Universidade de São Paulo
Instituto de Matemática e Estatística

Centro de Estatística Aplicada

Relatório de Análise Estatística

RAE-CEA–23P08

RELATÓRIO DE ANÁLISE ESTATÍSTICA SOBRE O PROJETO:

“Análise quimiométrica de seletividade nos sítios ativos de cisteíno proteases da família da papaína para o estudo de substâncias antiparasitárias e antineoplásicas”

Ian Tikkanen Belitsky

Juliana Tiemi Oda Kodono

Rafael Bassi Stern

Tamires Vieira Alves

São Paulo, julho de 2023

CENTRO DE ESTATÍSTICA APLICADA - CEA – USP

TÍTULO: Relatório de Análise Estatística sobre o Projeto: “Análise quimiométrica de seletividade nos sítios ativos de cisteíno proteases da família da papaína para o estudo de substâncias antiparasitárias e antineoplásicas”.

PESQUISADOR: Felipe Marçal Morgantini

ORIENTADOR: Prof. Dr. Andrei Leitão

INSTITUIÇÃO: Instituto de Química de São Carlos - IQSC

FINALIDADE DO PROJETO: Publicação de tese de doutorado

RESPONSÁVEIS PELA ANÁLISE: Ian Tikkanen Belitsky
Juliana Tiemi Oda Kodono
Rafael Bassi Stern
Tamires Vieira Alves

REFERÊNCIA DESTE TRABALHO: BELITSKY, I.T.; KODONO, J.T.O.; STERN, R.B.; ALVES, T.V. **Relatório de análise estatística sobre o projeto: “Análise quimiométrica de seletividade nos sítios ativos de cisteíno proteases da família da papaína para o estudo de substâncias antiparasitárias e antineoplásicas”.** São Paulo, IME-USP, 2023. (RAE–CEA-23P08)

FICHA TÉCNICA

REFERÊNCIAS BIBLIOGRÁFICAS:

ASSIS, D.M.; GONTIJO, V.S.; PEREIRA, I.O.; SANTOS, J.A.N.S.; CAMPS, I.; NAGEM, T.J.; ELLENA, J.; IZIDORO, M.A.; TERSARIOL, I.L.S.; BARROS, N.M.T.; DORIGUETTO, A.C.; SANTOS, M.H.; JULIANO, A.M. (2013) Inhibition of cysteine proteases by a natural biflavone: behavioral evaluation of fukugetin as papain and cruzain inhibitor. **Journal of Enzyme Inhibition and Medicinal Chemistry**, **28**, 661-670.

HARTIGAN, J.A.; WONG, M.A. (1979). Algorithm AS 136: A K-means clustering algorithm**. Applied Statistics**, 28, 100-108.

HOTHORN, T.; HORNIK, K.; ZEILEIS, A. (2012) Unbiased Recursive Partitioning: A Conditional Inference Framework. **Journal of Computational and Graphical Statistics, 15**, 651-674.

OTSUKI, N.; DANG, N.H.; KUMAGAI, E.; KONDO, A.; IWATA, S.; MORIMOTO, C. (2010) Aqueous extract of Carica papaya leaves exhibits anti-tumor activity and immunomodulatory effects. **Journal of Ethnopharmacology, 127**, 760-767.

ROUSSEEUW, P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, 20, 53-65.

PROGRAMAS COMPUTACIONAIS UTILIZADOS:

Microsoft Word *for Windows* (versão 2016);

R *for Windows* versão 4.1.2;

RStudio *for Windows* versão 2021.09.01+372.

TÉCNICAS ESTATÍSTICAS UTILIZADAS

Análise Descritiva Multidimensional (03:010)

Outros (06:990)

Outros (07:990)

ÁREA DE APLICAÇÃO

Outros (14:990)

Resumo

Por serem alvos terapêuticos para a Doença de Chagas e diversos tipos de câncer, as cisteíno proteases da família da papaína são estudadas com interesse no desenvolvimento de novos fármacos. No entanto, a alta similaridade entre os sítios ativos de diferentes classes enzimáticas da família da papaína dificulta o desenvolvimento de moléculas que interajam especificamente nos alvos enzimáticos desejados. Com isso podem ocorrer interações não desejadas com enzimas similares, aumentando os efeitos colaterais no uso de tais fármacos.

Assim, este estudo propõe uma forma de discriminar as classes enzimáticas consideradas na pesquisa (Cruzaína, Catepsina B, Catepsina K, Catepsina L e Catepsina S) a partir dos valores de interação energética entre a enzima e diferentes sondas que simulam partes do ligante.

Sumário

1. Introdução

As cisteíno proteases da família da papaína são enzimas amplamente estudadas por serem alvos terapêuticos validados para a Doença de Chagas (Assis et al., 2012) e diversos tipos de câncer (Otsuki et al., 2010), doenças que são estudadas no grupo de Química Orgânica e Biológica NEQUIMED, pertencente ao IQSC. Isto é, estas enzimas estão associadas a determinados processos destas doenças, de forma que, com o uso de um fármaco, ela pode ser inibida ou ativada, de modo a mudar o curso da doença de forma positiva.

Desta maneira, o estudo destas enzimas é de extremo interesse, já que contribui para o desenvolvimento de moléculas, que no futuro podem vir a ser novos fármacos para as doenças citadas. Mais especificamente, é de interesse o estudo dos sítios ativos de tais enzimas, locais em que há alta concentração de aminoácidos, que interagem com os fármacos.

Entretanto, o sítio ativo das diferentes classes enzimáticas da família da papaína apresenta apenas uma pequena diferença na natureza dos aminoácidos. Este fato dificulta o desenvolvimento de moléculas que interajam especificamente nos alvos enzimáticos desejados. Esta é uma característica relevante que deve ser considerada ao elaborar um novo fármaco. Por possivelmente aumentar os efeitos colaterais causados pela inibição promíscua, que é a interação simultânea entre a pequena molécula candidata à fármaco e as enzimas similares.

Neste contexto, foi proposto um estudo para a diferenciação das classes enzimáticas da família da papaína, através da medição da interação energética entre as enzimas e diversas sondas, para várias distâncias.

2. Objetivos

O objetivo deste estudo é encontrar pontos próximos aos sítios ativos das enzimas que possam discriminar uma enzima em relação às outras através das energias medidas, obtidas da interação entre as enzimas e as sondas. Um objetivo secundário seria a identificação de pontos discriminantes que estejam próximos um ao outro, de forma a obter regiões seletivas, além de pontos seletivos do objetivo principal.

3. Descrição do estudo

Para a obtenção dos dados do estudo, foram utilizados métodos *in-silico*, isto é, foram feitas simulações computacionais para modelar o processo que, naturalmente, seria feito no laboratório. Assim, para realizar as medições energéticas das interações entre a enzima e a sonda, foram extraídas as coordenadas atômicas das enzimas do RCSB Protein Data Bank (www.rcsb.org). Tais coordenadas atômicas foram inseridas no programa GRID, que efetivamente mapeou a interação energética entre cada sítio ativo e cada sonda química virtual, através da inserção destas sondas em uma malha tridimensional composta por 26000 coordenadas. As coordenadas desta malha têm 1 Angstrom de distância entre si.

Na base há 114 amostras de enzimas. Para cada amostra, as medições energéticas foram realizadas com 6 sondas diferentes, nas 26000 coordenadas da malha tridimensional.

4. Descrição das variáveis

As variáveis a serem analisadas neste estudo são referentes às energias medidas nas diversas coordenadas da malha, para cada combinação de amostra de enzima e sonda.

- **Identificação da amostra**
- **Classe enzimática da amostra**: Cruzaína, Catepsina B, Catepsina K, Catepsina L, Catepsina S

- **Sonda química**: Grupo Metila, Oxigênio Carbonílico, Água, Cátion Sódio, Nitrogênio Amídico, Ânion Cloreto
- **Energia (kcal/mol)**: energia resultante da interação entre a amostra enzimática e a sonda química. Valores negativos mostram que há atração entre a amostra e a sonda naquela coordenada, e valores positivos indicam que há repulsão. Já valores nulos indicam que não há interação entre a amostra e a sonda. Além disso, não são computados valores acima de 5 kcal/mol pelo programa GRID.
- **Coordenada**: valores que variam de 1 a 26000, identificando a coordenada da malha tridimensional
- **Eixo X da coordenada**
- **Eixo Y da coordenada**
- **Eixo Z da coordenada**

Como dito anteriormente, os valores energéticos foram medidos de maneira computacional, e, além disso, o programa utilizado não computa valores maiores que 5 kcal/mol. Por isso, para a análise descritiva feita, foi feita uma transformação dos valores 5 kcal/mol. em 0 kcal/mol, pois as energias resultantes em 5 kcal/mol, indicam alta repulsão, e podem ser pontos em que não há interesse no estudo por indicarem apenas que há muita proximidade entre a sonda e o sítio ativo da enzima. Logo, uma solução seria transformar as energias que resultaram em 5 kcal/mol em 0 kcal/mol, pois valores energéticos nulos também indicam pontos que não são de interesse de estudo, já que indicam que não há interação entre a sonda e o sítio ativo da enzima.

5. Análise descritiva

Nesta seção são feitas, para cada sonda utilizada, análises acerca da distribuição dos valores energéticos de sua interação molecular com os aminoácidos do sítio ativo de determinada classe enzimática. Ademais estudamos as coordenadas da malha em busca de pontos seletivos, ou seja, pontos nos quais há maior variabilidade entre os valores de interação de certo ligante com as diferentes enzimas. Encontrados tais pontos discriminantes, avaliamos se estes estão concentrados espacialmente, formando o que é denominado região seletiva.

5.1 Frequência das energias

Observada a base de dados notou-se grande presença de valores energéticos iguais a zero ou cinco. Informados pelos pesquisadores isto ocorre devido à distância entre a sonda e os aminoácidos do sítio ativo. O valor zero indica que a distância é muito grande para que haja interação molecular, já o valor cinco é o máximo positivo retornado pelo software e representa o domínio de forças repulsivas devido à proximidade da sonda em relação à enzima. Assim, nenhum destes valores são do interesse da pesquisa.

A fim de identificarmos com que frequência ocorrem tais cenários e o quanto de repulsão e atração temos na base de dados, foram feitos gráficos para cada par ligante/classe enzimática com os percentuais de valores negativos (forças atrativas), valores positivos diferentes de cinco (forças repulsivas), zeros e cincos.

Para as sondas C3 (Grupo Metila), N2 (Nitrogênio Amídico) e OH2 (Água) notamos poucos valores positivos diferentes de cinco (cerca de 1% ou 2%) e presença parecida de zeros, cincos e negativos, como retratado nas Figuras B.1, B.2 e B.3. Entre as classes enzimáticas não há muita diferença entre os percentuais observados.

As Figuras B.4 e B.5, referentes às sondas CL (Ânion Cloreto) e O (Oxigênio Carbonílico), por sua vez, apresentam interações de repulsão com maior frequência, tendo menos zeros e negativos, com exceção da classe enzimática Catepsina S, que apresenta um comportamento diferente em que se observa o predomínio de forças atrativas (57% e 58%, respectivamente).

Por fim, a sonda NA (Cátion Sódio), indicada na Figura B.6, em geral apresenta mais interações de atração, com o diferencial também em relação à Catepsina S, que aqui possui mais forças repulsivas (33%) em comparação às demais classes enzimáticas (entre 1% e 5%).

5.2 Histograma das energias

Além dos percentuais verificados na análise anterior, estudamos melhor a distribuição das energias de interação molecular por meio de histogramas, que estão indicados nas Figuras B.7 a B.12.

Podemos notar as maiores concentrações nos zeros e cincos, como esperado devido aos percentuais observados anteriormente. No entanto, a coluna zero aparece ainda maior do que imaginado por conter valores que, apesar de não serem exatamente zero, estão próximos a isto.

Para as sondas C3 (Grupo Metila), N2 (Nitrogênio Amídico) e OH2 (Água), retratadas nas Figuras B.7, B.8 e B.9, não vemos grandes diferenças entre as distribuições para cada classe enzimática. O mesmo comentário se aplica à sonda O (Oxigênio Carbonílico) e, ainda que os percentuais da análise anterior indicassem que para a Catepsina S teríamos mais valores energéticos positivos diferentes de 5, detalhando melhor estes valores na Figura B.10 nota-se que, apesar de não serem iguais a zero, estes valores positivos seriam muito próximos a isto, ficando representados na mesma barra referente ao valor 0.

Quanto à sonda CL (Ânion Cloreto), por sua vez, podemos verificar na Figura B.11 o mesmo comportamento diferente para a Catepsina S em que esta classe enzimática apresenta pouca presença em interações de repulsão diferentes daquela com valor 5 kcal/mol devido à proximidade. No entanto, aqui podemos notar ainda diferenças para a Catepsina K e a Catepsina L, as quais apresentam com maior frequência valores energéticos positivos diferentes de 5 em comparação às demais classes enzimáticas. Por fim, uma observação com relação à distribuição das energias de interação para o ânion cloreto independentemente da classe enzimática é que se trata da única sonda em que podemos notar com maior frequência valores positivos diferentes de cinco que não são basicamente zero.

