Estudo da Entropia de Tsallis para a Inferência de Redes Gênicas

Cassio Henrique dos Santos Amador

Dissertação de Mestrado

Programa de Pós-Graduação em Bioinformática
Orientador: Prof. Dr. Fabrício Martins Lopes

Cornélio Procópio, Agosto de 2020

CASSIO HENRIQUE DOS SANTOS AMADOR

ESTUDO DA ENTROPIA DE TSALLIS PARA A INFERÊNCIA DE REDES GÊNICAS

Study on Tsallis Entropy applied on Gene Network Inference

Dissertação apresentada como requisito para obtenção do título de Mestre em Bioinformática da Universidade Tecnológica Federal do Paraná (UTFPR).
Orientador: Fabrício Martins Lopes

CORNÉLIO PROCÓPIO

2022

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Cornélio Procópio

CASSIO HENRIQUE DOS SANTOS AMADOR

ESTUDO DA ENTROPIA DE TSALLIS PARA A INFERÊNCIA DE REDES GÊNICAS

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Bioinformática da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Bioinformática.

Data de aprovação: 21 de Dezembro de 2021

Prof Fabricio Martins Lopes, Doutorado - Universidade Tecnológica Federal do Paraná

Prof Fabio Fernandes Da Rocha Vicente, Doutorado - Universidade Tecnológica Federal do Paraná

Prof Ronaldo Fumio Hashimoto, Doutorado - Universidade de São Paulo (Usp)

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 21/12/2021.

Agradecimentos

Agradeço em primeiro lugar à minha esposa, que me incentivou e teve a feliz ideia que me levou a fazer este mestrado.

Deixo também meus agradecimentos à UTFPR e meus chefes, que me apoiaram durante toda a jornada.

Agradeço também ao programa de Pós-graduação em Bioinformática, que me aceitou como aluno, e proporcionou muito aprendizado. Em especial ao meu orientador, pelo conhecimento, paciência e pela proposta de um tema tão interessante.

Agradeço ao Laboratório Multiusuário Centro de Computação Científica e Tecnológica, do campus Cornélio Procópio (CCCT-CP) pela realização de muitos dos cálculos aqui utilizados.

Resumo

Amador, C. H. S. **Estudo da Entropia de Tsallis para a Inferência de Redes Gênicas**. 2020. 51 f. Proposta de Dissertação (Mestrado) - Programa de Pós-Graduação em Bioinformática, Universidade Tecnológica Federal do Paraná, Cornélio Procópio, 2020.

A quantidade de informação de um sistema pode ser medida pela entropia. Um caso particular de sistema é uma rede formada pela interação entre genes, conhecida como redes gênicas. Neste trabalho estuda-se como uma entropia não-extensiva, a entropia de Tsallis, pode fornecer a maior quantidade de informação para as redes gênicas, através da escolha do melhor parâmetro não-extensivo q. Mostra-se que é possível obter numericamente o melhor parâmetro, e que ele depende do número de graus de liberdade do sistema, no caso binário o melhor valor sendo aproximadamente $2,46$. Esse resultado é testado no contexto da inferências de redes gênicas, inicialmente com portas lógicas, seguido de redes gênicas artificiais e por último com dados experimentais obtidos no desafio DREAM4. Por fim, são comparados com resultados de trabalhos anteriores, indicando a adequação da entropia de Tsallis na inferência de redes gênicas.

Palavras-chave: redes gênicas, entropia de Tsallis, sistemas complexos.

Abstract

Amador, C. H. S. **Study on Tsallis Entropy applied on Gene Network Inference**. 2020. 51 f. Proposta de Dissertação (Mestrado) - Programa de Pós-Graduação em Bioinformática, Universidade Tecnológica Federal do Paraná, Cornélio Procópio, 2020.

The amount of information in a system can be measured by entropy. A particular case of a system is a network formed by the interaction between genes, known as gene networks. In this work we study how one type of non-extensive entropy, Tsallis entropy, can provide the greatest amount of information for gene networks, through the choice of the best non-extensive parameter q. It is shown that it is possible to obtain numerically the best parameter, and that it depends on the number of degrees of freedom of the system, in the binary case the best value being approximately 2.46. This result is tested in the context of gene network inferences, initially with logic gates, followed by artificial gene networks and finally with experimental data obtained from the DREAM4 challenge. At last, these results are compared with results from previous works, indicating the adequacy of Tsallis entropy for the inference of gene networks.

Keywords: genic network, tsallis entropy, complex systems.

Sumário

Lista de Figuras

Introdução

O ato de se medir é inerente à metodologia científica. Medidas permitem comparações e testes de hipóteses, sendo necessárias para se realizar predições. O próprio conceito de uma "grandeza física"está relacionada com uma quantidade que pode ser medida, seja ela distância, tempo, temperatura e tantas outras (NUSSENZVEIG (2018)). A medida de uma grandeza é feita por meio a comparação com outras quantidades, o que pode levar a diferentes escalas, dependendo do padrão utilizado, e mudar até a posição do zero na escala. Por exemplo, medir distância em centímetros ou polegadas, resultará em diferentes valores, porém uma distância de $0cm$ tem o mesmo valor em polegadas. Já com a temperatura, que atualmente pode ser medida em graus Celsius, graus Farenheit ou Kelvin, a diferença não ocorre somente no modo de se dividir a escala, mas também no que se constitui a temperatura 0.

O presente trabalho está relacionado com uma grandeza física, conhecida como Entropia, que possui muitas aplicações e implicações, e aqui será abordada em especial sua função para medir quantidade de informação. Essa grandeza já é conhecida e estudada desde o século XIX, dentro do domínio da Termodinâmica, surgindo no estudo de motores a combustão, e descrita pela 2ª Lei da Termodinâmica (NUSSENZVEIG (2018)), que diz que num sistema fechado, a entropia aumenta ou fica constante, nunca diminui. Posteriormente, no começo do século XX, foi descrita pela mecânica estatística, desenvolvida inicialmente por Boltzmann e Gibbs, para descrever o comportamento de sistemas com um número muito grande de partículas (CALLEN, 1985; REIF, 1965). Em 1948, Shannon mostrou que a entropia é uma medida de informação, permitindo mensurar matematicamente a quantidade de informação que um sinal transmite (SHANNON (1948)). Com isso, pode-se dizer que um sistema está mais organizado quanto menor for sua entropia, e, conforme a entropia aumenta, ele se torna mais desorganizado. Com isso, sistemas podem não somente ser classificados em organizados ou desorganizados, mas, sabendo que um sistema possui uma organização, é possível procurar as configurações que permitem-no ter organização máxima, e com isso, inferir a configuração de um sistema.

Um sistema com alto grau de organização, um número enorme de elementos e de

interações entre eles, ocorrem no interior de seres vivos. Os genes, produzidos a partir de trechos específicos de sequências de DNA, codificam produtos que realizam uma série de funções biológicas necessárias para a vida. A expressão gênica é o processo de sintetizar produtos a partir do código genético. Funções biológicas complexas são controladas por uma regulação rígida dessas expressões gênicas interdependentes. Os organismos mostram padrões com alta interação regulatória entre um grande número de genes, sendo um desafio inferir a rede de genes reguladores a partir de um pequeno número de perfis de expressão gênica(AALTO et al., 2020). Inúmeras ferramentas e estudos já foram feitos buscando a melhor forma de inferir essas redes.

O estudo de redes baseia-se no conceito de vértices e operadores que conectam esses vértices. Cada vértice representa um gene, e pode ser ativado ou inibido por um ou mais vértices, formando uma rede. Estes estudos se baseiam no estudo de genes utilizando operadores booleanos para representar as ligações entre os vértices (SHMULEVICH; DOUGHERTY; ZHANG (2002)). Entretanto, para inferir os modos como elas operam, existem poucos dados experimentais no tempo (por exemplo, análises feitas de hora em hora, por 3 horas), e muitas vezes tem-se poucas características fenotípicas expressas, que dependem de muitos genes(MERCATELLI et al., 2020).

Nos últimos anos, muitos estudos estão voltados para o estudo de grandes datasets, com milhares de redes. São analisados com técnicas de aprendizado de máquina (BILGEN; SARAC (2018)), Monte Carlo (LOW et al. (2014)), cadeias de Markov (RAM; CHETTY (2011)) ou separação de cluster (AUGUSTINE; JEREESH (2017)). Outras técnicas envolvem meta-heurísticas, como algoritmos genéticos (JIMENEZ; MARTINS; SANTOS (2014)), evolução diferencial (NOMAN; IBA (2007)) e programação genética(SAKAMOTO; IBA (2001). Também existem técnicas mistas de aprendizado de máquina com meta-heurísticas (XU; WUNSCH; FRANK (2007)). Uma revisão das técnicas de redução de complexidade do problema pode ser encontrada em PINDAH et al. (2015). Uma técnica possível é realizar a inferência de uma rede gênica procurando qual a configuração do sistema resultaria em menor valor para entropia, baseado nas observações disponíveis. O presente trabalho parte do trabalho de doutorado de Fabrício Lopes (LOPES (2011)), no qual estudou a inferência de redes gênicas através da entropia. Mais especificamente, mostrou que a entropia de Shannon não é a mais indicada para o estudo das redes gênicas, mas sim a entropia de Tsallis TSALLIS (1988).

Em 1988, o físico greco-brasileiro Constatino Tsallis sugeriu uma generalização da entropia TSALLIS (1988), numa teoria que ficou conhecida como "Entropia não-extensiva", ou "Entropia de Tsallis", formalizando o estudo da entropia para sistemas fora do equilíbrio. As diferenças entre estas entropias está descrita em detalhes no

capítulo 1. Tsallis mostrou que a generalização da entropia pode ser feita com um parâmetro q, que neste trabalho será chamado de parâmetro não-extensivo, o qual, no limite $q \leftarrow 1$, leva à entropia de Boltzmann-Shannon. Mudar o valor de q significa mudar a escala de entropia utilizada, o qual poderia, em princípio, alterar as situações onde a entropia está maximizada ou minimizada. Esta dissertação tenta responder à pergunta:

"Se a informação pode ser medida como uma grandeza física, e esta grandeza possui várias escalas diferentes, qual é a mais adequada para um determinado sistema?". A busca pela resposta para esta pergunta está no capítulo 2.

Com o melhor valor do parâmetro q para um dado sistema, é possível utilizar uma função critério para se inferir um sistema, como descrito no capítulo 3, o qual também possui um exemplo com portas lógicas digitais. O capítulo seguinte apresenta exemplos voltados para o estudo das redes gênicas, primeiramente com redes artificiais, como no trabalho de doutorado LOPES (2011), e publicado no artigo LOPES; OLIVEIRA; CESAR (2011), o qual obteve o mesmo valor de melhor q que encontramos analiticamente no presente trabalho. Finalmente, são utilizados dados processados do experimento DREAM4, onde conjuntos diferentes de genes são testados com discretização binária e ternária. Após estas análises, são feitas considerações sobre outras hipóteses levantadas para se analisar os resultados. Por fim, algumas conclusões e possibilidades de novas aplicações são discutidos em 4.2.3.

Capítulo 1

Entropia e Informação

1.1 Introdução à Entropia

A entropia surgiu no contexto da Teoria da Termodinâmica, inicialmente dentro do estudo de máquinas térmicas (NUSSENZVEIG (2018)). Sua definição original é:

$$S = \frac{dQ}{T}, \tag{1.1}$$

que mostra que a entropia S é a quantidade de calor infinitesimal dQ que foi transferida ou absorvida por um sistema à temperatura T. Esta grandeza surgiu no estudo de ciclos termodinâmicos, sejam eles de motores ou de refrigeradores, para classificar processos em reversíveis ou irreversíveis. Foi observado que, no caso de sistemas fechados, após um ciclo termodinâmico completo (ou seja, quando o sistema retorna ao estado inicial), a entropia somente pode aumentar ou permanecer a mesma. Matematicamente, isso pode ser escrito como (NUSSENZVEIG (2018)):

$$dS \leq 0 \tag{1.2}$$

Esta grandeza é aditiva, no caso de sistemas independentes. Por exemplo, ao somar um sistema A com um sistema B, a entropia do sistema resultante é:

$$S^{A+B} = S^A + S^B \tag{1.3}$$

Ao longo do séc. XIX, o estudo da entropia levou à 2ª Lei da Termodinâmica, enunciada por Clausius como (FONTANA; SANTOS (2016)):

> "É impossível para uma máquina auto-atuante, sem auxílio de um agente externo, transferir calor de um determinado corpo a outro de maior temperatura"

E por lorde Kelvin:

> "É impossível por meio de um agente inanimado, obter um efeito mecânico de qualquer quantidade de matéria através de seu resfriamento abaixo da temperatura do objeto de menor temperatura do entorno"

Ainda no século XIX, com os avanços da Teoria Cinética dos Gases, foi possível relacionar o movimento molecular com grandezas macroscópicas (NUSSENZVEIG (2018)). Por exemplo, considerando um modelo onde o gás, composto por uma grande quantidade de moléculas ou átomos (da ordem do número de Avogrado, $6,023.10^{23}$) em constante movimento, que colidem quando se encontram, trocando energia e *momentum* (momento linear, $\vec{p} = m\vec{v}$), de modo que haverá uma distribuição de velocidades das partículas, com pequenas quantidades em velocidade baixa, ou alta, e a maioria com velocidade em torno de uma média (v_{RMS}, *root mean square velocity*). Com isso, é possível mostrar que a temperatura T do gás é (NUSSENZVEIG (2018)):

$$T = \frac{mv_{RMS}^2}{3k} \tag{1.4}$$

onde m é a massa de cada partícula (átomo ou molécula), e $k \approx 1,38 \cdot 10^{-23} J/K$ é a constante de Boltzmann. Ou seja, a velocidade média das partículas de um gás define a temperatura macroscópica medida daquele gás. Na Figura (1.1) é possível observar diferentes distribuições para diferentes temperaturas.

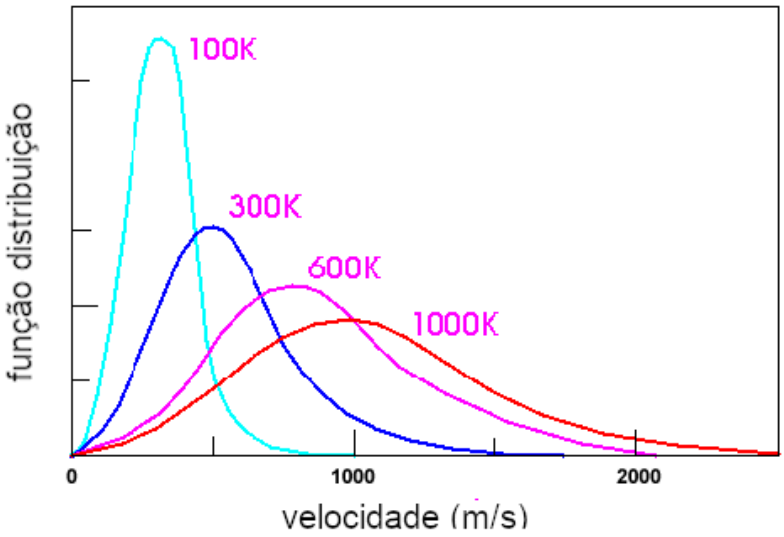


Figura 1.1: *Gráfico de funções distribuição em função da temperatura, para as temperaturas 100, 300, 600 e 1000 Kelvin.*

A distribuição de velocidades das partículas pode ser escrita por uma *função de distribuição* $F(\vec{v})$. Por exemplo, a porcentagem do número de partículas que possuem

velocidade entre $\vec{v}_0$ e $\vec{v}_0 + d\vec{v}$ é dado por $F(\vec{v}).d\vec{v}$. Pode-se inferir que:

$$\int_{-\infty}^{\infty} F(\vec{v})d\vec{v} = 1 \tag{1.5}$$

Entre o final do séc. XIX e começo do séc. XX, os físicos Ludwig Edward Boltzmann e Josiah Willard Gibbs, baseados nos trabalhos de James Clerk Maxwell, desenvolveram a base da chamada Mecânica Estatística (REIF (1965),CALLEN (1985)). Ela permite estudar a dinâmica de sistemas de muitos corpos de forma analítica, através do estudo da evolução dos *estados* termodinâmicos. Um *estado* pode ser definido como uma unidade de classificação de um determinado sistema, ou de um subconjunto dele, ou de um único constituinte, através de características em comum dos constituintes daquele estado. Por exemplo, pode-se utilizar as grandes pressão, temperatura e volume para se descrever o estado de um determinado gá, macroscopicamente, porém, no caso de um grupo de partículas do mesmo gás, é mais conveniente descrever um estado utilizando a distribuição de velocidades. Com isso, a distribuição $F(\vec{v})$ pode ser considerada como uma distribuição dos estados possíveis das partículas do sistema.

