UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE FÍSICA

Carlos Eduardo Gasparoni Santos

# ORDENAMENTO POR FUNÇÃO CUSTO DE REDES DE INTERAÇÕES PROTÉICAS: ESTUDO DA *ARABIDOPSIS THALIANA*

PORTO ALEGRE

2010

Carlos Eduardo Gasparoni Santos

# ORDENAMENTO POR FUNÇÃO CUSTO DE REDES DE INTERAÇÕES PROTÉICAS: ESTUDO DA *ARABIDOPSIS THALIANA*

Trabalho de Conclusão de Curso apresentado à Universidade Federal do Rio Grande do Sul como requisito parcial para obtenção do título de bacharel em Física — Pesquisa Básica

Orientador: Leonardo Gregory Brunnet

PORTO ALEGRE

2010

# SUMÁRIO

## RESUMO

Determinar com precisão o papel de cada agente bioquímico em todos os processos relevantes que ocorrem na célula se revelou uma tarefa complicada. Métodos locais, que se concentram apenas em um pequeno conjunto de componentes, não são capazes de explicar o funcionamento da célula em um nível que nos permita entender processos mais complexos, como o envelhecimento. Porém, a análise local sugere que o genoma pode ser considerado uma rede modular, onde genes integrantes de um mesmo módulo interagem mais fortemente entre si do que com genes integrantes de outros módulos. Este trabalho propõe a análise de um método que ordena uma lista de genes interagentes utilizando uma dinâmica de Monte Carlo, com o objetivo de agrupar unidimensionalmente os genes com interações mais fortes. A partir dessa lista ordenada, é possível perceber a separação em módulos funcionais da lista de genes e observar uma correlação entre esses módulos e a ativação dos genes durante variados processos biológicos realizados pela célula.

## ABSTRACT

Determining the precise role of every biochemical agent in every relevant process in the cell has turned out to be a complex task. Local methods, that focus just on a few components, are incapable of explaining the functioning of the cell in a level that allows us to understand more complex processes, such as aging. However, local analysis suggests that the genome can be regarded as a modular network, where genes in a module interact more strongly with one another in comparison to their interactions with genes in other modules. This work proposes the analysis of a method that orders a list of interacting genes utilizing Monte Carlo dynamics, with the objective of clustering in a line the strongly interacting genes. From that ordered list, it is possible to see the separation of the gene list in functional modules, and to observe a correlation between those modules and the activation of the genes during various biological processes performed by the cell.

# 1 INTRODUÇÃO

Todos os seres vivos, desde os mais simples até os mais complexos, possuem em seu código genético toda a informação necessária para seu desenvolvimento a partir da célula-ovo. Após a decodificação do genoma de vários organismos, tem sido objetivo da comunidade científica catalogar as moléculas e suas interações dentro de uma célula, para assim descrever o funcionamento dos mecanismos reguladores de um organismo vivo. No entanto, tal tarefa tem se revelado muito difícil, devido ao grande número de interações apresentado pelo conjunto de genes, proteínas e metabólitos em uma célula. O genoma pode ser visto como uma rede com caráter modular[1-3], ou seja, existem conjuntos de genes que interagem mais fortemente entre si do que com outros genes, mas ao mesmo tempo a célula é bastante integrada, de modo que mesmo genes participantes de módulos diferentes podem ter interferência um sobre o outro.

A idéia central deste trabalho é analisar o funcionamento de um método capaz de descrever a rede como um todo, atribuindo funções a grandes agrupamentos de genes, mas sem ignorar as já conhecidas interações locais existentes entre os mesmos.

Para tanto, este trabalho propõe o uso de um método de ordenamento de uma lista de genes, chamado de método de ordenamento por minimização de função custo do sistema, capaz de agrupá-los de tal maneira que seja possível observar a composição de módulos funcionais da rede. Este método é, de certo modo, descendente do ordenamento originalmente proposto por Barabási[2]. É utilizada uma dinâmica de Monte Carlo para organizar a rede de modo a aproximar genes que interajam mais fortemente entre si. Neste trabalho o método será aplicado em uma rede de interações de genes da *Arabidopsis thaliana*, que devido ao seu genoma relativamente simples, composto de cerca de 26 mil genes e um dos primeiros genomas de plantas a ser completamente sequenciado, é comumente utilizado como modelo para o estudo da genética e da biologia das plantas.

Após aplicar o ordenamento, é feita a análise da estatística da rede. A partir do conceito de modularidade, observamos a separação dos módulos funcionais. Utilizando bibliotecas de ontologia dos genes, esperamos observar uma relação entre a modularidade e a ativação dos genes, o que por sua vez indica uma relação entre um módulo funcional e um processo biológico específico da célula.

# 2 DESENVOLVIMENTO

## 2.1 ORDENAMENTO

Em primeiro lugar, a lista de genes interagentes da *Arabidopsis thaliana* é retirada da base de dados STRING[4,5]. Para tanto, selecionamos apenas aqueles pares de genes cuja probabilidade de ligação (um valor conhecido como o *score* da rede) é maior ou igual a $0{,}8$. Em outras palavras, consideramos dois genes interligados se em no mínimo $80\%$ dos processos biológicos conhecidos em que um dos genes é ativado, ele causa ou implica a ativação do outro. Com essa restrição, obtemos uma lista constituída por $4196$ genes e $77306$ interações.

A partir da lista, monta-se a matriz de interações. A cada um dos genes é associado um número, e na matriz de interações o sítio $M_{i,j}$ é definido como tendo valor igual a $1$ se os genes correspondentes aos números $i$ e $j$ interagem entre si, ou valor igual a $0$ no caso contrário. Com isso temos que a matriz de interações é uma matriz quadrada simétrica de tamanho $n \times n$, onde $n$ é o número de genes do organismo, composta apenas pelos valores $0$ e $1$, com a diagonal principal nula. É importante ressaltar que a lista de genes retirada da base de dados é apresentada em uma ordem aleatória e, portanto, nenhuma informação sobre as características globais da rede pode ser retirada da matriz de interações antes da aplicação do método de ordenamento.

Em seguida, é definido um valor chamado de função custo da matriz, que é simplesmente a soma do custo de todos os sítios que a compõem. O custo de um sítio é definido como sendo proporcional à menor distância daquele sítio até a diagonal principal da matriz e à diferença do valor daquele sítio e dos seus quatro primeiros vizinhos. Matematicamente, esse custo é dado pela equação $(1)$.

$$S_{ij} = d_{ij} \cdot \left[\left|M_{i,j} - M_{i,j+1}\right| + \left|M_{i,j} - M_{i,j-1}\right| + \left|M_{i,j} - M_{i+1,j}\right| + \left|M_{i,j} - M_{i-1,j}\right|\right] \qquad (1)$$

Deste modo, o custo da matriz completa é dado pela equação $(2)$.

$$S = \sum_i \sum_j d_{ij} \cdot \left[\left|M_{i,j} - M_{i,j+1}\right| + \left|M_{i,j} - M_{i,j-1}\right| + \left|M_{i,j} - M_{i+1,j}\right| + \left|M_{i,j} - M_{i-1,j}\right|\right] \qquad (2)$$

Onde o termo da distância à diagonal principal é dado pela equação $(3)$.

$$d_{ij} = \sqrt{|i^2 - j^2|} \qquad (3)$$

Em seguida, aplica-se a dinâmica de Monte Carlo no problema. Este método consiste em escolher aleatoriamente dois genes da rede e trocar sua posição. Em seguida, calcula-se o custo da nova matriz. Se ele for menor que o custo da matriz anterior à troca, esta é aceita. Isso garante que, com o tempo, o sistema tenda a um mínimo da função custo do sistema.

Porém, apenas com essa restrição o sistema pode ficar preso em um mínimo local da função custo. Para impedir que isso aconteça, utilizamos um processo chamado *annealing* simulado, que consiste em definir uma quantidade, chamada de temperatura, que representa uma probabilidade de que o sistema aceitará uma troca que aumente o valor da função custo. É importante ressaltar que essa nomenclatura da temperatura não possui significado físico intrínseco, e remete apenas à idéia original do método, que vem do processo de *annealing* da metalurgia.

Definimos um valor inicial de temperatura de $T = 10^7$, que é um valor propositalmente grande, utilizado para garantir que no início do processo haja probabilidade muito alta de que o sistema aceite trocas que aumentem o seu custo. Além disso, a temperatura é reduzida a um quarto de seu valor a cada $300$ passos de Monte Carlo (MCS, *Monte Carlo step*). Um passo de Monte Carlo consiste em um número de iterações igual ao tamanho do sistema, ou seja, no caso da *Arabidopsis thaliana* um passo de Monte Carlo representa $4196$ trocas aleatórias.

Tendo definido o *annealing*, consideramos agora que uma troca de genes que aumente o custo da matriz em um valor $\Delta S$ tem probabilidade $e^{-\frac{\Delta S}{T}}$ de ser aceita, ou seja, selecionamos um número aleatório entre 0 e 1 e a troca é aceita se o número sorteado é menor ou igual a $e^{-\frac{\Delta S}{T}}$. Isso faz com que, durante o início do processo, a grande maioria das trocas seja aceita devido ao alto valor de $T$, o que na prática age para colocar o sistema em uma ordem aleatória, assim eliminando qualquer possível influência que o estado inicial da rede poderia ter sobre o seu estado final, antes de começar o processo de minimização da função custo propriamente dito.

Devido à definição utilizada para a função custo, após um grande número de passos de Monte Carlo, a matriz agrupa os genes de acordo com suas interações. O termo da distância à diagonal principal na função custo garante que as ligações tendem a ficar próximas à diagonal da matriz, enquanto o termo da diferença entre um sítio e seus vizinhos garante que as ligações se agrupam de modo que uma

interação tende a estar cercada de outras interações e um sítio nulo tende a estar cercado de outros sítios nulos.

Após a realização de 9 mil passos de Monte Carlo, o que significa que foram feitas mais de 30 milhões de trocas aleatórias, o sistema já se apresenta bem próximo do mínimo da função custo, ao ponto em que é possível reparar na própria matriz o agrupamento das interações entre os genes em torno da diagonal principal. Para tanto, podemos observar a caráter de comparação a matriz antes e depois da aplicação do ordenamento, respectivamente, nas figuras 1 e 2.

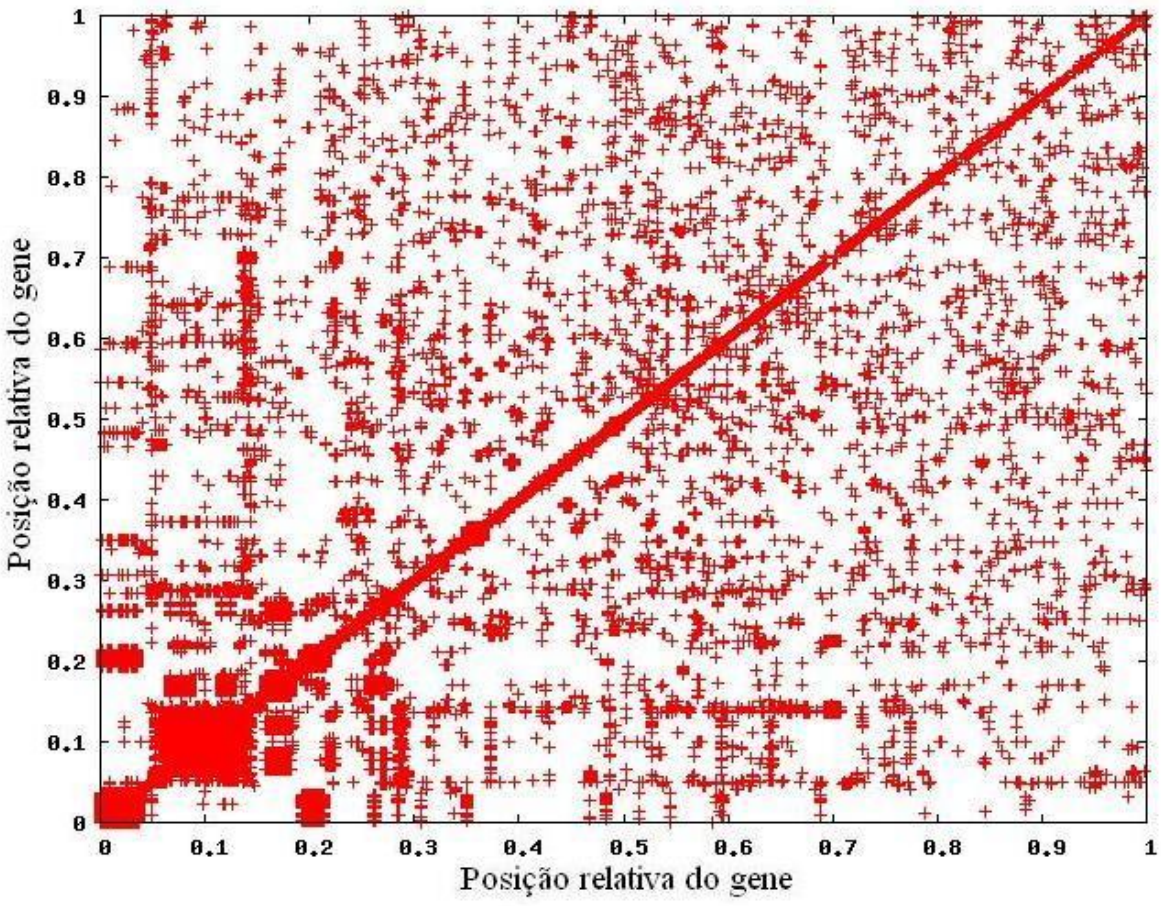


**Figura 1: Configuração inicial da matriz de interações**

Na matriz inicial, pode-se perceber uma grande quantidade de ligações espalhadas nos cantos distantes da diagonal principal e, com exceção do agrupamento no início da rede, a presença de poucos clusters em torno da diagonal principal. Pela natureza do método utilizado, que enumera os genes na ordem em que eles são apresentados na lista, não é surpresa o aparecimento desse cluster. Porém, como já comentado anteriormente, o processo do *annealing* simulado garante que esse cluster será desmontado no intervalo de tempo inicial do

ordenamento e que sua presença no estado inicial no sistema não terá qualquer influência na configuração do estado final do mesmo.

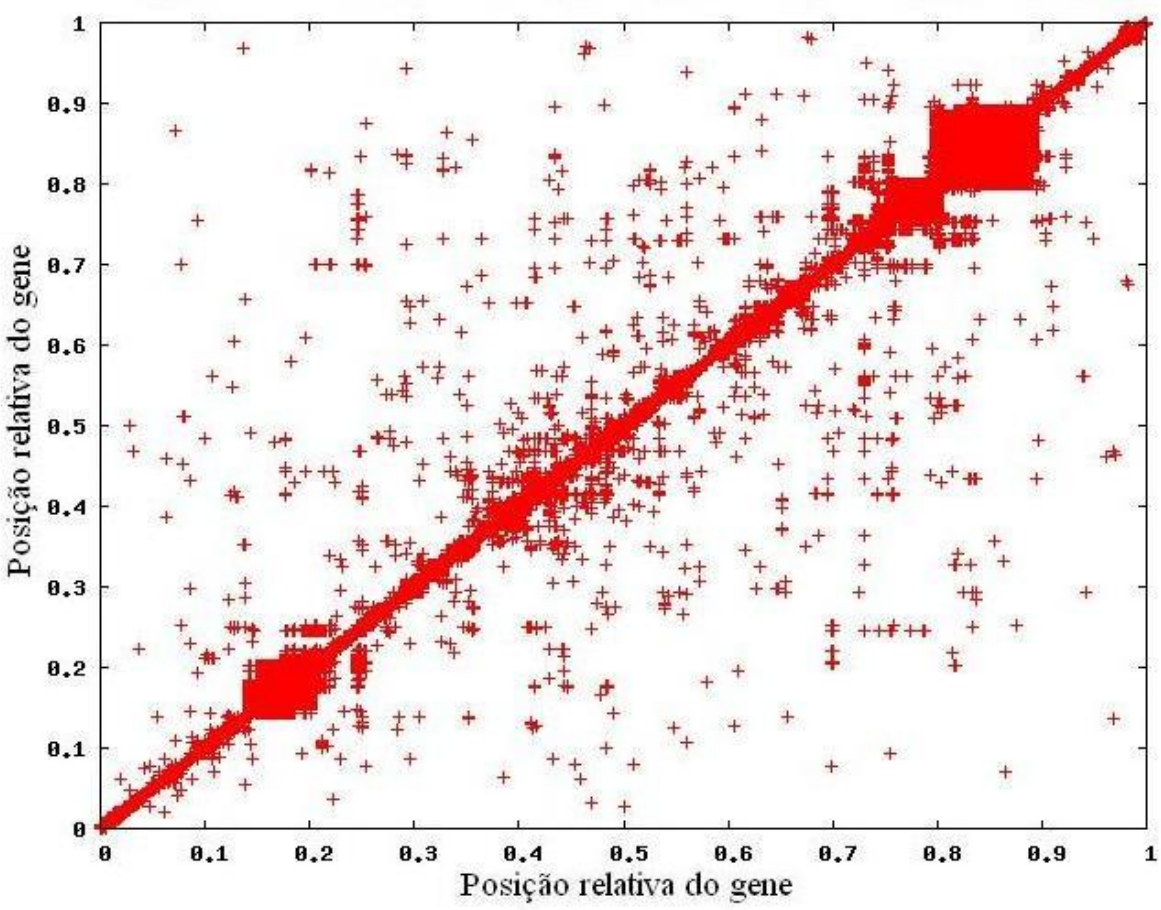


**Figura 2: Configuração final da matriz de interações**

É importante ressaltar que, como já mencionado anteriormente, o sistema é representado por uma matriz de ordem maior que 4 mil. Logo, a resolução da matriz é maior do que as figuras 1 e 2 podem demonstrar claramente; de fato, todos os sítios da diagonal principal em ambos os casos são nulos, e se tivéssemos resolução maior, isso poderia ser facilmente confirmado.

