



UNIVERSIDADE ESTADUAL PAULISTA
“JÚLIO DE MESQUITA FILHO”
Câmpus de São José do Rio Preto

Laís Ozelin de Lima Pimentel

Papel da Hidrofobicidade na Evolução de Proteínas

São José do Rio Preto
2015

Laís Ozelin de Lima Pimentel

Papel da Hidrofobicidade na Evolução de Proteínas

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestra em Biofísica Molecular, junto ao Programa de Pós-Graduação em Biofísica Molecular, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Campus de São José do Rio Preto

Orientador: Prof. Dr. Vitor Barbanti
Pereira Leite

SÃO JOSÉ DO RIO PRETO

2015

Pimentel, Laís Ozelin de Lima.

Papel da hidrofobicidade na evolução de proteínas / Laís Ozelin de Lima Pimentel. -- São José do Rio Preto, 2015

67 f. : il., gráfs., tabs.

Orientador: Vitor Barbanti Pereira Leite

Dissertação (mestrado) – Universidade Estadual Paulista “Júlio de Mesquita Filho”, Instituto de Biociências, Letras e Ciências Exatas

1. Biologia molecular. 2. Biofísica. 3. Proteínas - Evolução.
4. Interações hidrofóbicas e hidrofílicas. 5. Dobramento de proteína.
I. Leite, Vitor Barbanti Pereira. II. Universidade Estadual Paulista "Júlio de Mesquita Filho". Instituto de Biociências, Letras e Ciências Exatas.
III. Título.

CDU – 577.112

Ficha catalográfica elaborada pela Biblioteca do IBILCE
UNESP - Câmpus de São José do Rio Preto

Laís Ozelin de Lima Pimentel

Papel da Hidrofobicidade na Evolução de Proteínas

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestra em Biofísica Molecular, junto ao Programa de Pós-Graduação em Biofísica Molecular, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Campus de São José do Rio Preto

Orientador: Prof. Dr. Vitor Barbanti Pereira Leite

Banca examinadora

Prof. Dr. Vitor Barbanti Pereira Leite
Professor Doutor
UNESP – São José do Rio Preto
Orientador

Prof. Dr. Ronaldo Junio de Oliveira
Professor Doutor
UFTM – Uberaba

Prof^a. Dra. Fátima Pereira de Souza
Professora Doutora
UNESP – São José do Rio Preto

SÃO JOSÉ DO RIO PRETO

2015

AGRADECIMENTOS

Agradeço aos meus pais e meus irmãos por todo apoio e carinho.

Agradeço ao professor Vitor pela orientação, aos colegas de grupo pela companhia e ajuda.

Aos professores que participaram e contribuíram com a minha formação. Aos funcionários do IBILCE, principalmente do departamento de Física.

Agradeço aos amigos por toda ajuda e por todos os momentos compartilhados. Em especial, agradeço ao Marcelo por sempre ser a pessoa mais disposta a ajudar e por ser meu grande amigo.

Agradeço a Nayara por sempre me apoiar e incentivar.

Agradeço ao Instituto de Biociências, Letras e Ciências exatas por todos os anos de minha formação.

Por fim, agradeço a CAPES pelo auxílio financeiro.

**“Our lives are just a fragment
Of the universe and all
There may be more then we can see”**

Kamelot

RESUMO

As proteínas são as moléculas mais abundantes nas células, sendo de crucial importância para o funcionamento das mesmas, desempenhando atividades tanto regulatórias quanto estruturais. Uma proteína se torna biologicamente ativa ao passar por um processo chamado enovelamento e, assim, alcançar seu estado nativo e funcional. Dentre vários fatores que influenciam no enovelamento da proteína, um que possui grande relevância é a hidrofobicidade da proteína, que ocorre devido às propriedades físicas das cadeias laterais dos aminoácidos. A hidrofobicidade contribui para a formação de um núcleo estável e compacto, além de participar na formação de contatos. Em trabalhos anteriores (J. Chem. Phys. 125, 084904, 2006), foi mostrado que perturbar uma proteína através de uma mutação pode causar efeitos diferentes dependendo de sua hidrofobicidade média. Proteínas com baixa hidrofobicidade mostraram-se sensíveis a mutações, resultando em uma taxa de enovelamento mais lenta, enquanto as proteínas de alta hidrofobicidade permaneceram consideravelmente inalteradas. Utilizando quatro proteínas reais de hidrofobicidades diferentes, este trabalho visa mostrar a influência da hidrofobicidade média na variabilidade sequencial de proteínas e, posteriormente, correlacionar com os valores Φ e com o acoplamento direto entre os aminoácidos que formam contato.