É possível relacionar as grandezas macro com as microscópicas, assim como no caso da temperatura explicado anteriormente: a pressão pode ser relacionada com a troca de *momentum* pelas colisões entre as partículas e a parede do recipiente que contém o gás, e o volume pode ser obtido pelas dimensões do mesmo recipiente. Para a entropia, existe uma relação com os estados das partículas microscópicas, a qual é possível de se obter utilizando um raciocínio similar ao de Boltzmann (NUSSENZVEIG (2018)), a partir da eq. (1.1). Do ponto de vista da Mecânica Estatística, a temperatura está relacionada com a distribuição de estados das partículas do sistema. Já o calor representa a transferência de energia do sistema, fenômeno que altera, macroscopicamente, dois ou mais parâmetros de estado (por exemplo, pressão e temperatura). Microscopicamente, as alterações dos parâmetros macroscópicos indicam uma alteração na função distribuição $F(\vec{v})$.

A outra componente da entropia é dada pela eq. (1.2), que diz que esta grandeza sempre aumenta, o que representa, microscopicamente, que as alterações dos estados são direcionadas. Unindo os conceitos, a entropia representa a possibilidade de alterações de estados num sistema que tem uma determinada distribuição. Um sistema com alta entropia representaria um sistema que qualquer estado é possível, enquanto uma entropia baixa representa um sistema com poucas alterações possíveis.

Para descrever a entropia matematicamente, serão supostos dois sistemas, A e B, com quantidade diferente de partículas N_A e N_B, respectivamente, e cada partícula pode estar em qualquer um dos W_i estados possíveis, sendo W estados permitidos. Também considera-se que as interações entre partículas são de curto alcance ou ine-

xistentes (por exemplo, moléculas num gás que se chocam ocasionalmente), e que o número de partículas é muito grande, a probabilidade de se encontrar uma partícula do sistema A num estado W_i é igual para qualquer estado. Nesse caso, o número total de possibilidades de configuração do sistema é:

$$W_A = \prod^{N_A} W = W^{N_A} \tag{1.6}$$

Com isso, a probabilidade de se encontrar o sistema em um determinado estado F_A, nessas condições, é:

$$F_A = \left(\frac{1}{W}\right)^{N_A} \tag{1.7}$$

O mesmo ocorre com o sistema B. Soma-se as partículas dos sistemas A e B, criando-se um novo sistema, que terá probabilidade:

$$F_{A+B} = \left(\frac{1}{W}\right)^{N_A} \cdot \left(\frac{1}{W}\right)^{N_B} = \left(\frac{1}{W}\right)^{N_A+N_B} \tag{1.8}$$

Relacionando essa alteração na probabilidade com as propriedades da entropia, o físico Ludwig Boltzmann propôs uma expressão matemática para explicar como a adição de dois sistemas independentes causa uma multiplicação na probabilidade total, enquanto a entropia é somada (eq. 1.3), utilizando a função logarítmica[1]:

$$S = k \ln W \tag{1.9}$$

onde k é a constante de Boltzmann. Esta expressão permite utilizar o conceito de entropia em diversos sistemas, além de gases, como inicialmente feito no contexto da termodinâmica de ciclos de motores térmicos. Inclusive, ela funciona tanto para sistemas com uma distribuição contínua de estados possíveis (como no caso da velocidade de gases), como no caso de estados discretos (por exemplo, um sistema onde queremos saber se determinada partícula está do lado direito ou esquerdo de uma caixa). Este trabalho lidará com sistema com um estados discretizados. O caso descrito acima, da eq. (1.6), não é o caso mais geral, pois as probabilidades não precisam ser iguais de se encontrar as partículas num determinado estado. Independente da situação, sempre teremos a relação de normalização:

$$\sum_i^W p_i = 1 \tag{1.10}$$

[1] Pode-se utilizar um logaritmo qualquer (log) ou o logaritmo natural (ln).

Nesse caso geral, a fórmula da entropia precisa ser ponderada pela probabilidade de se encontrar cada estado:

$$S = -k \sum_{i}^{W} p_i \ln p_i \tag{1.11}$$

Pode ser observado que no caso em que todos os estados são igualmente possíveis ($p_i = 1/W$), retorna-se à eq. (1.9). No estudo da entropia é comum utilizar uma formulação adimensional, ou seja, a grandeza é calculada como S/k, o que será utilizado neste trabalho. Como exemplo, se é estudado um sistema binário, existe uma probabilidade p_0 de encontrar uma partícula no estado 0, e uma probabilidade p_1 de encontrá-la no estado 1. O número de estados discretos é $W = 2$, e pode-se calcular a entropia de um conjunto de 0s e 1s pela expressão acima:

$$S = -\left(p_0 \ln p_0 + p_1 \ln p_1\right) \tag{1.12}$$

com a condição:

$$p_1 = 1 - p_0 \tag{1.13}$$

Em um conjunto muito grande de elementos binários criado de forma totalmente aleatória, espera-se que a distribuição de probabilidades seja $p_0 \approx p_1 \approx 1/W = 50\%$, que resulta numa entropia:

$$S = -\left(\frac{1}{2}\ln\frac{1}{2} + \frac{1}{2}\ln\frac{1}{2}\right) = \ln 2 \tag{1.14}$$

Para qualquer outra distribuição diferente destas probabilidades, a entropia do sistema será menor, como pode ser observado na Figura (1.2). Ou seja, a configuração totalmente aleatória representa a entropia máxima do sistema, que é o mesmo resultado da eq. (1.9) ($lnW = ln2$). Se utilizarmos uma sequência de números binários para representar uma informação, como ocorre na informática, esperar-se-á numa mensagem que possua informações ordenadas que a probabilidade dos estados 0 e 1 sejam diferentes de 50%. Ou seja, é possível associar uma entropia menor ou maior com um sistema com maior ou menor quantidade de informação, respectivamente. Isto foi mostrado matematicamente por Shannon, em 1948 (SHANNON (1948)).

1.2 Entropia de Tsallis

As expressões mostradas até o momento para a entropia apresentam condições específicas para serem aplicadas(HANEL; THURNER, 2011):

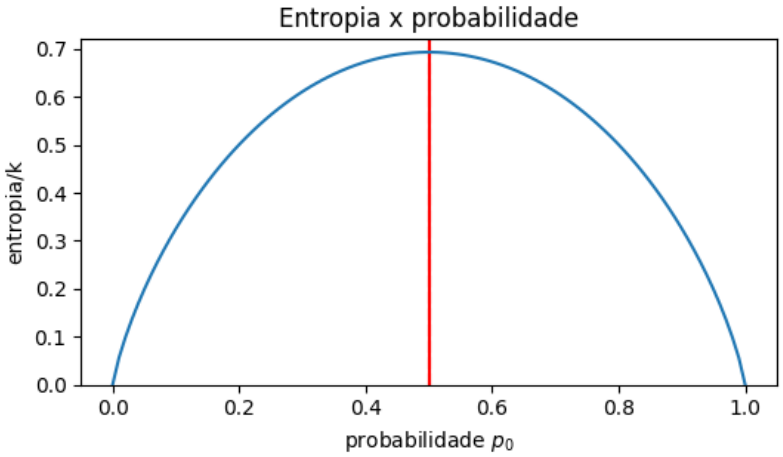


Figura 1.2: *Gráfico da entropia em função da probabilidade p_0, para um sistema binário. A linha vermelha indica o caso $p_0 = 50\%$.*

1. Número muito grande de partículas, para que pelo menos ocorra uma partícula em todos os estados possíveis.

2. Para ocorrer a condição acima, o sistema precisa estar no equilíbrio, ou próximo dele, para que todos os estados possíveis sejam ocupados conforme o tempo passa, e o sistema chegue à entropia máxima possível.

3. Não podem ocorrer interações de longo alcance, pois nesse caso a relação entre entropia e probabilidades não se comportaria como um logaritmo.

Existem muitos sistemas que não atendem às condições acima: No caso do primeiro e segundo itens, sistemas ordenados com regras de ocupação de estados (por exemplo, supondo um gás numa caixa com um membrana que só permite o movimento em uma direção, ou, sistemas biológicos), como descrito nos primeiros capítulos do livro (TSALLIS, 2009a). Sobre o terceiro item, existem sistemas físicos com interações de longo alcance, por exemplo, sistemas gravitacionais ou com interação elétrica (incluindo sistemas quânticos com potencial elétrico), onde a força depende de $1/r^2$, com r sendo o raio, o que não permite calcular a entropia do sistema utilizando a formulação de Boltzmann(KODAMA et al., 2005). Para incluir esses casos, algumas generalizações foram propostas ao longo das décadas (AMIGÓ; BALOGH; HERNÁNDEZ, 2018), e uma que se destacou foi criada pelo físico greco-brasileiro Constantino Tsallis em 1988 (TSALLIS (1988)). Sua formulação matemática pode ser obtida iniciando-se pela definição da função exponencial, de Euler, para uma exponencial de x:

$$\lim_{h\to\infty}\left(1+\frac{x}{h}\right)^h \equiv e^x \tag{1.15}$$

Pode-se chamar a função dentro do limite de uma função de h:

$$e_h^x = \left(1 + \frac{x}{h}\right)^h \tag{1.16}$$

Trocando o parâmetro h por $h = 1/(1-q)$, obtemos a *q-exponencial*:

$$e_q^x = (1 + (1-q)x)^{\frac{1}{1-q}} \tag{1.17}$$

de modo que a definição (1.15) pode ser reescrita como:

$$\lim_{q \to 1} (1 + (1-q)x)^{\frac{1}{1-q}} \equiv e^x \tag{1.18}$$

O inverso da função q-exponencial é o *q-logaritmo*:

$$\log_q x = \frac{1 - x^{1-q}}{1-q} \tag{1.19}$$

Alterando-se o valor do parâmetro q, obtém-se diferentes funções, e quando $q \to 1$, volta-se às conhecidas funções exponencial e logaritmo, as quais foram então generalizadas. Substituindo a eq. (1.19) na eq. (1.11), encontramos a função conhecida como Entropia de Tsallis:

$$S_q = \frac{1 - \sum_{i=1}^{W} {p_i}^q}{1-q} \tag{1.20}$$

de modo que $S_q(q \to 1) = S$. É possível ver o comportamento desta função para diferentes valores do parâmetro q, no caso de um sistema binário, na Figura (1.3).

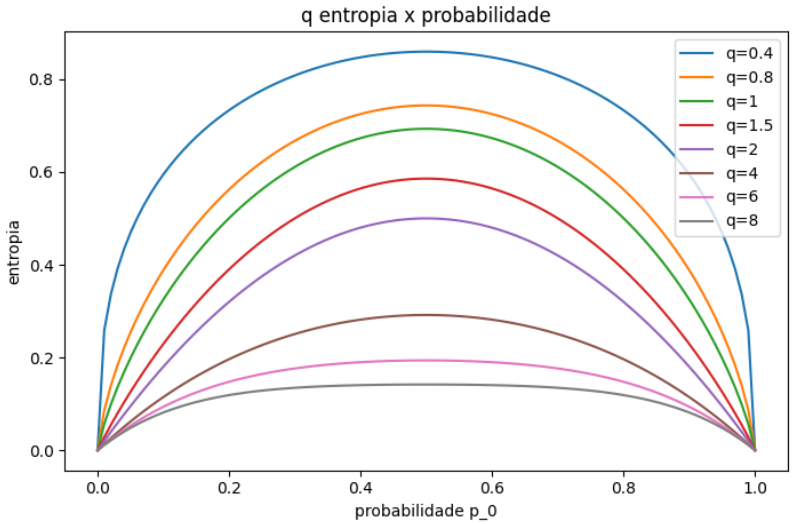


Figura 1.3: *Entropia em função da probabilidade p_0, para diferentes valores de q.*

Matematicamente, é uma possível generalização da entropia de Boltzmann, mas

restaria a dúvida se está relacionada à teoria da informação e mecânica estatística. Nas últimas décadas esta relação já foi amplamente demonstrada, com vários exemplos, como descrito no livro TSALLIS (2009a). Um exemplo de sua capacidade de generalizar, ou seja, expandir os possíveis sistemas estudados com entropia, pode ser encontrado no trabalho de KODAMA et al. (2005), onde foi mostrado que sistemas com interações cuja força age com $1/r^2$, podem ser estudados com entropia de Tsallis e a q-exponencial, o que inclui sistemas gravitacionais e quânticos (por terem interação elétrica) (AMADOR; ZAMBRANO (2010)). Mais recentemente, em TSALLIS (2019), foi feito um review sobre as aplicações de suas teorias em vários ramos da Física.

No artigo de HANEL; THURNER (2011), é explicado que uma entropia precisa satisfazer os 4 axiomas de Shannon-Khinchin, mas que as generalizações alteram o quarto axioma, que é da extensividade, ou seja, a eq. (1.3). No caso da entropia de Tsallis, ao se somar dois sistemas que possuem inicialmente entropias S_q^A e S_q^B, a entropia do sistema resultante deve ser calculada como(HANEL; THURNER, 2011):

$$S_q^{A+B} = S_q^A + S_q^B + (1-q)S_q^A S_q^B \tag{1.21}$$

de modo que no caso $q \to 1$ retorna para a eq. (1.3).

1.3 Medida de Informação

A relação entre entropia termodinâmica, entropia da informação, e quantidade de informação não é clara em muitos textos. Apesar de não ser o foco deste trabalho, nesta seção tentaremos explicar melhor esta relação. Alguns trabalhos tentam analisar em pormenores esta relação, como o de BÉRUT et al. (2012) e MARUYAMA; NORI; VEDRAL (2009). Primeiramente, podemos definir alguns termos[2]:

- *Ordem (de um sistema)*: consideramos que um dado sistema, composto por várias componentes, está ordenado quando existem relações entre as componentes que obedecem a alguma lei, ou dinâmica, específica, e o sistema teve tempo para que cada componente esteja em um determinado estado que não irá se alterar.

- *Desordem (de um sistema)*: um sistema desordenado pode ser um sistema onde as componentes não possuem interações entre si, portanto podem estar em qualquer estado possível, de forma aleatória, ou também pode ser um sistema que possui interações, mas que as componentes ainda não tiveram tempo de atingir uma configuração de equilíbrio.

[2]procurou-se a definição destes termos em artigos e livros, mas não foram encontrados de forma satisfatória. Optou-se por descrevê-lo com as próprias palavras do autor, para que fiquem claros nas próximas seções, quando utilizados

- *Quantidade de Informação (de um sistema)*: tem duas interpretações, complementares. Uma delas é que informação é a quantidade de estados diferentes que um determinado sistema possui (por exemplo, informação em um sistema que 10 componentes são a letra "A"é diferente da informação de um sistema onde 8 componentes são "A"e 2 são "B", com as posições de cada letra indicando diferentes informações). A outra interpretação é, baseado nas informações anteriores, qual a probabilidade de se encontrar a próxima componente em um determinado estado[3].

Pelas definições consideradas aqui, pode-se analisar se um sistema está ordenado ou não, pela quantidade de informação que este mostra. Se um sistema possui componentes distribuídas aleatoriamente, consideramos que ele está mais desordenado do que um sistema que apresenta apresenta componentes ordenadas. Qual dos dois casos transmite mais ou menos informação? Pela definição da entropia, quanto maior a distribuição de estados possíveis, maior é a entropia. Pelo exemplo de uma sequência de letras, quanto maior a distribuição de todas as letras possíveis em tal sequência, indicaria uma distribuição mais próxima da aleatória, ou seja desordem, e numericamente, resultaria um valor maior de entropia. Já uma sequência de letras que envolvesse regras de escrita, teria poucas variações entre as letras, indicando a existência de uma ordem, e numericamente, menor entropia.

No próximo capítulo será explorado como utilizar a entropia de Tsallis para se inferir a configuração de um sistema.

[3]Essa interpretação está relacionado com a fórmula de Bayes(PUGA; KRZYWINSKI; ALTMAN (2015)), que não será tratada neste trabalho.