Finalizando com a sonda NA (Cátion Sódio), os comentários são o oposto do verificado para o CL (Ânion Cloreto). Na Figura B.12 observa-se que com a Catepsina S há maior presença de interações de repulsão diferentes de 5 kcal/mol (como vimos nos

percentuais, mesmo que no histograma nota-se que são basicamente zero) e com a Catepsina K e Catepsina L temos uma menor presença nesses valores.

5.3 Gráficos de variância 3D

Para a construção destes gráficos utilizamos a base com transformação de 5 em 0. Para cada sonda seguiu-se o seguinte processo: primeiramente foi calculada a média amostral dos valores energéticos para cada classe enzimática em cada ponto da malha para então ser calculada a variância destas médias por ponto. O resultado pode ser visto nas Figuras B.13 a B.18, nos quais podemos ver os pontos com maior variância, ou seja, maior seletividade entre as classes enzimáticas (pontos seletivos).

Para as sondas C3 (Grupo Metila) e CL (Ânion Cloreto) os pontos seletivos estão mais dispersos pela malha, diferentemente do que podemos ver para a sonda NA (Ânion Sódio) em que claramente formam-se três regiões seletivas próximas às coordenadas (2, 36, 19), (14, 21, 19) e (13, 25, 37).

Em relação às sondas N2 (Nitrogênio Amídico), O (Oxigênio Carbonílico) e OH2 (Água) notamos que os pontos seletivos formam pequenos aglomerados não tão claros quanto para o Ânion Sódio, porém há uma mesma concentração para estes três ligantes próximo às coordenadas (2, 26, 40).

Os gráficos interativos podem ser acessados ao baixar os arquivos no link: https://drive.google.com/drive/u/1/folders/11yyt986ZbwI72-69aEPGbUo4axLymfV1.

Na Tabela A.1 seguem todas as coordenadas destacadas com maior variância para cada sonda.

6. Análise estatística

Nesta seção, são obtidos os clusters por sonda formados pelas coordenadas de maiores variâncias energéticas, considerando-se como variáveis para a formação de tais clusters os eixos X, Y e Z, assim como a energia média de cada classe enzimática nestas coordenadas. Também são geradas regras utilizando as energias médias das

amostras em cada um destes clusters, a fim de classificá-las nas diferentes classes enzimáticas.

6.1 Análise de clusters

Com o objetivo final de realizar uma análise de árvore de inferência condicional, é de interesse diminuir a quantidade de variáveis explicativas do modelo. Para isso foi feito uma análise de clusters aos pontos de alta variância.

Estes pontos foram obtidos, para cada sonda, a partir da variância dos valores energéticos de cada unidade amostral em determinado ponto. Foram selecionados 0,05% dos pontos (de maior variabilidade) em cada sonda para prosseguir a análise.

Em seguida, os pontos selecionados foram clusterizados considerando a média energética nestes pontos para cada classe enzimática e as coordenadas espaciais dadas pelos eixos X, Y e Z da malha.

O número de clusters a ser formado para cada classe é decidido a partir da comparação computacional entre as diferentes opções pelo coeficiente de silhueta (Rousseeuw, 1987). O número total de clusters é de 24; 7 da sonda Grupo Metila; 3 da sonda Oxigênio Carbonílico; 4 da sonda Água; 3 da sonda Cátion Sódio; 4 da sonda Nitrogênio e 3 da sonda Ânion Cloreto.

A clusterização é realizada pelo algoritmo kmeans (Hartigan e Wong, 1979). Como a clusterização tem como variáveis as médias energéticas por classe enzimática, além das coordenadas geométricas, o resultado da clusterização pode ser diferente de um agrupamento intuitivo somente pela visão geométrica.

As representações gráficas desses clusters podem ser vistas nas Figuras B.19 até B.24.

Nas Tabelas A.2 até A.7 encontram-se as coordenadas equivalentes a cada cluster por sonda.

Note que pode haver diferença entre os pontos aqui selecionados e aqueles apresentados na Seção 5.3, uma vez que lá foi calculada a média energética entre as

unidades amostrais de cada classe enzimática para, a partir destes valores médios, ser obtida a variância (assim como nas demais análises descritivas, o intuito daquela seção foi analisar e comparar as classes enzimáticas em cada sonda, fazendo mais sentido calcular a variância considerando um valor energético médio por classe enzimática e não unitário, já que temos tamanhos amostrais distintos para cada classe enzimática). Nesta seção, por sua vez, a variância foi calculada a partir dos valores energéticos das unidades amostrais, e a média por classe enzimática foi utilizada apenas na clusterização.

6.2 Árvores de inferência condicional

Dados os clusters obtidos na Seção 6.1, foram calculadas as médias energéticas de cada amostra em cada um destes 24 clusters. Desta forma, novas variáveis foram obtidas:

- **Energia média no cluster da sonda química**: são 24 energias médias obtidas dos 7 clusters formados na sonda Grupo Metila; 3 da sonda Oxigênio Carbonílico; 4 da sonda Água; 3 da sonda Cátion Sódio; 4 da sonda Nitrogênio Amídico; e 3 da sonda Ânion Cloreto.

Assim, estas médias energéticas foram usadas como covariáveis para a árvore de inferência condicional (Hothorn et al., 2012). O objetivo desta árvore é encontrar as melhores quebras dos valores das covariáveis construídas, de forma a conseguirmos discriminar as amostras nas diferentes classes enzimáticas. A árvore foi construída usando 69% das amostras da base como dados de treino, sendo que as outras 31% foram utilizadas para testar a qualidade da árvore ajustada. A distribuição destas bases quanto à classe enzimática das amostras pode ser vista na Tabela A.8.

Primeiramente, foi ajustada uma árvore utilizando todas as 24 variáveis de energia média dos clusters das sondas químicas. O resultado pode ser visto na Figura B.25. Observe que nesta árvore, as quebras são determinadas por um cluster da sonda NA e da sonda OH2, e dois clusters da sonda N2.

Enquanto esta árvore não produz nenhum erro de classificação com relação à base de treino, ao utilizarmos as regras definidas com a base de treino na base de

teste, a acurácia é de 97,14%, e há apenas um erro de classificação, como é possível observar na Tabela A.9. A amostra que foi classificada erroneamente é a que tem identificação 2aim, que pertence à classe enzimática Cruzaína e foi classificada como Catepsina S.

Posteriormente, para o conhecimento da capacidade de classificação de cada sonda com relação às classes enzimáticas, foi realizada uma árvore de inferência condicional para cada sonda, considerando os clusters de referência.

Para a sonda Grupo Metila, foi ajustada uma árvore utilizando as variáveis de energia média dos 7 clusters. O resultado pode ser visto na Figura B.26. Observe que os clusters 1 e 7 não foram utilizados nesta árvore. Além disso, mesmo com a base de treino há 6 erros de classificação desta árvore, sendo duas amostras da Catepsina B, duas da Catepsina K e duas da Cruzaína.

Já ao utilizarmos as regras definidas com a base de treino na base de teste, esta árvore resulta em 9 erros de classificação e acurácia de 74,29%, como é possível observarmos na Tabela A.10.

Na árvore obtida com relação à sonda Oxigênio Carbonílico, foi ajustada uma árvore com variáveis de energia média dos 3 clusters. O resultado pode ser visto na Figura B.27. Observe que mesmo com a base de treino há 2 erros de classificação da Catepsina B.

Ao utilizarmos as regras definidas com a base de treino na base de teste, as regras de classificação desta árvore resultam em 5 erros de classificação e acurácia de 85,71%, como é possível observarmos na Tabela A.11. Observe que todos os erros de classificação foram classificados como Catepsina L.

Com relação à sonda Água, foi ajustada uma árvore utilizando as variáveis de energia média dos 4 clusters. O resultado pode ser visto na Figura B.28. Observe que com a base de treino há apenas um erro de classificação, de uma amostra da Catepsina L, classificada como Catepsina S.

Desta árvore, ao utilizarmos as regras definidas com a base de treino na base de teste, se obtém 3 erros de classificação e acurácia de 91,43%, como é possível observar na Tabela A.12.

Para a sonda Cátion Sódio, foi ajustada uma árvore utilizando as variáveis de energia média dos 3 clusters. O resultado pode ser visto na Figura B.29. Observe que com a base de treino há 6 erros de classificação, sendo duas amostras da Catepsina B, 3 da Catepsina L e uma da Catepsina S. Além disso, veja que apenas 4 grupos finais foram gerados, sendo que há 5 classes enzimáticas, e o cluster 2 não foi utilizado.

Com esta árvore, ao utilizarmos as regras definidas com a base de treino na base de teste, há 7 erros de classificação e acurácia de 80%, como vemos na Tabela A.13.

Para a sonda Nitrogênio Amídico, foi ajustada uma árvore utilizando as variáveis de energia média dos 4 clusters. O resultado pode ser visto na Figura B.30. Observe que com a base de treino há 3 erros de classificação, pertencentes à Catepsina B, Catepsina K e Catepsina S.

Esta árvore, ao utilizarmos as regras definidas com a base de treino na base de teste, resulta em 3 erros de classificação e acurácia de 91,43%, como podemos ver na Tabela A.14.

E por último, para a sonda Ânion Cloreto, foi ajustada uma árvore utilizando as variáveis de energia média dos 3 clusters. O resultado pode ser visto na Figura B.31, e observe que o cluster 3 não foi utilizado. Além disso, veja que com a base de treino há 4 erros de classificação, sendo duas amostras da Catepsina L e duas da Catepsina S.

Esta árvore, ao utilizarmos as regras definidas com a base de treino na base de teste, resulta em apenas um erro de classificação e acurácia de 97,14%, como podemos ver na Tabela A.15. A amostra que foi classificada erroneamente é a que tem como identificação 5mae. Ela pertence à classe enzimática Catepsina L, e foi classificada como Catepsina S.

As acurácias citadas acima estão dispostas na Tabela A.16, assim como as acurácias por classe enzimática em cada uma das árvores de decisão construídas.

Pode-se notar que, entre as árvores que consideram apenas uma sonda, a maior acurácia geral foi obtida na árvore da sonda Ânion Cloreto, cuja acurácia (97,14%) foi a mesma da árvore construída com todas as sondas. Já a menor acurácia geral foi observada na árvore da sonda Metila (74,29%), que, apesar de classificar corretamente em teste 100,00% das amostras pertencentes às classes Catepsina K, Catepsina L e Catepsina S, obteve acurácia baixa para a classe Cruzaína (33,33%).

Devido ao tamanho amostral pequeno, os valores de acurácia para algumas classes enzimáticas, em especial para a Catepsina B, devem ser interpretados com cautela. Neste caso, há apenas duas unidades amostrais utilizadas em treino e uma reservada para teste, de modo que reduzimos a acurácia a um resultado binário (100,00% em caso de acerto e 0,00% caso classifique erroneamente).

As acurácias mais baixas foram observadas nos seguintes pares: classe enzimática Cruzaína na árvore da sonda Metila (33,33%), classe Catepsina S na árvore da sonda Cátion Sódio (40,00%), e classe Catepsina K na árvore da sonda Água (50,00%).

7. Conclusões

A seleção de pontos com maior variabilidade energética considerando certa sonda destaca locais em que esta interage de forma distinta entre as classes enzimáticas. Além disso, seguindo para as árvores de inferência, em geral foram obtidas acurácias altas, indicando que as árvores conseguiram classificar as unidades amostrais nas respectivas classes enzimáticas de forma satisfatória.

APÊNDICE A

Tabelas

Tabela A.1 Identificação das coordenadas com maiores variações de energias medidas por sonda.

Sonda	Coordenada
C3	736, 1064, 1656, 2755, 3104, 3144, 3327, 6510, 7136, 7173, 7258, 8208, 8616, 10010, 10340, 11011, 11049, 15311, 17502, 19384
O	148, 449, 488, 633, 1487, 6859, 7253, 7785, 7859, 8657, 8664, 17296, 18336, 19255, 20215, 20255, 21175, 21215, 21256, 21343, 22216, 22217, 22256, 22257, 22343, 23177, 23217, 23218, 23258
OH2	27, 448, 449, 487, 488, 656, 1488, 1787, 1827, 1828, 1865, 3475, 3487, 7574, 20215, 20255, 21215, 21256, 22216, 22256, 23216, 23217, 23218, 23258, 23298, 23299, 24217, 24258
NA	28, 58, 68, 69, 108, 109, 148, 188, 616, 656, 657, 734, 1058, 1062, 1068, 1148, 1656, 1696, 1734, 2781, 3782, 17181, 18179, 18180, 18187, 18227, 19187, 19227, 19262, 19303, 20303, 20304, 21186, 22225
N2	27, 487, 488, 617, 656, 696, 736, 788, 1488, 1656, 1827, 1828, 2475, 2656, 2787, 2788, 3475, 3487, 3772, 4434, 4474, 5112, 6614, 6615, 6733, 7573, 7574, 7613, 7614, 7615, 7654, 7655, 7694, 8558, 8657, 10010, 11009, 12009, 12011, 16138, 16139, 16179, 18136, 20215, 20255, 21215, 21256, 22216, 22256, 22296, 22336, 23177, 23217, 23218, 23296, 23297, 23298, 24217, 24218, 24257, 24258, 25259

CL	64, 65, 108, 148, 737, 1064, 1084, 2104, 3144, 3327, 5005, 6005, 7173, 7174, 7218, 7258, 8657, 9039, 16031, 16032

Tabela A.2 Identificação das coordenadas equivalentes aos clusters da sonda Grupo Metila.