Palavras-chave: Proteínas, Variabilidade Sequencial, Hidrofobicidade, Evolução.

ABSTRACT

Proteins are the most abundant molecules in the cell, being of crucial importance for your functioning, performing regulatory functions both as structural. A protein becomes biologically active passing through a process called folding and then achieves your native state. Among several factors that influence the protein folding, one that has a great participation is the protein hydrophobicity, which occurs because of physical properties of side chains of amino acids. The hydrophobicity contributes to form a compact and stable core, also participates of contacts formation. In previous work (J. Chem. Phys. 125, 084904, 2006), was showed that inserting a mutation in a protein can cause different effects depending on its global hydrophobicity. Proteins with low hydrophobicity showed sensibility to mutations, resulting in a slower folding rate, while proteins with high hydrophobicity remain almost the same. This work will present the influence of global hydrophobicity in sequence variability throughout evolution. This work aims to show the influence of mean hydrophobicity in protein sequence variability and subsequently correlate with Φ values and direct coupling between aminoacids forming a contact.

Keywords: *Proteins, Sequence variability, Hydrophobicity, Evolution*

Sumário

I.	Lista de Figuras	11
II.	Lista de Tabelas.....	13
1.	INTRODUÇÃO	14
1.1.	Proteínas	14
1.2.	Enovelamento de Proteínas	14
1.3.	Evolução das Proteínas.....	16
1.4.	Hidrofobicidade.....	17
1.5.	Hidrofobicidade no enovelamento de proteínas	18
1.6.	Motivação	19
PARTE I		21
2.	OBJETIVOS ESPECÍFICOS.....	21
3.	METODOLOGIA.....	21
3.1.	Proteínas estudadas	21
3.1.1.	Lisozima.....	21
3.1.2.	Mioglobina.....	22
3.1.3.	Citocromo C	22
3.1.4.	Histona H3.....	23
3.2.	Seleção das Sequências	23
3.3.	Escalas de Hidrofobicidade.....	24
3.3.1.	Método do Particionamento	24
3.3.2.	Método da área acessível ao solvente	25
3.3.3.	Métodos cromatográficos.....	25
3.3.4.	Mutagênese de sítio dirigido	25
3.3.5.	Método das Propriedades Físicas	25
3.4.	Divisão em Grupos.....	26
3.5.	Análise.....	27
3.5.1.	Alinhamentos e o cálculo da dissimilaridade	27
3.5.2.	Cálculo do RMSD (Root Mean Square Deviation).....	28
4.	RESULTADOS.....	29
4.1.	Análise Sequencial	30
4.2.	Análise Estrutural.....	32
5.	CONCLUSÃO.....	33
PARTE II		35

6. OBJETIVOS ESPECÍFICOS.....	35
7. METODOLOGIA.....	35
7.1. Direct Coupling Analysis.....	35
7.2. Cálculo da Informação Direta	36
7.3. Valores Φ	38
7.3.1. Cálculo dos valores Φ	40
8. RESULTADOS.....	41
9. CONCLUSÃO.....	44
10. PERSPECTIVAS	45
11. REFERÊNCIAS	46
APÊNDICE A.....	52
PARTE I	52
A.1. Análise Sequencial	52
A.2. Análise Estrutural.....	59
PARTE II.....	62
A.3. Gráficos Citocromo C.....	62
A.4. Gráficos Mioglobina	63
APÊNDICE B.....	65
Filogenética dos Vertebrados.....	65