Capítulo 2

Entropia Relativa

No capítulo anterior foi apresentada brevemente uma relação entre entropia e informação. Um sistema onde todos os estados são igualmente possíveis apresenta entropia máxima, independente do valor do parâmetro q, o que ocorreria numa distribuição infinita e aleatória. Dada uma sequência, qualquer estrutura organizada que ali existir irá alterar a probabilidade dos estados, e a entropia será menor do que no caso totalmente aleatório. Um sistema com regras, ou seja, ordenado, terá entropia menor do que um sistema desordenado. Esse resultado pode ser usado de forma inversa: dada uma sequência de dados, como utilizar a entropia para ali encontrar possíveis estruturas? Uma possibilidade é supor determinados padrões esperados de ordem do sistema, e calcular aquele padrão que possui menor entropia, a qual será consequentemente a configuração mais ordenada. Esse processo é chamado de *inferência*, e implica algumas propriedades do sistema (COX, 2006).

2.1 Inferência da Configuração de um Sistema

Um *sistema* consiste em um conjunto de elementos, iguais ou diferentes, que podem possuir interações uns com os outros. As *interações* são um conjunto de regras que relacionam os elementos do sistema, por exemplo:

- Em um gás ideal, a interação entre átomos é de curto alcance, ocorrendo somente durante a colisão de uma partícula com outra, ou com a parede (NUSSENZVEIG, 2018). Num simulador, a regra associada poderia ser descrita como "quando dois átomos estiverem numa distância X, resolver equações de conservação de momentum linear";

- Em um ambiente intracelular, pode existir a interação entre o DNA e uma proteína. O processo pode ocorrer com diferentes regras, uma delas sendo a tradução

do DNA em RNA, e a transcrição do RNA em aminoácidos, e finalmente estes se unem para formar uma proteína (ALBERTS, 2008);

- Na linguagem portuguesa moderna escrita, a interação entre letras é baseada em regras ortográficas. Por exemplo, é possível existir a combinação de letras "ça", porém a combinação "çs" não existe.

Os exemplos acima, que possuem regras, e portanto, possuem interações entre elementos do sistema, não são os únicos na natureza, onde ocorrem muitos sistemas que não têm interações entre os elementos, por exemplo, uma sequência de números aleatórios (COX, 2006, Cap.1). Essas interações também podem ser classificadas como de curto ou de longo alcance. Nos exemplos acima, o primeiro item seria uma interação de curto alcance, pois ocorre somente durante a colisão. Já o segundo item seria uma de longo alcance, pois mesmo as proteínas, assim como qualquer outro elemento no caminho entre DNA e proteína, pode influenciar nas traduções daquele trecho específico de DNA. Outro exemplo de interação de longo alcance, na física, é a força gravitacional.

Se *inferir* significa encontrar informações sobre um sistema, a *inferência estatística* (COX, 2006) é o ato de procurar informações do sistema a partir de um conjunto de dados menor do que o total existente. Isso implica que ele possua alguma ordem ou organização, e que os dados disponíveis sejam uma amostra representativa das interações entre os elementos. Como exemplo, supõe-se que se encontra um pedaço de papel, com uma sequência de símbolos:

$$\triangleright \quad \zeta \oslash \triangleright \quad \bowtie \triangleright \Uparrow\Uparrow \triangleright \quad \aleph \ddagger \quad \amalg \oslash \aleph \oslash \quad \aleph \mp \oslash \quad \bowtie \triangleright \Uparrow \quad \aleph \oslash \mp \quad \models \triangleright \prec \ddagger$$

Observa-se que existem espaços separando grupos de elementos, e alguns símbolos aparecem mais do que outros, o que poderia indicar que a sequência apresentada possui uma ordem. Não se sabe, a princípio, se essa sequência mostra todos os elementos e combinações possíveis daquele sistema.

Para inferir se existe alguma informação significativa, pode-se supor que a ordenação dos elementos obedece a regras ortográficas e gramaticais de alguma língua. O próximo passo é fazer suposições sobre qual conjunto de regras (ou seja, de interações), esta sequência obedece. Após escolher uma língua, pode-se testar se a sequência possuiria combinações comuns naquela língua, o que indicaria uma possível ordem. Outras línguas poderiam ser testadas, e a que apresentasse uma maior ordem, poderia indicar a correta.

Para realizar a análise, pode-se supor que se trata de uma frase escrita na língua portuguesa, de modo que é possível tentar inferir letras, sílabas e palavras para encontrar o significado completo. Devido à pequena amostra de símbolos, dificilmente

será possível inferir todas as letras da frase, mas ainda é possível buscar a melhor inferência, ou seja, procura-se uma configuração que forneça mais informação sobre o sistema.

A configuração de maior desordem (maior entropia), onde todas as letras estejam aleatórias, não é o mais provável, visto que, como já apontado, algumas letras se repetem, e a frase precisa estar dentro das normas e padrões da língua portuguesa. Regras gramaticais e ortográficas também impõe limitações aos estados possíveis. Por exemplo, na sequência acima proposta, o símbolo sozinho (inicial) só poderia representar as letras "a", "e" ou "o". Na quarta palavra, com duas letras, só pode iniciar com as letras "a", "e", "o", "d", "m" ou "n". Como esses símbolos aparecem mais vezes em outras posições, poderia se buscar em palavras com estrutura similar, até encontrar aquelas que garantam uma maior informação, e consequentemente encontrar a configuração correta (ou as configurações corretas, dependendo do resultado) para o significado da frase. Símbolos que só aparecem uma vez, como os presentes na última palavra, disponibilizam pouca informação pela sua escassez, e aumentam a entropia do sinal.

Outra língua poderia ser testada, como inglês, ou russo. Encontrando a língua correta, seria possível até mesmo prever quais combinações de símbolos não poderiam ser observadas. Ao final do processo, neste exemplo, a inferência passou por duas grandes etapas: encontrar a língua correta, que garantiria maior ordem, e encontrar a associação de símbolos e letras que possui maior significado. Uma outra forma de expressar essa constatação, mais geral, é dizer que se busca qual a melhor métrica (no caso do exemplo, a língua) para se obter a informação de um sistema, e quais interações ocorrem entre os elementos que poderiam prever futuros comportamentos. Numerica e analiticamente, uma função que permite obter valores mensuráveis da quantidade de informação, como a entropia, é uma candidata para se buscar a inferência de um sistema.

2.2 Entropia Relativa

Para a entropia de Tsallis, a métrica vem da escolha do parâmetro q, portanto uma possibilidade para se escolher o melhor q seria procurar aquele que resulta em menor entropia dos sistema (maior ordem). Entretanto, não é possível obter resultados com a comparação da aplicação direta da equação (1.20), como pode ser observado na Figura 1.3. Independente da configuração de probabilidades, o valor da entropia é sempre maior ou sempre menor para diferentes valores de q. Logo, entre entropias com diferentes valores de q, não se aplicaria o conceito de que uma entropia menor representa um sistema mais organizado do que um sistema com maior entropia. Para

permitir a comparação entre entropias com diferentes valores de q, uma das propostas é utilizar variações da entropia, como a entropia condicional, como será descrito no cap. 3, que também não possui uma forma única de ser aplicada A. TEIXEIRA A. SOUTO (2014).

Encontrar a melhor inferência do sistema significa encontrar a configuração do sistema, com o melhor parâmetro q, que permita obter a maior quantidade de informação. Essa última sentença, "Obter a melhor quantidade de informação", implica saber qual seria o caso com pior quantidade de informação, que seria o caso totalmente aleatório. Utilizando uma analogia com "sinal/ruído", melhorar a inferência significa diferenciar de forma mais acurada o sinal do ruído. O sinal, neste caso, é a entropia medida para um dado conjunto de probabilidades, como escrita em 1.20. O ruído máximo ocorreria quando todas as possibilidades de um sistema estão presentes, ou, em outras palavras, quando as probabilidades de todos os estados são as mesmas $(1/W)$. A entropia máxima (ruído) pode ser escrita como:

$$S_q^{max} = \frac{1 - \sum_{i=1}^{W} W^{-q}}{1-q} = \frac{1 - W^{1-q}}{1-q} = \log_q W \tag{2.1}$$

Neste trabalho, propomos a utilização de uma outra grandeza, a qual será chamada de *Entropia Relativa*, representada pelo símbolo S_q^{rel}, definida como a razão entre a entropia da configuração e a entropia máxima:

$$S_q^{rel} = \frac{S_q}{S_q^{max}} = \frac{1 - \sum_{i=1}^{W} {p_i}^q}{1 - W^{1-q}} \tag{2.2}$$

Esta função é válida para $q \neq 1$. No caso $q = 1$:

$$S_1^{rel} = \frac{\sum_{i=1}^{W} p_i \ln p_i}{\ln W} \tag{2.3}$$

Encontrando o valor de q que minimiza essa função (se este mínimo existir), é possível ter uma melhor inferência para aquele sistema.

2.3 Encontrando o melhor valor de q

Antes de procurar o valor de q para minimizar a função descrita acima, é útil introduzir uma notação para cada sistema, ou, em outras palavras, para cada configuração de probabilidades. Será atribuído o nome de *função de distribuição discreta* $F_W = (p_0, p_1, ..., p_W)$, para um sistema com W estados discretos possíveis, sempre obedecendo à regra $\sum_i^W p_i = 1$. Por exemplo, um sistema binário pode ser caraterizado como uma função $F_2 = (p_0, p_1)$.

É necessário cuidado quando se trabalha com estas funções q-logarítmicas ou q-exponenciais, pois numericamente elas divergem quando q tende a 1, mas convergem analiticamente, de modo que é necessário colocar mecanismos que troquem as funções para valores próximos de $q \approx 1$, por exemplo, estabelecendo um ϵ (usualmente $\epsilon < 10^{-3}$) tal que quando $|q-1| < \epsilon$, a função utilizada é trocada da (2.2) para a (2.3). Outra forma de garantir o resultado, se houver, é procurar o valor de q tal que minimize a função (2.2) para uma dada distribuição F_W, com técnicas diferentes, e compará-las. Os métodos escolhidos foram:

1. Calcular numericamente S_q^{rel} para diferentes F_W, e, caso exista um mínimo, procurar o valor de q, chamado de $\mathbf{q_{min}}$;

2. Calcular a derivada de S_q^{rel} em função do parâmetro q e igualar a zero:

$$\frac{dS_q^{rel}}{dq} = 0 \tag{2.4}$$

A derivada resulta em:

$$\begin{aligned} 0 &= \frac{-\sum_{i=1}^{W} {p_i}^q \ln\left(p_i\right)}{1-W^{1-q}} + \frac{\left(-W^{1-q}\ln(W)\right)\left(1-\sum_{i=1}^{W} {p_i}^q\right)}{\left(1-W^{1-q}\right)^2} \\ 0 &= \left(\sum_{i=1}^{W} {p_i}^q \ln\left(p_i\right)\right)\left(W^{q-1}-1\right) + \left(1-\sum_{i=1}^{W} {p_i}^q\right)\ln(W) \end{aligned} \tag{2.5}$$

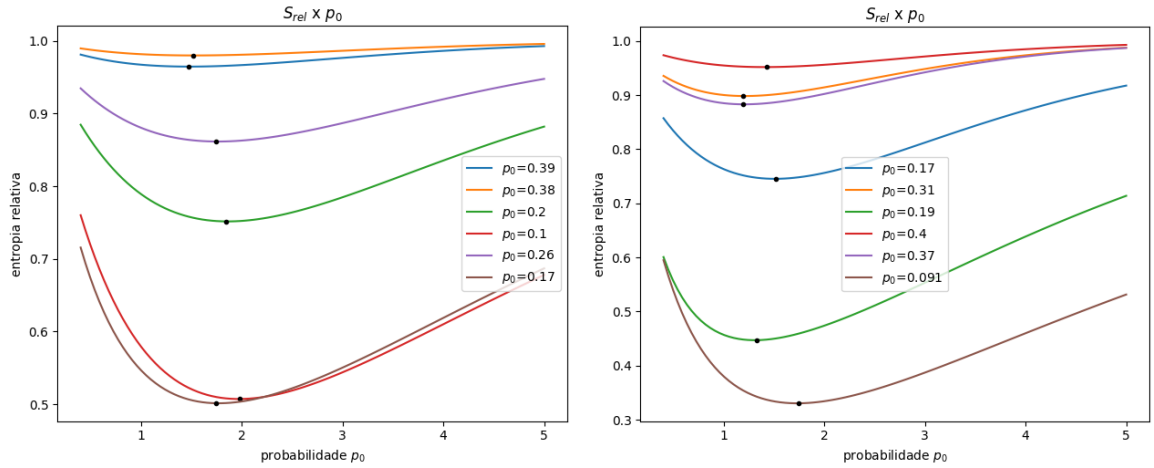
Para cada F_W diferente, pode-se então buscar um q_{min}, utilizando ferramentas de busca de raízes de funções não lineares. A análise será dividida em dois casos, um para W qualquer e outro para o caso específico de 2 estados possíveis (sistema binário).

2.3.1 Para $W > 2$

Para se estudar o q_{min} para um $W > 2$, foram criadas diferentes configurações aleatórias $F_N = (p_0, p_1, ..., p_N)$, com o seguinte método[1]:

```
p₀ = aleatório entre 0 e 0,5
p₁ = aleatório entre 0 e p₀
p₂ = aleatório entre 0 e p₁
```

[1] Outro método possível consiste em sortear N número aleatórios entre 0 e 1, utilizando a ferramenta "random.random()" do Python (VAN ROSSUM (2020)), somar todos estes número e dividir pela soma, para garantir a condição da eq. (1.10).



(a) $W = 3$ **(b)** $W = 4$

Figura 2.1: *Entropia relativa em função de q, para diferentes configurações de probabilidade. Os pontos pretos marcam o mínimo de cada curva.*

```
...
pN = 1 - soma (p0,p1,...,pN-1)
```

Como citado anteriormente, uma forma de se obter o valor de q_{min} é calcular a entropia relativa de uma determinada distribuição F_N,para vários valores de q, e procurar o menor deles. Nas Figuras 2.1a e 2.1b foram plotadas algumas curvas de diferentes F_N, para os casos $W = 3$ e $W = 4$. Cada curva está anotada pelo valor de seu p_0. Observa-se que o mínimo da entropia relativa pode variar para diferentes F_N, porém estão numa mesma faixa.

Com isso, outro método foi utilizado para encontrar um valor médio de q_{min} para cada W. Para um dado W, foram simuladas 1000 configurações, e para cada uma resolveu-se numericamente a equação (2.2). Após, foi calculada a média dos valores encontrados, e repetia-se o processo para outros valores de W. Isto foi realizado para $2 \leq W < 30$.

A solução numérica da equação (2.2) foi obtida pelo pacote Scipy (Virtanen et al. (2020)), no qual foram testadas várias funções para se obter o q_{min}. Buscou-se uma função com as características:

- Robustez (pouca diferença de valores obtidos, após ser executada várias vezes);
- Convergência (alguns métodos não convergiram para $W > 8$);
- Não depender de valores iniciais propostos de q_{min} (alguns métodos, para valores altos de W, não alteravam a resposta final do valor testado inicialmente, ou seja, os valores não se alteravam ao longo das iterações de busca).