Cluster	Coordenada
1	6276, 6315, 7275, 8198, 8276, 9158, 10039, 10118, 11158
2	3327, 4045, 6005, 6095, 6258, 7133, 7173, 7213, 7218, 7258, 7370, 8218, 9176, 9258, 10010, 11010, 11011
3	15066, 15351, 16391, 19384, 19468, 22193, 22232, 22233, 23020
4	736, 775, 1448, 4814, 7614, 8696, 9496
5	66, 145, 1066, 1145, 2106, 3185, 4144, 4185
6	4392, 4393, 4432, 4433, 6431, 6510, 6589, 6590, 7508, 8429, 9392, 9430, 10430
7	8340, 8501, 10340, 10380, 10387, 11347, 11386, 11388, 12347, 14388, 14502, 14504, 15502

Tabela A.3 Identificação das coordenadas equivalentes aos clusters da sonda Oxigênio Carbonílico.

Cluster	Coordenada
1	9820, 17420, 17421, 18296, 18336, 18420, 18421, 18422, 19013, 19255, 19383, 19422, 20215, 20255, 20383, 21215, 21343, 22216, 22217, 22257, 22343, 24139, 24178, 24219

2	447, 448, 487, 488, 489, 1365, 1448, 1487, 1488, 1527, 2448, 2488, 3287, 3368, 3408, 3448, 4287, 4328, 4368, 4407, 4408, 4447, 4448, 4488, 5368, 5369, 5408, 5409, 5448, 5449, 6368, 6369, 6408, 6409, 6410, 6448, 7370, 7410, 7656, 7859, 8697, 8736
3	40, 68, 80, 159, 237, 238, 277, 633, 671, 672, 673, 711, 751, 790, 829, 830, 869, 870, 1080, 1237, 1238, 1277, 1711, 1787, 1788, 1789, 1828, 1829, 2237, 2277, 3392, 4861, 4862, 9118

Tabela A.4 Identificação das coordenadas equivalentes aos clusters da sonda Água.

Cluster	Coordenada
1	18012, 18296, 19013, 19053, 21343, 22216, 22256, 22305, 23217, 23257, 23298, 23299, 24218, 24259
2	237, 238, 277, 278, 1237, 1238, 1277, 1278, 2237, 2277, 2278, 3237, 3277, 3392, 10117, 11118
3	447, 448, 487, 488, 489, 527, 528, 567, 1448, 1487, 1527, 2448, 2488, 2527, 3408, 3448, 4368, 4407, 4408, 4448, 4488, 5368, 5407, 5408, 5409, 5448, 5488, 5489, 6368, 6369, 6408, 6409, 6448, 6449, 6531
4	630, 671, 711, 751, 790, 865, 1711, 1751, 1787, 1865, 2787, 2821, 2822

Tabela A.5 Identificação das coordenadas equivalentes aos clusters da sonda Cátion Sódio.

Cluster	Coordenada
1	616, 617, 618, 619, 656, 657, 1617, 1656, 1657, 1696, 1735, 2617, 2780, 2781, 3618, 3781, 3782, 4732, 4733, 6615
2	5038, 8053, 15099, 16140, 16262, 16306, 17099, 17139, 17179, 17181, 17303, 18099, 18138, 18139, 18140, 18141, 18179, 18180, 18186, 18187, 18217, 18218, 18219, 18220, 18221, 18227, 18303, 19179,

	19180, 19181, 19186, 19187, 19221, 19222, 19262, 19302, 19303, 19343, 20186, 20227, 20261, 20262, 20263, 20266, 20267, 20302, 20303, 20304, 20306, 21028, 21186, 21226, 21263, 21303, 22185, 22186, 22224, 22225, 22226, 22264, 23068, 24025, 25104
3	575, 576, 629, 734, 1414, 1453, 1491, 1531, 1616, 1655, 1695, 1734, 2413, 2532, 3170, 3531, 3578, 3733, 4693, 4781, 5452, 5575, 5615, 5654, 5693, 6574, 6614, 6654, 11012, 12012, 13011

Tabela A.6 Identificação das coordenadas equivalentes aos clusters da sonda Nitrogênio.

Cluster	Coordenada
1	198, 237, 238, 277, 278, 1238, 1277, 1278, 2277, 2278, 3237, 3277, 3278, 3392, 11118
2	448, 487, 488, 489, 1448, 1487, 1488, 2448, 2488, 3448, 4287, 4368, 4407, 4408, 4448, 4488, 5368, 5407, 5408, 5448, 5488, 5529, 6137, 6368, 6408, 6409, 6410, 6449, 6489, 6530, 6531, 6614, 7410, 7450, 7533, 7613, 8091, 12009, 12011
3	17180, 18012, 19013, 19053, 20342, 20345, 22216, 22266, 22296, 22305, 23177, 23217, 23265, 23298, 23339
4	617, 630, 656, 671, 696, 736, 1711, 1751, 2820, 2821, 3821

Tabela A.7 Identificação das coordenadas equivalentes aos clusters da sonda Ânion Cloreto.

Cluster	Coordenada
1	10387, 11105, 12386, 13031, 14026, 14066, 15032, 15351, 16031, 16032, 16391, 17030, 17037, 18274, 19384, 20383, 21270, 21344,

	21345, 21383, 22232, 23020
2	3, 44, 488, 489, 737, 1093, 1448, 1488, 2093, 3327, 4328, 4368, 5005, 5733, 6005, 6045, 6134, 6178, 6258, 7133, 7134, 7173, 7213, 7218, 7253, 7254, 7258, 7381, 7421, 7422, 8091, 8132, 8218, 8576, 8624, 8657, 8664, 8696, 9091, 9176, 9258, 9618, 10340, 11010, 11299, 12010
3	25, 66, 108, 145, 148, 198, 1040, 1063, 1066, 1104, 1145, 1238, 2064, 2066, 2106, 3392, 3393, 4112, 4237, 4392, 4393, 4394, 4433, 6238, 6276, 6431, 8198, 8276, 9039, 9118, 9158, 10039, 10118

Tabela A.8 Quantidade de amostras de cada classe enzimática nas bases de treino e de teste para a realização da árvore e inferência condicional.

Classe enzimática	Quantidade de amostras na base de treino	Quantidade de amostras na base de teste
Catepsina B	2	1
Catepsina K	32	10
Catepsina L	16	7
Catepsina S	18	5
Cruzaína	11	12

Tabela A.9 Tabela de confusão utilizando as amostras destinada ao teste da árvore de inferência condicional ajustada com todas as 24 variáveis de média energética.

	Classe predita

		Catepsina B	Catepsina K	Catepsina L	Catepsina S	Cruzaína
Classe verdadeira	Catepsina B	1	0	0	0	0
	Catepsina K	0	10	0	0	0
	Catepsina L	0	0	7	0	0
	Catepsina S	0	0	0	5	0
	Cruzaína	0	0	0	1	11

Tabela A.10 Tabela de confusão utilizando as amostras destinada ao teste da árvore de inferência condicional ajustada com as 7 variáveis de média energética relacionadas a sonda Grupo Metila.

		Classe predita				
		Catepsina B	Catepsina K	Catepsina L	Catepsina S	Cruzaína
Classe verdadeira	Catepsina B	0	0	0	1	0
	Catepsina K	0	10	0	0	0
	Catepsina L	0	0	7	0	0
	Catepsina S	0	0	0	5	0
	Cruzaína	0	0	5	3	4

Tabela A.11 Tabela de confusão utilizando as amostras destinada ao teste da árvore de inferência condicional ajustada com as 3 variáveis de média energética relacionadas a sonda Oxigênio Carbonílico.

		Classe predita				
		Catepsina B	Catepsina K	Catepsina L	Catepsina S	Cruzaína
Classe verdadeira	Catepsina B	0	0	1	0	0
	Catepsina K	0	9	1	0	0
	Catepsina L	0	0	7	0	0
	Catepsina S	0	0	0	5	0
	Cruzaína	0	0	3	0	9

Tabela A.12 Tabela de confusão utilizando as amostras destinada ao teste da árvore de inferência condicional ajustada com as 4 variáveis de média energética relacionadas a sonda Água.

		Classe predita				
		Catepsina B	Catepsina K	Catepsina L	Catepsina S	Cruzaína
Classe verdadeira	Catepsina B	1	0	0	0	0
	Catepsina K	2	8	1	0	2
	Catepsina L	0	0	7	0	0
	Catepsina S	0	0	0	5	0
	Cruzaína	0	0	0	1	11

Tabela A.13 Tabela de confusão utilizando as amostras destinada ao teste da árvore de inferência condicional ajustada com as 3 variáveis de média energética relacionadas a sonda Cátion Sódio.

		Classe predita				
		Catepsina B	Catepsina K	Catepsina L	Catepsina S	Cruzaína
Classe verdadeira	Catepsina B	0	0	0	0	1
	Catepsina K	0	10	0	0	0
	Catepsina L	0	0	6	1	0
	Catepsina S	0	0	3	2	0
	Cruzaína	0	0	0	2	10

Tabela A.14 Tabela de confusão utilizando as amostras destinada ao teste da árvore de inferência condicional ajustada com as 4 variáveis de média energética relacionadas a sonda Nitrogênio Amídico.

		Classe predita				
		Catepsina B	Catepsina K	Catepsina L	Catepsina S	Cruzaína
Classe verdadeira	Catepsina B	1	0	0	0	0
	Catepsina K	0	7	1	0	2
	Catepsina L	0	0	7	0	0
	Catepsina S	0	0	0	5	0
	Cruzaína	0	0	0	0	12

Tabela A.15 Tabela de confusão utilizando as amostras destinada ao teste da árvore de inferência condicional ajustada com as 3 variáveis de média energética relacionadas a sonda Ânion Cloreto.

		Classe predita				
		Catepsina B	Catepsina K	Catepsina L	Catepsina S	Cruzaína
Classe verdadeira	Catepsina B	1	0	0	0	0
	Catepsina K	0	10	0	0	0
	Catepsina L	0	0	6	1	0
	Catepsina S	0	0	0	5	0
	Cruzaína	0	0	0	0	12

Tabela A.16 Acurácias observadas em teste para as árvores de decisão nas visões geral e por classe enzimática.

	Sondas consideradas na árvore de decisão						
	Todas	C3	O	OH2	Na	N2	Cl
Geral	97,14%	74,29%	85,71%	82,86%	80,00%	91,43%	97,14%
Catepsina B	100,00%	0,00%	0,00%	100,00%	0,00%	100,00%	100,00%
Catepsina K	100,00%	100,00%	90,00%	50,00%	100,00%	70,00%	100,00%
Catepsina L	100,00%	100,00%	100,00%	100,00%	85,71%	100,00%	85,71%
Catepsina S	100,00%	100,00%	100,00%	100,00%	40,00%	100,00%	100,00%
Cruzaína	91,67%	33,33%	75,00%	91,67%	83,33%	100,00%	100,00%

APÊNDICE B

Figuras

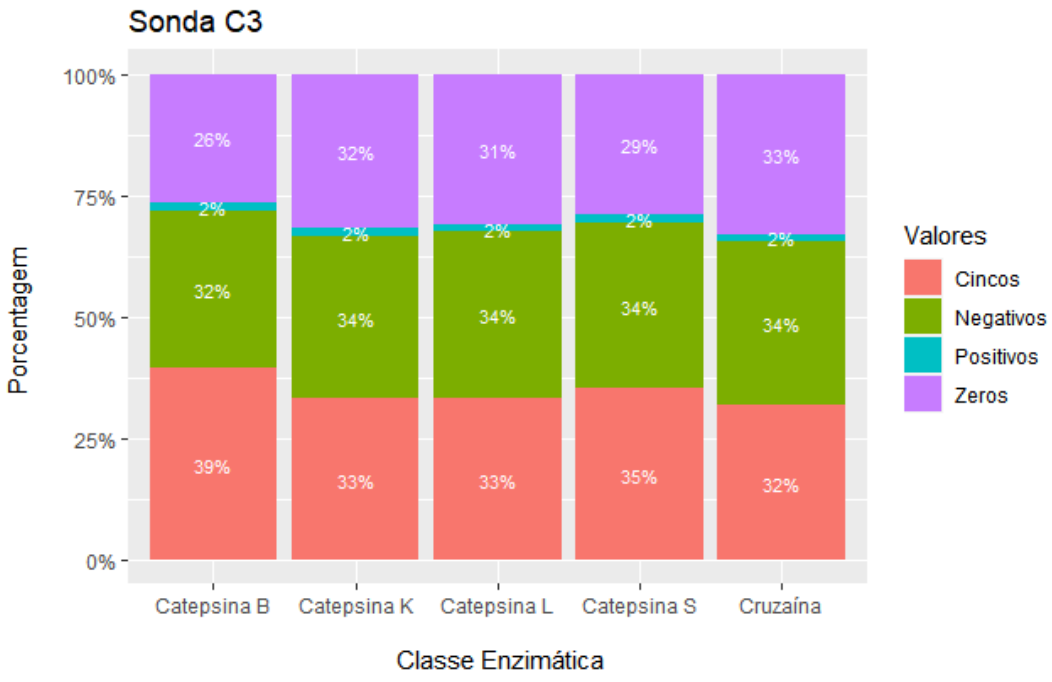

Figura B.1:

Gráfico de frequência de energias medidas com a sonda C3 para as 5 classes enzimáticas, com valores positivos, iguais a 5, negativos e nulos.