Após a realização dos 9 mil passos de Monte Carlo podemos observar que os valores não-nulos da matriz de interações se agrupam fortemente em torno da diagonal principal. O número de ligações longe da mesma diminuiu significantemente, mas pela natureza aleatória do método, seria necessário um tempo infinito para alcançar o mínimo global do sistema, então é de se esperar que ainda haja ligações fora dos agrupamentos da diagonal. Também podemos perceber que as interações formam, de modo geral, agrupamentos aproximadamente quadrados em torno da diagonal.

## 2.2 ANÁLISE DA REDE

A partir da configuração final da matriz, somos capazes de iniciar a análise da estatística da rede. Esta análise é feita principalmente a partir de três conceitos: a modularidade, a conectividade e a clusterização dos genes da rede.

A modularidade de um gene é um valor que mede quão próximos daquele gene estão, na configuração da rede, os genes com os quais ele interage. Para fazer esse cálculo para um dado gene, centramos no gene escolhido uma janela de tamanho apropriado ao organismo em questão e calculamos a quantidade de genes com os quais ele interage dentro da janela, relativo à quantidade total de genes com os quais ele interage na rede. Por causa dessa definição, a modularidade de um gene é sempre um valor normalizado, ou seja, entre $0$ e $1$. Matematicamente, se temos uma janela de tamanho $2w + 1$ centrada em um gene $i$, a modularidade é calculada conforme a equação $(4)$.

$$m_i = \frac{\sum_{j=i-w}^{i+w} M_{i,j}}{\sum_{j=1}^{4196} M_{i,j}} \tag{4}$$

A conectividade de um gene é definida simplesmente como o número de genes com o qual ele interage. Porém, para a análise da rede, não utilizaremos o valor da conectividade de um gene, mas sim associaremos àquele gene a conectividade média de todos os genes dentro de uma janela de tamanho $2w + 1$ centrada nele. Matematicamente, os valores são calculados conforme a equação $(5)$.

$$\begin{aligned} k_i &= \sum_{j=1}^{4196} M_{i,j} \\ K_i &= \frac{\sum_{j=i-w}^{i+w} k_j}{2w+1} \end{aligned} \tag{5}$$

A clusterização de um gene é uma medida de quão interligados estão os vizinhos de um gene. Ela é definida como o número de pares de genes interligados a um gene $i$ que possuem ligação entre si, em relação ao número máximo de pares de genes interligados ao gene $i$. Logo, a clusterização de um gene, assim como a modularidade, é um valor normalizado, variando entre $0$ e $1$. Porém, analogamente ao cálculo da conectividade, o que será utilizado para a análise da rede não será a clusterização de um gene, mas sim a clusterização média dos genes em uma janela

de tamanho $2w + 1$ centrada nele. Se tomarmos $v_{in}$ como o $n$-ésimo vizinho do gene de número $i$, a clusterização é dada pela equação $(6)$.

$$c_i = \frac{\sum_{n=1}^{k_i} \sum_{m=n+1}^{k_i} M_{v_{in},v_{im}}}{\frac{(k_i)(k_i - 1)}{2}}$$
$$C_i = \frac{\sum_{j=i-w}^{i+w} c_j}{2w + 1} \qquad (6)$$

Centrar uma janela de tamanho $2w + 1$ em um dado gene $i$ significa observar o próprio gene $i$, os $w$ genes imediatamente à sua esquerda e os $w$ genes imediatamente à sua direita. Para o cálculo dessas três quantidades, utiliza-se o mesmo tamanho, que no caso da *Arabidopsis thaliana* é de $251$. A escolha desse tamanho é feita observando o gráfico da modularidade feito a partir de diferentes tamanhos de janela e escolhendo aquele que melhor representa a separação entre os módulos, ou seja, aquele que possui separação mais distinta entre seus picos e vales. Além disso, para esses cálculos são utilizadas no sistema condições de contorno periódicas.

O resultado desses cálculos é apresentado na figura $3$.

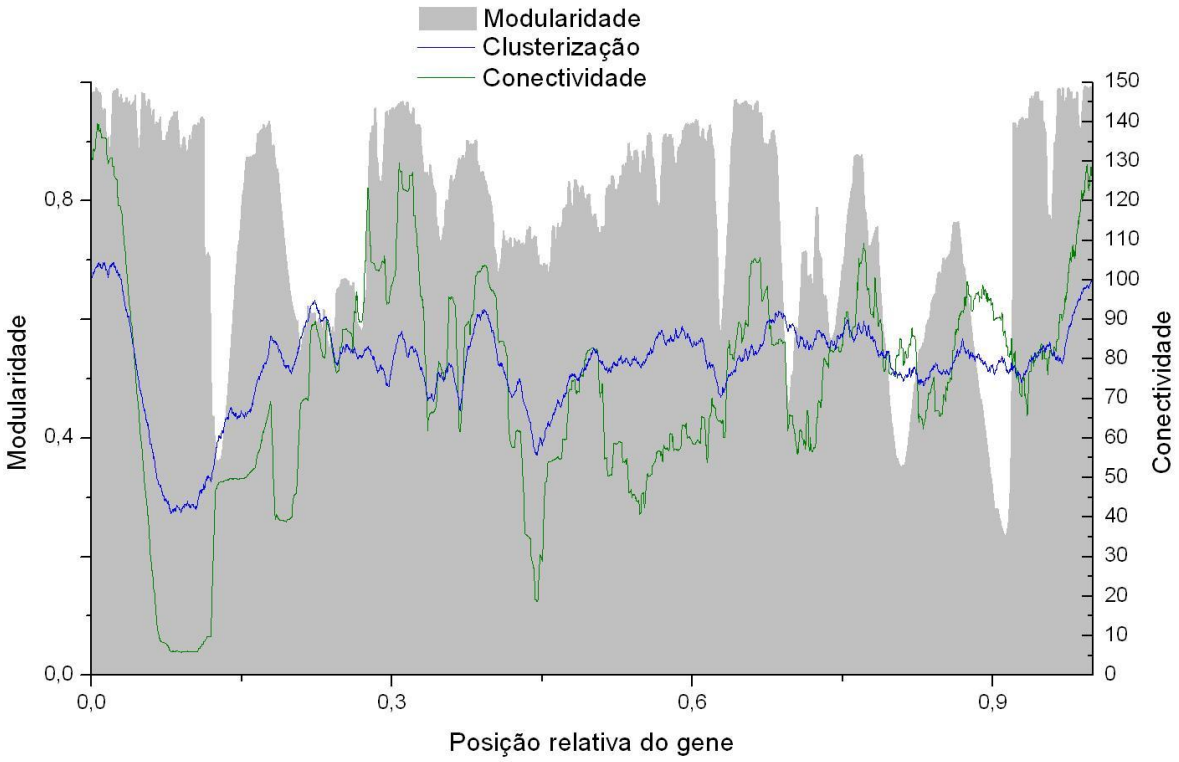


**Figura 3: Estatística da rede**

A partir desse gráfico, é feita a separação da rede em módulos funcionais. Parte-se da idéia de que os vales do gráfico são os pontos que separam os

módulos; a modularidade é o valor primário, mas a conectividade e a clusterização também devem ser usadas como base para definir as separações. Em outras palavras, em geral definimos os mínimos locais de modularidade como os pontos que separam dois módulos, mas em regiões em que a modularidade não varia muito, se houver um mínimo local de conectividade ou clusterização, é feita uma separação.

É importante ressaltar neste ponto que essa separação da rede em módulos, apesar de seguir as restrições apresentadas anteriormente, não possui uma fórmula matemática exata para ser realizada. Isso significa que, realizando mais de uma análise do sistema, os pontos de separação entre módulos obtidos podem variar um pouco. Porém, como será mencionado adiante, isso influencia pouco na escolha de funções biológicas para representar os módulos. Uma das possibilidades futuras para aprimoramento do método é exatamente a criação de um algoritmo computacional capaz de realizar a separação automaticamente e do modo mais preciso possível.

No caso do algoritmo aplicado à *Arabidopsis thaliana*, foi feita a separação em módulos funcionais representada na figura 4.

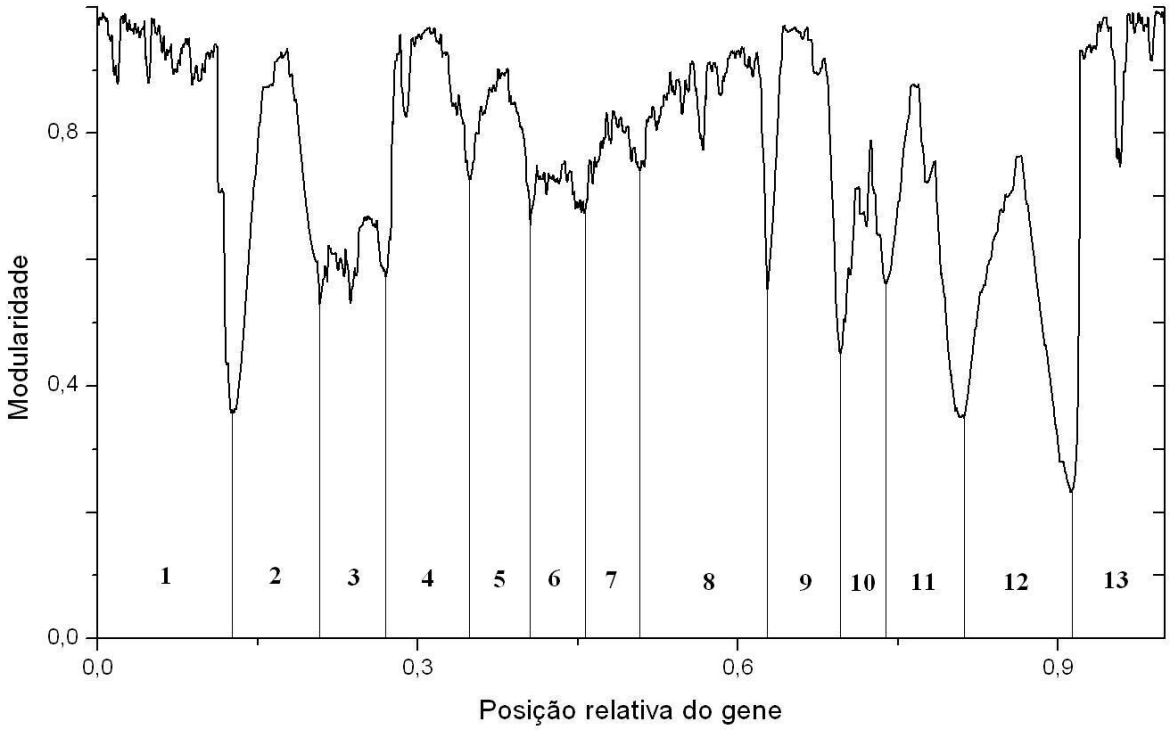


**Figura 4: Separação da rede em módulos funcionais**

Como a lista original listava os genes pelo nome, feita a separação da rede em módulos o que possuímos são as listas dos genes que fazem parte de cada um dos módulos observados.

## 2.3 FUNÇÕES BIOLÓGICAS

Com essa informação em mãos, é possível atribuir a cada módulo do ordenamento funções biológicas do organismo que podem estar relacionadas com aquele agrupamento particular de genes. Para tanto, em primeiro lugar, é utilizada a Base de Dados para Anotação, Visualização, e Descoberta Integrada[6,7] (DAVID, *Database for Annotation, Visualization and Integrated Discovery*), cuja interface é apresentada na figura 5.


**Figura 5: Base de dados do DAVID**

A base de dados do DAVID é capaz de, tendo como entrada uma lista de genes de um dado organismo que definimos como sendo aqueles que constituem um módulo funcional, listar um grande número de funções e processos biológicos daquele dado organismo que podem estar associados com a lista de genes apresentada, ou seja, com o módulo em questão.

Para tanto, selecionamos a opção de anotação funcional (*functional annotation*), inserimos a lista de genes de entrada e selecionamos o identificador TAIR_ID, normalmente utilizado para a *Arabidopsis thaliana*. Filtramos os resultados

para a exibição apenas do termo GOTERM_BP_ALL, que significa que serão exibidos apenas os resultados dos genes relacionados a processos biológicos do organismo. Selecionando a opção tabela de anotação funcional (*functional annotation chart*), é apresentado um grande número de funções biológicas, junto de vários valores que usaremos como critério para selecionar os melhores candidatos a representarem o módulo.

Os critérios mais importantes a serem utilizados para a escolha de uma função são o número de genes integrantes do módulo que participam daquela função, o número total de genes que participam daquela função e um valor probabilístico conhecido como Benjamini, calculado através do procedimento de Benjamini-Hochberg, que na prática pode ser interpretado como a probabilidade de que aquela função específica corresponda ao módulo selecionado.

É importante que o número de genes integrantes do módulo que participam da função seja o maior possível, enquanto que o número total de genes que participam da função seja o menor possível, pois selecionar uma função composta de muitos genes invariavelmente implica que esta função estará espalhada em vários módulos pela rede e não pode ser atribuída a apenas um agrupamento específico de genes.