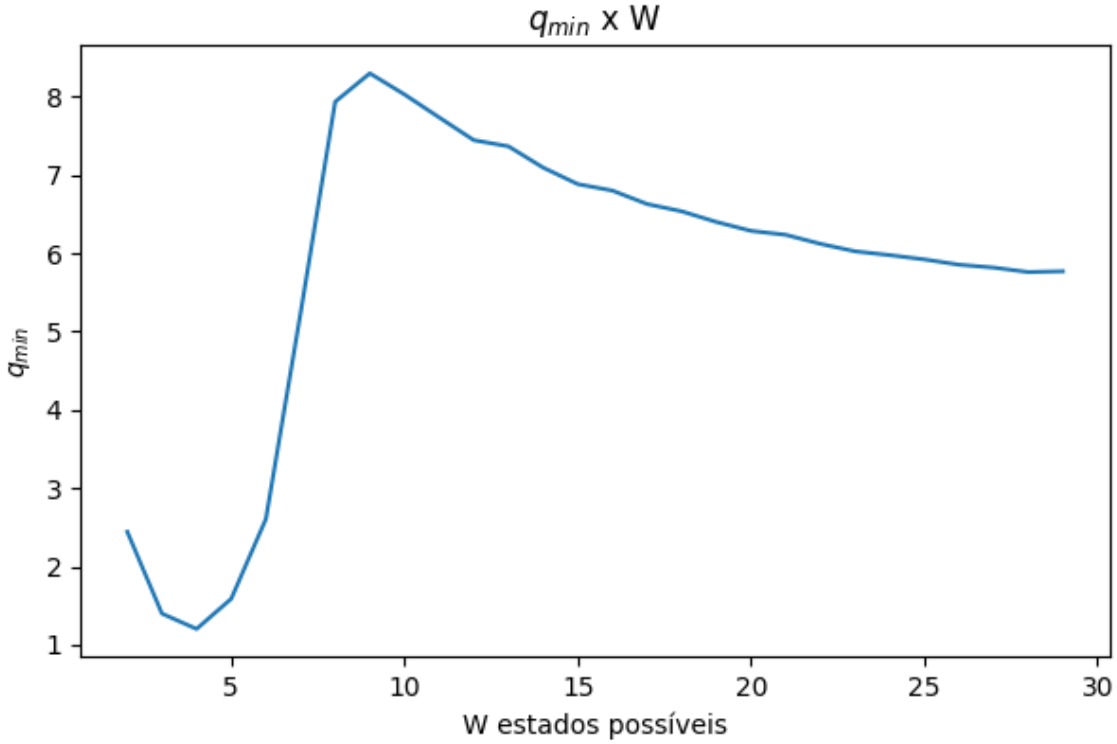


Figura 2.2: *Média do valor de q_{min} para sistemas com $2 \leq W < 30$ estados possíveis.*

Verificou-se que a função *minimize*, com o método "BFGS" (método quase Newtoniano de Broyden, Fletcher, Goldfarb, e Shannon (Nocedal; Wright (2006))) apresentava as características buscadas. O resultado pode ser observado na Figura 2.2. Alguns valores médios obtidos estão listados na Tabela 2.1. Ressalta-se que conforme os valores de W aumentam, mais configurações F_N seriam necessárias para se obter um valor médio mais confiável, por isso o próprio conceito de um valor médio pode não ser adequado, mas está mantido no presente trabalho para uma comparação.

W	q_{min}
3	1,52
4	1,22
8	8,16
15	6,91
20	6,30
29	5,73

Tabela 2.1: *Valores de q_{min} para sistemas com alguns valores de W estados possíveis.*

2.3.2 Para $W = 2$

No caso $W = 2$, o espaço de configurações possível é simétrico, devido à relação $p_0 + p_1 = 1$, e para qualquer $F_2 = (p_0, 1-p_0) = (1-p_1, p_1)$, portanto pode-se analisar a probabilidade p_0 no intervalo $0 < p_0 < 0.5$, obtendo-se todas as configurações possíveis. Do mesmo modo que no caso anterior, inicia-se a análise calculando-se a entropia relativa de algumas distribuições F_2 para diferentes valores de q, separando cada curva em função de p_0. O resultado está na Figura (2.3), que mostra que o mínimo para $W = 2$ oscila em torno de valor, $q_{min} = 2,46$.

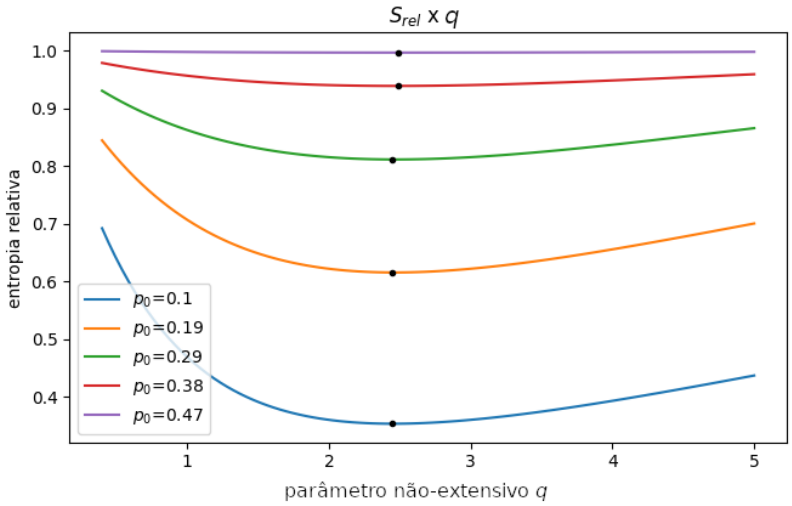


Figura 2.3: *Entropia relativa em função de q, para algumas configurações de probabilidade em sistemas binários. Os pontos pretos marcam o mínimo de cada curva.*

Utilizando a equação (2.5), para $W = 2$, é possível escrever explicitamente uma expressão em função somente de p_0:

$$0 = \left(2^{q-1} - 1\right)\left[{p_0}^q \ln\left(p_0\right) + (1-p_0)^q \ln\left(1-p_0\right)\right] + \ln(2)\left[1 - {p_0}^q - (1-p_0)^q\right] \quad (2.6)$$

Esta equação é não linear, e por isso foi resolvida numericamente. Entretanto, alguns valores extremos podem ser estudados analiticamente:

- Se $p_0 = 0$, ou se $p_1 = 1 - p_0 = 0$, os logaritmos do primeiro termo divergem para $-\infty$, portanto somente o caso $q = 1$ é possível numericamente.
- Se $p_0 \approx p_1 \approx 0,5$, a equação acima é nula para qualquer valor de q, como esperado, já que é o caso de entropia máxima.

Para os demais casos, foi utilizado um método numérico, diferente do método "BFGS", utilizado anteriormente, mais especificamente o pacote *curve_fit* do Scipy,

com o método "trf" (*Trust Region Reflective*) de mínimos quadrados BRANCH; COLEMAN; LI (1999). Ele apresentou vantagens por não exigir um valor inicial de busca, e sim uma faixa de valores para encontrar a solução. Além disso, para $W = 2$, ele apresentou resultados mais robustos do que o *minimize*. Foi escolhida a faixa $0,1 \leq q_{min} \leq 8$, faixa que se suspeita estar o mínimo.

Foi possível obter o valor de q_{min} para qualquer configuração $F_2 = (p_0, 1 - p_0)$ no intervalo $0,001 < p_0 < 0,499$, como mostrado na Figura 2.4. Para valores $p_0 > 0.47$, o valor de q_{min} sobe rapidamente, pois a configuração se aproxima numericamente do caso $p_0 \approx p_1 \approx 0,5$, devido à dificuldade inerente de cálculo de raízes em funções exponenciais, como é o caso da eq. (2.6). Considerando o intervalo $0,001 \leq p_0 \leq 0,47$, o gráfico mostra que, para sistemas binários, o valor médio é $2,460$, e q_{min} está na faixa:

$$\boxed{2.423 < q_{min} < 2,472 \quad \text{para } W = 2} \tag{2.7}$$

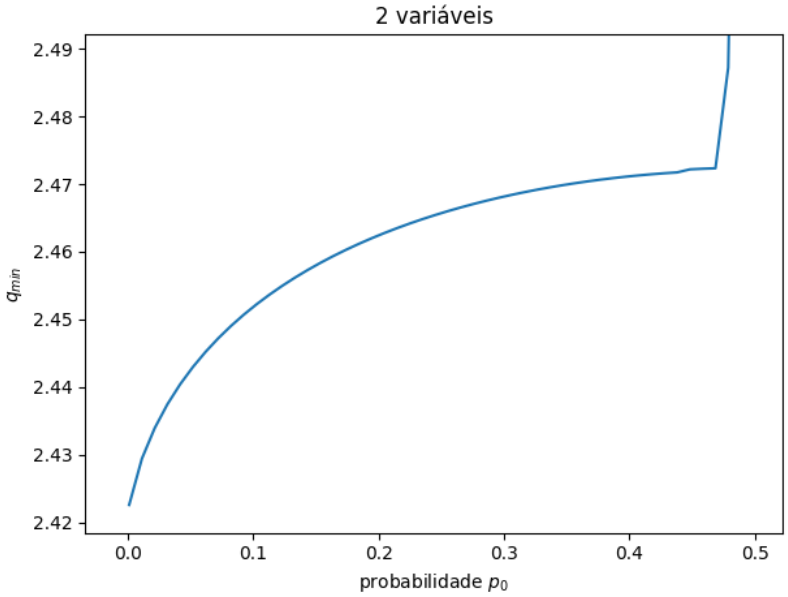


Figura 2.4: *Valor de q_{min} para diferentes configurações de probabilidade de um sistema binário. Os valores de $p_0 > 0,5$ são simétricos.*

Esse resultado também pode ser observado pela inclinação da curva de entropia por probabilidade, como mostrado na figura 1.3. Na figura 2.5 é mostrado um gráfico similar, com a probabilidade indo de $0,01$ até $0,5$, e a entropia normalizada com o maior valor. Também é mostrada a derivada (inclinação) das curvas, e é possível observar que numa faixa grande de valores, $0,15 < p_0 < 0,5$, a curva com maior valor para a inclinação é a $q = 2.5$. Isto indica que uma pequena diferença de probabilidade causa uma maior variação da entropia, o que a torna mais indicada para a inferência,

onde busca-se uma entropia que quaisquer valores que estejam afastados do aleatório ($p_0 = 0,5$) tenham uma grande diferença no valor da entropia, para indicar mais facilmente a presença de uma estrutura ordenada.

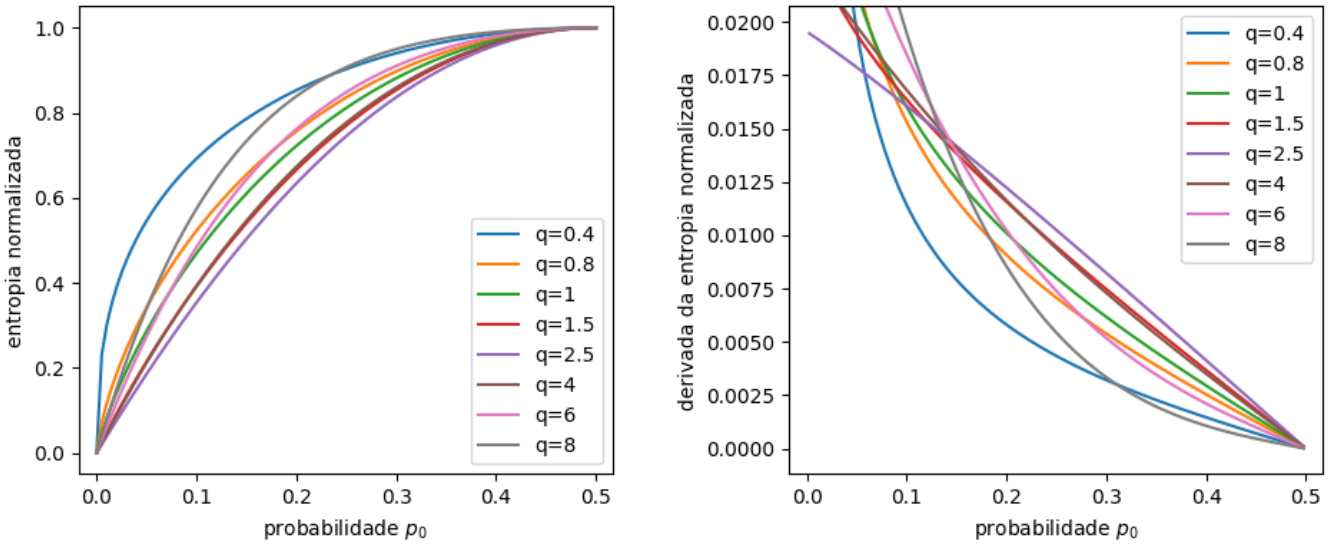


Figura 2.5: *Esquerda: entropia normalizada. Direita: derivada da entropia normalizada. Ambos em função da probabilidade* p_0.

Nos próximos capítulos os resultados obtidos nesta seção serão comparados com sistemas de interesse para redes gênicas, primeiramente com sistemas booleanos e após com um número maior de estados discretizados.

Capítulo 3

Aplicações em Inferência de Sistemas Booleanos

Os resultados obtidos no capítulo anterior podem ser testados em sistemas simulados, ou sistemas reais dos quais a dinâmica (relação entre elementos) é conhecida. Nesta seção será descrito como efetivamente a inferência pode ser feita utilizando-se uma função critério, que tem por base a entropia, mas que precisa levar em consideração alguns casos especiais. Neste capítulo, primeiramente será explicada a função critério que utilizaremos, seguido de como a interação de elementos pode ser modelada em um sistema binário, através das *portas lógicas*. Posteriormente será analisado um resultado do estudo de redes gênicas artificiais, que também utiliza uma modelagem booleana.

3.1 Inferência e Entropia Condicional

Sistemas binários são também conhecidos como booleanos, nome usado em homenagem a George Boole, que definiu um sistema de lógica algébrica (LIPSCHUTZ (2004)). Os estados booleanos são o *VERDADEIRO* e o *FALSO* (comumente descritos em inglês *True* e *False*), que podem ser representados, respectivamente, pelos números 1 e 0. Dado um conjunto com N elementos (onde cada elemento será chamado de $E_1, E_2, ..., E_N$), cada elemento consiste de uma sequência de 0s e 1s.

Pode-se utilizar o resultado do capítulo anterior para mensurar o nível de organização de um determinado sistema booleano. Por "sistema", como dito anteriormente, entende-se um conjunto de elementos, relacionados por regras, que podem ou não ser fixas no tempo. Com isso, para analisar o grau de organização de um sistema, pode ser realizado o seguinte procedimento:

- Testam-se diferentes configurações do sistema, por exemplo, supondo-se que os

E_1	0	0	1	1	0	0	1	1
E_2	0	1	0	1	0	1	0	1
E_3	0	0	0	0	1	1	1	1

Tabela 3.1: *Possíveis configurações para elementos E_1 e E_2 atuando sobre o elemento E_3*

.

		E_3	
E_1	E_2	0	1
0	0	f_{000}	f_{001}
0	1	f_{010}	f_{011}
1	0	f_{100}	f_{101}
1	1	f_{110}	f_{111}

Tabela 3.2: *Frequências das possíveis configurações*

valores expressos em um determinado elemento está relacionado com o estado de outros elementos;

- Utiliza-se uma *função critério*, baseada em entropia, que calcula um valor da organização daquela possível configuração;
- Se o valor daquela configuração estiver abaixo de um certo limiar, aquela configuração tem chance de representar uma estrutura organizada.

A relação entre elementos pode ser estudada medindo-se inicialmente as frequências de estados expressos em uma mesma posição. Por exemplo, supondo que a mistura de dois elementos (E_1, E_2) resultou num terceiro (elemento alvo E_3), todos com N estados observados, significa que procuramos conexões como mostrado na tabela 3.1, que também podem ser representadas como ($E_1E_2E_3$) (por exemplo (000), (010) etc.).

O número de eventos observados de cada configuração equivale à frequência (f_i), e podemos construir uma tabela de frequências, como mostrada na tabela 3.2. A soma de todas as frequências deve ser igual a N:

$$\sum f_i = f_{000} + f_{001} + f_{010} + f_{011} + f_{100} + f_{101} + f_{110} + f_{111} = N \tag{3.1}$$

Supondo que os elementos E_1 e E_2 são preditores do comportamento elemento E_3, a função entropia pode ser calculada sobre uma probabilidade condicional[1]. Essa

[1]É importante entender as semelhanças e diferenças entre os termos "frequência" e "probabilidade", pois estes influenciam na interpretação dos resultados obtidos pelo uso da função entropia. A frequência de uma determinada configuração é uma contagem do número de ocorrências, dividido pelo número total de ocorrências observadas. Somando-se todas as frequências, e divivindo-se pelo total, resulta na unidade, como esperado. No caso da probabilidade, a soma de todas elas também resulta na unidade, porém ela pode ser interpretada como a expectativa de uma ocorrência ser observada.