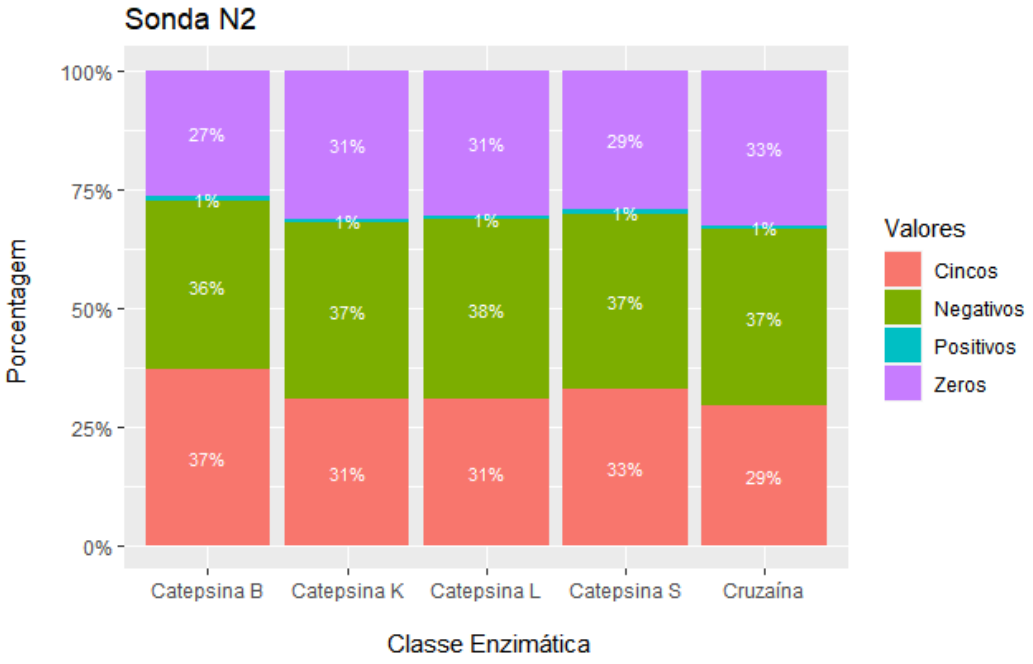

Figura B.2:

Gráfico de frequência de energias medidas com a sonda N2 para as 5 classes enzimáticas com valores positivos, iguais a 5, negativos e nulos.

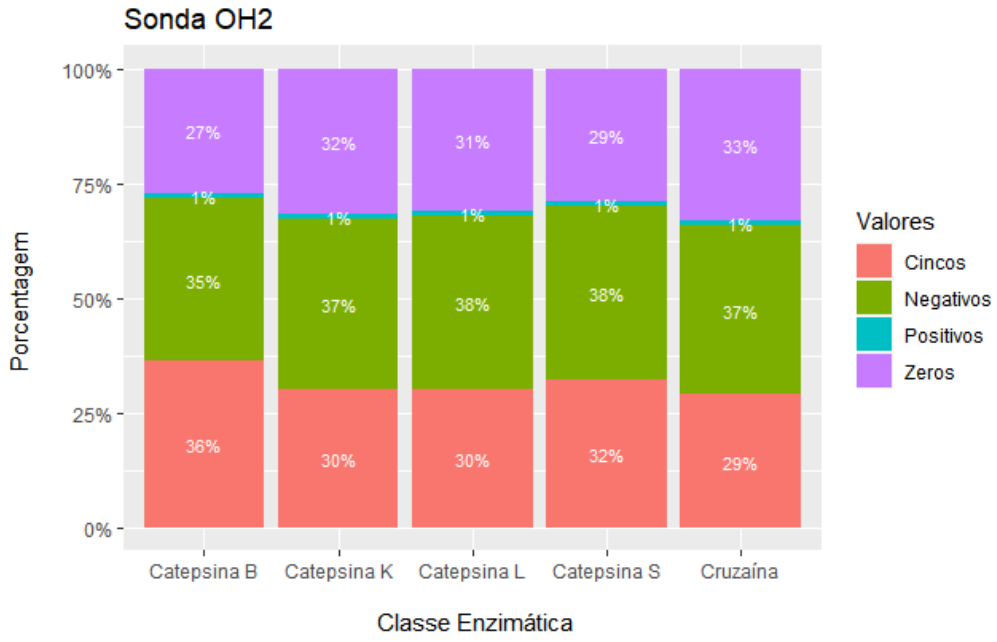

Figura B.3: Gráfico de frequência de energias medidas com a sonda OH2 para as 5 classes enzimáticas com valores positivos, iguais a 5, negativos e nulos.

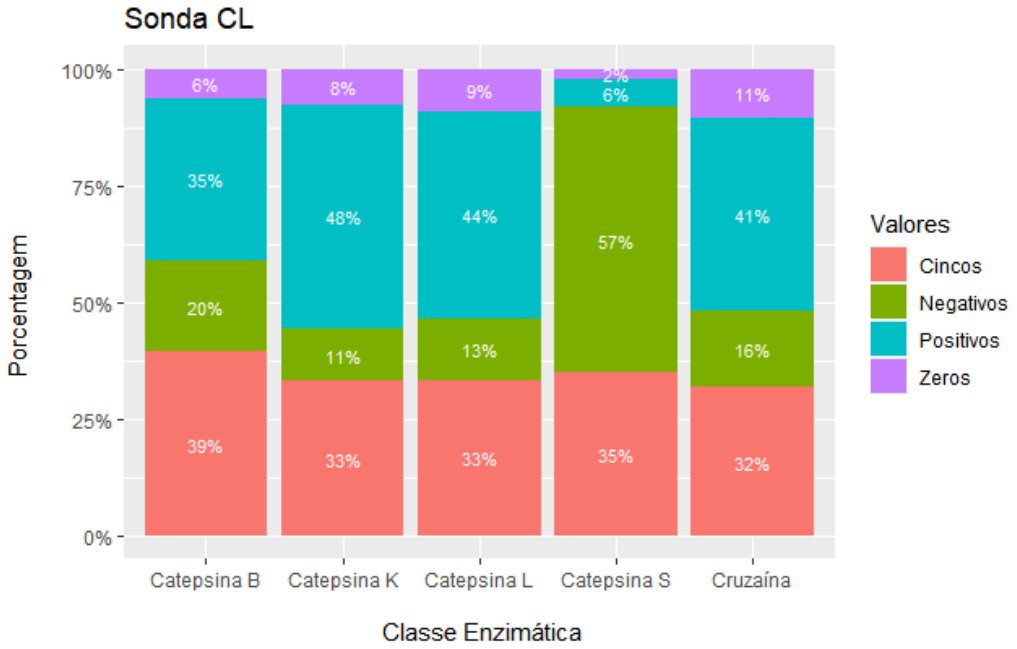

Figura B.4: Gráfico de frequência de energias medidas com a sonda CL para as 5 classes enzimáticas com valores positivos, iguais a 5, negativos e nulos.

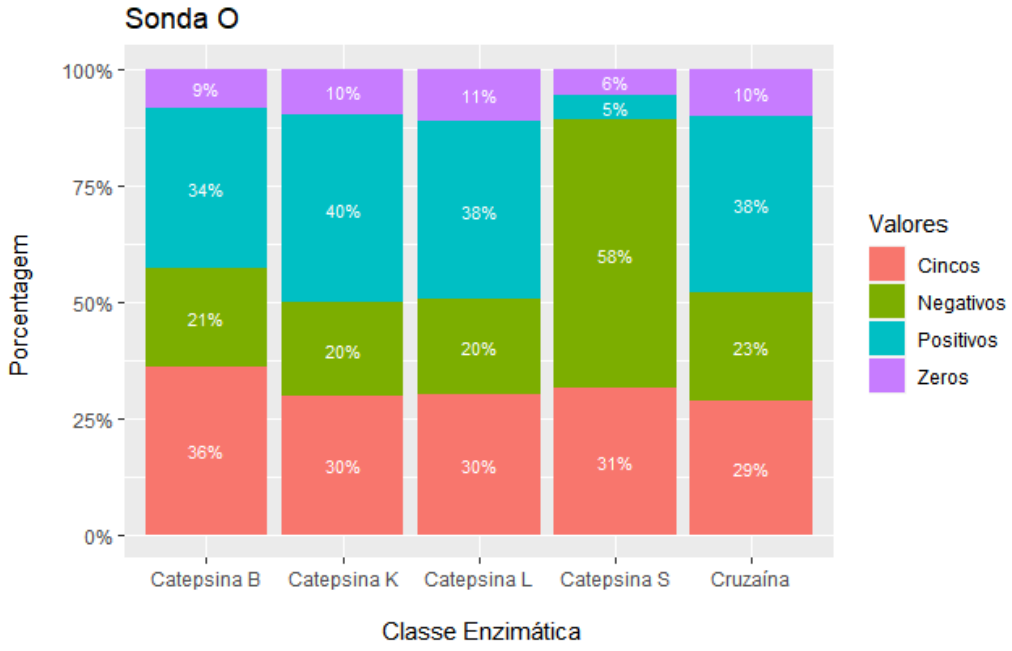

Figura B.5: Gráfico de frequência de energias medidas com a sonda O para as 5 classes enzimáticas com valores positivos, iguais a 5, negativos e nulos.

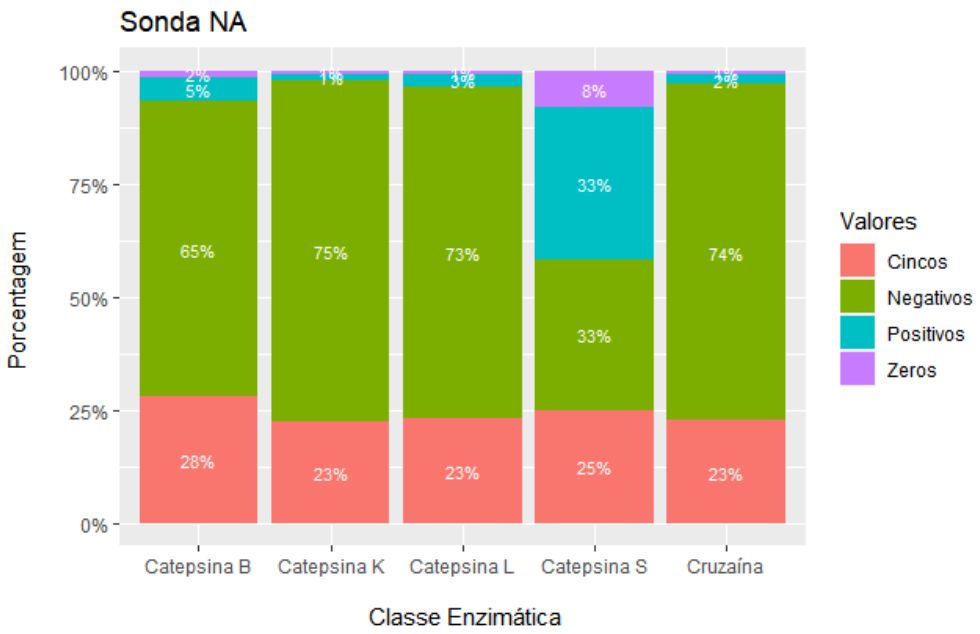

Figura B.6: Gráfico de frequência de energias medidas com a sonda NA para as 5 classes enzimáticas com valores positivos, iguais a 5, negativos e nulos.

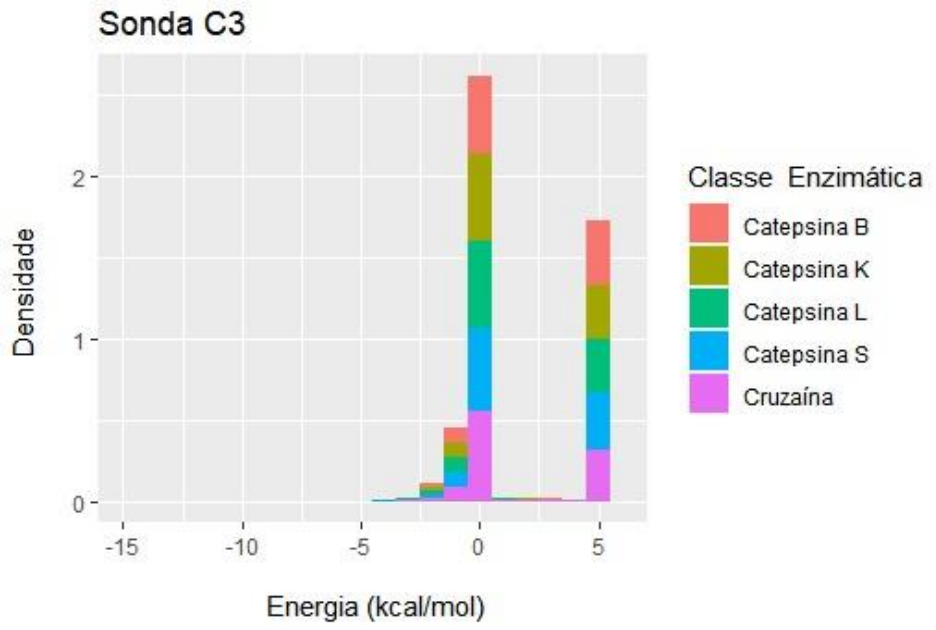

Figura B.7: Histograma das energias medidas com a sonda C3 para as 5 classes enzimáticas.