Utilizando as restrições acima para a escolha das funções mais prováveis de representar cada módulo, obtemos os resultados apresentados na tabela 1.

| **Módulo** | **Função** | **Identificador** |
|---|---|---|
| 1 | Processo catabólico de proteínas dependente de ubiquitina | GO:0006511 |
| | Processo catabólico de macromoléculas | GO:0009057 |
| | Processo catabólico de proteínas | GO:0030163 |
| | Proteólise envolvida no processo catabólico de proteínas | GO:0051603 |
| 2 | Geração de metabólitos e energia | GO:0006091 |
| | Derivação de energia por oxidação de componentes orgânicos | GO:0015980 |
| | Processo metabólico de aminas | GO:0044106 |
| 3 | Processo metabólico de glucose | GO:0006006 |
| | Processo de biossíntese de compostos nitrogenados | GO:0044271 |
| | Processo catabólico de alcoóis | GO:0046164 |
| 4 | Resposta a espécimes oxigenados | GO:0000302 |
| | Processo metabólicos de oxigênio e espécimes oxigenados | GO:0006800 |
| | Resposta ao peróxido de hidrogênio | GO:0042542 |
| 5 | Processo metabólico de ácidos orgânicos | GO:0006082 |
| | Processo metabólico de ácidos carboxílicos | GO:0019752 |
| | Processo metabólico de cetonas | GO:0042180 |
| | Processo metabólico de oxiácidos | GO:0043436 |
| | Redução de oxidação | GO:0055114 |

| Módulo | Função | Identificador |
|---|---|---|
| 6 | Caminhos de sinalização intracelular | GO:0007242 |
| | Desenvolvimento de órgãos | GO:0048513 |
| | Desenvolvimento de sistemas | GO:0048731 |
| 7 | Processo metabólico de ácidos orgânicos | GO:0006082 |
| | Resposta a estímulos abióticos | GO:0009628 |
| | Processo metabólico de ácidos carboxílicos | GO:0019752 |
| | Processo metabólico de cetonas | GO:0042180 |
| | Resposta a estímulos químicos | GO:0042221 |
| | Processo metabólico de oxiácidos | GO:0043436 |
| 8 | Processo metabólico de ácidos orgânicos | GO:0006082 |
| | Geração de metabólitos e energia | GO:0006091 |
| | Processo metabólico de ácidos carboxílicos | GO:0019752 |
| | Processo metabólico de cetonas | GO:0042180 |
| | Processo metabólico de oxiácidos | GO:0043436 |
| | Processo de biossíntese de compostos nitrogenados | GO:0044271 |
| | Redução de oxidação | GO:0055114 |
| 9 | Processo metabólico de alcoóis | GO:0006066 |
| | Processo metabólico de nucleosídeos de fosfato | GO:0006753 |
| | Processo metabólico de nucleotídeos | GO:0009117 |
| | Processo metabólico de nucleobases, nucleosídeos e nucleotídeos | GO:0055086 |
| 10 | Processo metabólico de monossacarídeos | GO:0005996 |
| | Processo metabólico de compostos de enxofre | GO:0006790 |
| | Processo metabólico de hexoses | GO:0019318 |
| | Redução de oxidação | GO:0055114 |
| 11 | Processo metabólico de ácidos orgânicos | GO:0006082 |
| | Processo metabólico de ácidos carboxílicos | GO:0019752 |
| | Processo metabólico de cetonas | GO:0042180 |
| | Processo metabólico de oxiácidos | GO:0043436 |
| 12 | Resposta a estímulos abióticos | GO:0009628 |
| | Resposta a substâncias inorgânicas | GO:0010035 |
| | Processo metabólico de cetonas | GO:0042180 |
| | Processo de biossíntese de compostos nitrogenados | GO:0044271 |
| 13 | Resposta a estímulos abióticos | GO:0009628 |
| | Processo metabólico de compostos heterocíclicos | GO:0046483 |
| | Redução de oxidação | GO:0055114 |

**Tabela 1: Funções biológicas relacionadas aos módulos funcionais**

Devido à grande variedade de resultados do DAVID, diferentes definições das separações entre os módulos funcionais influenciam pouco no resultado final, pois um grande número de funções é listado, e essas tendem a aparecer mesmo que a correspondência entre o módulo e a função não seja exata.

Em seguida, para verificar a relação existente entre essas funções e os genes da lista ordenada, devemos consultar a base de dados Ontologia dos Genes[8] (*The Gene Ontology*), cuja interface é apresentada na figura 6.

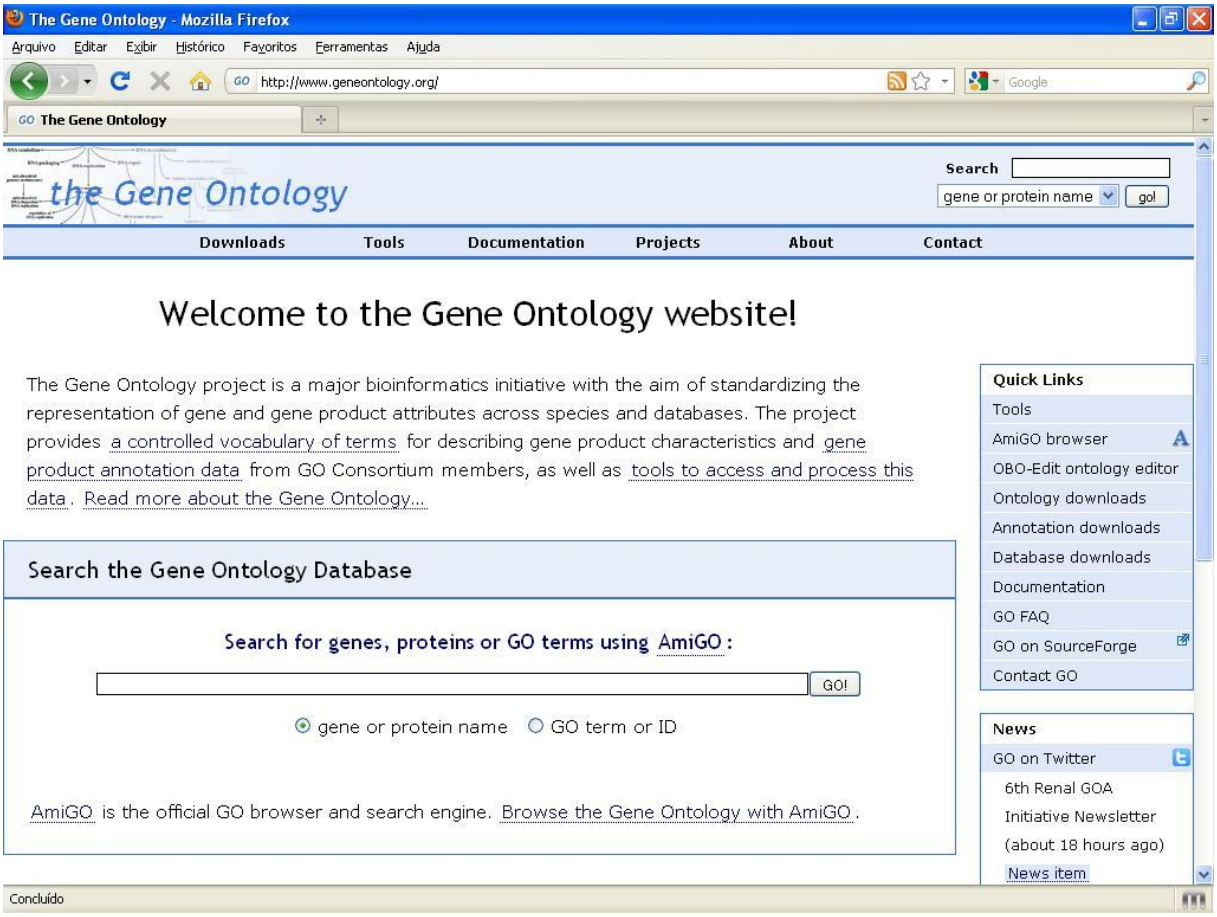


**Figura 6: Base de dados do The Gene Ontology**

A ontologia é definida na ciência da informação como uma representação de um conjunto de dados e das relações esses dados. No caso dos genes, a ontologia representa a ativação dos genes nas funções biológicas e as relações pelas quais um gene causa a ativação de outro.

A função da base de dados do Gene Ontology é, em geral, o oposto daquela do DAVID, que nos apresentava, dada uma lista dos genes em um módulo, as funções que podem corresponder àquele módulo; ela apresenta, dada uma função do organismo, a lista de genes daquele organismo que faz parte daquela função.

Em primeiro lugar utilizamos o identificador biológico de cada função, obtido do DAVID, para buscar uma lista de genes que correspondem àquela função, filtrando os resultados para a exibição apenas dos pertencentes à *Arabidopsis thaliana*. A lista resultante é apresentada através dos nomes dos identificadores dos genes. Esta é traduzida através de um dicionário obtido na base de dados do EnsemblPlants[9] e, então, comparada com a lista de genes do ordenamento final.

Para analisar uma função biológica, a cada gene atribuímos o valor 1 se ele faz parte da lista do Gene Ontology correspondente à função, ou o valor 0 se ele não

aparece na lista. Esse valor representa a ativação ou não do gene na função. Em seguida, atribuímos a cada gene o valor médio da ativação dos genes em a uma janela de tamanho $2w + 1$ centrada nele. Matematicamente, os cálculos são definidos conforme a equação $(7)$. Como já mencionado anteriormente, no caso da *Arabidopsis thaliana* foi utilizado o tamanho da janela de $251$.

$$a_i = \begin{cases} 1, se\ o\ gene\ pertence\ \text{à}\ lista \\ 0, se\ o\ gene\ não\ pertence\ \text{à}\ lista \end{cases}$$
$$A_i = \frac{\sum_{j=i-w}^{i+w} a_j}{2w + 1} \qquad (7)$$

O valor resultante dessa média de ativação da rede é comparado ao valor da modularidade da rede. A presença de um pico no gráfico de ativação na mesma região de um módulo funcional implica que os genes pertencentes àquele módulo são os responsáveis por aquela função e, portanto, podemos estabelecer uma relação direta entre o módulo e a função biológica em questão. Para facilitar a visualização, é conveniente normalizar os gráficos da ativação média de acordo com seu valor máximo. Os resultados obtidos são apresentados nas figuras $7$ a $38$.