		E_3	
E_1	E_2	0	1
0	0	$\frac{f_{000}}{(f_{000}+f_{001})} = P(0\|00)$	$\frac{f_{001}}{(f_{000}+f_{001})} = P(1\|00)$
0	1	$\frac{f_{010}}{(f_{010}+f_{011})} = P(0\|01)$	$\frac{f_{011}}{(f_{010}+f_{011})} = P(1\|01)$
1	0	$\frac{f_{100}}{(f_{100}+f_{101})} = P(0\|10)$	$\frac{f_{101}}{(f_{100}+f_{101})} = P(1\|10)$
1	1	$\frac{f_{110}}{(f_{110}+f_{111})} = P(0\|11)$	$\frac{f_{111}}{(f_{110}+f_{111})} = P(1\|11)$

Tabela 3.3: *Probabilidades condicionais $P(y|x)$ das possíveis configurações.*

probabilidade responde à seguinte pergunta: *dada uma configuração x dos elementos preditores, qual a chance de ocorrer um determinado estado y no elemento alvo?* Esta probabilidade condicional é escrita como $P(y|x)$, ou seja, a probabilidade de um estado do alvo ser y caso o(s) preditor(es) esteja(m) num estado x. Na tabela 3.3, estão representadas as probabilidades condicionais para cada configuração.

Com essa probabilidade condicional, é possível calcular uma entropia condicional, baseada na generalização de Tsallis. Essa entropia não possui uma forma única, e várias propostas foram feitas ao longo dos anos(A. TEIXEIRA A. SOUTO, 2014). A versão escolhida para este trabalho é aquela utilizada pelo próprio Tsallis (TSALLIS, 2009b) (e que também foi utilizada no trabalho de LOPES; OLIVEIRA; CESAR (2011)), onde primeiramente calculamos a entropia de um conjunto de configurações sobre todos os estados possíveis da sequência alvo (Y) para uma dada configuração do preditor x:

$$S_q(Y|x) = \frac{(1 - \sum_{y \subset Y} P(y|x)^q)}{q-1} \tag{3.2}$$

E a entropia condicional $H_q(Y|X)$ é calculada utilizando a probabilidade de se encontrar uma dada configuração do preditor como peso para a entropia daquele conjunto de configurações:

$$H_q(Y|X) = \sum_{x \subset X} P(x) S_q(Y|x) \tag{3.3}$$

onde a soma em x representa todas as possibilidades de configurações dos preditores, por exemplo, a probabilidade de ocorrer a configuração ($x = 01$), ou seja, o elemento $E_1 = 0$ e $E_2 = 1$, é calculada como se segue:

$$P(01) = \frac{f_{010} + f_{011}}{N} \tag{3.4}$$

3.2 Função Critério

Com esta entropia condicional, é possível construir uma função critério, inclusive uma que possamos comparar diferentes valores de q. Essa equação tem uma importante propriedade onde, dadas duas configurações $x = a$ e $x = b$, é possível ter $H_q'^a(Y|X) > H_q'^b(Y|X)$ e $H_q"^a(Y|X) < H_q"^b(Y|X)$, para dois q quaisquer ($q' \neq q"$), ou seja, entropias de diferentes valores de q não apresentam sempre valores maiores ou menores, como ocorreria se não usasse a probabilidade condicional, conforme mostrado na Figura 1.3.

Observando-se a Tabela 3.2, pode ocorrer casos onde os preditores não expressam todos os estados possíveis, e com isso uma determinada configuração dos preditores não está associado com nenhum estado do alvo. Ou seja, para uma determinada configuração dos preditores x_i, tem-se $P(y|x_i) = 0$ para qualquer y, o que resultaria em entropia condicional nula, alterando o resultado buscado, já que uma baixa entropia indicaria uma alta correlação entre os elementos, o que não é o caso. Muitos sistemas onde se procura inferir a organização apresentam poucos dados, então uma frequência nula pode significar que não foram obtidos dados suficientes, ou que outros fatores podem estar influenciando aquele estado do alvo. Independente do motivo, uma frequência nula indica uma falta de informação sobre aquela possibilidade, portanto precisa ser penalizada, com um aumento de entropia.

Para corrigir esse problema, é adotado uma entropia penalizante para estes casos, proporcional àquela do alvo(LOPES; OLIVEIRA; CESAR, 2011). Ou seja, usa-se a distribuição da sequência alvo completa $F_2 = (p_0, p_1)$ para calcular essa entropia penalizante $H_q(Y)$. A proporção é implementada através de um parâmetro α, que também altera a probabilidade de se observar uma configuração $P(X)$:

$$P(X) = \frac{(f_n + \alpha)}{\alpha M + d} \tag{3.5}$$

onde f_n é a frequência observada de uma determinada configuração X, M é o número de configurações possíveis (incluindo observadas e não observadas), e d é o número de amostras total disponíveis (tamanho do sinal). Finalmente, a função critério, incluindo o termo de pnealização, é escrita como:

$$C_q(Y|X) = \frac{\alpha(M - n)}{\alpha M + d} H_q(Y) + H_q(Y|X) \tag{3.6}$$

onde n é o número total de configurações i observadas. Será adotado $\alpha = 1$ para este trabalho. No caso de sistemas com 2 estados discretos possíveis ($W = 2$), foi mostrado anteriormente que, excetuando-se valores próximos de $p_0 = 0,5$, as demais distribuições de probabilidade $F_2 = (p_0, 1 - p_0)$ possíveis tem o valor de q_{min} bem

definido. Com isso, a função critério é calculada pelos seguintes passos:

- Ler as sequências de cada elemento, agrupando preditores e alvos;
- Mota-se a tabela de frequências;
- Calcula-se os P(y|x);
- Utilizando a equação (2.5), encontra-se o melhor valor de q;
- Calcula-se a entropia condicional com aquele valor de q.

3.3 Portas Lógicas

As interações entre elementos de uma rede booleana podem ser modeladas por portas lógicas, que realizam as operações lógicas. Por exemplo, se um elemento Verdadeiro e outro elemento no estado Falso interagem com uma operação lógica *AND*, resultarão num elemento Falso. No caso da interação se dar pela porta lógica *OR*, o resultado é um estado Verdadeiro. Algumas operações lógicas fornecem resultados triviais, como qualquer combinação resultar em Verdadeiro, ou em Falso, e serão desconsideradas neste trabalho. As portas aqui consideradas estão descritas abaixo, em formato texto e como tabela-verdade (Figura 3.1):

- *NOT*: para 1 elemento, o qual tem seu estado invertido como resultado;
- *AND*: 2 elementos precisam ser Verdadeiros para o resultado ser Verdadeiro, senão, é Falso;
- *OR*: dados 2 elementos, pelo menos um precisa ser Verdadeiro, para o resultado ser Verdadeiro;
- *NAND*: inverso do AND, ou seja, quando pelo menos um elemento é Falso, o resultado é Verdadeiro;
- *XOR*: ou OR exclusivo, os 2 elementos precisam ser diferentes para resultar em Verdadeiro

Como exemplo, são mostradas 3 sequências com 10 elementos aleatórios, na Figura 3.4, e uma sequência alvo, que é suposto ser o resultado da aplicação de alguma das portas lógicas descritas acima. A inferência do sistema ocorre quando se descobre quais 2 sequências originaram o alvo, ou seja, quando se encontra as sequências preditoras do alvo. Para isso monta-se uma tabela de probabilidades (Figura 3.2). Calculando a

NOT	
Elemento 1	Resultado
0	1
1	0

AND		
Elemento 1	Elemento 2	Resultado
0	0	0
0	1	0
1	0	0
1	1	1

NAND		
Elemento 1	Elemento 2	Resultado
0	0	1
0	1	1
1	0	1
1	1	0

OR		
Elemento 1	Elemento 2	Resultado
0	0	0
0	1	1
1	0	1
1	1	1

XOR		
Elemento 1	Elemento 2	Resultado
0	0	0
0	1	1
1	0	1
1	1	0

Figura 3.1: *Tabelas-verdade das portas lógicas NOT, AND, OR, NAND e XOR.*

E_1	0	0	0	1	1	1	0	1	0	0
E_2	1	0	1	0	1	0	1	0	1	0
E_3	0	1	0	1	0	0	1	1	0	1
$E_4(alvo)$	1	0	1	1	1	1	1	1	1	0

Tabela 3.4: *Sequências binárias simuladas, com o resultado de sua aplicação em uma portal lógica OR.*

função critério, com dois valores de q, 1 e 2,46 (valor médio para o caso binário) para todas as combinações possíveis de elementos, obtém-se o resultado da tabela 3.5. Com isso, infere-se que a sequência alvo resultou da combinação das sequências 1 e 2 através de alguma porta lógica (o tipo da porta, por este método, não é possível inferir).

		Estado do Alvo	
E_1	E_2	0	1
0	0	2	0
0	1	0	4
1	0	0	3
1	1	0	1

		Estado do Alvo	
E_1	E_3	0	1
0	0	1	1
0	1	1	3
1	0	0	2
1	1	0	2

		Estado do Alvo	
E_2	E_3	0	1
0	0	1	1
0	1	1	2
1	0	0	2
1	1	0	3

Figura 3.2: *Tabela com frequências de estados entre possíveis sequências preditoras e o alvo.*

Outro teste foi realizado, com mais casos, no qual geramos 80 sequências diferentes, cada uma com 100 componentes binários escolhidos aleatoriamente. As 40 primeiras foram escolhidas para gerar uma sequência de alvos, com 50% de chance de ser uma cópia e 50% de passar pela porta NOT. As demais 40 sequências geradas foram combinadas em pares, a cada interação uma porta lógica AND, OR, NAND ou XOR sendo sorteadas, totalizando 20 sequências alvo. Para o resultado não ser totalmente previsível, todas sequências alvos passaram por um embaralhador, que sorteou e inverteu 10

Combinação	Função Critério ($q = 1$)	Função Critério ($q = 2.46$)
1,2	0,00	0,00
1,3	0,364	0,217
2,3	0,330	0,203

Tabela 3.5: *Valores da função critério para o exemplo da porta OR.*

componentes do alvo[2]. Após, cada sequência alvo foi analisada com quatro possíveis preditores, um deles sendo a correta. A Tabela confusão 3.3, contém os resultados dos testes, descritos pelos parâmetros *True Negative (TN), False Positive (FP), False Negative (FN)* e *True Positive (TP)*, além da *precision*, *recall* e *accuracy* (MARBACH et al., 2012), calculados pelas equações:

$$precision = \frac{TP}{TP + FP} \tag{3.7}$$

$$recall = \frac{TP}{TP + FN} \tag{3.8}$$

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{3.9}$$

A função critério foi utilizada com 3 valores diferentes de q, para comparação:

- $q = 1$ (Entropia de Boltzmann-Shannon);
- $q_m = 2,46$ (valor médio que foi encontrado para sistemas com estados binários);
- q_v variável (o valor é obtido para cada distribuição de probabilidades, pela eq. (2.6));

Quantidade	q=1	q=$q_{médio}$	q variável
TN	288	298	298
FP	12	2	2
FN	15	2	2
TP	85	98	98
precision	0,88	0,98	0,98
recall	0,85	0,98	0,98
accuracy	0,93	0,99	0,99
F1	0,86	0,98	0,98

Figura 3.3: *Tabela confusão para teste com inferência de sequências alvo e 100 sequencias preditoras, utilizando a função critério com diferentes valores do parâmetro q.*

Este exemplo mostra a vantagem de se utilizar a entropia não extensiva, porém, não houve diferença entre utilizar q_m ou q_v (valor médio e variável do parâmetro q,

[2]Cada porta lógica, por ser determinística, apresentaria sempre o mesmo valor de entropia

respectivamente), provavelmente pelo valor variável estar numa faixa pequena, somente $\Delta q = 0.025$ em torno do valor médio. Ou seja, para medir se um sistema binário está organizado ou não, pode ser usado o valor $q_m = 2,46$, e não é necessário calcular a equação (2.6) para todas as diferentes distribuições.

Este resultado também foi obtido, experimentalmente, para sistemas biológicos, como descrito no artigo LOPES; OLIVEIRA; CESAR (2011). Nele, mostraram que ao se utilizar a entropia de Tsallis para inferência de redes gênicas, obtiveram melhores resultados para $2,5 < q < 3,5$, observando que o valor de $q > 1$ reduziu o número de falsos negativos.

3.4 Topologia de uma rede

A topologia de uma rede indica a estrutura de conexões entre os elementos, geralmente considerando a média do número de conexões. Alguns modelos teóricos são muito utilizados, e nomeados a partir de seus proponentes. Quando se busca inferir a estrutura de uma rede, alguns métodos supõe uma topologia *apriori* para aquela rede, e utilizam os resultados da função critério conforme esperam determinada estrutura na rede. Por exemplo, uma rede onde os elementos podem ter um número aleatório de conexões com outros, é conhecida como rede de topologia aleatório de Erdös-Rényi (ER) (ERDöS; RéNYI (1958)). Caso a rede possua elementos *hub*, ou seja, poucos elementos estão ligados com muitos, e muitos elementos com poucos, é conhecida como topologia *small-world*, de Watts-Strogatz (WS) (WATTS; STROGATZ (1998)). Em especial, neste trabalho, serão utilizadas duas topologias, uma delas sendo a rede livre de escala (*scale-free*), de Barabási-Albert (BA) (BARABÁSI; ALBERT (1999)), supõe que exista uma lei de potência na distribuição de suas conexões:

$$P(k) \approx k^{-\gamma} \tag{3.10}$$

O valor de k indica o número médio de elementos conexões com um dado elemento, também conhecido como sua cardinalidade. Ou seja, considera que existe uma grande quantidade de elementos com $k = 1$, uma quantidade menor com $k = 2$, ainda menor com $k = 3$, e assim por diante, até que a chance de encontrar elementos com grande número de conexões é praticamente nula. Neste trabalho, assim como em LOPES; OLIVEIRA; CESAR (2011), foi considerado $\gamma = 2.5$, valor este próximo do encontrado na literatura para alguns sistemas biológicos, como $\gamma = 2.2$ para a rede metabólica da *Escherichia coli* (JEONG et al. (2000)), e $\gamma = 2.4$ para a rede de uma proteína de levedura (BOCCALETTI et al. (2006); JEONG et al. (2000)).

3.5 Algoritmo SFFS-BA

Em uma rede com N elementos, um alvo pode ter até $N-1$ preditores. Computacionalmente não é viável calcular a função critério para todas as combinações possíveis, e mesmo uma rede com N na casa de dezenas já é proibitivo calcular todas as possibilidades para 4 ou mais preditores. Para uma rede com $N > 100$, calcular todas as possibilidades para 3 preditores pode durar alguns dias, em computadores atuais (não paralelizados). É necessário um algoritmo para se fazer uma busca em tempo adequado, e neste trabalho será utilizado o método SFFS-BA (*Sequential Forward Floating Selection - Barabási-Albert*), proposto em LOPES; OLIVEIRA; CESAR (2011). Este algoritmo considera que a rede analisada tem uma topologia que segue uma lei de potência em sua distribuição de conexões (*scale-free*). São procuradas conexões entre preditores e alvos que indicam uma organização no sistema, por um método que pode ser resumido da seguinte forma, considerando-se um determinado alvo de uma rede com N elementos:

1. Dada a função critério (FC), estabelece-se um valor mínimo FC_{min}.
2. Nível 0: Analisa-se se o alvo é constante ao longo de toda expressão da rede. Caso positivo, ele não é considerado como alvo, pois fornece baixa informação sobre o sistema;
3. Nível 1: Calcula-se FC considerando-se todos os demais elementos como possíveis preditores. Aqueles que resultarem em $FC < FC_{min}$, são considerados como possíveis preditores, e não são considerados para busca nas próximas etapas;
4. Pela lei de escala (eq. 3.10), com $k = 1$, separa-se os $k^{-\gamma}$ menores valores de FC para se continuar a busca, com uma nova cardinalidade.
5. Nível 2: Forma-se uma dupla entre cada um dos elementos separados no passo anterior, com todos os demais elementos disponíveis na rede, e calcula-se o FC. Assim como no nível 1, caso alguma das duplas tenha $FC < FC_{min}$, ela é considerada possível preditora, e desconsiderada para os próximos níveis.
6. Repete-se o passo 4, para $k = 2$, separando as melhores duplas;
7. Nível 3: É acrescentado um terceiro elemento à dupla que obteve menor valor FC no passo anterior, mas que ainda não teve $FC < FC_{min}$, exaustivamente dentre os demais elementos possíveis. Os trios com $FC < FC_{min}$ são considerados preditores. Caso um dos elementos acrescentados diminua o valor de FC, mas não o suficiente para ser menor do que FC_{min}, aquele trio é guardado.