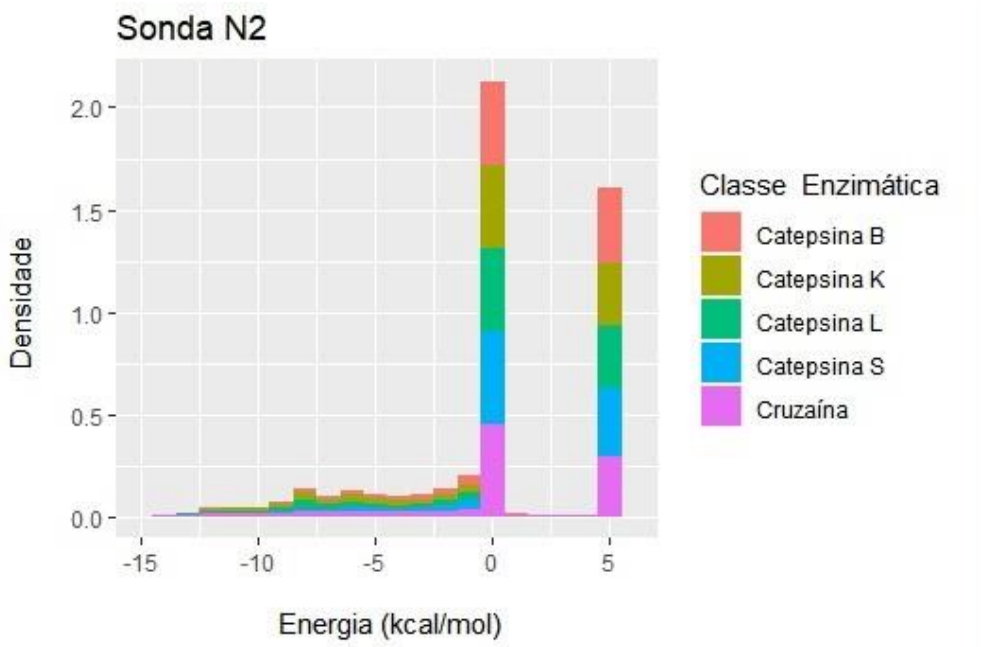

Figura B.8: Histograma das energias medidas com a sonda N2 para as 5 classes enzimáticas.

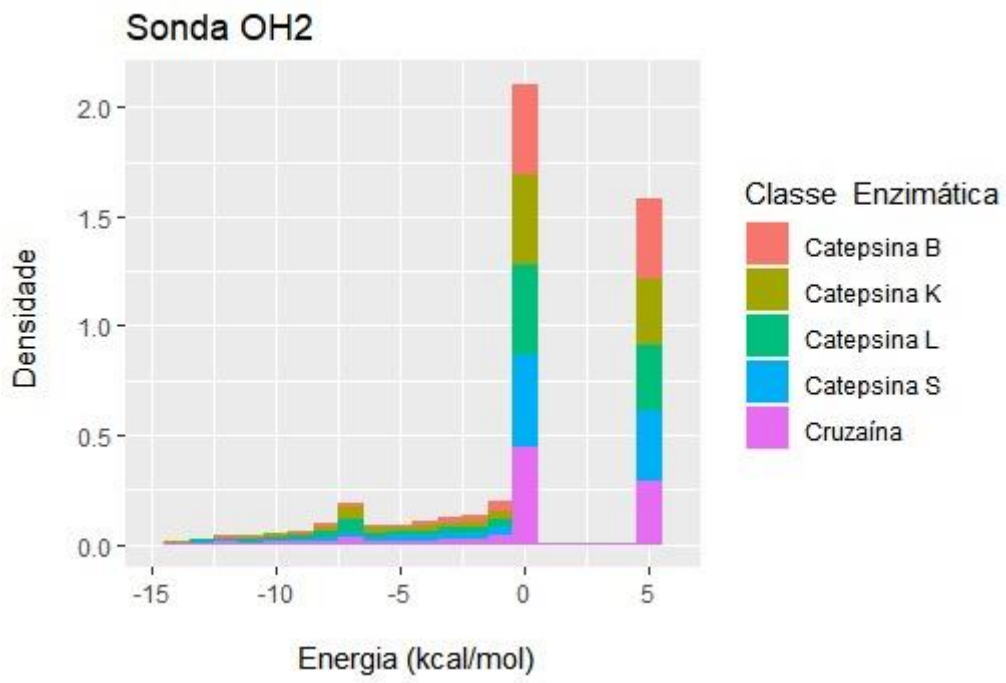

Figura B.9: Histograma das energias medidas com a sonda OH2 para as 5 classes enzimáticas.

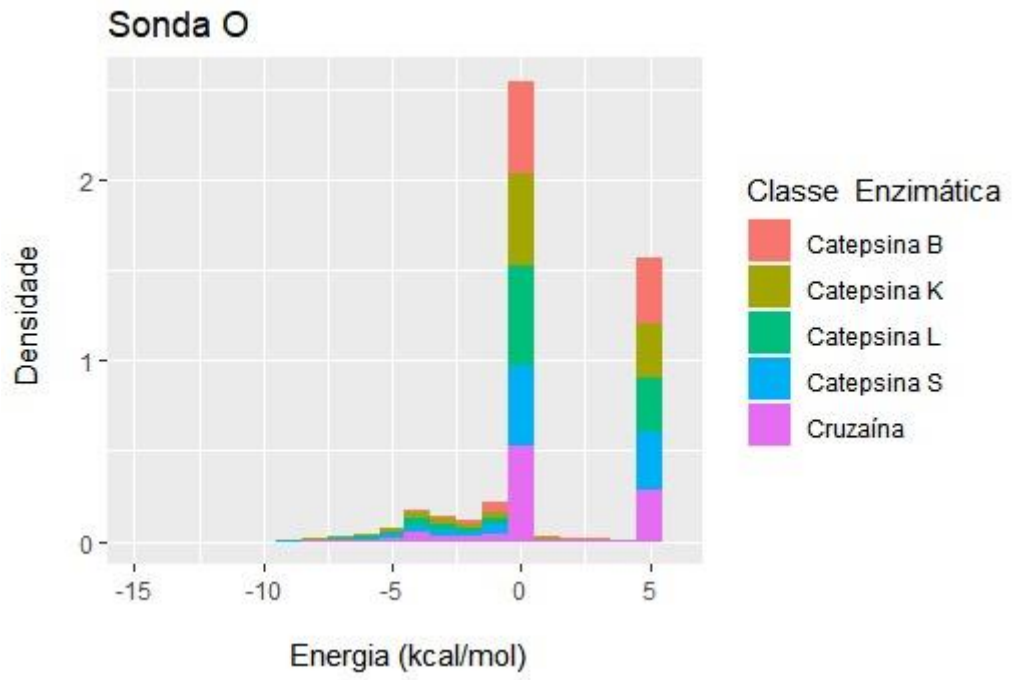

Figura B.10: Histograma das energias medidas com a sonda O para as 5 classes enzimáticas.

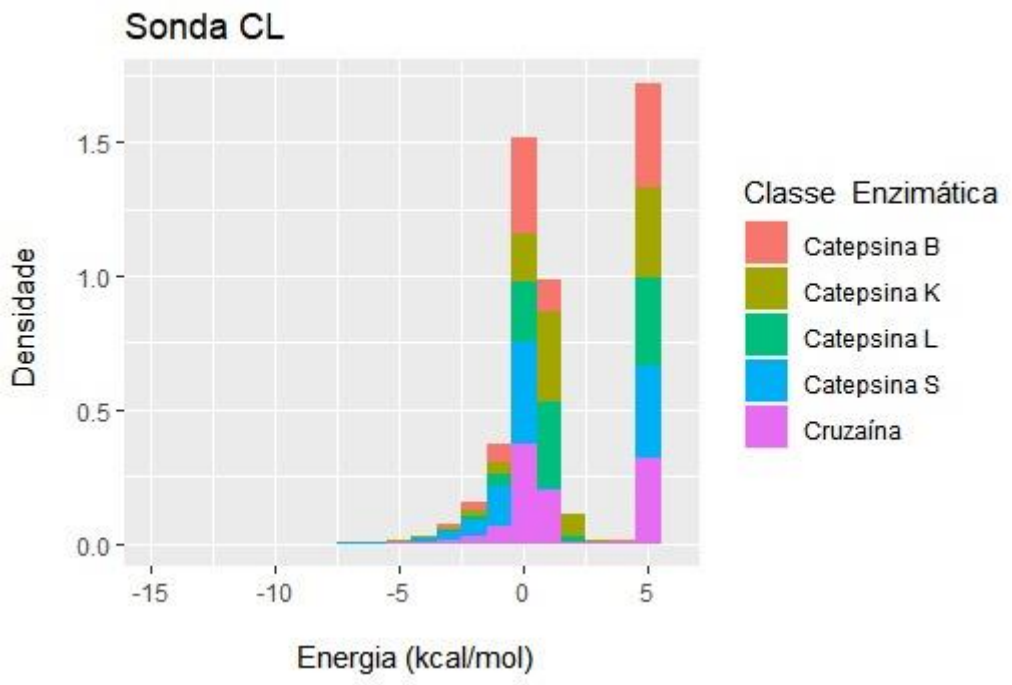

Figura B.11: Histograma das energias medidas com a sonda CL para as 5 classes enzimáticas.

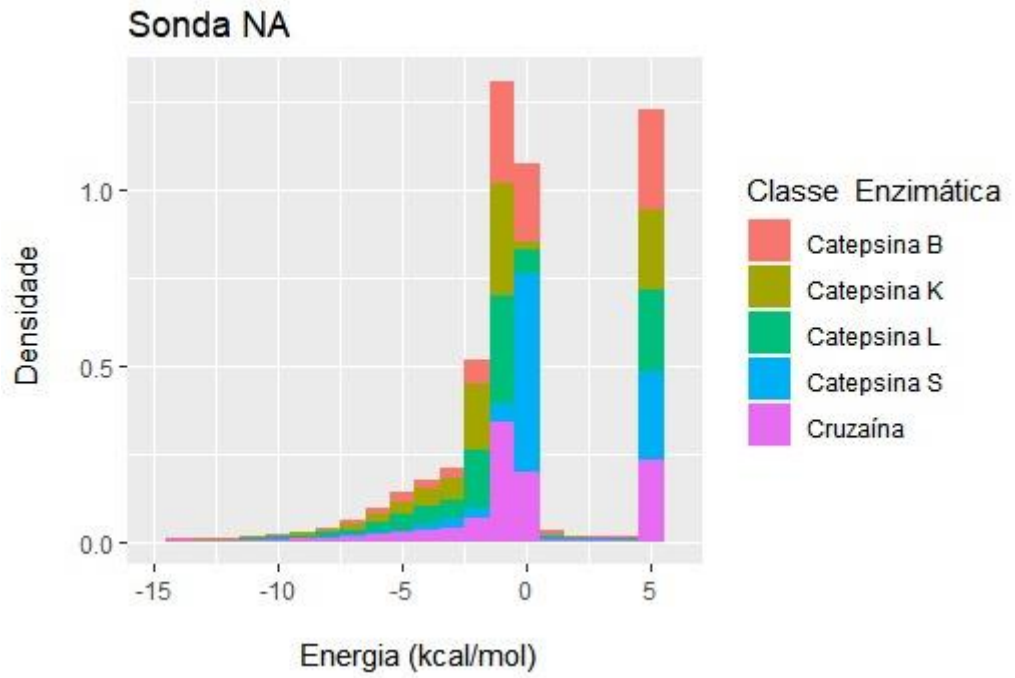

Figura B.12: Histograma das energias medidas com a sonda NA para as 5 classes enzimáticas.

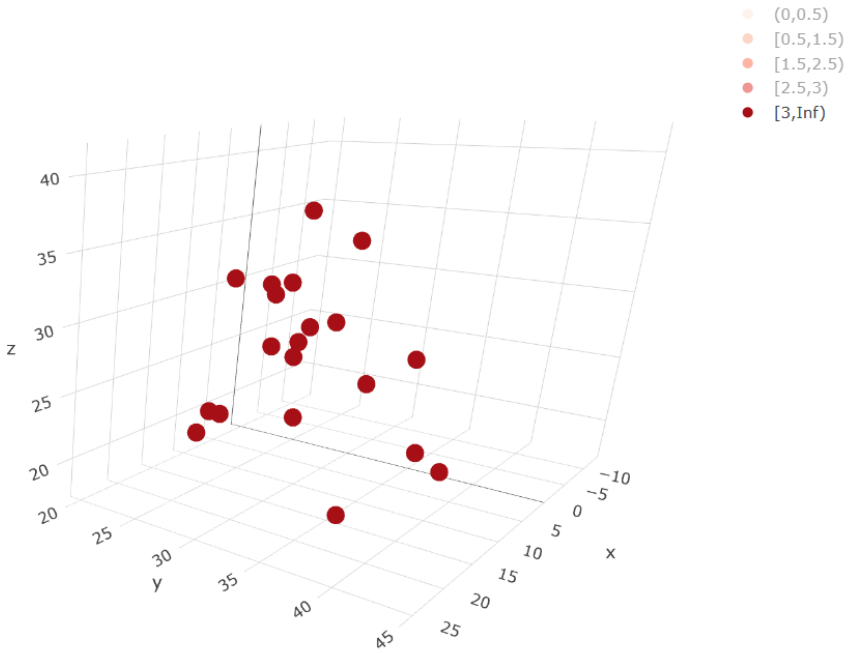

Figura B.13: Gráfico 3D da variância por ponto das energias medidas com a sonda C3 para as 5 classes enzimáticas.