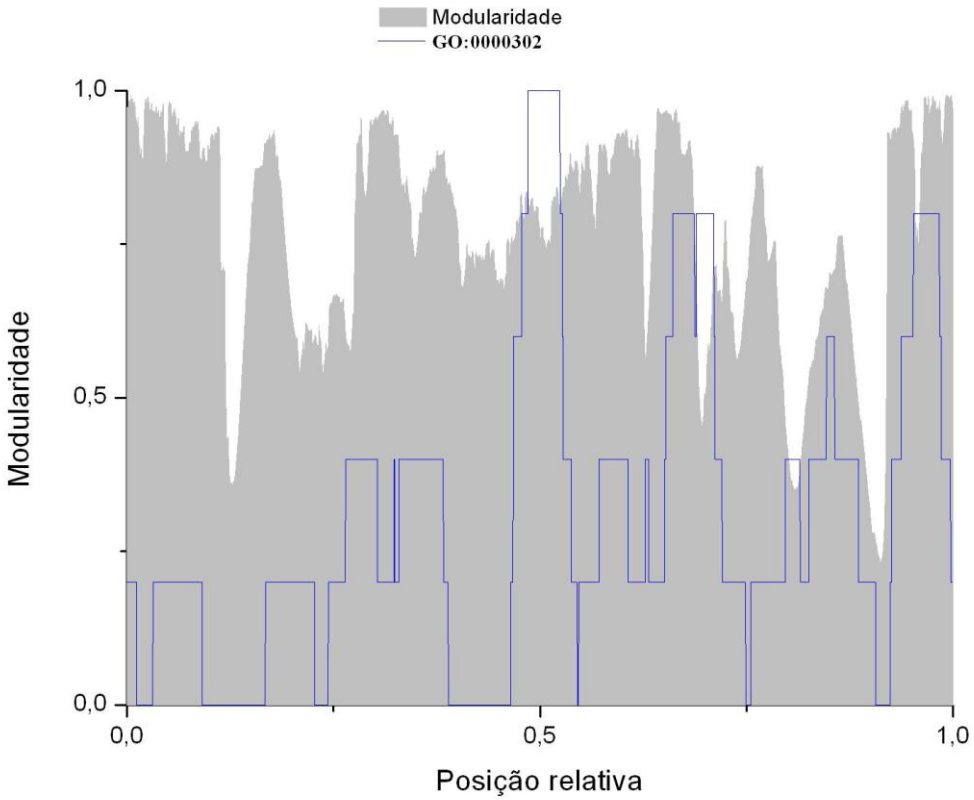


**Figura 7: Função GO:0000302**

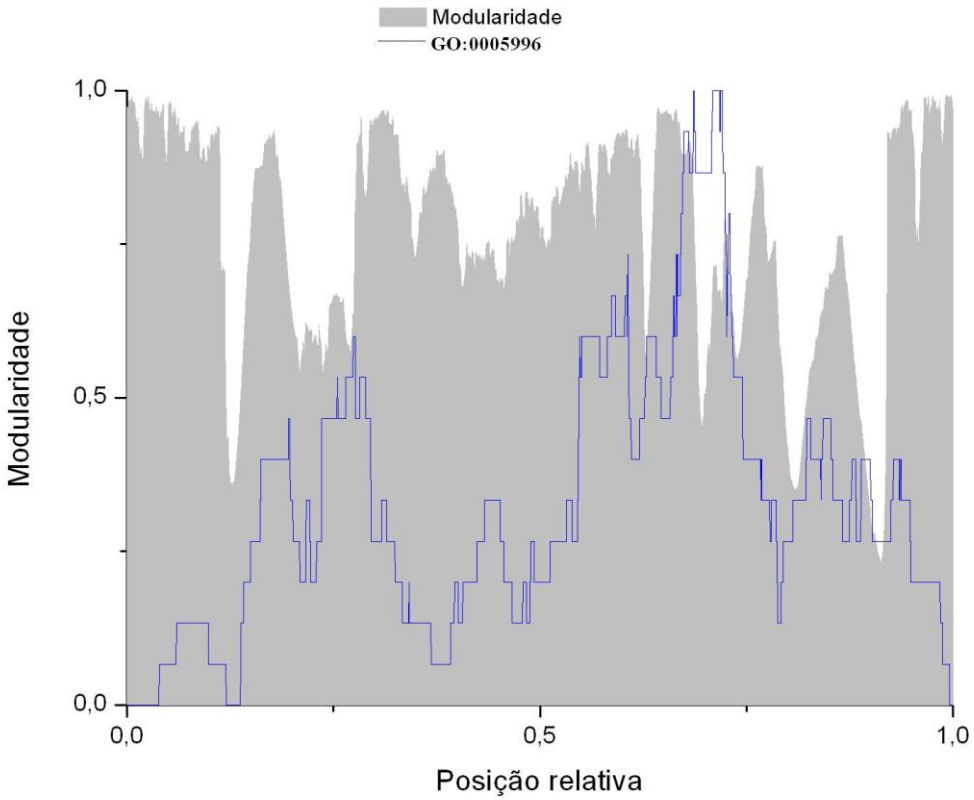


Figura 8: Função GO:0005996

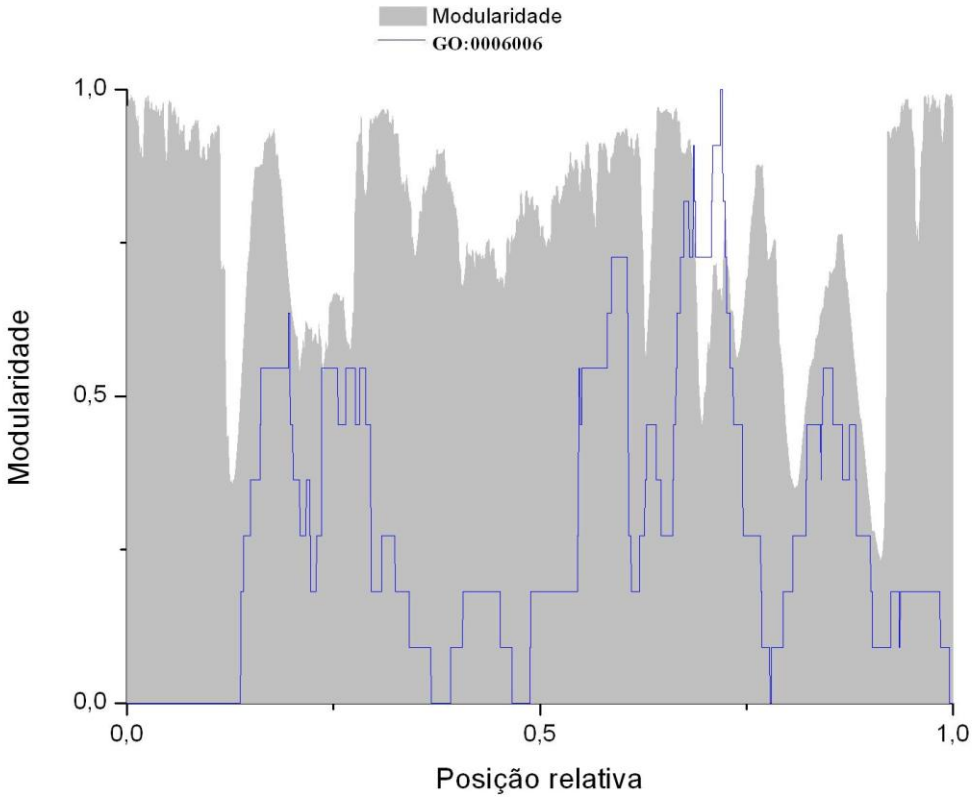


Figura 9: Função GO:0006006

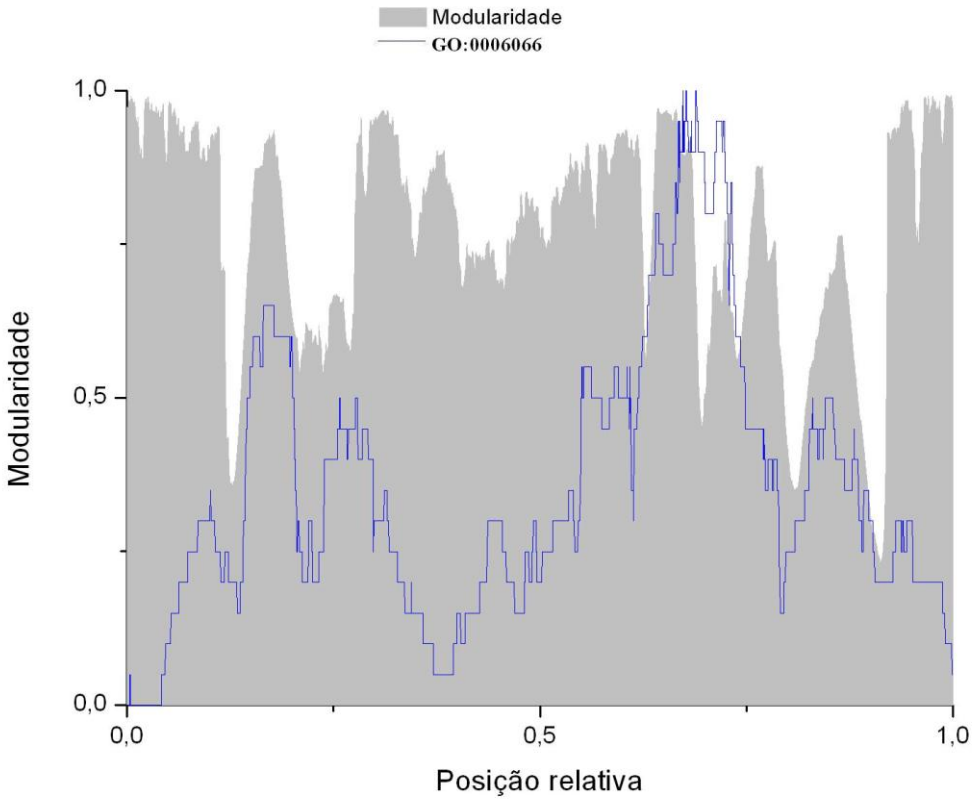


Figura 10: Função GO:0006066

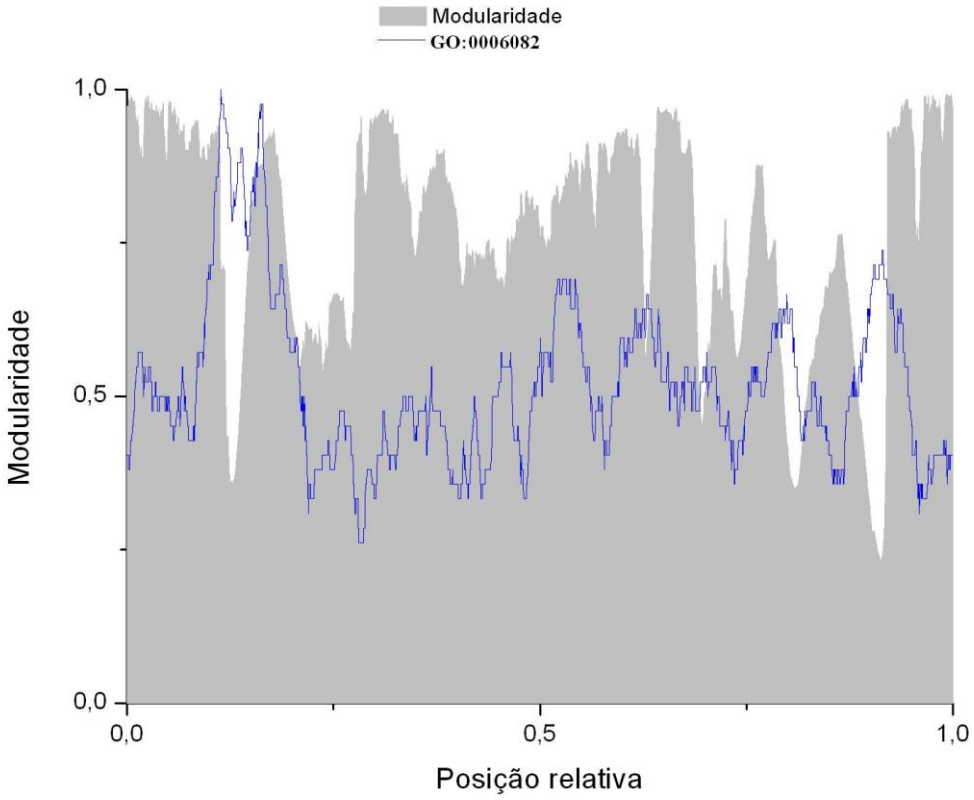


Figura 11: Função GO:0006082

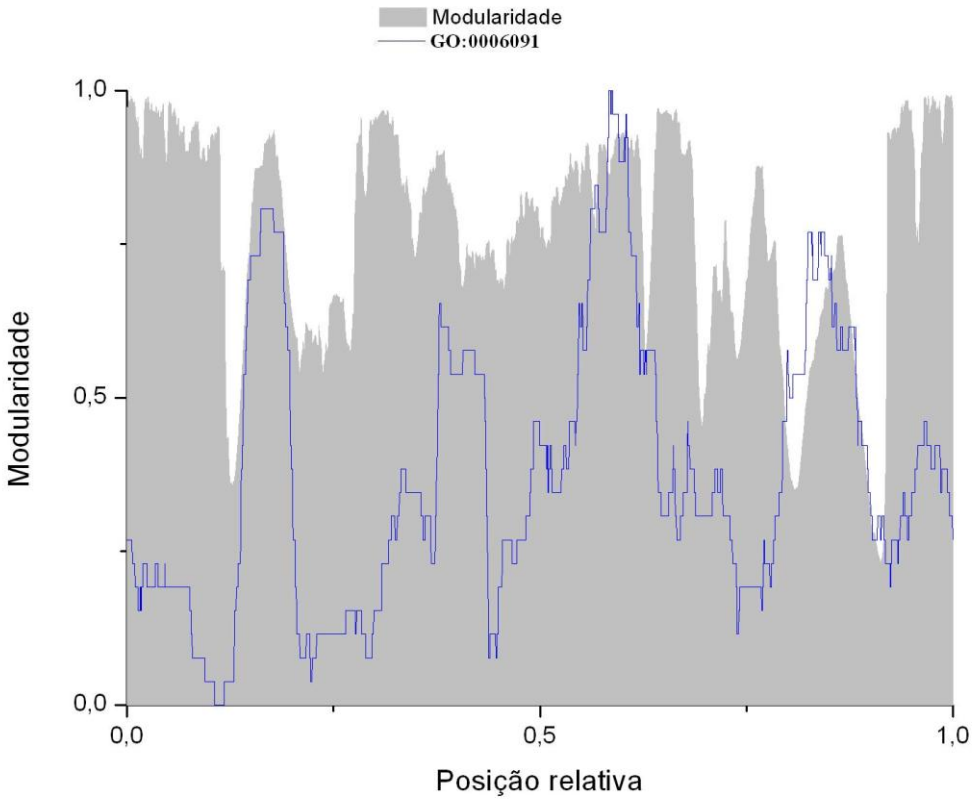


Figura 12: Função GO:0006091

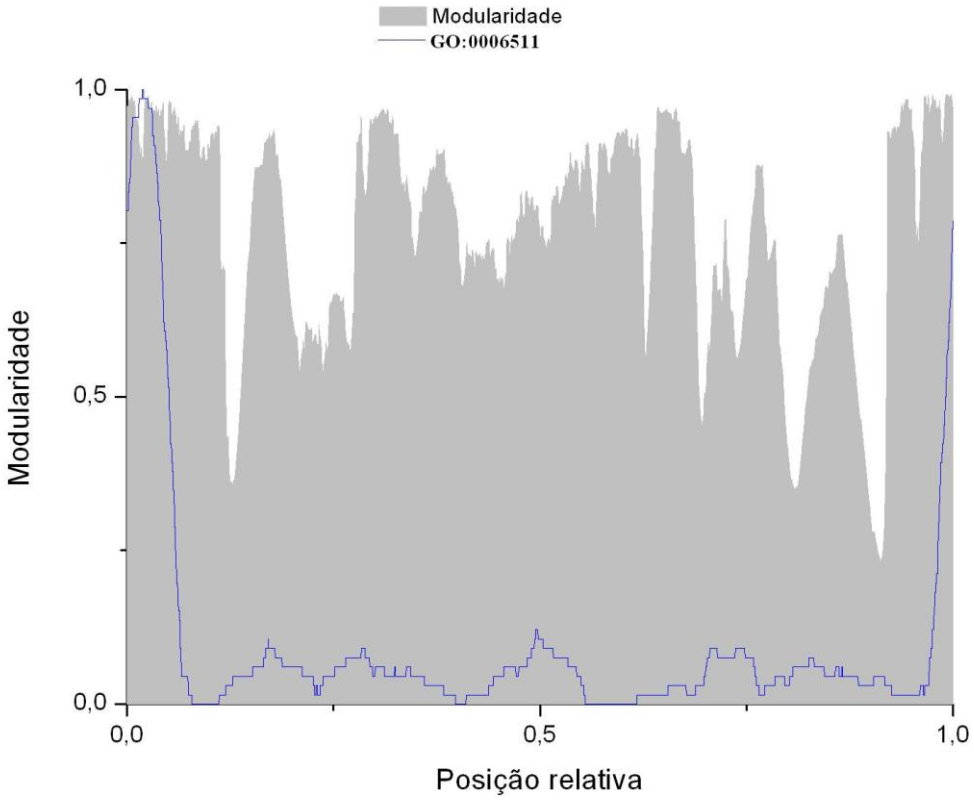


Figura 13: Função GO:0006511

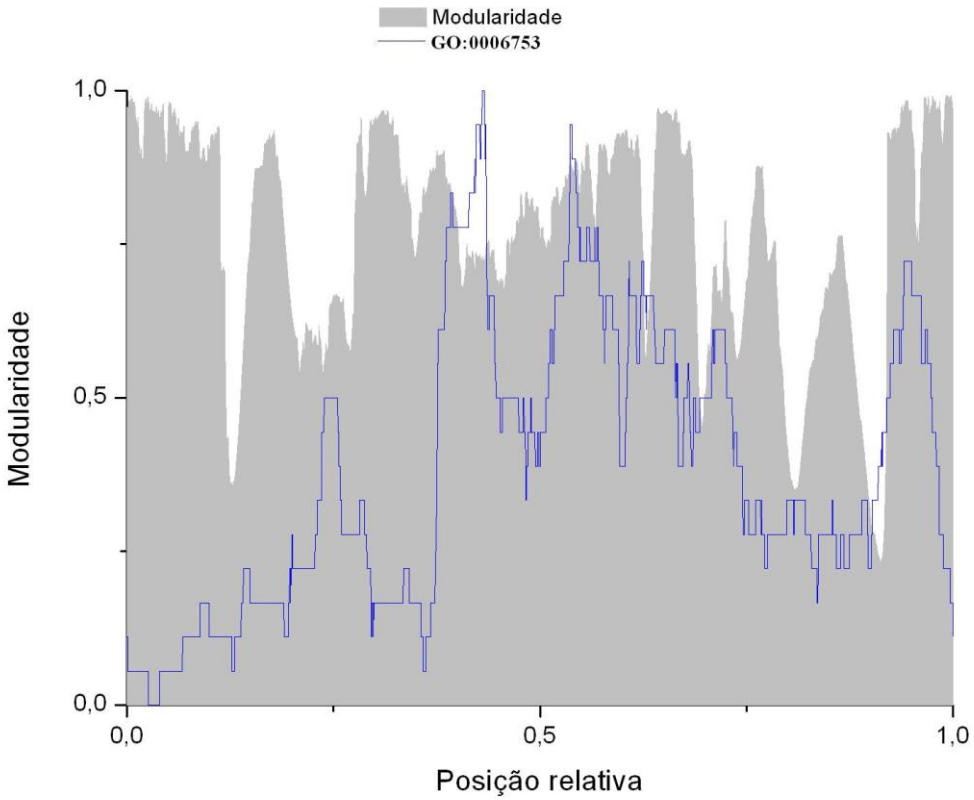


Figura 14: Função GO:0006753

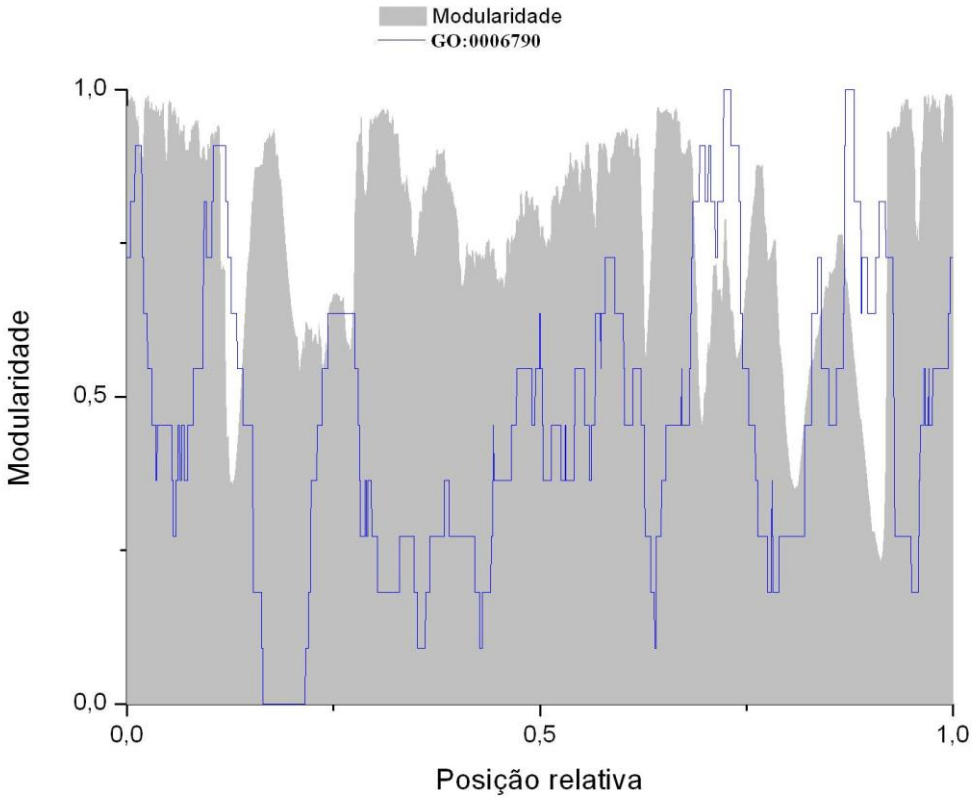


Figura 15: Função GO:0006790

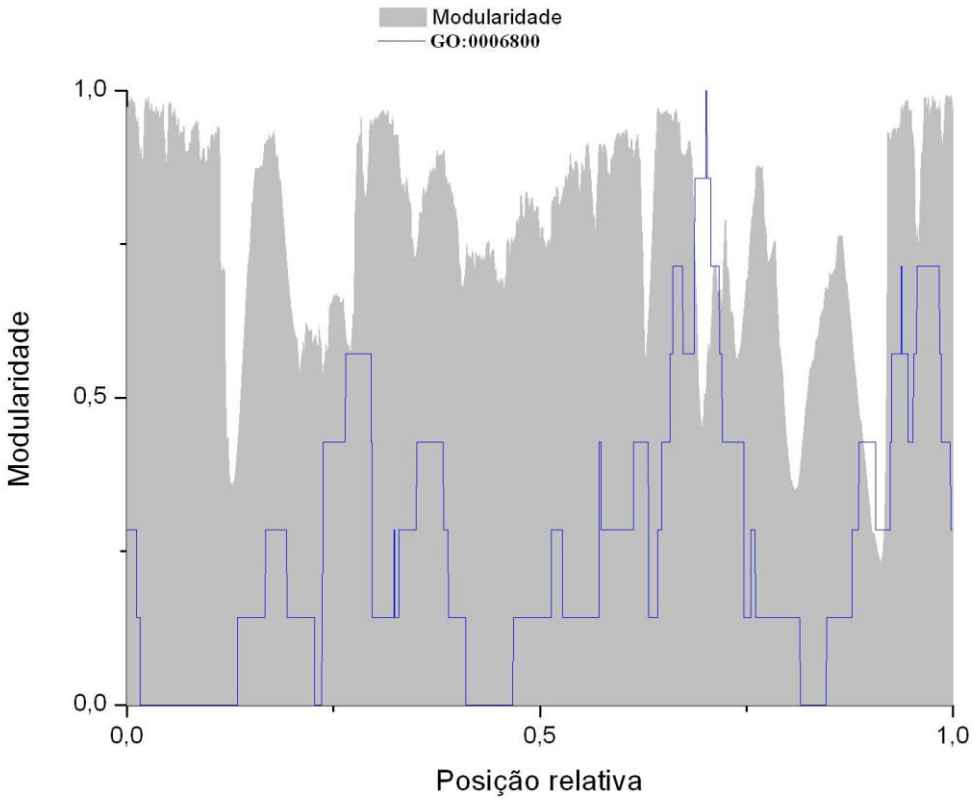


Figura 16: Função GO:0006800

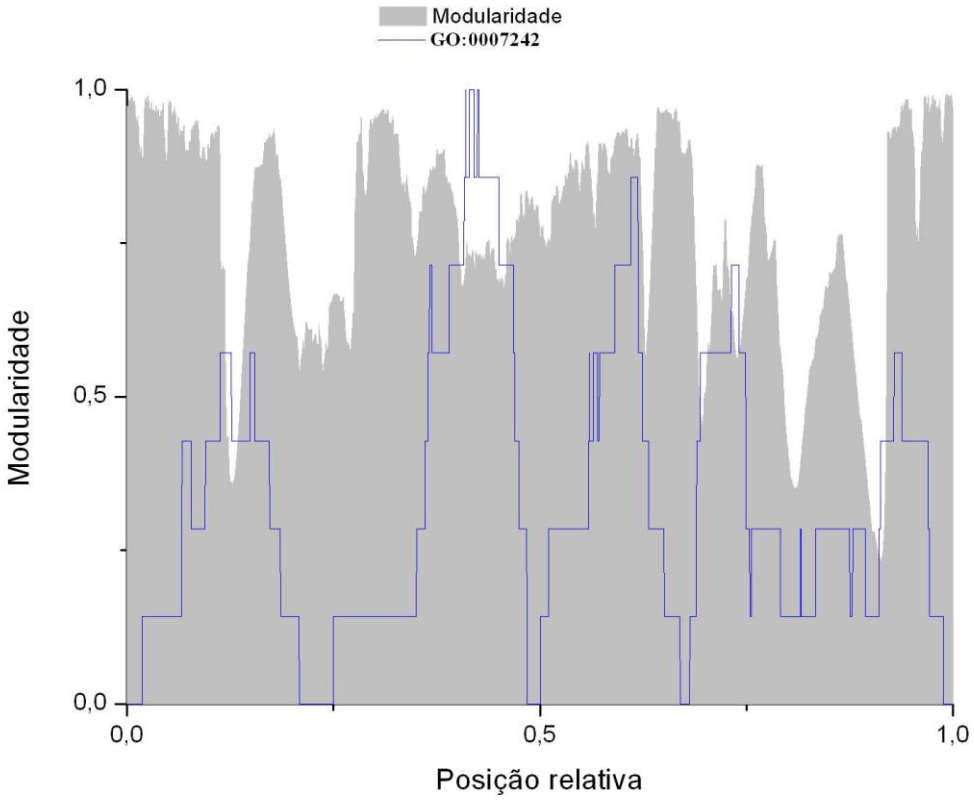


Figura 17: Função GO:0007242

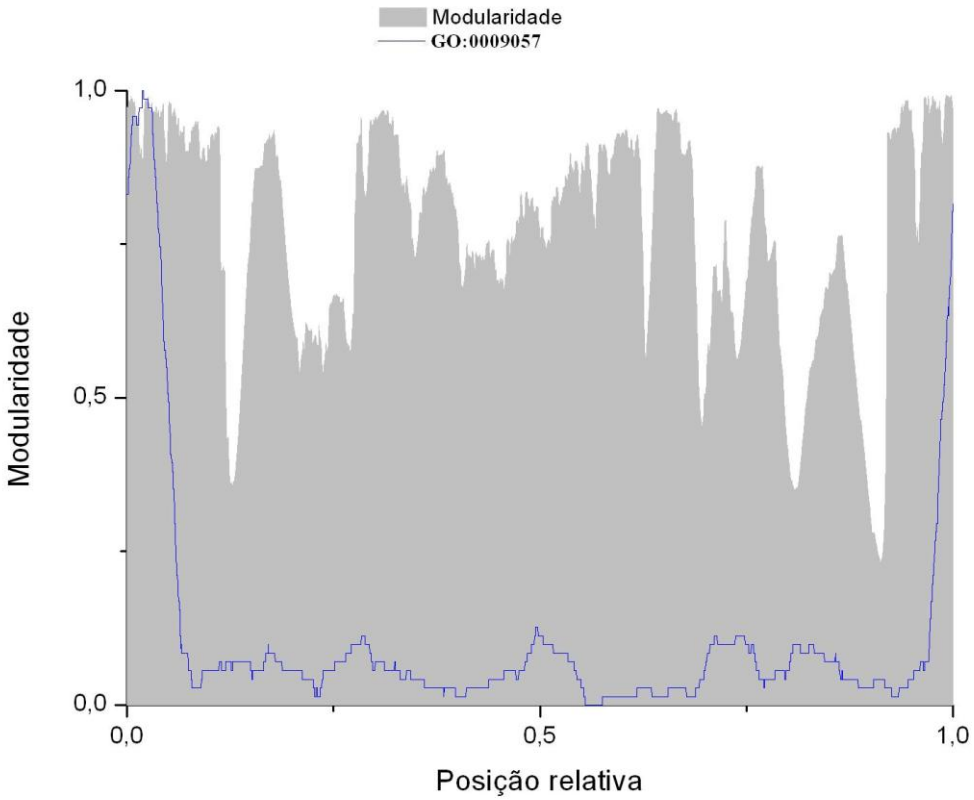


Figura 18: Função GO:0009057

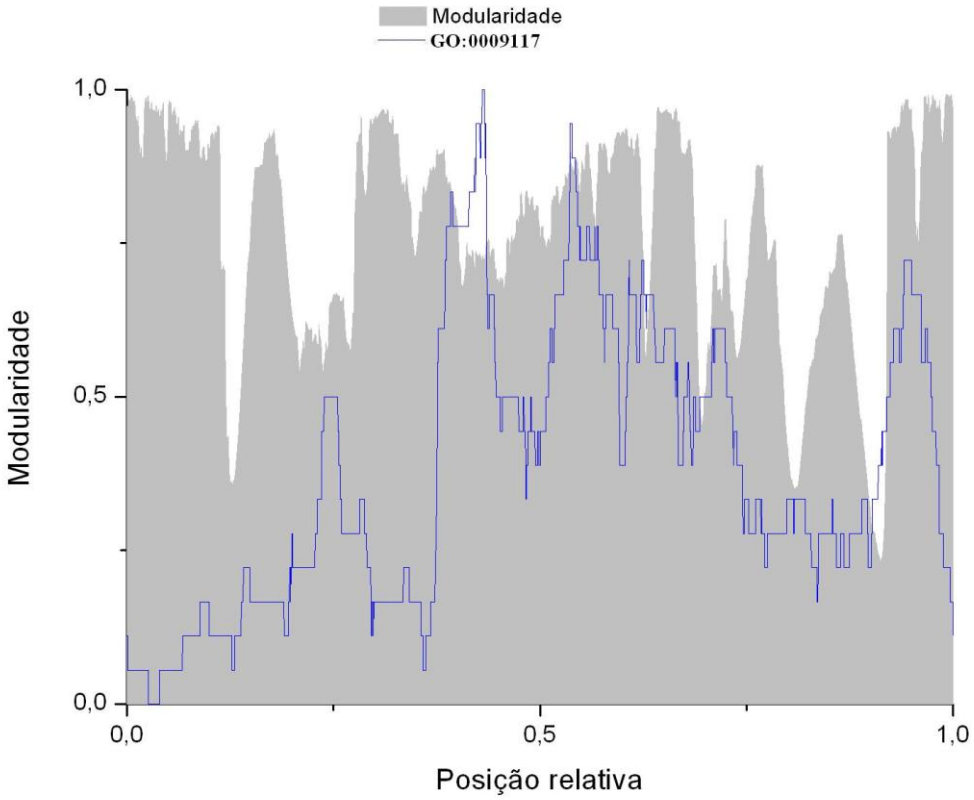


Figura 19: Função GO:0009117

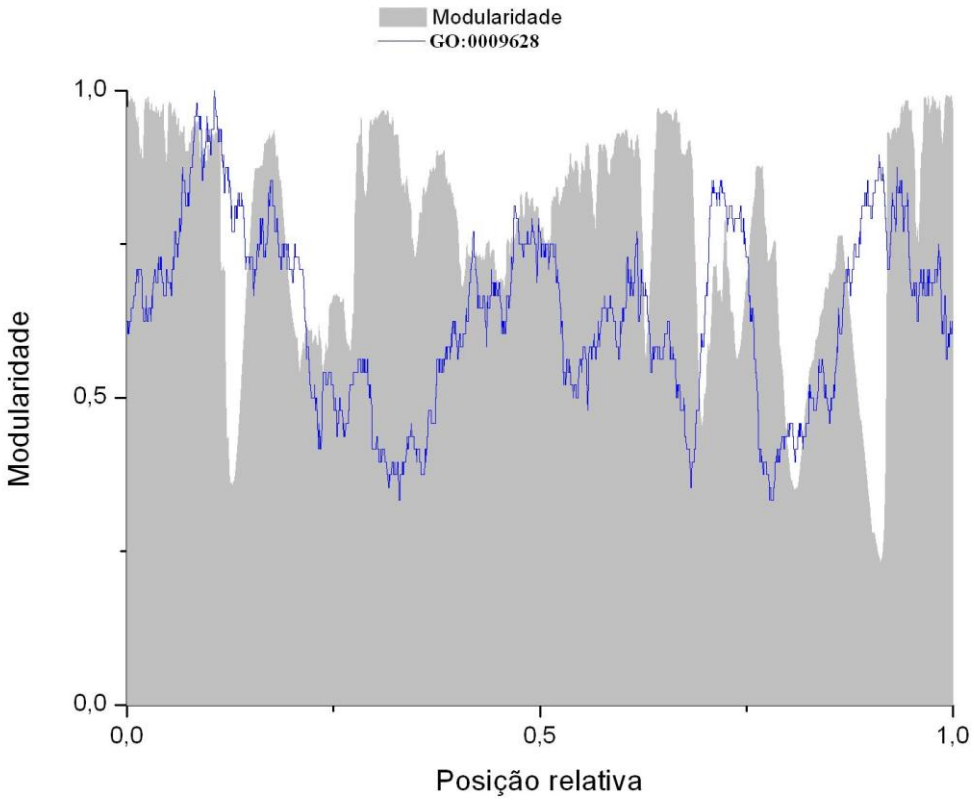


Figura 20: Função GO:0009628

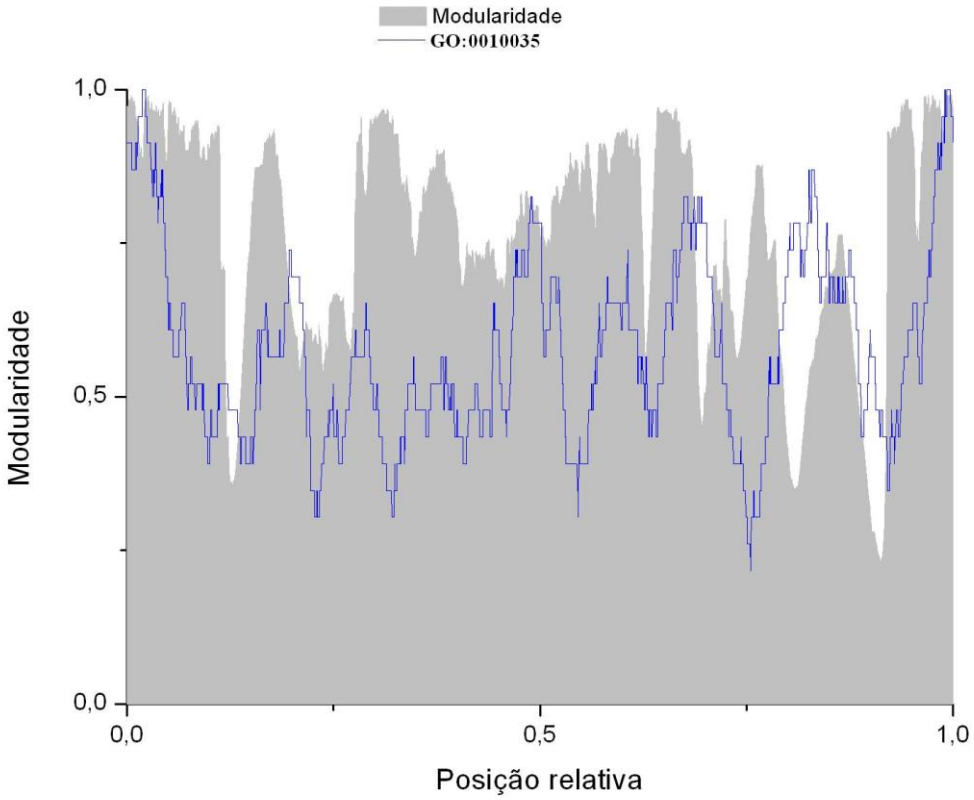


Figura 21: Função GO:0010035

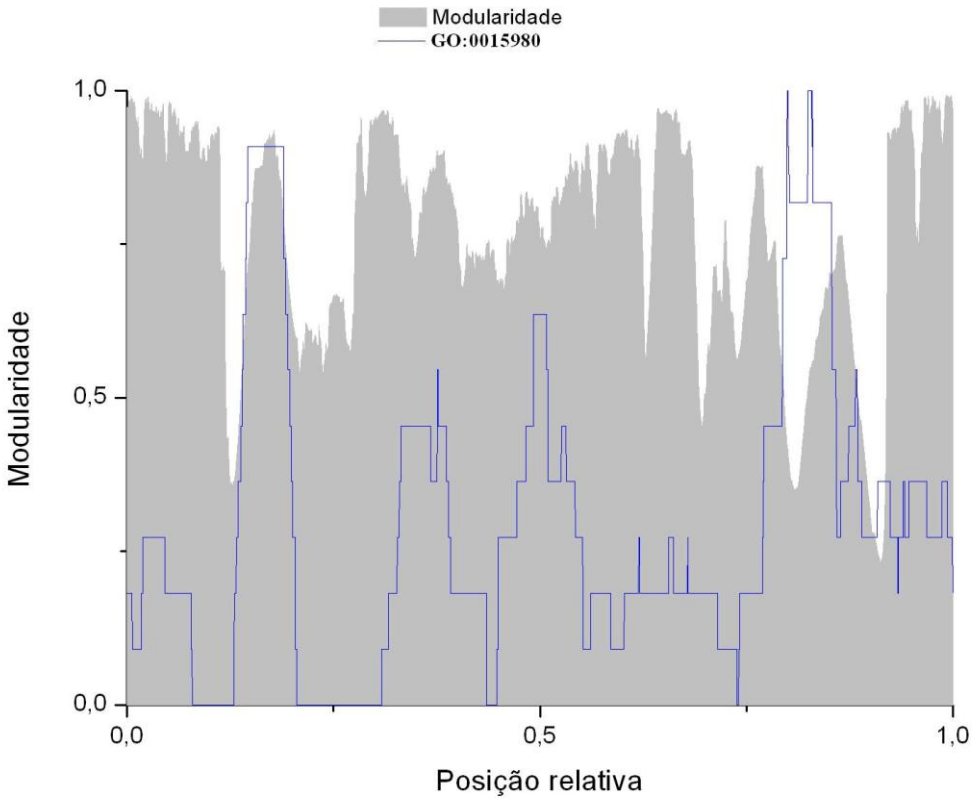


Figura 22: Função GO:0015980

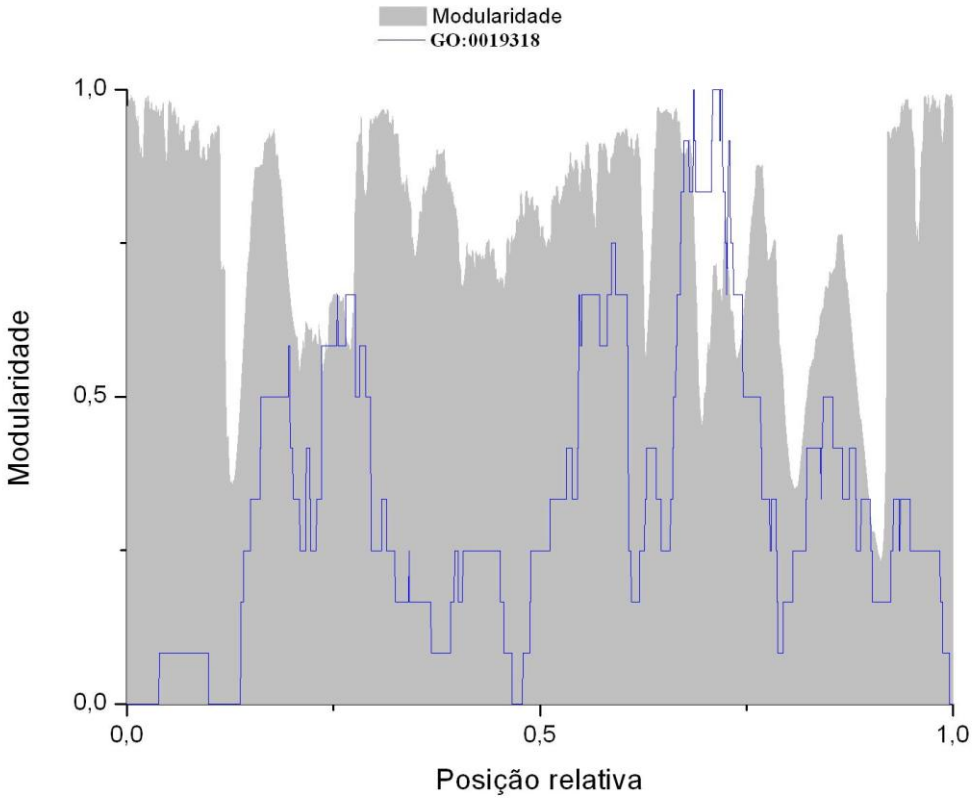


Figura 23: Função GO:0019318

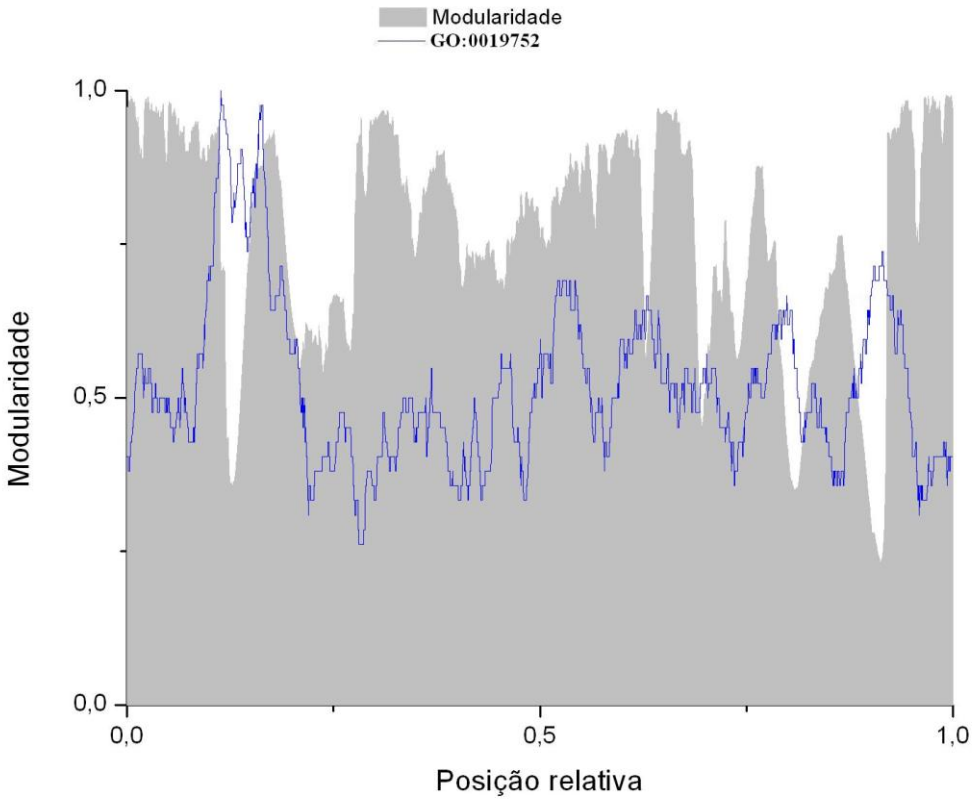


Figura 24: Função GO:0019752

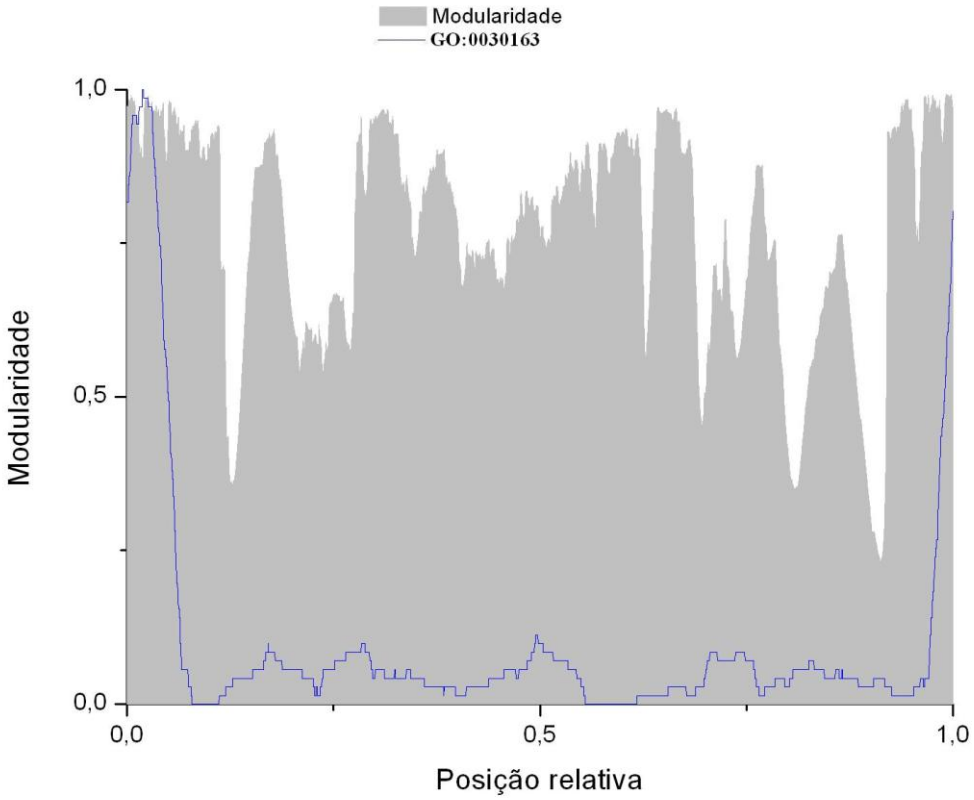


Figura 25: Função GO:0030163

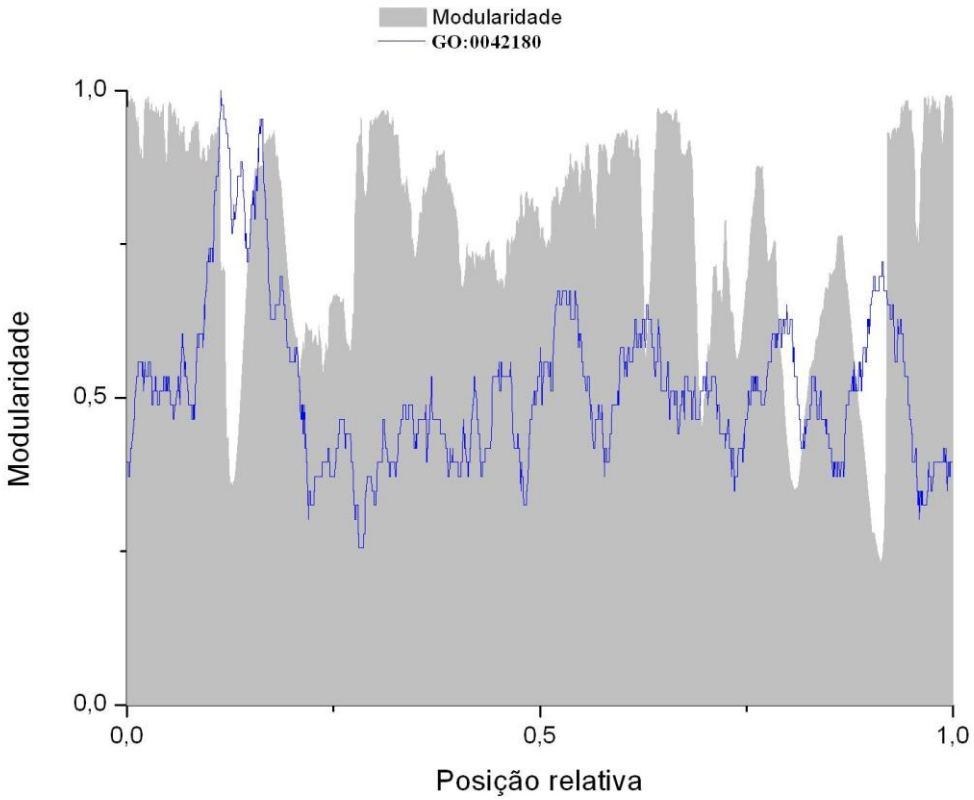


Figura 26: Função GO:0042180

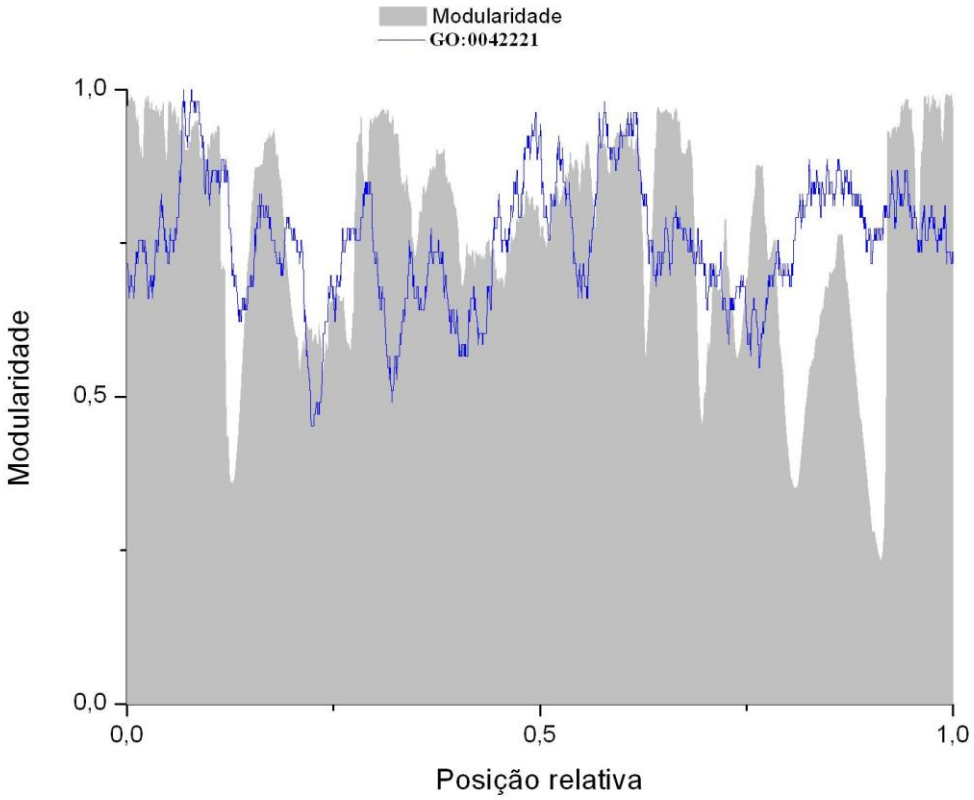


Figura 27: Função GO:0042221

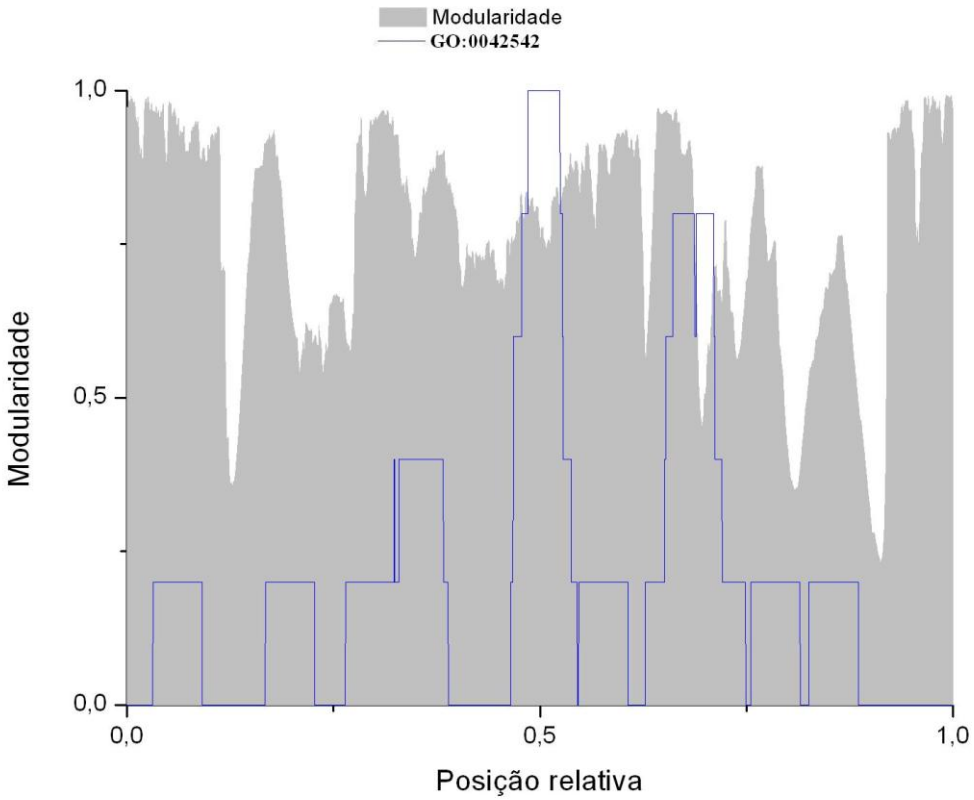


Figura 28: Função GO:0042542

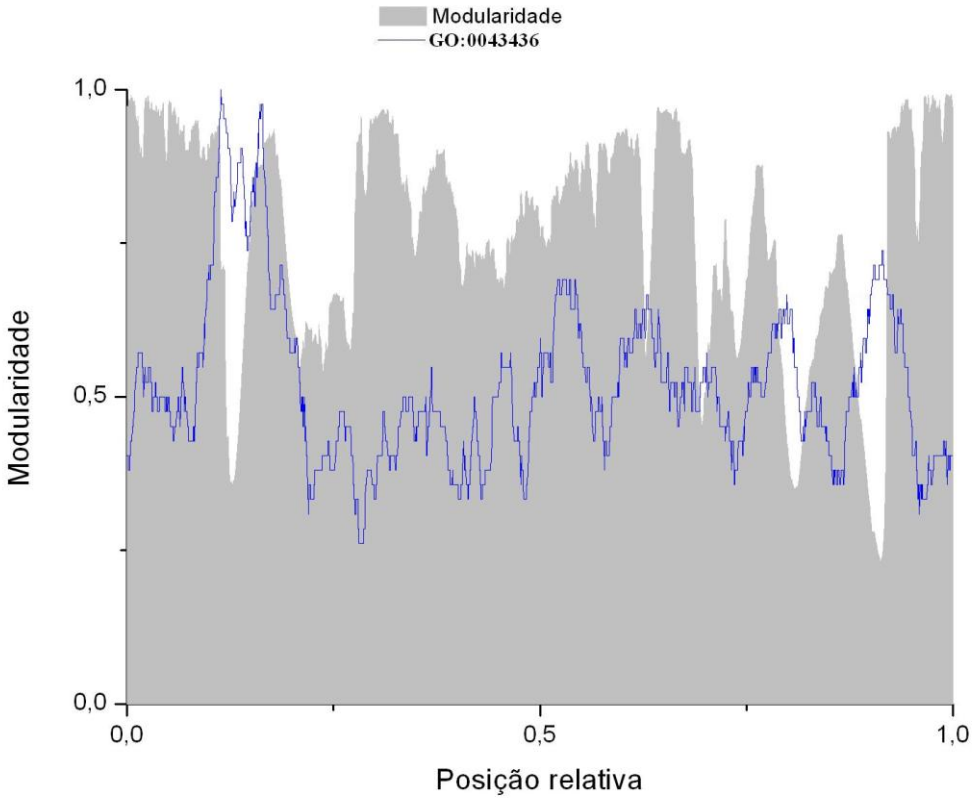


Figura 29: Função GO:0043436

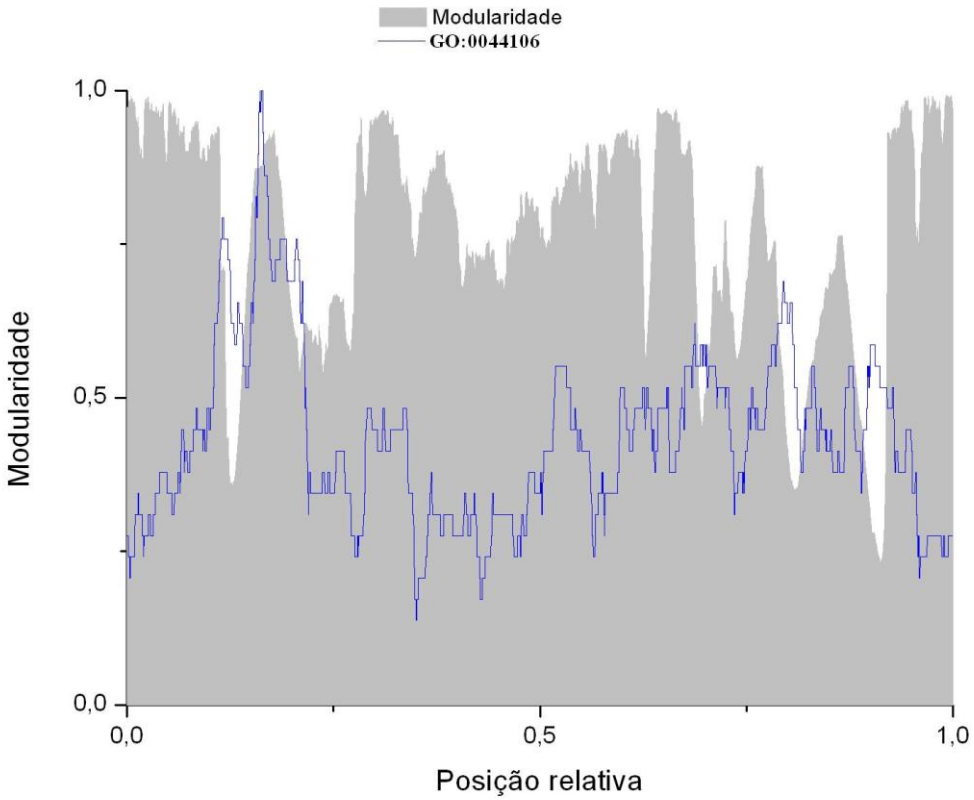


Figura 30: Função GO:0044106

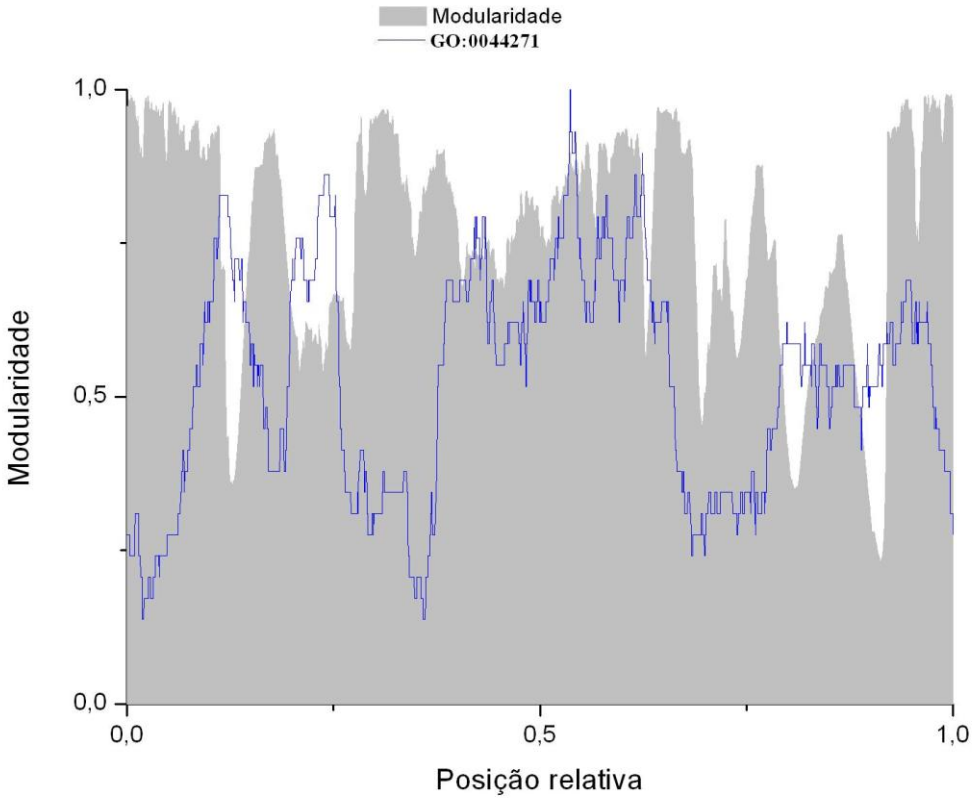


Figura 31: Função GO:0044271

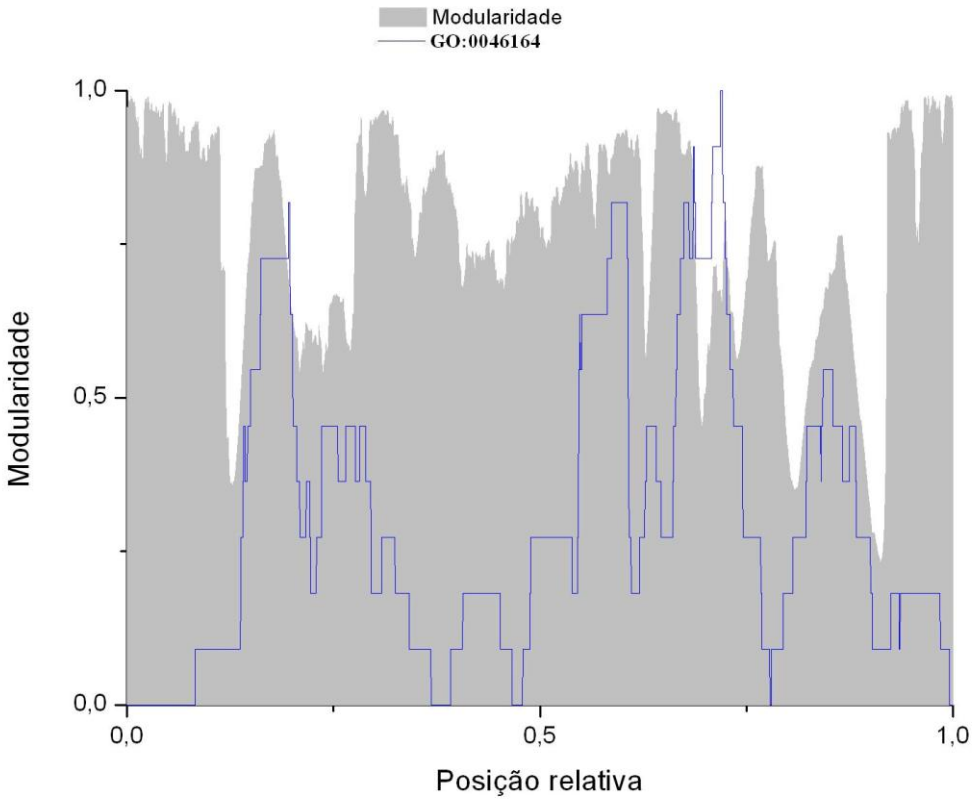


Figura 32: Função GO:0046164

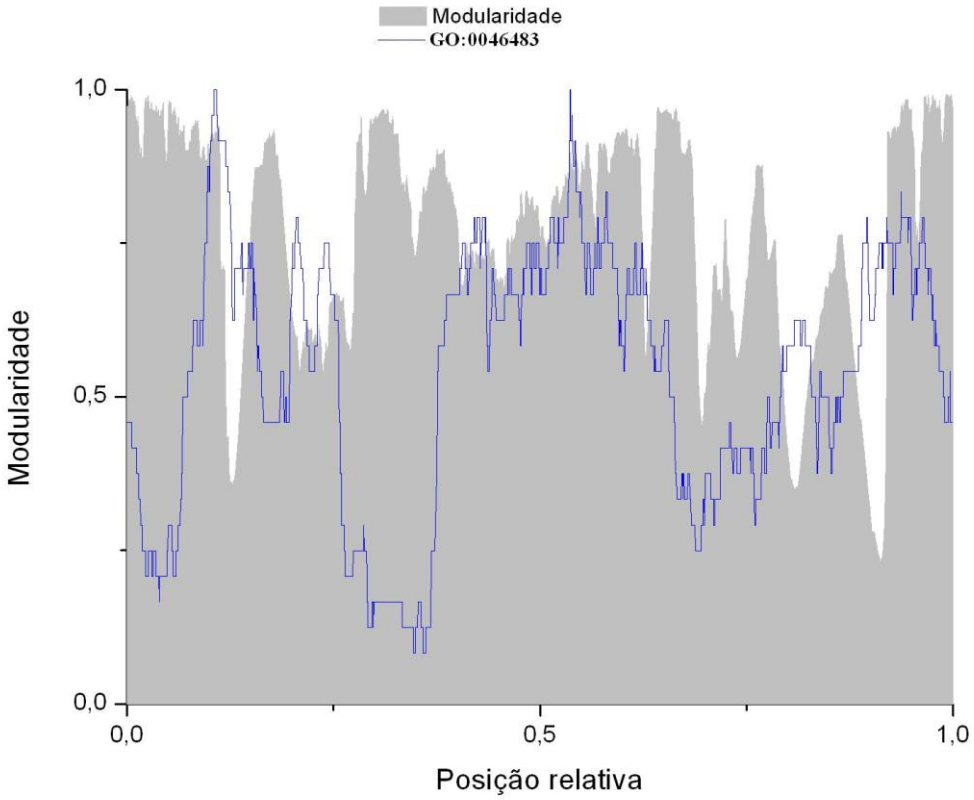


Figura 33: Função GO:0046483

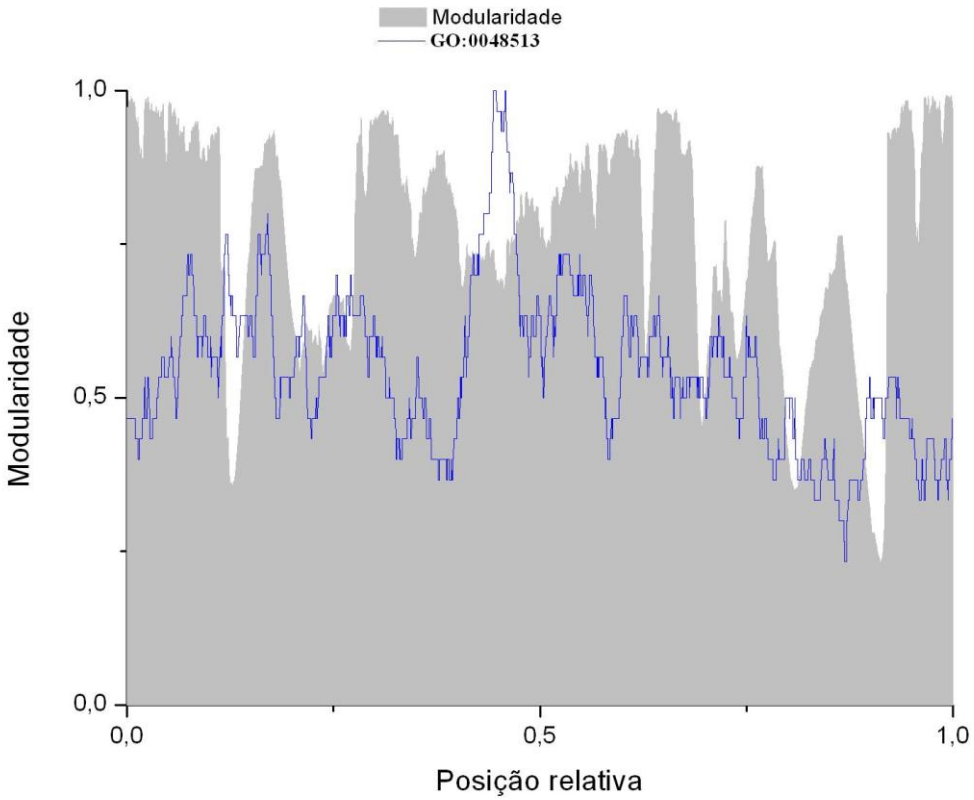