8. Repete-se o passo 4, para $k = 3$;

9. Nível 4 em diante: O trio com menor FC é testado com outros elementos, com procedimento similar ao do passo anterior. Isso pode ser estendido para outros níveis até não ocorrer mais a diminuição de FC pelo acréscimo de um novo elemento. Quando isso ocorre, o algoritmo é finalizado. Também é possível encerrar num nível arbitrário, por exemplo no terceiro, devido ao custo computacional proibitivo para se estudar muitas combinações.

10. Em todos os passos, é considerado que a função FC teve alteração com o acréscimo de um novo elemento se o valor teve alteração superior a um $\Delta = 0,05$, para evitar aumentos de cardinalidade por pequenas flutuações no valor.

Este algoritmo permite realizar uma busca para uma quantidade variada de número de preditores, de forma direcionada e com maiores chances de encontrar bons candidatos, sem realizar uma busca extensiva em todas as combinações possíveis, para todas as cardinalidades. Em casos com alvos com predição intrinsecamente multivariada (IMP), ou seja, aqueles alvos que o acréscimo ou remoção de somente um elemento preditor influencia pouco no valor de FC, pode ocorrer do método SFFS-BA explorar somente um subespaço dos possíveis preditores, e nesse caso só é possível descobrir todos os preditores por busca exaustiva (LOPES; OLIVEIRA; CESAR (2011)).

Uma variação pode ser implementada, na qual somente os melhores resultados (menores valores de entropia) são mantidos entre cada etapa. Qual função critério utilizar ainda é um campo de estudo aberto.

No próximo capítulo iremos explorar como aplicar esses resultados e ferramentas em redes gênicas.

Capítulo 4

Inferência de Redes Gênicas

Em sistemas biológicos, a relação entre diversos genes pode ser modelada como uma rede, chamada de rede gênica. Neste capítulo, explicaremos como essas redes podem ser construídas, e aplicaremos o estudo de inferência de sistemas com estados discretos em um sistema de redes gênicas, iniciando com um caso booleano, simulado, e posteriormente aplicando em dados experimentais, com um número $W \neq 2$ de estados discretos.

4.1 Redes Gênicas com Estados Discretos

Genes, produzidos a partir de trechos específicos de sequências de DNA (ácido desoxirribonucléico), codificam produtos que realizam uma série de funções biológicas necessárias para a vida ALBERTS (2008). A expressão gênica é o processo de sintetizar produtos a partir do código genético. Esse código é composto pelos nucleotídeos adenina, tinina, citosina e guanina (representados pelas letras A, T, C e G, respectivamente), cuja sequência, no interior das células de todos os seres vivos, possuem as informações para construir e regular todas as estruturas que compõe a célula. Trechos dessa sequência podem ser transcritos para outras estruturas, os chamados RNA (ácido ribonucléico), que são compostos dos nucleotídeos A,C, G e U (uracila). Por sua vez, os RNA podem ser traduzidos, ou seja, a informalção contida neles pode ser convertida em aminoácidos, e estes, podem formar proteínas. As proteínas podem ter variados tamanhos e funções, inclusive produzir mais aminoácidos, nucleotídeos e outros componentes necessários para o funcionamento da célula.

Em um dado organismo vivo, as estruturas que foram expressas a partir do trecho de um DNA (passando pela transcrição e tradução), são conhecidas como genes. Dependendo das necessidades de uma célula, pode-se encontrar determinados genes expressos. A expressão de determinados genes pode afetar outros, por exemplo inibindo

a produção de outros, ou incentivando. Funções biológicas complexas são controladas por uma regulação rígida dessas expressões gênicas interdependentes, e organismos mostram padrões com alta interação regulatória entre um grande número de genes (facilmente chegando aos milhares), sendo um desafio inferir a rede de genes reguladores a partir de um pequeno número de perfis de expressão gênica (muitas vezes algumas unidades). Várias ferramentas e estudos já foram feitos buscando a melhor forma de inferir essas redes (MARKOWETZ; SPANG (2007)). Os métodos de inferência podem ser classificados em 3 grandes grupos (BROECK et al. (2020)): aprendizado de máquina, correlação entre genes e dinâmica bayesiana (que é a base para a função critério utilizada neste trabalho). Dentre os métodos bayesianos, alguns utilizam parâmetros extensivos, como a entropia, e a grande maioria dos trabalhos que o fazem utilizam entropia de Boltzmann-Shannon, uma exceção já citada sendo o trabalho de LOPES; OLIVEIRA; CESAR (2011).

Uma rede é um sistema cujos elementos são os vértices e a interação é realizada por conectores. Em uma rede gênica, cada vértice representa um gene, chamado de gene alvo, o qual pode ser ativado ou inibido por um ou mais vértices. Por sua vez, este alvo pode ser o preditor de outros genes, formando uma rede. Por exemplo, na Figura 4.1 temos uma rede com poucos genes.

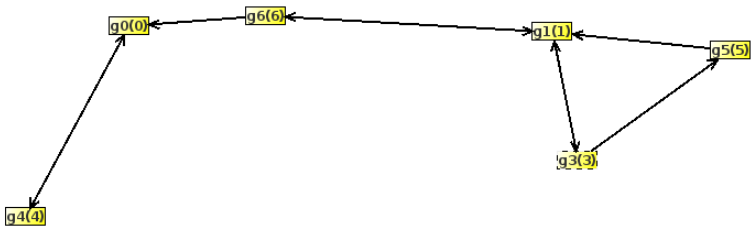


Figura 4.1: *Exemplo de rede com 6 genes (nós). Observe que algumas ligações (arestas) são de via dupla.*

Os dados experimentais podem ser obtidos, por exemplo, pela técnica de microarray (figura 4.2). Nele, uma placa contém vários receptores dispostos numa grade, e cada receptor é construído de uma forma que se ligue em trechos específicos de códigos genéticos. Foi desenvolvido no começo dos anos 80 (FLOYD; DELEO; THOMPSON (1983)), inicialmente com poucas divisões, e atualmente existem placas com centenas e até milhares de divisões. Uma amostra á retirada do objeto estudado, colocada em toda a grade, a qual é limpa de modo a retirar qualquer material genético que não tenha se ligado aos receptores. Ao final do processo, uma marcação de cores mostra o quanto o material procurado em cada pedaço da grade estava presente na amostra (por exemplo,

vermelho para pouco expresso, verde para muito expresso). Digitalizando uma imagem dessa grade permite utilizar os valores da coloração como uma medida de expressão gênica.

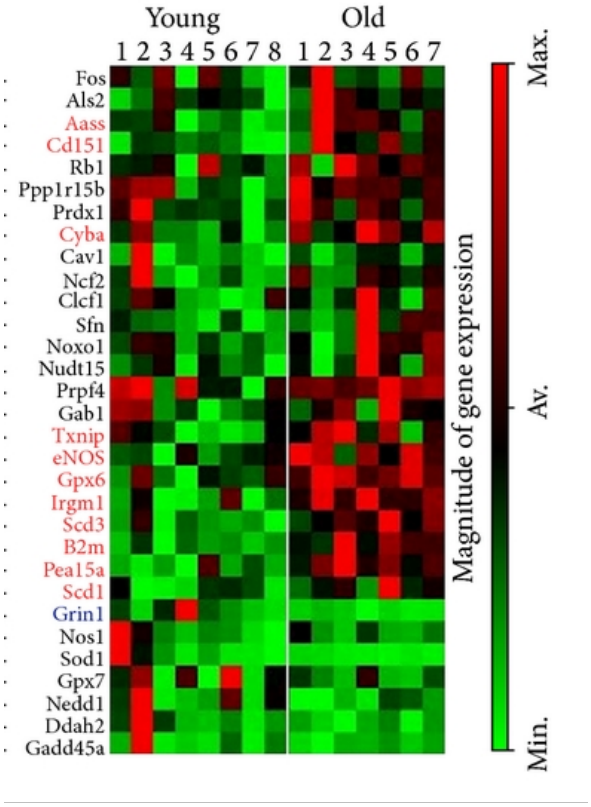


Figura 4.2: *Exemplo de placa de microarray, para estudo de expressão gênica de genes presentes em células de ratos novos e velhos. Adaptado de CHEPKOVA; SCHöNFELD; SERGEEVA (2015).*

Com isso tem-se a expressão gênica para aqueles genes de interesse, de uma amostra que foi retirada num determinado instante de tempo e condição do objeto estudado (seja ele planta, animal ou bactéria). Do ponto de vista da rede, uma medida de micro-array equivale a somente um dado para cada gene, outros sendo necessários para se estudar a rede. Outras medidas podem ser realizadas em indivíduos diferentes, ou no mesmo indivíduo, em situações diferente. Por exemplo, analisar uma célula de uma folha de uma planta em condição normal de humidade, e outra análise da mesma planta em um clima seco, ou então, retirar uma amostra num instante temporal, esperar 30 minutos, e retirar outra. O objetivo é ter diferentes expressões gênicas, que fornecerão diferentes medidas no micro-array, e com isso pode-se tentar inferir uma rede entre aqueles genes.

Cada medida dessa tem um custo elevado, por isso muitos estudos apresentam poucas medidas, ou então placas de microarray com poucos genes. Existem outros métodos mais atuais para se obter a expressão gênica, porém ainda de alto custo, conhecidas como *"Next Generation Sequencing" (NGS)*, com alguns estudos comparando com o micro-array (ver RAI et al. (2017); RAO et al. (2019) e sites das empresas que fabricam os equipamentos, como a ILLUMINA (2021); OTOGENETICS (2021)). No RNA-Seq, um exemplo de NGS (STARK; GRZELAK; HADFIELD (2019)), não é necessário saber previamente quais genes estão sendo analisados, sendo possível saber a expressão gênica de todos os genes expressos na amostra. Entretanto, o micro-array tem vantagens, como um custo menor por gene medido, e métodos de análise são mais simples do que os NGS (SMITH (20215)).

4.2 Inferência de Redes Gênicas Artificiais

Para testar métodos de inferência de redes gênicas, é necessário medir a expressão dos genes (seja por microrray ou por RNA-Seq) de sistemas biológicos já conhecidos e bem estudados, para comparar os resultados obtidos. Entretanto, poucos sistemas possuem estas características, ou devido ao custo de se obter os dados, ou pelo fato da própria interação entre muitos genes não ser conhecida. Por isso, uma alternativa é utilizar redes artificiais para testar métodos de inferência, ou seja, pode ser usado um método *in silico* ao invés do *in vitro*.

Neste trabalho, para fins de comparação, foram analisados 2 tipos de conjuntos de dados, que já foram estudados no artigo LOPES; OLIVEIRA; CESAR (2011), um gerado pelo software jAGN (LOPES; CESAR; COSTA (2011)) e o outro pelo desafio DREAM4, explicados a seguir. Cada conjunto de dados foi analisado de forma similar:

- Os dados de expressão gênica foram discretizados, caso já não estivessem;
- Todos os genes foram escolhidos como genes alvo, e foi aplicado o algoritmo SFFS-BA 3.5, utilizando a função critério 3.6;
- Os resultados obtidos foram comparados com as conexões conhecidas, e foram obtidos os valores de TP, FN, FP e TN, *precisão*, *especificadade* e *similaridade*.

Cada um desses parâmetros mostra diferentes aspectos do classificador[1]. Como estudaremos redes *in vitro*, poderemos utilizar a *similaridade* como principal indicador:

[1]Em redes gênicas não artificiais (*in vitro*), espera-se que a grande maioria dos genes não interajam entre si, por isso o valor de TN é muito alto, e parâmetros derivados, normalmente não são utilizados, o modelo interessante sendo aquele que acerta mais positivos. Uma discussão interessante sobre a diferença entre alguns testes para classificadores gerais encontra-se em: DöRING (2018).

$$SIM = \sqrt{\frac{TP}{TP+FP} * \frac{TN}{TN+FP}} \tag{4.1}$$

4.2.1 Rede gerada pelo jAGN

Nessas redes, a expressão gênica pode ser convertida em Verdadeiro (gene expresso) ou Falso (gene não expresso), e as interações podem ser modeladas como portas lógicas (SHMULEVICH; DOUGHERTY; ZHANG (2002)). Podem ocorrer combinações entre as portas, por exemplo dois genes interagirem por uma porta OR, e sua saída está ligada a um terceiro gene por uma porta AND. As redes booleanas podem ser separadas em dois tipos, as que tem conexões fixas, chamadas de "boolean networks"(BN), e as "probabilistic boolean network"(PBN), onde portas lógicas apresentam uma chance, muito pequena, de serem alteradas a cada passo temporal.

As PBNs são úteis para se considerar fatores não presentes na rede analisada. Em dados *in vitro*, a rede gênica é maior do que o número de genes cuja expressão é medida. Por exemplo, numa placa de microarray podem-se medir centenas ou milhares de genes, enquanto a rede completa pode conter dezenas de milhares. Por isso, é preciso considerar que existem genes que afetam a rede considerada, mas que estão fora dela. Também pode ocorrer fatores externos às medidas, por exemplo alterações em elementos na célula medida, como mudança em temperatura, umidade e outros, que podem causar alterações na expressão gênica por fatores externos. Por último, mas não menos importante, podem ocorrer problemas na captação dos dados, e na interação dos genes expressos com os reagentes(HARDO; BAKSHI (2021) e KIM; ZAKHARKIN; ALLISON (2010)). Por todos esses fatores, é importante ter um elemento de aleatoriedade nos dados gerados *in silico*, mesmo que seja um valor baixo.

Na rede simulada, a expressão dos genes num determinado tempo t_i é causada pela interação dos estados dos genes preditores no tempo t_{i-1}. Inicialmente as conexões entre os genes são sorteadas, ou seja, o simulador já possui a rede gênica. Cada conexão representa uma porta lógica que dita as regras da interação do sistema. Um estado inicial é sorteado para todos os genes da rede simulada, e o estado seguinte é composto pelas interações dos genes preditores com as portas lógicas. Os próximos passos temporais são simulados da mesma forma. O resultado é apresentado como uma sequência temporal de 0s e 1s para cada gene, que simulam as medidas de expressão gênica.

Para simular as redes utilizadas neste trabalho, foi utilizado o software jAGN LOPES; CESAR; F. COSTA (2008), que permite escolher a topologia da rede (neste trabalho foi escolhida o tipo Erdös-Rényi, também conhecida como aleatória (ERDöS; RéNYI (1958))), o número de genes e amostras temporais. Foi escolhida uma PBN (Probabilistic Boolean Network) com 3% de chance de alterar a porta a cada passo

temporal. O software disponibiliza como resultado a tabela temporal da rede e as relações entre alvos e preditores, escrito como uma matriz de adjacência. Além disso, ele cria representações gráficas das estruturas da redes (figuras 4.3 e 4.4).