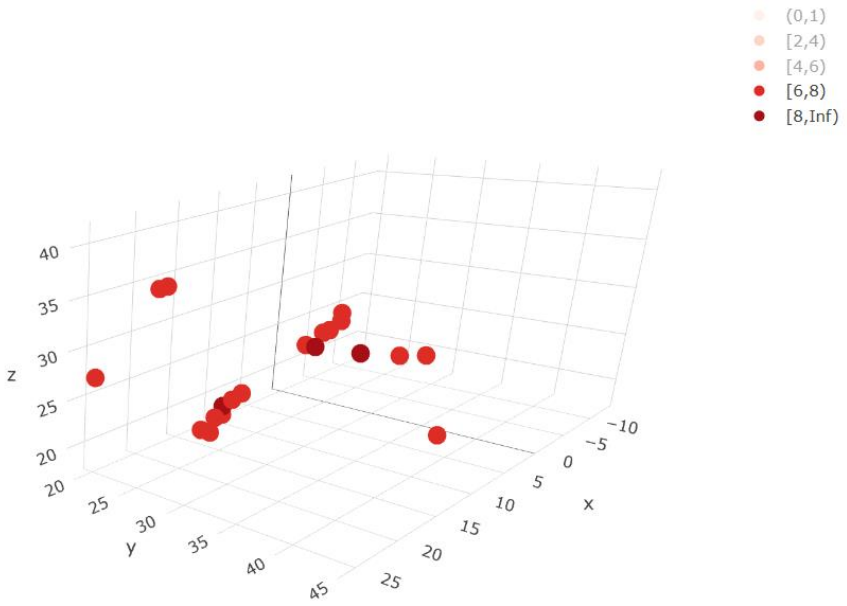

Figura B.14: Gráfico 3D da variância por ponto das energias medidas com a sonda CL para as 5 classes enzimáticas.

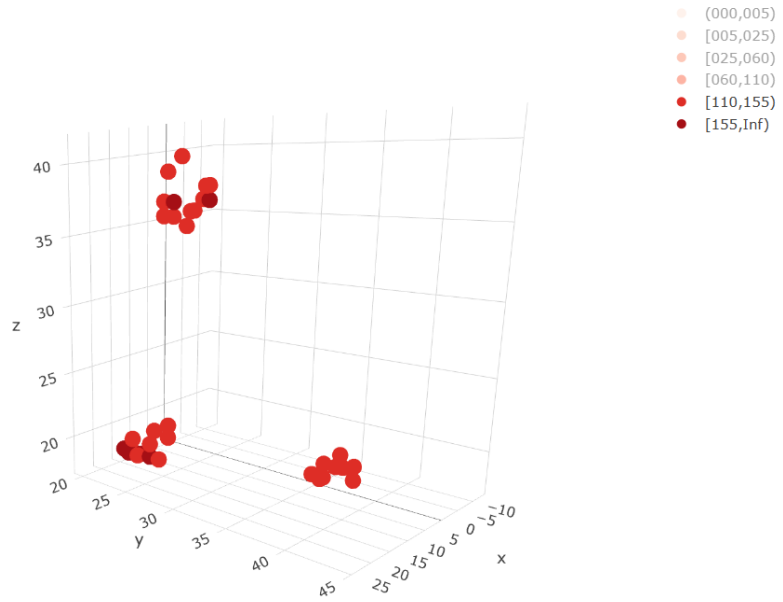

Figura B.15: Gráfico 3D da variância por ponto das energias medidas com a sonda NA para as 5 classes enzimáticas.

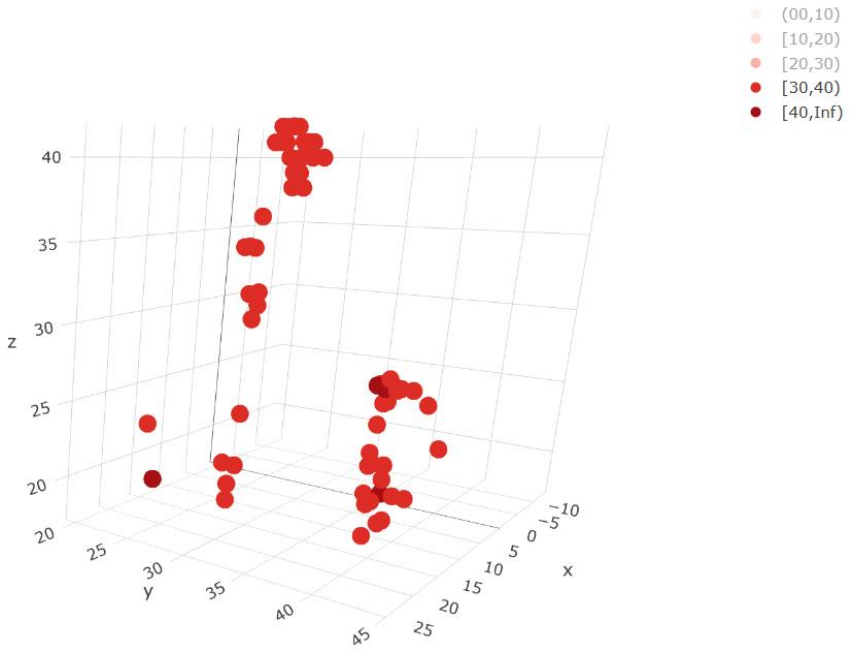

Figura B.16: Gráfico 3D da variância por ponto das energias medidas com a sonda N2 para as 5 classes enzimáticas.

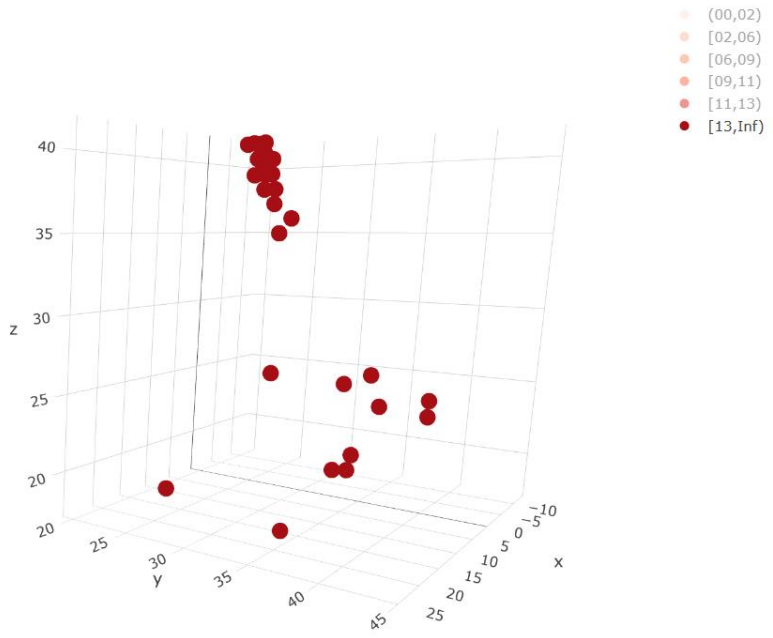

Figura B.17: Gráfico 3D da variância por ponto das energias medidas com a sonda O para as 5 classes enzimáticas.

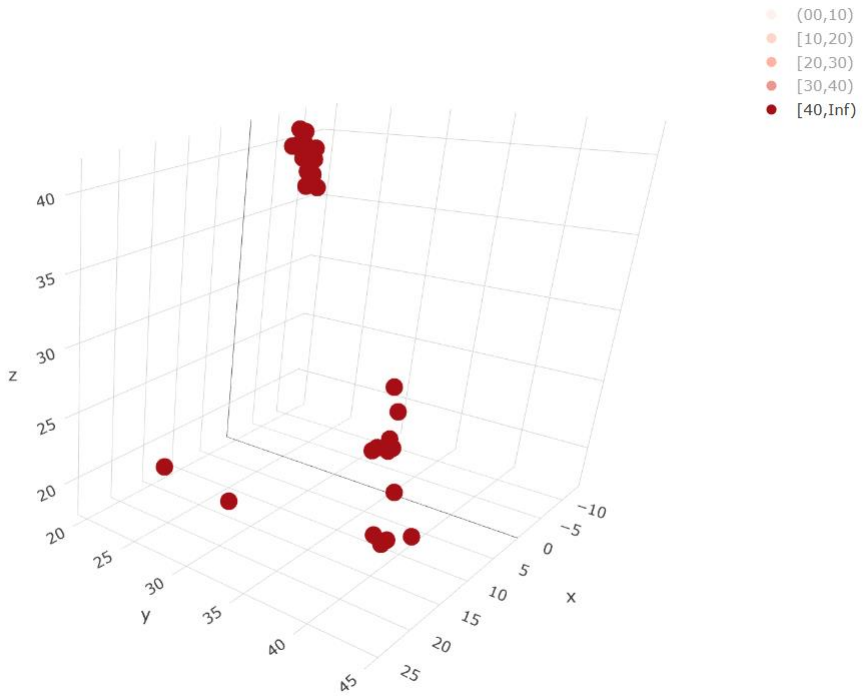

Figura B.18: Gráfico 3D da variância por ponto das energias medidas com a sonda OH2 para as 5 classes enzimáticas.

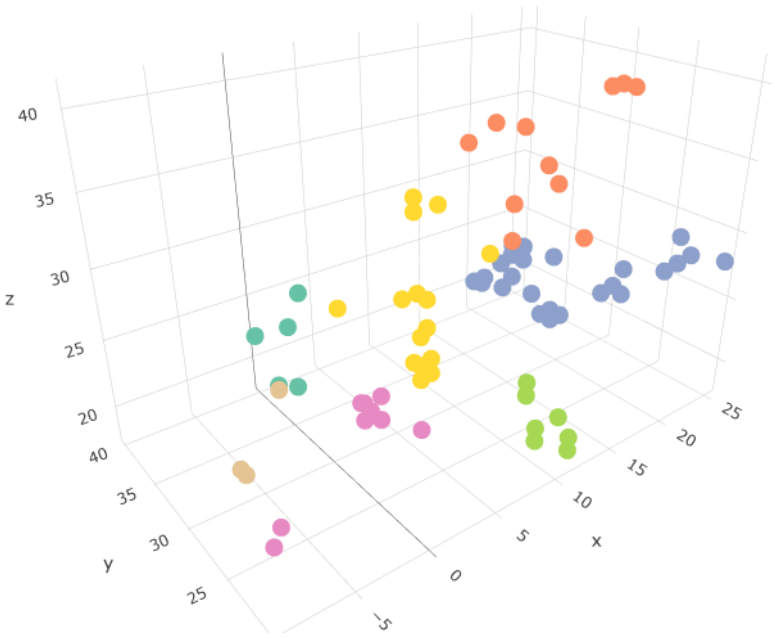

Figura B.19: Gráfico 3D dos 7 clusters formados por coordenadas considerando a sonda C3.

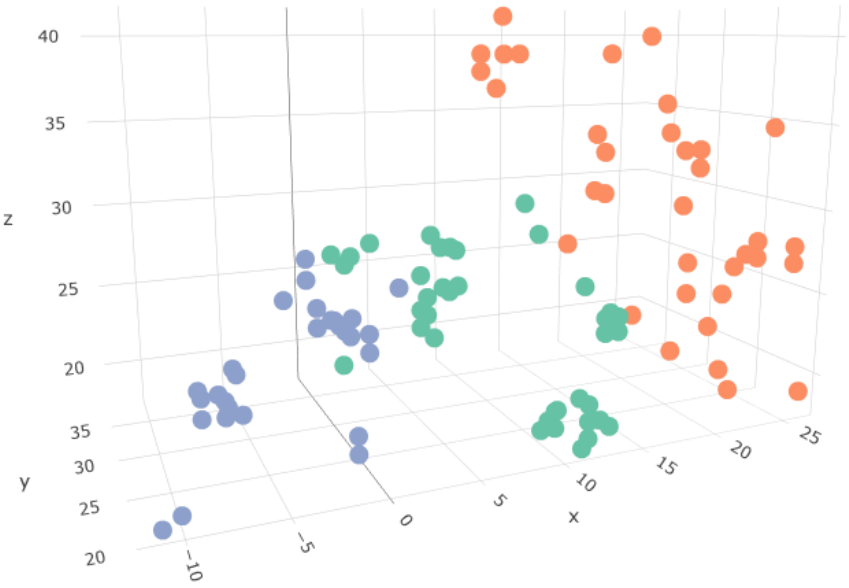

Figura B.20: Gráfico 3D dos 3 clusters formados por coordenadas considerando a sonda CL.