Figura 34: Função GO:0048513

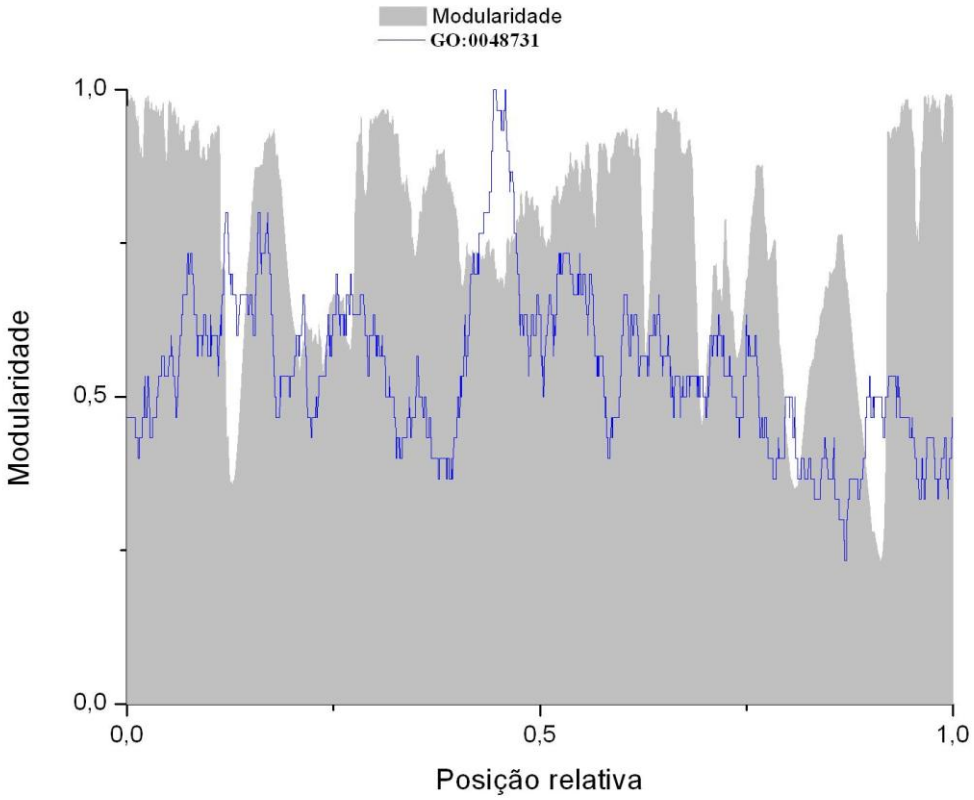


Figura 35: Função GO:0048731

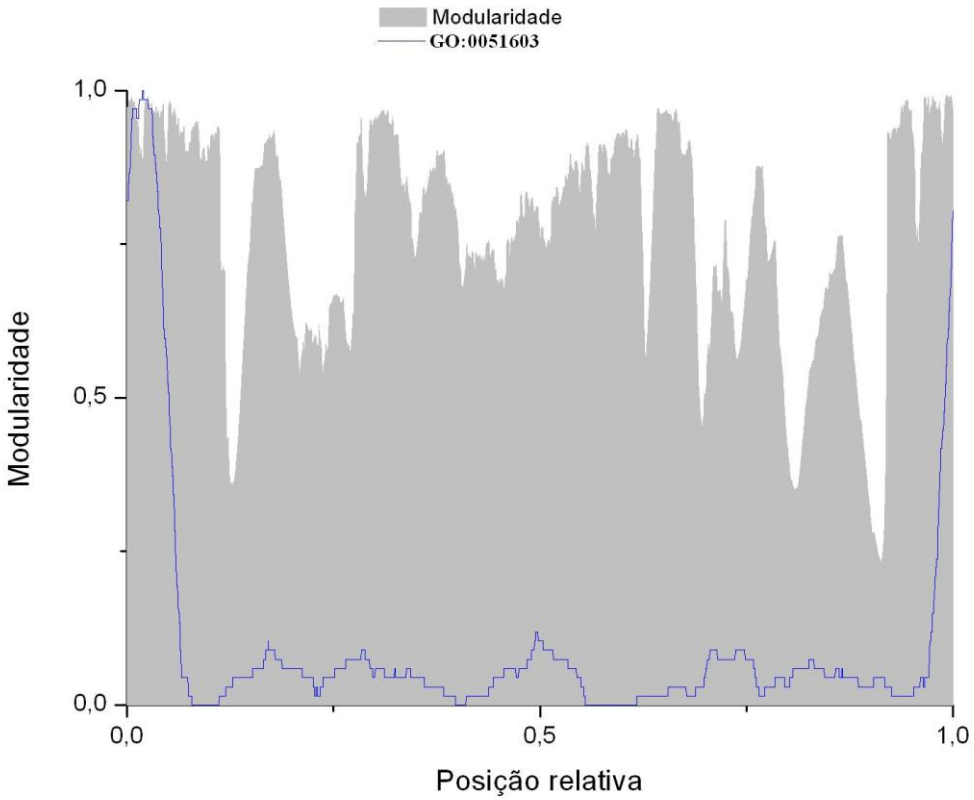


Figura 36: Função GO:0051603

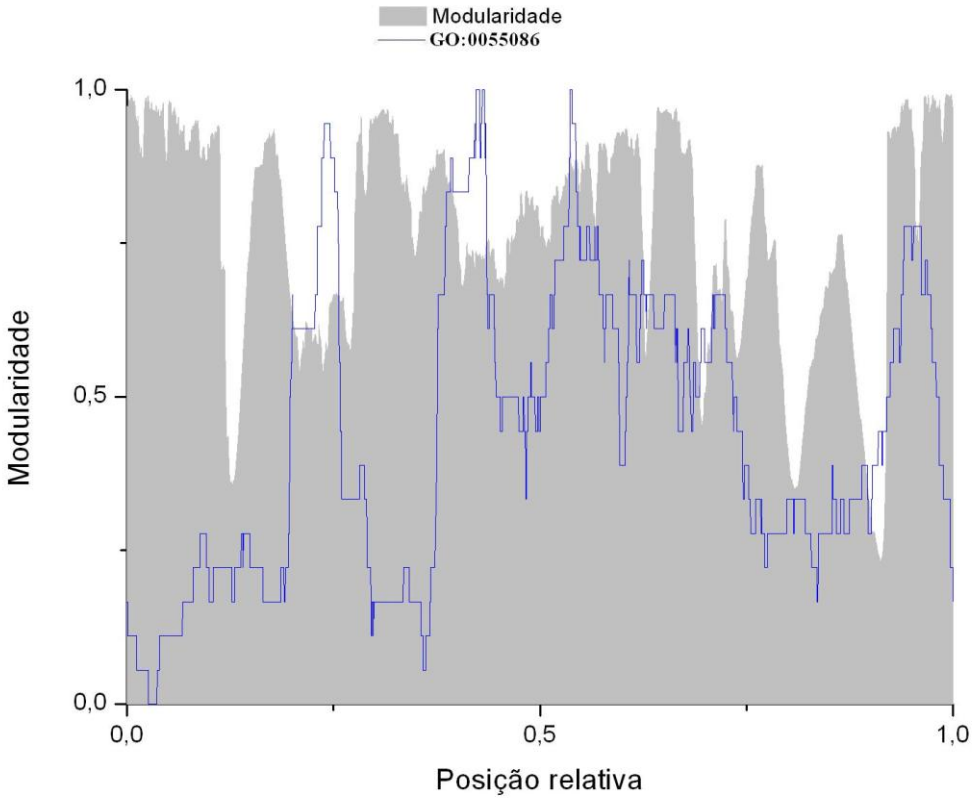


Figura 37: Função GO:0055086

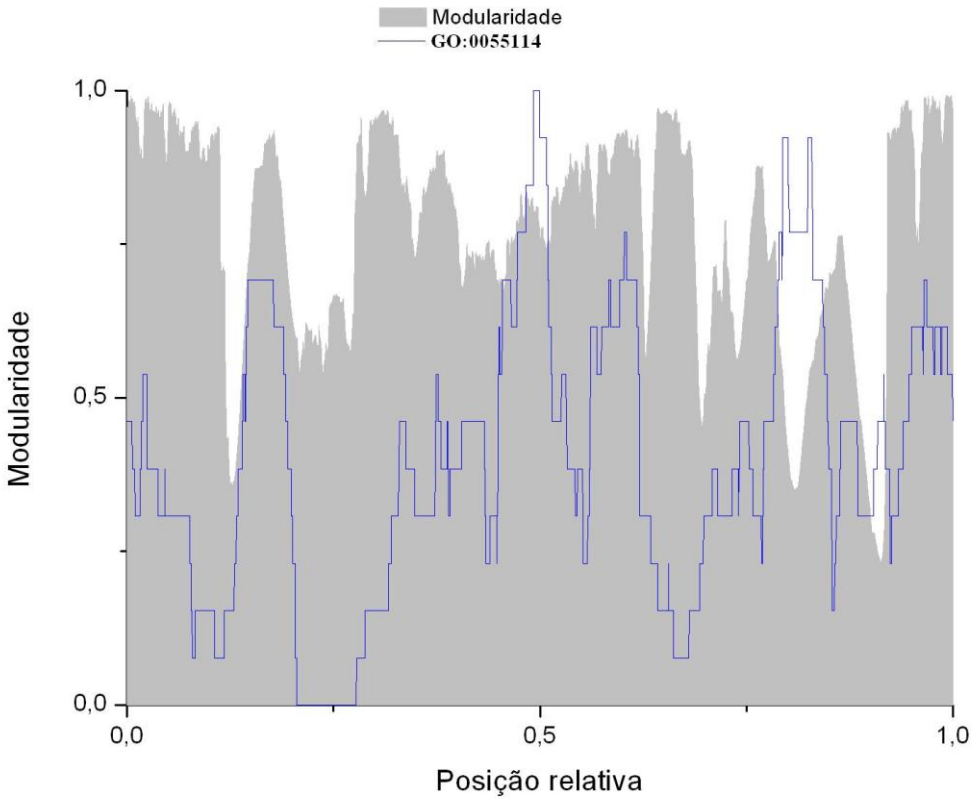


**Figura 38: Função GO:0055114**

Com esses resultados, podemos observar agora as relações existentes entre os módulos funcionais e os processos biológicos em questão.

## 2.4 RESULTADOS

Para o módulo 1, é possível observar nas figuras 13, 18, 25 e 36 algumas funções predominantemente realizadas pelos genes ali agrupados. Porém, também nota-se que a ativação dos genes nessas funções restringe-se a uma parte do módulo, ao invés de englobá-lo por completo. Se observássemos esse resultado a partir de apenas uma função, seria possível supor que aquela função realmente corresponde apenas a uma fração do módulo. Porém, como esse comportamento repete-se, podemos deduzir que provavelmente há uma separação entre módulos funcionais dentro daquela região, ou seja, aquele agrupamento que chamamos de módulo 1 é melhor descrito como sendo composto de 2 ou mais módulos. Seria possível verificar a veracidade desse fato através de uma análise da modularidade com um tamanho diferente de janela, possivelmente menor que o utilizado.

Para o módulo 2, é possível observar nas figuras 11, 24, 26, 29 e acima de tudo na figura 30 funções predominantemente realizadas pelos genes ali presentes.

Para o módulo 3, apenas a figura 37 apresenta um pico de ativação dos genes de tamanho considerável relacionado àquele agrupamento particular. Porém, essa mesma figura possui picos correspondentes a outros módulos, o que indica que o módulo 3 tem uma participação importante na realização daquela função, mas não a rege por completo.