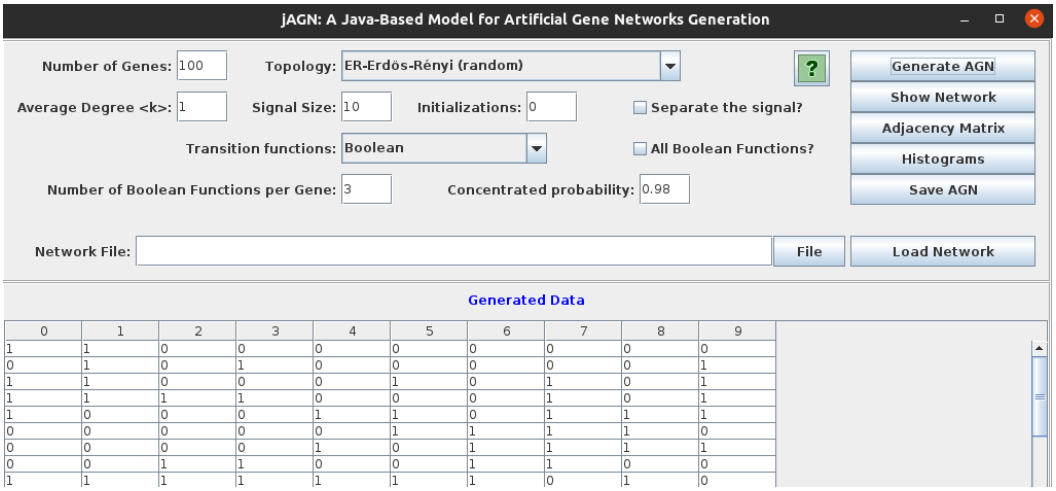


Figura 4.3: *Interface do software jAGN.*

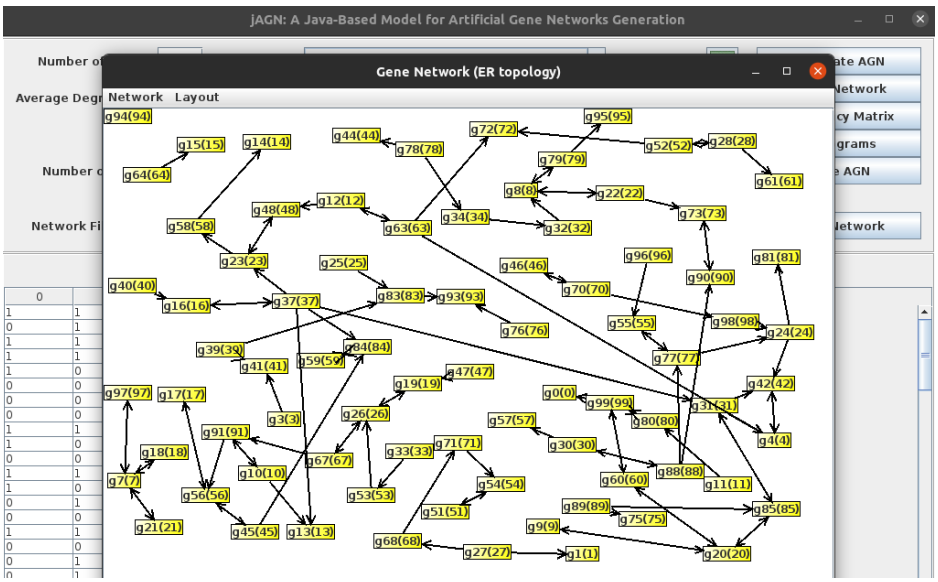


Figura 4.4: *Representação gráfica da rede simulada.*

Utilizando este software, é possível testar o método de inferência do capítulo anterior. Escolhe-se um gene como possível alvo, e 1, 2 ou 3 outros genes como preditores (mais do que 3 tem um custo computacional proibitivo), monta-se uma tabela de frequências, similar à 3.2, e calcula-se o valor da função critério para um determinado valor de q. Caso o valor esteja abaixo de um limiar, é considerado que aquela possibilidade de preditores e alvo representa uma aresta da rede gênica. Realizando essa conta com vários alvos, e possuindo os dados da criação da rede. é possível calcular

a precisão, *recall* e acurácia, e comparar os resultados entre diferentes valores de q. O trabalho de LOPES (2011), já citado anteriormente, obteve a Figura 4.5, na qual é possível observar que para vários casos estudados, o valor que permitiu melhor inferência das redes gênicas artificiais foi para q próximo a $2,5$, o que corrobora o resultado obtido neste trabalho, no qual mostramos que que o estudo da informação de sistemas binários é mais eficiente com entropia de Tsallis para $q \approx 2,46$.

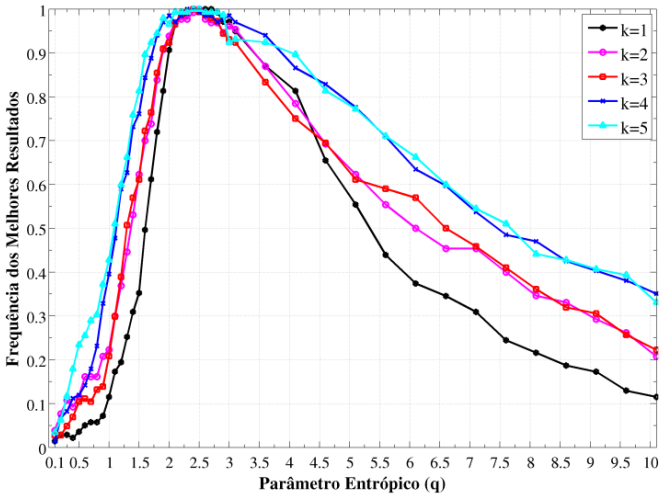


Figura 4.5: *Gráfico da similaridade para redes gênicas artificiais, extraído do trabalho de LOPES (2011), mostrando resultados para o caso da rede Barabási-Albert, cada curva para um valor k diferente.*

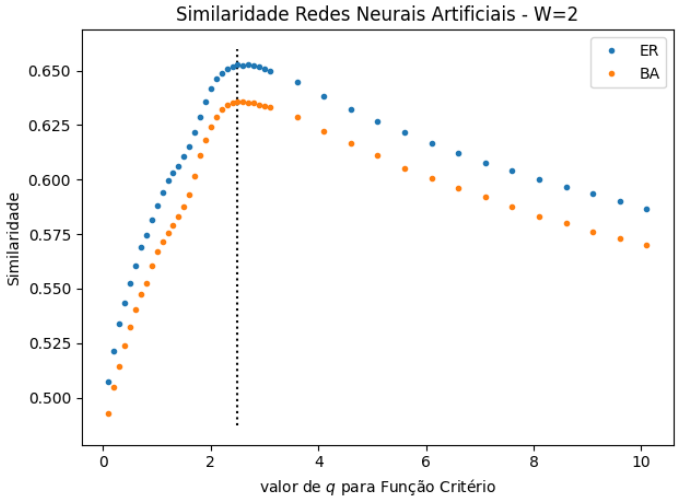


Figura 4.6: *Gráfico da similaridade da média dos resultados obtidos para as redes simuladas de Erdös-Rényi e de Barabási-Albert (mesmos dados da figura 4.5. A reta tracejada representa a posição $q = 2.5$.*

4.2.2 Rede do desafio DREAM4

Como mencionado anteriormente, para testar métodos de inferência de redes gênicas, é necessário ter tanto as medidas de expressão de uma grande quantidade de genes, como também interações já conhecidas entre aqueles genes. Para testar as ferramentas descritas neste trabalho, foi escolhido o desafio DREAM 4 (*Dialogue on Reverse Engineering Assessment and Methods*), que disponibilizou uma grande quantidade de dados de redes *in silico*, ou seja, simuladas, cujos dados de expressão gênica podem ser obtidos no trabalho MARBACH et al. (2010, 2009); PRILL et al. (2010). Nele foram apresentadas 2 tipos de redes, uma com 10 genes e outra com 100, cada uma com 5 redes diferentes. Os dados disponibilizados estão em *float*, ou seja, não discretizados, num formato de série temporal de expressão gênica, e também foram divulgadas as arestas de ligação entre genes, para comparação da inferência.

Os dados da rede com 10 genes foram discretizados binariamente ($W = 2$), e a rede de 100 genes com $W = 3$, como feito no trabalho LOPES; OLIVEIRA; CESAR (2011), pois com esta discretização obtiveram melhores resultados. O método SFFS foi aplicado, considerando somente os menores valores em cada nível analisado, com a função critério calculado para vários valores de q. Os resultados da similaridade para o DREAM4, após aplicação da função critério, estão mostrados nas figuras 4.7 e 4.8. Para comparação com os resultados da tabela 2.1, foi desenhado um tracejado indicando o valor médio para q_{min} para os casos $W = 2$ e $W = 3$, e mostram que realmente a similaridade maior ocorre naqueles pontos.

4.2.3 Análise dos resultados

Nesta seção, iremos fazer algumas considerações sobre os resultados obtidos.

Como mencionado anteriormente, ao longo do trabalho, a inferência busca encontrar a informação contida no sistema. Para se medir essa informação, é preferível uma escala que diferencie o estado de "ordem"de estado de "desordem", com a menor quantidade de dados disponíveis. Os resultados obtidos mostram que, para sistemas onde os estados possíveis estão discretizados, a métrica mais indicada para realizar a inferência não está relacionada com o grau de complexidade daquele sistema, mas sim com o número de estados possíveis. Como mencionado ao final da seção 2.3.2, o valor de q obtido para cada sistema permite separar de forma mais eficaz configurações do sistema que indicam uma estrutura organizada (entropia menor) de uma estrutura desorganizada (entropia maior).

O efeito deste resultado na função critério ainda é um trabalho em aberto, principalmente por utilizar a entropia condicional, que poderia alterar a análise da entropia.

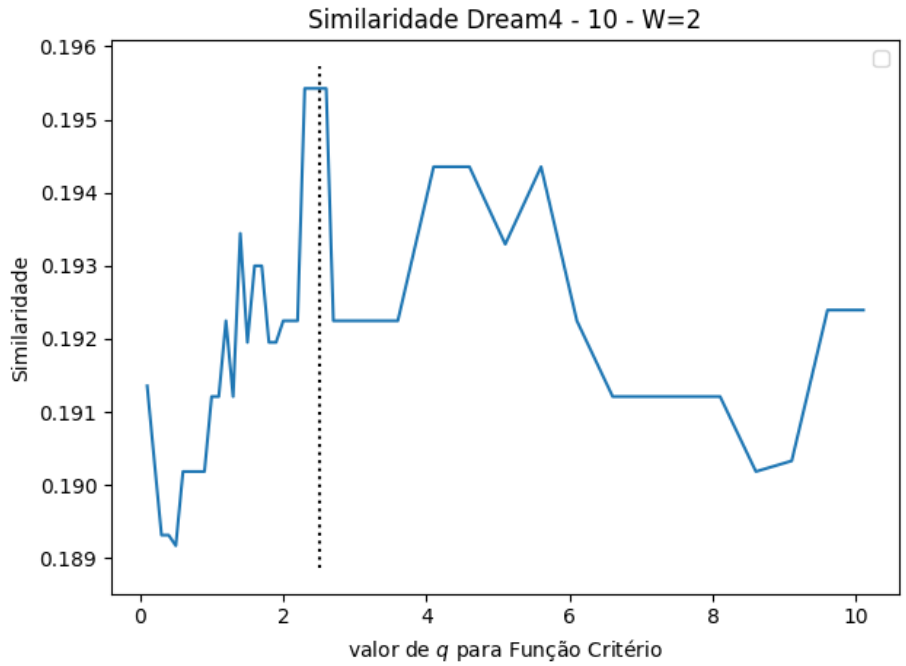


Figura 4.7: *Similaridade em função de q, para o conjunto com 10 genes, do DREAM4.*

Por exemplo, num sistema binário, considerando dois preditores para um alvo, os estados possíveis são somente 0 e 1, ou 000, 001, 010 etc.? Isto alteraria os estados possíveis de 2 para 8, porém, já sabemos que o valor continua sendo $q \approx 2,46$, diferente do caso $W = 8$. Principalmente, é preciso se analisar e compreender como unir a entropia condicional quando os valores de W são diferentes de 2, pois neste caso, para cada configuração de probabilidades considerada, o melhor valor de q seria diferente. No caso $W = 2$, como a maioria dos valores está próximo da média, essa questão não influencia tanto quando em outros casos.

Outra parte considerada, que pode ser de interesse para outros pesquisadores, são as diferentes rotas de pensamento testadas antes de se chegar no resultado aqui apresentado. Durante a pesquisa para esta dissertação, já eram conhecidos os resultados das redes gênicas (como pode ser visto nos artigos originais deLOPES (2011) e LOPES; OLIVEIRA; CESAR (2011)). Aquele resultado indicava não só que a inferência era melhor com a entropia de Tsallis do que a entropia de Boltzmann, como que ter um valor específico de q, que permitia melhor inferência da rede gênica, poderia indicar uma estrutura no sistema relacionada com aquele valor. A ideia principal era que cada parte da rede influenciaria a entropia da rede como um todo. Algumas hipóteses foram levantadas, tais como: efeito da extensidade na soma dos trechos das redes; topologia da rede; quantidade de genes preditores e genes *hub*.

O primeiro fator considerado, envolvia o uso da equação 1.21, para medir a "exten-

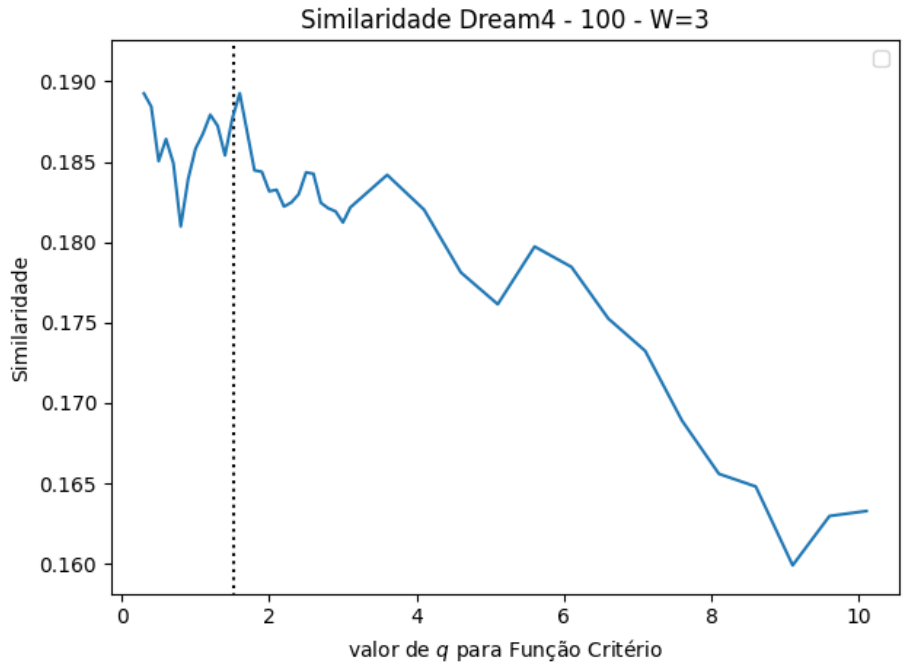


Figura 4.8: *Similaridade em função de q, para o conjunto com 100 genes, do DREAM4.*

sividade"de cada trecho da rede, e com isso encontrar o valor de q da rede completa. Por exemplo, supondo que uma rede pode ser subdividida em dois grupos menores, e a entropia de cada trecho é calculada. Seria procurado qual valor de q resultaria na igualdade $S_q^{A+B} = S_q^A + S_q^B + (1-q)S_q^A S_q^B$. A topologia da rede teria influência direta no cálculo da entropia, pois redes *small world*, *scale free* e aleatórias teriam distribuições diferentes de genes com poucas e com muitas ligações. Poderia ocorrer de ter poucos ou muitos genes preditores, e alguns genes teriam muitas conexões, tanto como preditor quanto como alvo (genes "hub"). É de se imaginar, por este método, que cada topologia deveria resultar em diferentes valores de q.

Entretanto, esta hipótese não se confirmou, redes simuladas com diferentes topologias apresentam o mesmo valor para o melhor q. Além disso, dividir uma rede em subgrupos não é uma tarefa simples, pois a forma de se dividir uma rede é arbitrária. Outro fator importante, que levou ao estudo das portas lógicas, é que seguindo o argumento de que a rede poderia ser subdividida em redes menores, estas sub-redes também poderiam ser subdivididas, sucessivamente, até que cada subgrupo seria composto por somente um preditor e um gene alvo. Isto foi um indicativo de que o valor encontrado de q está relacionado com alguma característica básica do sistema, o que se mostrou estar relacionado com a discretização utilizada. Ou seja, as outras hipóteses levantadas aqui, sobre o efeito da topologia, e formato da rede, podem influenciar outros resultados da inferência, por exemplo melhorar a eficiência dos algoritmos de

busca de relações entre genes, mas não estão relacionados com o valor de q encontrado.

Conclusão

A entropia é uma grandeza que ainda elude muitos cientistas. Conceitos como energia e simetria, pilares básicos da natureza, os quais somos expostos desde pouca idade, parecem mais próximos e poderia-se dizer que "concretos" do que a entropia. Para muitos estudantes de física, a relação entre a entropia da termodinâmica e a entropia de teoria da informação parece desconexa, mesmo sabendo trabalhar com as equações. Espera-se que esse trabalho contribua para melhorar a compreensão do conceito de entropia, e em especial como demonstração das possibilidades de uso da entropia não extensiva.

Na estatística e áreas correlatas, a entropia possui papel importante como medida de informação do sistema. Uma generalização, como é a entropia de Tsallis, permite o estudos de sistemas que não podiam ser analisados, ou que apresentavam baixos resultados se aplicada a entropia de Boltzmann-Shannon. Apesar das novas possibilidades, um obstáculo é saber qual valor correto do parâmetro q deve ser escolhido para o melhor estudo de um determinado sistema, os pesquisadores dispensando muito tempo em busca dos parâmetros mais adequados.