I. Lista de Figuras

Figura 1- Representação de aminoácido.....	14
Figura 2- Representação da superfície de energia proposto por Leopold [8].....	15
Figura 3- Funil de enovelamento representando o Molten Globule.....	18
Figura 4 - a) Parâmetro de enovelamento T_f / T_g em função do numero de contatos frustrados f na conformação nativa e da hidrofobicidade ¹ .b)Tempo mínimo de enovelamento em função da frustração f^1	19
Figura 5 - Relação entre Hidrofobicidade e Dissimilaridade referentes à Escala de Eisenberg. Resultado referente à Tese de Doutorado de Ricardo H. T. Silva [17]. Legenda de cores: Ciano – Histona, Vermelho – Citocromo C, Verde – Mioglobina, Azul – Lisozima.	20
Figura 6- Estrutura de Raio-X da lisozima da clara do ovo de peru. Fonte:Protein Data Bank.....	22
Figura 7- Estrutura e refinamento da oximioglobina ligada ao oxigênio. Fonte: Protein Data Bank	22
Figura 8- Estrutura do Citocromo C do <i>Paracoccus denitrificans</i> . Fonte: Protein Data Bank.....	23
Figura 9- Histona H3 da galinha. Fonte: Protein Data Bank.....	23
Figura 10-Relação entre Hidrofobicidade e Dissimilaridade. a) Grupos: Primatas e Aves. b) Grupos: Aves e Peixes. c)Grupos: Primatas e Peixes.....	30
Figura 11- a) Coeficiente angular dos gráficos Hidrofobicidade x Dissimilaridade para cada escala de Hidrofobicidade. Pr – Primatas; Pe – Peixes; A- Aves. b) Coeficiente de correlação entre as variáveis Hidrofobicidade e Dissimilaridade	31
Figura 12- Relação entre Hidrofobicidade e RMSD em Angstrom.....	32
Figura 13- a) Coeficiente angular dos gráficos Hidrofobicidade x Dissimilaridade para cada escala de Hidrofobicidade. b) Coeficiente de correlação entre as variáveis Hidrofobicidade e RMSD(em Angstrom) para todas as escalas de hidrofobicidade.	33
Figura 14 - Esquema do processo de utilização do método da Análise do Acoplamento Direto.	38
Figura 15- Variação da barreira de energia ao inserir uma mutação [39].....	39

Figura 16 - Gráfico Acoplamento Direto (DCA) x Phi x Hidrofobicidade.....	42
Figura 17 - Gráficos referentes à Proteína Lisozima. a) Hidrofobicidade x DCA. b) Valores phi x DCA.....	43
Figura 18 - Gráfico Acoplamento Direto (DCA) x Phi x Hidrofobicidade.....	43
Figura 19 - Gráficos referentes à Proteína Histona. a) Hidrofobicidade x DCA. b) Valores phi x DCA.....	44
Figura 20- Relação entre Hidrofobicidade e Dissimilaridade para os grupos dos Primatas e das Aves. Escala Chothia.	52
Figura 21- Relação entre Hidrofobicidade e Dissimilaridade para os grupos dos Primatas e dos Peixes. Escala Chothia.	52
Figura 22- Relação entre Hidrofobicidade e Dissimilaridade para os grupos dos Primatas e dos Peixes. Escala Chothia.	53
Figura 23- Relação entre Hidrofobicidade e Dissimilaridade para os grupos dos Primatas e das Aves. Escala Wolfenden.....	53
Figura 24 - Relação entre Hidrofobicidade e Dissimilaridade para os grupos dos Primatas e dos Peixes. Escala Wolfenden.....	54
Figura 25- Relação entre Hidrofobicidade e Dissimilaridade para os grupos das Aves e dos Peixes. Escala Wolfenden.	54
Figura 26- Relação entre Hidrofobicidade e Dissimilaridade para os grupos dos Primatas e das Aves. Escala Miyazawa.	55
Figura 27- Relação entre Hidrofobicidade e Dissimilaridade para os grupos dos Primatas e dos Peixes. Escala Miyazawa.	55
Figura 28 - Relação entre Hidrofobicidade e Dissimilaridade para os grupos das Aves e dos Peixes. Escala Miyazawa.	56
Figura 29- Relação entre Hidrofobicidade e Dissimilaridade para os grupos dos Primatas e das Aves. Escala Roseman.	56
Figura 30 - Relação entre Hidrofobicidade e Dissimilaridade para os grupos dos Primatas e dos Peixes. Escala Roseman.	57
Figura 31- Relação entre Hidrofobicidade e Dissimilaridade para os grupos das Aves e dos Peixes. Escala Roseman.....	57
Figura 32- Relação entre Hidrofobicidade e Dissimilaridade para os grupos dos Primatas e das Aves. Escala Kyte-Doolittle.....	58
Figura 33- Relação entre Hidrofobicidade e Dissimilaridade para os grupos dos Primatas e dos Peixes. Escala Kyte-Doolittle.....	58