Para o módulo 4, nenhuma das figuras acima apresenta um pico de tamanho considerável relacionado àquele agrupamento de genes. Isso não significa que nenhuma função pode ser atribuída a esse módulo; significa apenas que nenhuma das funções analisadas corresponde bem a ele. É importante relembrar que foram escolhidas para análise poucas dezenas de funções entre centenas existentes e apresentadas pelas bases de dados.

Para o módulo 5, novamente não foi possível identificar nenhuma função fortemente relacionada àqueles genes. Porém, a mesma lógica indicada acima se aplica a esse caso.

Para o módulo 6, podemos observar nas figuras 34 e 35 funções realizadas predominantemente por aquele agrupamento de genes. Além disso, as figuras 14, 17, 19 e 37 apresentam funções em que há a participação ativa do módulo 6 e de outros módulos.

Para o módulo 7, as figuras 7 e 28 apresentam funções predominantemente realizadas por aqueles genes, enquanto a figura 38 apresenta uma função com forte participação daqueles genes, mas com ativação também de outros módulos.

Para o módulo 8, a figura 12 apresenta uma função fortemente relacionada com aqueles genes, enquanto as figuras 14 e 19 apresentam funções regidas por outros módulos mas com forte participação dos genes do módulo 8. É possível perceber que na figura 12 o pico aparece antes do mínimo local de modularidade existente no módulo, enquanto nas figuras 14 e 19 o pico aparece depois do mesmo. Analogamente ao caso do módulo 1, isso significa que é possível que esse módulo seja melhor descrito como dois ou mais módulos separados e que ele possa ser melhor analisado com uma janela de tamanho menor.

Para o módulo 9, as figuras 8, 9, 10, 23 e 32 apresentam picos posicionados acima tanto da região do módulo 9 quanto da região do módulo 10. Isso indica que

talvez essa região possa ser melhor descrita como um módulo único, embora também seja possível apenas que as funções escolhidas são realizadas por mais de um módulo.

Para o módulo $10$, além das já mencionadas funções representadas pelas figuras $8$, $9$, $10$, $23$ e $32$, a figura $15$ apresenta uma função com um pico bem definido naquela região, mas também com a participação de outros módulos.