O presente trabalho apresentou um método para se obter o valor de q que forneceria mais informações para um determinado conjunto de probabilidades. Em especial, para sistemas binários, o sistema apresenta maior quantidade de organização se estudado com a entropia não extensiva para $q \approx 2,46$. Este resultado é válido para qualquer sistema binário, e não somente para os casos mostrados aqui. O estudo de sistemas complexos é um campo ainda em aberto, com novos métodos sendo propostos e testados continuamente. Especialmente em redes gênicas, mesmo que aproximadas para sistemas booleanos, o número de variáveis é altíssimo, com poucas amostras temporais disponíveis, o que exige modelos bem testados e fundamentados. Previamente, o trabalho de LOPES (2011) havia encontrado esse valor de q realizando um estudo experimental, com vários casos diferentes. Obtivemos neste trabalho, a partir de uma equação, o mesmo valor. Citando uma pergunta presente no trabalho LOPES; OLIVEIRA; CESAR (2011): "*what is the best entropy function for the inference of GRNs?*", é possível dizer que neste trabalho foi encontrada a resposta.

O método apresentado permite também estudos com diferentes discretizações, o que foi comparado com os resultados obtidos para se estudar o conjunto de dados disponibilizado pelo desafio DREAM4, com os valores previstos para q_{min} permitindo melhores resultados para a inferência.

Os resultados aqui obtidos mostram-se promissores, e apesar das dificuldades, outros sistemas podem ser estudados pelos métodos aqui descritos, e mais estudos podem ser realizados, por exemplo, com modificações na função critério. Além disso, a equação geral para quaisquer valores de estados discretos abre a possibilidade para estudos de sistemas com diferentes discretizações, como estudos envolvendo nucleotídeos A,T,C,G. Estes, como outros sistemas biológicos, serão analisados futuramente pelo método descrito neste trabalho.

Referências Bibliográficas

A. TEIXEIRA A. SOUTO, L. A. **Entropias condicionais de Tsallis**. Accessed: 2020-08-01, http://rss.di.fc.ul.pt/tools/tsallis-entropy. 16, 25

AALTO, A.; VIITASAARI, L.; ILMONEN, P.; MOMBAERTS, L.; GONÇALVES, J. Gene regulatory network inference from sparsely sampled noisy data. **Nature Communications**, [S.l.], v.11, n.1, July 2020. 2

ALBERTS, B. **Molecular biology of the cell**. New York: Garland Science, 2008. 14, 33

AMADOR, C.; ZAMBRANO, L. Evidence for energy regularity in the Mendeleev periodic table. **Physica A: Statistical Mechanics and its Applications**, [S.l.], v.389, n.18, p.3866–3869, 2010. 11

AMIGÓ, J.; BALOGH, S.; HERNÁNDEZ, S. A Brief Review of Generalized Entropies. **Entropy**, [S.l.], v.20, n.11, p.813, Oct. 2018. 9

AUGUSTINE, J.; JEREESH, A. S. Gene regulatory network inference: a semi-supervised approach. In: INTERNATIONAL CONFERENCE OF ELECTRONICS, COMMUNICATION AND AEROSPACE TECHNOLOGY (ICECA), 2017., 2017. **Anais...** IEEE, 2017. 2

BARABÁSI, A.-L.; ALBERT, R. Emergence of Scaling in Random Networks. **Science**, [S.l.], v.286, n.5439, p.509–512, Oct. 1999. 30

BÉRUT, A.; ARAKELYAN, A.; PETROSYAN, A.; CILIBERTO, S.; DILLENSCHNEIDER, R.; LUTZ, E. Experimental verification of Landauer's principle linking information and thermodynamics. **Nature**, [S.l.], v.483, n.7388, p.187–189, Mar. 2012. 11

BILGEN, I.; SARAC, O. S. Gene regulatory network inference from gene expression dataset using autoencoder. In: SIGNAL PROCESSING AND COMMUNICATIONS APPLICATIONS CONFERENCE (SIU), 2018., 2018. **Anais...** IEEE, 2018. 2

BOCCALETTI, S.; LATORA, V.; MORENO, Y.; CHAVEZ, M.; HWANG, D. Complex networks: structure and dynamics. **Physics Reports**, [S.l.], v.424, n.4-5, p.175–308, Feb. 2006. 30

BRANCH, M. A.; COLEMAN, T. F.; LI, Y. A Subspace, Interior, and Conjugate Gradient Method for Large-Scale Bound-Constrained Minimization Problems. **SIAM Journal on Scientific Computing**, [S.l.], v.21, n.1, p.1–23, Jan. 1999. 21

BROECK, L. V. den; GORDON, M.; INZÉ, D.; WILLIAMS, C.; SOZZANI, R. Gene Regulatory Network Inference: connecting plant biology and mathematical modeling. **Frontiers in Genetics**, [S.l.], v.11, May 2020. 34

CALLEN, H. **Thermodynamics and an introduction to thermostatistics**. New York: Wiley, 1985. 1, 6

CHEPKOVA, A.; SCHöNFELD, S.; SERGEEVA, O. Age-Related Alterations in the Expression of Genes and Synaptic Plasticity Associated with Nitric Oxide Signaling in the Mouse Dorsal Striatum. **Neural plasticity**, [S.l.], v.2015, p.458123, 03 2015. v, 35

COX, D. R. **Principles of statistical inference**. Cambridge New York: Cambridge University Press, 2006. 13, 14

DöRING, M. **The Case Against Precision as a Model Selection Criterion**. Accessed: 2021-02-20, https://www.datascienceblog.net/post/machine-learning/specificity-vs-precision/. 36

ERDöS, P.; RéNYI, A. **On random graphs I.** Accessed: 2021-02-13, https://www.renyi.hu/~p_erdos/1959-11.pdf. 30, 37

FLOYD, E. T.; DELEO, J. M.; THOMPSON, E. B. Sequential Comparative Hybridizations Analyzed by Computerized Image Processing Can Identify and Quantitate Regulated RNAs. **DNA**, [S.l.], v.2, n.4, p.309–327, Dec. 1983. 34

FONTANA, R.; SANTOS, I. dos. Os enunciados da segunda lei da termodinâmica: uma possível abordagem. **Revista Brasileira de Ensino de Física**, [S.l.], v.38, n.1, Mar. 2016. 4

HANEL, R.; THURNER, S. A comprehensive classification of complex statistical systems and an axiomatic derivation of their entropy and distribution functions. **EPL (Europhysics Letters)**, [S.l.], v.93, n.2, p.20006, jan 2011. 8, 11

HARDO, G.; BAKSHI, S. Challenges of analysing stochastic gene expression in bacteria using single-cell time-lapse experiments. **Essays in Biochemistry**, [S.l.], v.65, n.1, p.67–79, Apr. 2021. 37

ILLUMINA. **Advantages of RNA-Seq technology**. Accessed: 2021-02-13, https://www.illumina.com/science/technology/next-generation-sequencing/microarray-rna-seq-comparison.html. 36

JEONG, H.; TOMBOR, B.; ALBERT, R.; OLTVAI, Z. N.; BARABÁSI, A.-L. The large-scale organization of metabolic networks. **Nature**, [S.l.], v.407, n.6804, p.651–654, Oct. 2000. 30

JIMENEZ, R. D.; MARTINS, D. C.; SANTOS, C. S. Gene Networks Inference through One Genetic Algorithm Per Gene. In: IEEE INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND BIOENGINEERING, 2014., 2014. **Anais...** IEEE, 2014. 2

KIM, K.; ZAKHARKIN, S. O.; ALLISON, D. B. Expectations, validity, and reality in gene expression profiling. **Journal of Clinical Epidemiology**, [S.l.], v.63, n.9, p.950–959, Sept. 2010. 37

KODAMA, T.; ELZE, H.-T.; AGUIAR, C. E.; KOIDE, T. Dynamical correlations as origin of nonextensive entropy. **Europhysics Letters (EPL)**, [S.l.], v.70, n.4, p.439–445, may 2005. 9, 11

LIPSCHUTZ, S. **Teoria e problemas de matemática discreta**. Porto Alegre (RS: Bookman, 2004. 23

LOPES, F. M. **Redes complexas de expressão gênica**: síntese, identificação, análise e aplicações. 2011. Tese (Doutorado em Bioinformática) — . vi, 2, 3, 39, 41, 44

LOPES, F. M.; CESAR, R. M.; COSTA, L. D. F. Gene Expression Complex Networks: synthesis, identification, and analysis. **Journal of Computational Biology**, [S.l.], v.18, n.10, p.1353–1367, Oct. 2011. 36

LOPES, F. M.; CESAR, R. M.; F. COSTA, L. da. AGN Simulation and Validation Model. In: ADVANCES IN BIOINFORMATICS AND COMPUTATIONAL BIOLOGY, 2008, Berlin, Heidelberg. **Anais...** Springer Berlin Heidelberg, 2008. p.169–173. 37

LOPES, F. M.; OLIVEIRA, E. A. de; CESAR, R. M. Inference of gene regulatory networks from time series by Tsallis entropy. **BMC Systems Biology**, [S.l.], v.5, n.1, May 2011. 3, 25, 26, 30, 31, 32, 34, 36, 40, 41, 44

LOW, S. T.; MOHAMAD, M. S.; OMATU, S.; CHAI, L. E.; DERIS, S.; YOSHIOKA, M. Inferring gene regulatory networks from perturbed gene expression data using a dynamic Bayesian network with a Markov Chain Monte Carlo algorithm. In: IEEE INTERNATIONAL CONFERENCE ON GRANULAR COMPUTING (GrC), 2014., 2014. **Anais...** IEEE, 2014. 2

MARBACH, D.; ; COSTELLO, J. C.; KüFFNER, R.; VEGA, N. M.; PRILL, R. J.; CAMACHO, D. M.; ALLISON, K. R.; KELLIS, M.; COLLINS, J. J.; STOLOVITZKY, G. Wisdom of crowds for robust gene network inference. **Nature Methods**, [S.l.], v.9, n.8, p.796–804, July 2012. 29

MARBACH, D.; PRILL, R. J.; SCHAFFTER, T.; MATTIUSSI, C.; FLOREANO, D.; STOLOVITZKY, G. Revealing strengths and weaknesses of methods for gene network inference. **Proceedings of the National Academy of Sciences**, [S.l.], v.107, n.14, p.6286–6291, Mar. 2010. 40

MARBACH, D.; SCHAFFTER, T.; MATTIUSSI, C.; FLOREANO, D. Generating Realistic In Silico Gene Networks for Performance Assessment of Reverse Engineering Methods. **Journal of Computational Biology**, [S.l.], v.16, n.2, p.229–239, Feb. 2009. 40

MARKOWETZ, F.; SPANG, R. Inferring cellular networks – a review. **BMC Bioinformatics**, [S.l.], v.8, n.S6, Sept. 2007. 34

MARUYAMA, K.; NORI, F.; VEDRAL, V. Colloquium: the physics of maxwell's demon and information. **Reviews of Modern Physics**, [S.l.], v.81, n.1, p.1–23, Jan. 2009. 11

MERCATELLI, D.; SCALAMBRA, L.; TRIBOLI, L.; RAY, F.; GIORGI, F. M. Gene regulatory network inference resources: a practical overview. **Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms**, [S.l.], v.1863, n.6, p.194430, June 2020. 2

Nocedal, J.; Wright, S. J. **Numerical Optimization**. [S.l.]: Springer New York, 2006. 19

NOMAN, N.; IBA, H. Inferring Gene Regulatory Networks using Differential Evolution with Local Search Heuristics. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, [S.l.], v.4, n.4, p.634–647, oct 2007. 2

NUSSENZVEIG, H. M. **Curso de física básica 2** : fluidos, oscilações e ondas, calor. Brasil: Blucher, 2018. 1, 4, 5, 6, 13

OTOGENETICS. **RNA Sequencing VS Microarray**. Accessed: 2021-02-13, https://www.otogenetics.com/rna-sequencing-vs-microarray/. 36

PINDAH, W.; NORDIN, S.; SEMAN, A.; SAID, M. S. M. Review of dimensionality reduction techniques using clustering algorithm in reconstruction of gene regulatory networks. In: INTERNATIONAL CONFERENCE ON COMPUTER, COMMUNICATIONS, AND CONTROL TECHNOLOGY (I4CT), 2015., 2015. **Anais...** IEEE, 2015. 2

PRILL, R. J.; MARBACH, D.; SAEZ-RODRIGUEZ, J.; SORGER, P. K.; ALEXOPOULOS, L. G.; XUE, X.; CLARKE, N. D.; ALTAN-BONNET, G.; STOLOVITZKY, G. Towards a Rigorous Assessment of Systems Biology Models: the DREAM3 challenges. **PLoS ONE**, [S.l.], v.5, n.2, p.e9202, Feb. 2010. 40

PUGA, J. L.; KRZYWINSKI, M.; ALTMAN, N. Bayes' theorem. **Nature Methods**, [S.l.], v.12, n.4, p.277–278, Mar. 2015. 12

RAI, M. F.; TYCKSEN, E. D.; SANDELL, L. J.; BROPHY, R. H. Advantages of RNA-seq compared to RNA microarrays for transcriptome profiling of anterior cruciate ligament tears. **Journal of Orthopaedic Research**, [S.l.], Aug. 2017. 36

RAM, R.; CHETTY, M. A Markov-Blanket-Based Model for Gene Regulatory Network Inference. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, [S.l.], v.8, n.2, p.353–367, mar 2011. 2

RAO, M. S.; VAN VLEET, T. R.; CIURLIONIS, R.; BUCK, W. R.; MITTELSTADT, S. W.; BLOMME, E. A. G.; LIGUORI, M. J. Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From

Short-Term Rat Toxicity Studies. **Frontiers in Genetics**, [S.l.], v.9, p.636, 2019. 36

REIF, F. **Fundamentals of statistical and thermal physics**. New York: McGraw-Hill, 1965. 1, 6

SAKAMOTO, E.; IBA, H. Inferring a system of differential equations for a gene regulatory network by using genetic programming. In: CONGRESS ON EVOLUTIONARY COMPUTATION (IEEE CAT. NO.01TH8546), 2001., 2001. **Proceedings...** IEEE, 2001. 2

SHANNON, C. E. A Mathematical Theory of Communication. **Bell System Technical Journal**, [S.l.], v.27, n.3, p.379–423, July 1948. 1, 8

SHMULEVICH, I.; DOUGHERTY, E.; ZHANG, W. From Boolean to probabilistic Boolean networks as models of genetic regulatory networks. **Proceedings of the IEEE**, [S.l.], v.90, n.11, p.1778–1792, nov 2002. 2, 37

SMITH, C. **DNA Microarrays**: a trusted tool keeps evolving. Accessed: 2021-02-13, https://www.biocompare.com/Editorial-Articles/172195-DNA-Microarrays-A-Trusted-Tool-Keeps-Evolving/. 36

STARK, R.; GRZELAK, M.; HADFIELD, J. RNA sequencing: the teenage years. **Nature Reviews Genetics**, [S.l.], v.20, n.11, p.631–656, July 2019. 36

TSALLIS, C. Possible generalization of Boltzmann-Gibbs statistics. **Journal of Statistical Physics**, [S.l.], v.52, n.1-2, p.479–487, jul 1988. 2, 9

TSALLIS, C. **Introduction to nonextensive statistical mechanics** : approaching a complex world. New York, NY: Springer, 2009. 9, 11

TSALLIS, C. **Introduction to Nonextensive Statistical Mechanics**: approaching a complex world. [S.l.]: Springer, 2009. 25

TSALLIS, C. Beyond Boltzmann–Gibbs–Shannon in Physics and Elsewhere. **Entropy**, [S.l.], v.21, n.7, p.696, July 2019. 11

VAN ROSSUM, G. **The Python Library Reference, release 3.8.2**. [S.l.]: Python Software Foundation, 2020. 17

Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Jarrod Millman, K.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C.; Polat, İ.; Feng, Y.; Moore, E. W.; Vand erPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; Contributors, S. . . SciPy 1.0: fundamental algorithms for scientific computing in python. **Nature Methods**, [S.l.], v.17, p.261–272, 2020. 18

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world' networks. **Nature**, [S.l.], v.393, n.6684, p.440–442, June 1998. 30

XU, R.; WUNSCH, D.; FRANK, R. Inference of Genetic Regulatory Networks with Recurrent Neural Network Models Using Particle Swarm Optimization. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, [S.l.], v.4, n.4, p.681–692, oct 2007. 2