength (mas). *Journal of Computational Biology*, 12(7):1029–1045, 2005.
- [65] George S Davidson, Brian N Wylie, and Kevin W Boyack. Cluster stability and the use of noise in interpretation of clustering. In *infovvis*, pages 23–30, 2001.
- [66] Homin K Lee, Amy K Hsu, Jon Sajdak, Jie Qin, and Paul Pavlidis. Coexpression analysis of human genes across many microarray data sets. *Genome research*, 14(6):1085–1094, 2004.
- [67] Brynn H Voy, Jon A Scharff, Andy D Perkins, Arnold M Saxton, Bhavesh Borate, Elissa J Chesler, Lisa K Branstetter, and Michael A Langston. Extracting gene networks for low-dose radiation using graph theoretical algorithms. *PLoS Comput Biol*, 2(7):e89, 2006.
- [68] Despina Sanoudou, Judith N Haslett, Alvin T Kho, Shaoqiang Guo, Hanna T Gazda, Steven A Greenberg, Hart GW Lidov, Isaac S Kohane, Louis M Kunkel, and Alan H Beggs. Expression profiling reveals altered satellite cell numbers and glycolytic enzyme transcription in nemaline myopathy muscle. *Proceedings of the National Academy of Sciences*, 100(8):4666–4671, 2003.

- [69] Masaru Moriyama, Yujin Hoshida, Motoyuki Otsuka, ShinIchiro Nishimura, Naoya Kato, Tadashi Goto, Hiroyoshi Taniguchi, Yasushi Shiratori, Naohiko Seki, and Masao Omata. Relevance network between chemosensitivity and transcriptome in human hepatoma cells1. *Molecular Cancer Therapeutics*, 2(2):199–205, 2003.
- [70] Masataka Maruyama, Tamihide Matsunaga, Eri Harada, and Shigeru Ohmori. Comparison of basal gene expression and induction of cyp3as in hepg2 and human fetal liver cells. *Biological and pharmaceutical bulletin*, 30(11):2091–2097, 2007.
- [71] Eoghan ODuibhir, Philip Lijnzaad, Joris J Benschop, Tineke L Lensstra, Dik van Leenen, Marian JA Groot Koerkamp, Thanasis Margaritis, Mariel O Brok, Patrick Kemmeren, and Frank CP Holstege. Cell cycle population effects in perturbation studies. *Molecular systems biology*, 10(6):732, 2014.
- [72] Atul J Butte and Isaac S Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Pac Symp Biocomput*, volume 5, pages 418–429, 2000.
- [73] Carlos Prieto, Alberto Risueño, Celia Fontanillo, and Javier De Las Rivas. Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles. *PloS one*, 3(12):e3911, 2008.
- [74] Raya Khanin and Ernst Wit. Construction of malaria gene expression network using partial correlations. In *Methods of Microarray Data Analysis V*, pages 75–88. Springer, 2007.
- [75] Meeta Mistry, Jesse Gillis, and Paul Pavlidis. Meta-analysis of gene coexpression networks in the post-mortem prefrontal cortex of patients with schizophrenia and unaffected controls. *BMC neuroscience*, 14(1):1, 2013.
- [76] Feng Luo, Yunfeng Yang, Jianxin Zhong, Haichun Gao, Latifur Khan, Dorothea K Thompson, and Jizhong Zhou. Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. *BMC bioinformatics*, 8(1):1, 2007.
- [77] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9, 2006.

- [78] Recep Colak, Flavia Moser, Jeffrey Shih-Chieh Chu, Alexander Schönhuth, Nansheng Chen, and Martin Ester. Module discovery by exhaustive search for densely connected, co-expressed regions in biomolecular interaction networks. *PloS one*, 5(10):e13348, 2010.
- [79] Mingzhu Zhu, Xin Deng, Trupti Joshi, Dong Xu, Gary Stacey, and Jianlin Cheng. Reconstructing differentially co-expressed gene modules and regulatory networks of soybean cells. *BMC genomics*, 13(1):1, 2012.
- [80] Zhiao Shi, Catherine K Derow, and Bing Zhang. Co-expression module analysis reveals biological processes, genomic gain, and regulatory mechanisms associated with breast cancer progression. *BMC systems biology*, 4(1):1, 2010.
- [81] Yoshiyuki Ogata, Hideyuki Suzuki, Nozomu Sakurai, and Daisuke Shibata. Cop: a database for characterizing co-expressed gene modules with biological information in plants. *Bioinformatics*, 26(9):1267–1268, 2010.
- [82] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):1, 2008.
- [83] Yuanyuan Zhang, Shudong Wang, Meixi Yang, Dashun Xu, and Dazhi Meng. A new method of determining threshold of gene network based on phenotypes. *Journal of Biological Systems*, 19(04):607–616, 2011.
- [84] Andrei Grigoriev. A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage t7 and the yeast *saccharomyces cerevisiae*. *Nucleic acids research*, 29(17):3513–3519, 2001.
- [85] Nadya Kondrashov, Aya Pusic, Craig R Stumpf, Kunihiro Shimizu, Andrew C Hsieh, Shifeng Xue, Junko Ishijima, Toshihiro Shiroishi, and Maria Barna. Ribosome-mediated specificity in hox mrna translation and vertebrate tissue patterning. *Cell*, 145(3):383–397, 2011.
- [86] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.
- [87] Lei Nie, Gang Wu, and Weiwen Zhang. Correlation between mrna and protein abundance in *desulfovibrio vulgaris*: a multiple regression to identify sources of variations. *Biochemical and biophysical research communications*, 339(2):603–610, 2006.

- [88] William D Shannon, Mark A Watson, Arie Perry, and Keith Rich. Mantel statistics to correlate gene expression levels from microarrays with clinical covariates. *Genetic Epidemiology*, 23(1):87–96, 2002.
- [89] Jari Oksanen, Roeland Kindt, Pierre Legendre, Bob OHara, M Henry H Stevens, Maintainer Jari Oksanen, and MASS Suggests. The vegan package. *Community ecology package*, 10, 2007.
- [90] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.
- [91] Matthew J Brauer, Curtis Huttenhower, Edoardo M Airoidi, Rachel Rosenstein, John C Matese, David Gresham, Viktor M Boer, Olga G Troyanskaya, and David Botstein. Coordination of growth rate, cell cycle, stress response, and metabolic activity in yeast. *Molecular biology of the cell*, 19(1):352–367, 2008.
- [92] Edoardo M Airoidi, Darach Miller, Rodoniki Athanasiadou, Nathan Brandt, Farah Abdul-Rahman, Benjamin Neymotin, Tatsu Hashimoto, Tayebah Bahmani, and David Gresham. Steady-state and dynamic gene expression programs in *saccharomyces cerevisiae* in response to variation in environmental nitrogen. *Molecular biology of the cell*, 27(8):1383–1396, 2016.
- [93] Aravind Subramanian, Heidi Kuehn, Joshua Gould, Pablo Tamayo, and Jill P Mesirov. Gsea-p: a desktop application for gene set enrichment analysis. *Bioinformatics*, 23(23):3251–3253, 2007.
- [94] Yu Liu, Mehmet Koyutürk, Jill S Barnholtz-Sloan, and Mark R Chance. Gene interaction enrichment and network analysis to identify dysregulated pathways and their interactions in complex diseases. *BMC systems biology*, 6(1):1, 2012.
- [95] Jigang Zhang, Jian Li, and Hong-Wen Deng. Identifying gene interaction enrichment for gene expression data. *PloS one*, 4(11):e8064, 2009.