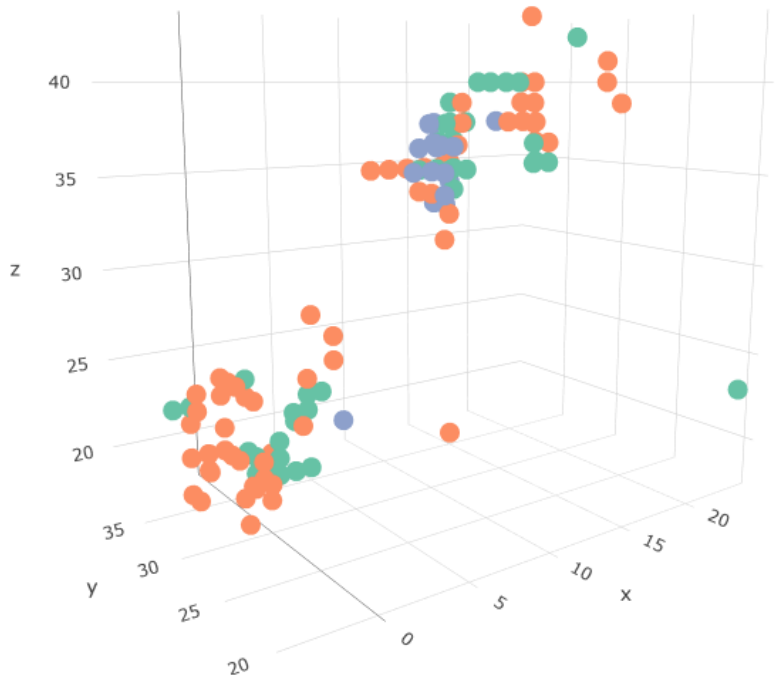

Figura B.21: Gráfico 3D dos 3 clusters formados por coordenadas considerando a sonda NA.

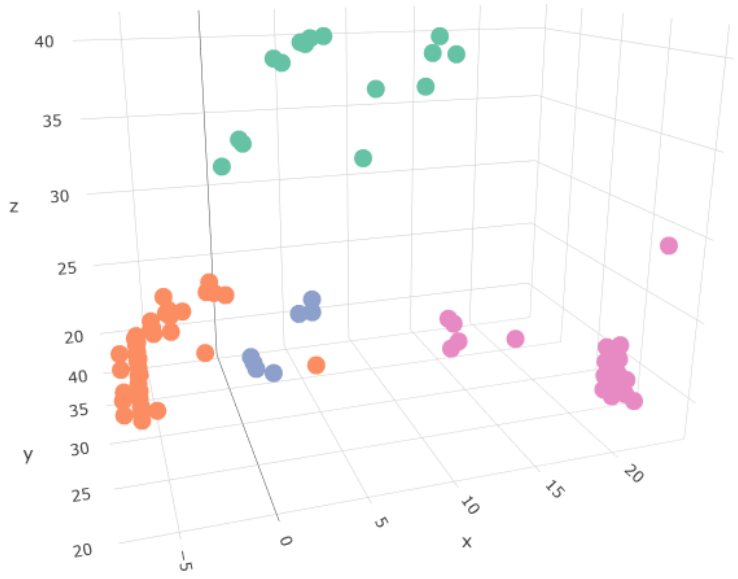

Figura B.22: Gráfico 3D dos 4 clusters formados por coordenadas considerando a sonda N2.

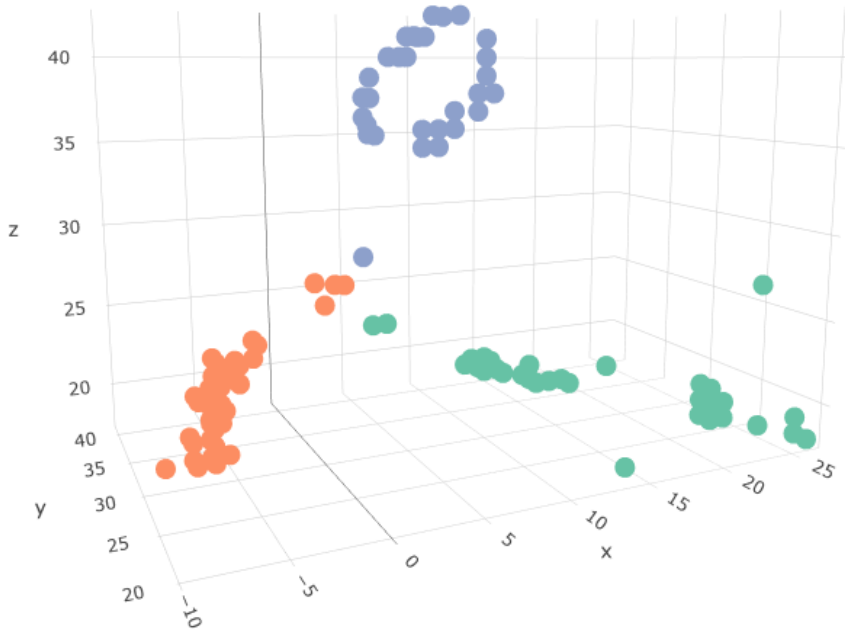

Figura B.23: Gráfico 3D dos 3 clusters formados por coordenadas considerando a sonda O.

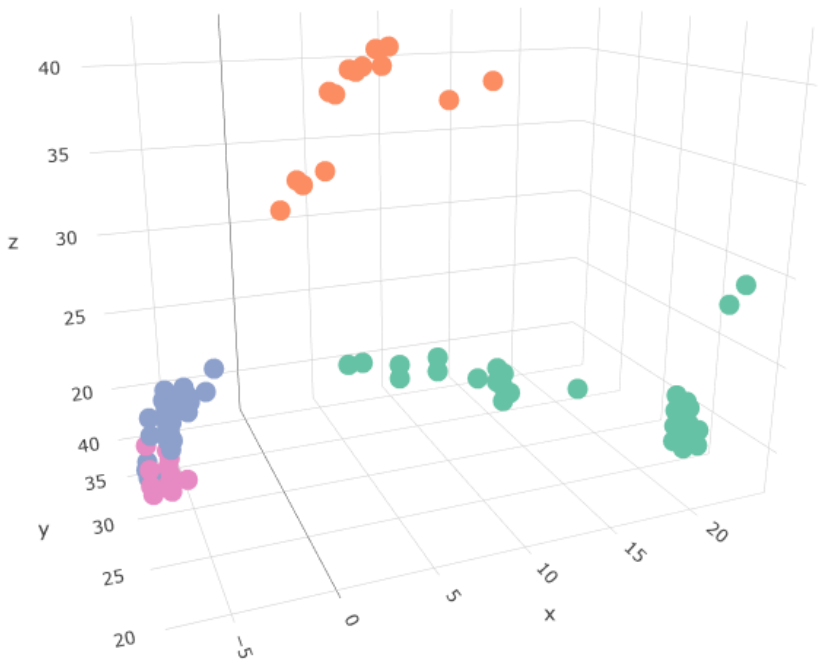

Figura B.24: Gráfico 3D dos 4 clusters formados por coordenadas considerando a sonda OH2.

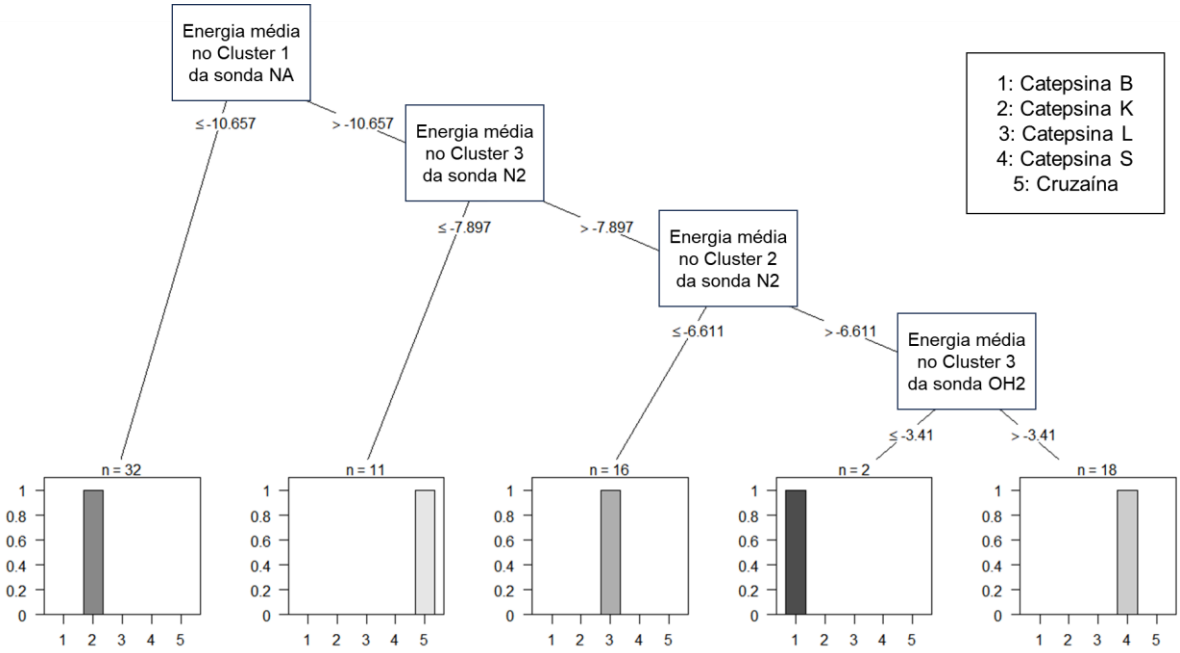

Figura B.25: Árvore de inferência condicional para todas as 24 médias energéticas de cada cluster de todas as sondas, considerando as amostras dadas como treino.

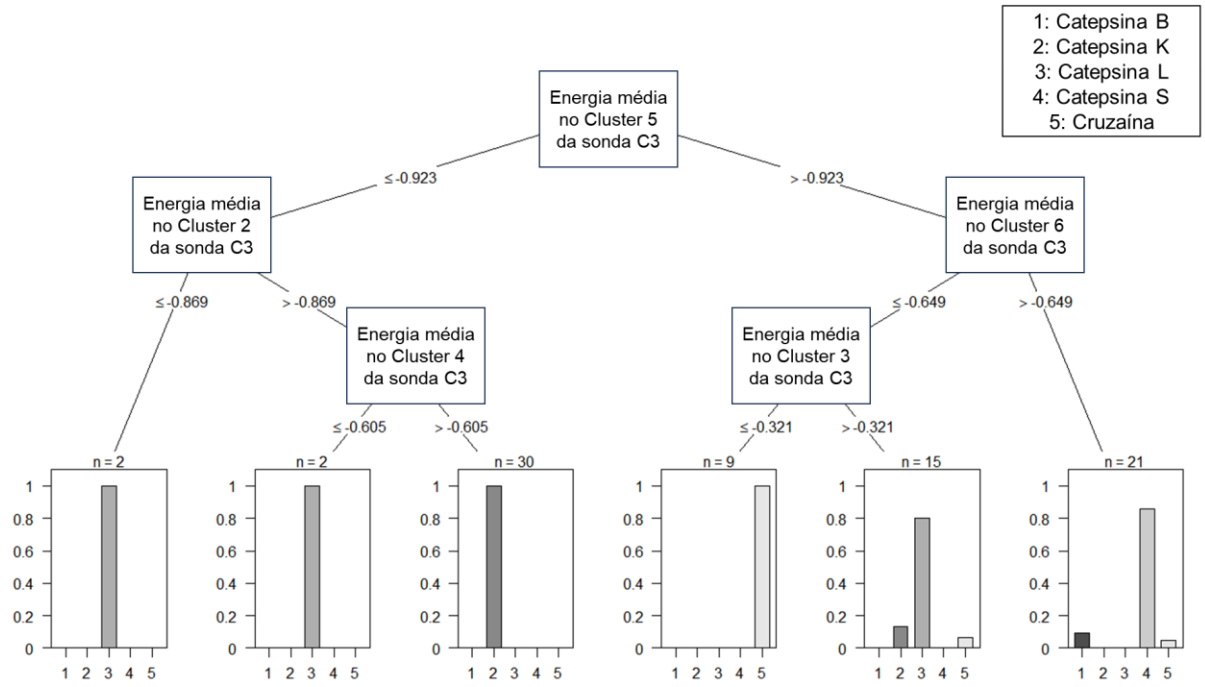

Figura B.26: Árvore de inferência condicional para as 7 médias energéticas de cada cluster da sonda Grupo Metila, considerando as amostras dadas como treino.