Para o módulo $11$, assim como para os módulos $4$ e $5$, os resultados não apresentam nenhuma função que pode ser fortemente relacionada com aquele agrupamento de genes.

Para o módulo $12$, as figuras $12$ e $22$ apresentam funções relacionadas a esse módulo, mas onde também existe participação forte de outros módulos.

Finalmente, para o módulo $13$, relembrando que os cálculos são feitos utilizando condições de contorno periódicas, a figura $21$ apresenta uma função relacionada mais fortemente com os genes daquele agrupamento, mas ainda assim bem distribuída ao longo do ordenamento.

# 3 CONCLUSÃO

O objetivo deste trabalho era propor e analisar um método capaz de separar uma lista de genes interligados em agrupamentos chamados de módulos e observar uma correlação entre esses módulos e as funções biológicas realizadas pelo organismo sendo estudado, com o fim de explicar o funcionamento da célula além do caráter local.

Após concluída a análise, observa-se que nem todos os módulos foram bem relacionados com funções biológicas do organismo e, ao mesmo tempo, nem todas as funções foram mapeadas com precisão em um ou mais módulos.

Porém, pode-se ver que foi possível observar correlações bem estabelecidas entre vários módulos e funções biológicas, ou seja, os agrupamentos de genes feitos pelo método de ordenamento são capazes de explicar, pelo menos em alguns casos específicos, o comportamento da rede em um caráter mais global.

É importante ressaltar que o método utilizado é novo e com certeza existem inúmeros aprimoramentos que podem ser feitos no futuro, um dos quais, como já mencionado, é a criação de um algoritmo capaz de automatizar a separação dos módulos funcionais.

Além disso, nesse trabalho o método foi utilizado em sua forma mais simples, considerando para o cálculo da função custo do sistema apenas os primeiros vizinhos de um sítio e utilizando um termo linear da distância, e por limitações humanas foi examinado um número pequeno de funções relativo ao total. Juntamente com algoritmos que automatizem a análise biológica do sistema, a utilização de cálculos mais complexos, que considerem mais do que a vizinhança imediata de um sítio, ou que utilizem termos não-lineares para a distância, com certeza são capazes de aprimorar o método e os resultados que ele apresenta.

De modo geral, observa-se que existe uma relação entre a modularidade da rede ordenada pelo método aqui apresentado e a ativação dos genes durante a realização de diversas funções biológicas pelo organismo. Portanto, é possível concluir que o método de ordenamento por minimização de função custo do sistema é o primeiro passo para o desenvolvimento de um método de ordenamento da rede capaz de produzir resultados úteis para o entendimento da célula e de seus processos como um todo.

## 4 REFERÊNCIAS

1. Rives,A.W. & Galitski,T. Modular organization of cellular networks. Proceedings of the National Academy of Sciences of the United States of America 100, 1128-1133 (2003).

2. Ravasz,E., Somera,A.L., Mongru,D.A., Oltvai,Z.N. & Barabasi,A.L. Hierarchical Organization of Modularity in Metabolic Networks. Science 297, 1551-1555 (2002).

3. Hintze,A. & Adami,C. Evolution of Complex Modular Biological Networks. PLoS Comput Biol 4, e23 (2008).

4. Jensen,L.J. et al. STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res. 37, D412-D416 (2009).

5. Disponível na internet no endereço http://string.embl.de/

6. Huang,d.W., Sherman,B.T. & Lempicki,R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 4, 44-57 (2009).

7. Disponível na internet no endereço http://david.abcc.ncifcrf.gov/

8. Disponível na internet no endereço http://www.geneontology.org/

9. Disponível na internet no endereço http://plants.ensembl.org/biomart/martview/728c7712201527ba5e03e19effe86215

10. Rybarczyk,J.L., Castro,M.A.A., Dalmolin,R.J.S., Moreira,J.C.F., Brunnet,L.G. & de Almeida,R.M.C. Towards a genome-wide transcriptogram: the *Saccharomyces cerevisiae* case. Nucleic Acid Research (2010)