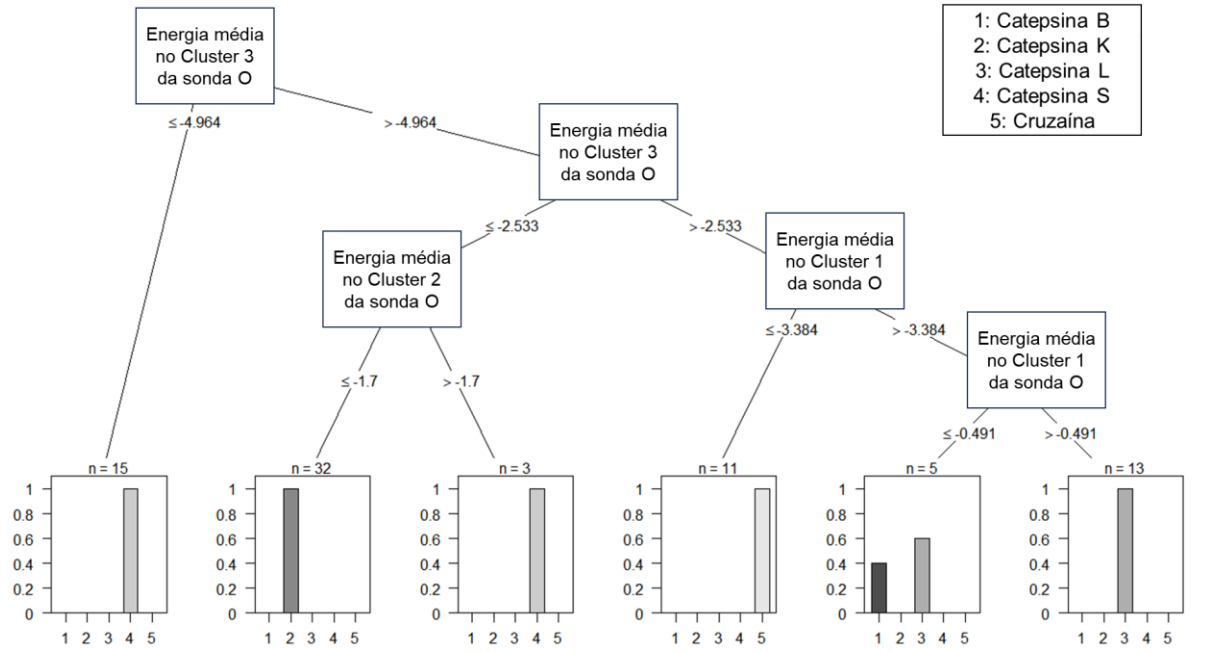

Figura B.27: Árvore de inferência condicional para as 3 médias energéticas de cada cluster da sonda Oxigênio Carbonílico, considerando as amostras dadas como treino.

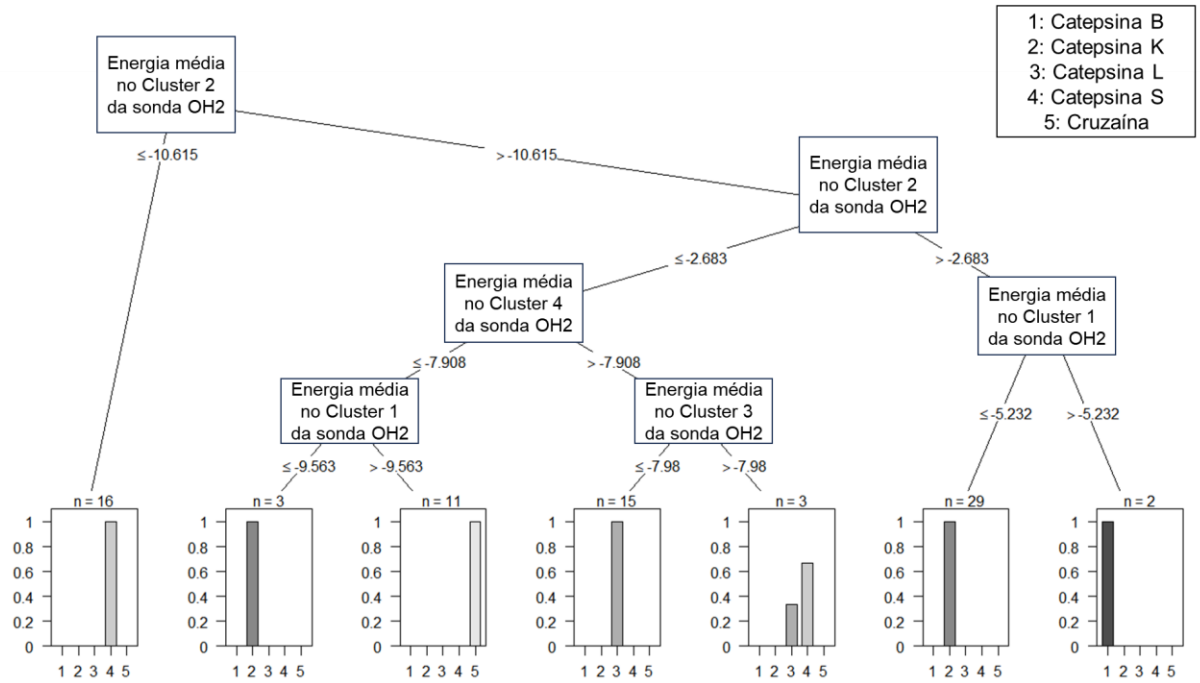

Figura B.28: Árvore de inferência condicional para as 4 médias energéticas de cada cluster da sonda Água, considerando as amostras dadas como treino.

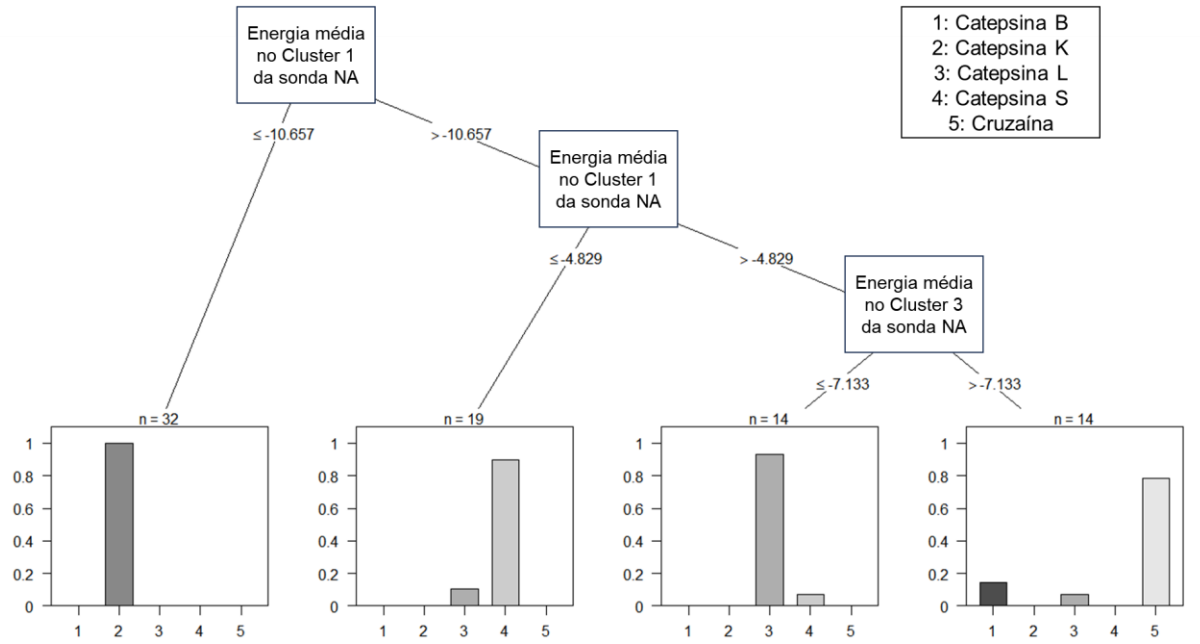

Figura B.29: Árvore de inferência condicional para as 3 médias energéticas de cada cluster da sonda Cátion Sódio, considerando as amostras dadas como treino.

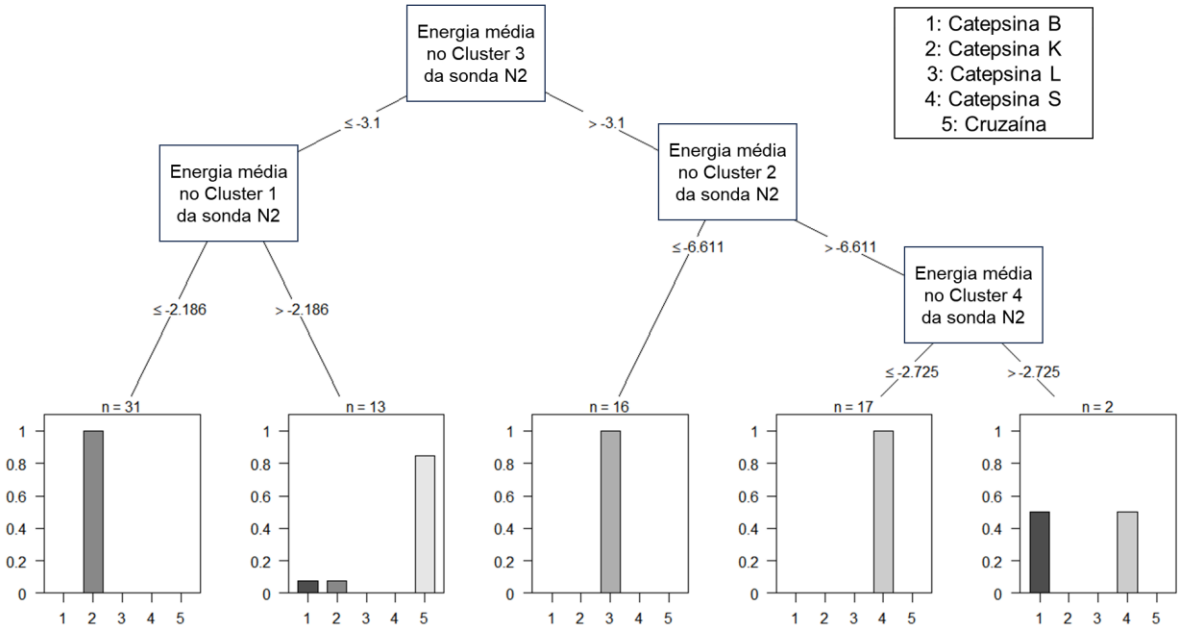

Figura B.30: Árvore de inferência condicional para as 4 médias energéticas de cada cluster da sonda Nitrogênio Amídico, considerando as amostras dadas como treino.

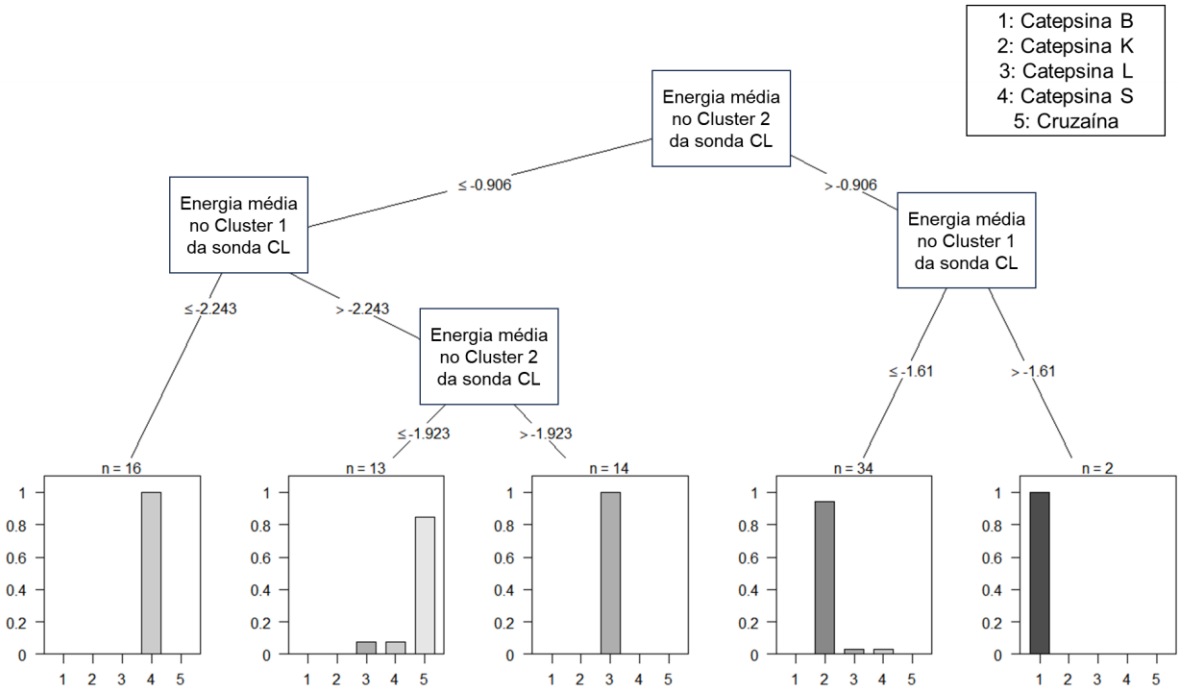

Figura B.31: Árvore de inferência condicional para as 3 médias energéticas de cada cluster da sonda Ânion Cloreto, considerando as amostras dadas como treino.