Figura 34 - Relação entre Hidrofobicidade e Dissimilaridade para os grupos das Aves e dos Peixes. Escala Kyte-Doolittle.	59
Figura 35- Relação entre Hidrofobicidade e RMSD para todos os grupos. Escala Chothia.	59
Figura 36- Relação entre Hidrofobicidade e RMSD para todos os grupos. Escala Wolfenden.....	60
Figura 37- Relação entre Hidrofobicidade e RMSD para todos os grupos. Escala Miyazawa.....	60
Figura 38 - Relação entre Hidrofobicidade e RMSD para todos os grupos.Escala Rose man.	61
Figura 39- Relação entre Hidrofobicidade e RMSD para todos os grupos. Escala KD.	61
Figura 40 - Gráfico Acoplamento Direto (DCA) x Phi x Hidrofobicidade	62
Figura 41- Gráfico Hidrofobicidade x DCA referente ao Citocromo C	62
Figura 42 - Gráfico Hidrofobicidade x Phi referente ao Citocromo C	63
Figura 43- Gráfico Acoplamento Direto (DCA) x Phi x Hidrofobicidade	63
Figura 44- Gráfico Hidrofobicidade x DCA referente a Mioglobina	64
Figura 45- Gráfico Hidrofobicidade x Phi referente a Mioglobina	64
Figura 46 - Relações Filogenéticas entre os principais grupos de vertebrados vivos ⁵³	66
Figura 47 - Relação filogenética entre as espécies envolvidas no trabalho. Construída utilizando o software Winclada com Ri=63 e Ci=76. As cores indicam os grupos: Azul – Peixes, Amarelo-Aves, Vermelho-Primatas.	67

II. Lista de Tabelas

Tabela 1-Escala consensus de hidrofobicidade para cada resíduo	26
Tabela 2- Espécies analisadas e o código UNIPROT	27
Tabela 3- Espécies analisadas e o código PDB	29

1. INTRODUÇÃO

1.1. Proteínas

As proteínas são as moléculas mais abundantes das células e estão envolvidas em todos os processos biológicos mais importantes. Todas as proteínas são formadas por aminoácidos, que são ligados de forma covalente em sequências lineares por ligação peptídica. Na natureza, existem 20 tipos de aminoácidos compostos por um carbono tetraédrico, o carbono alfa, ligado a um grupo amina, um grupo carboxila, um átomo de hidrogênio e a uma cadeia lateral específica. Cada aminoácido possui uma cadeia lateral (R) que determina suas propriedades físicas e químicas [2], sendo o diferencial entre um aminoácido e outro. De acordo com sua cadeia lateral, os aminoácidos podem ser classificados como apolares, polares não carregados, polares carregados e aromáticos.

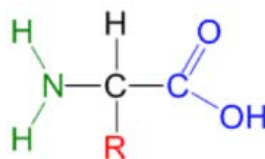


Figura 1- Representação de aminoácido

As proteínas possuem um sistema de descrição formado por quatro níveis estruturais, são eles: estrutura primária, estrutura secundária, estrutura terciária e estrutura quaternária.

1.2. Enovelamento de Proteínas

Todas as proteínas possuem uma função específica, sendo esta associada a uma estrutura única, chamada de estado nativo. Anfinsen descobriu, por meios de experimentos físico-químicos, que a proteína se enovela, pois o estado nativo é o estado termodinâmico mais estável, obtendo o mínimo global da energia livre de Gibbs [3,4], e esta estrutura é única por ser determinada pela sequência de aminoácidos que compõe a proteína. Então, a proteína se enovelaria a fim de buscar o estado mais estável e faria isso espontaneamente, sem que outra molécula intervisse na busca. Esta ficou conhecida como a hipótese termodinâmica do enovelamento.

O Paradoxo de Levinthal diz que, pela hipótese termodinâmica, a escala de tempo do enovelamento seria astronômica. Para solucionar o problema, Levinthal sugeriu a hipótese cinética, onde a proteína não busca aleatoriamente seu estado nativo, mas percorre um caminho de enovelamento bem definido [5], com estados intermediários.

A teoria mais aceita para descrever o processo do enovelamento é a teoria do funil (Energy Landscape)[6-8]. Esta teoria é baseada na ideia de que o estado nativo da proteína é único. A superfície de energia representa, por meio de um funil, as possíveis conformações acessíveis da proteína. A largura do topo do funil está relacionada ao conjunto das conformações da proteína no início do enovelamento, representando estruturas com um alto nível energético e alta entropia. À medida que a superfície vai se afinilando, a energia livre diminui, assim como o número de conformações, de forma que ao chegar ao final do funil é encontrado o estado nativo da proteína, um único estado com o mínimo de energia (Figura 2). O caminho que a proteína segue durante o enovelamento não é único, existindo, portanto, inúmeras rotas que levam ao estado nativo.

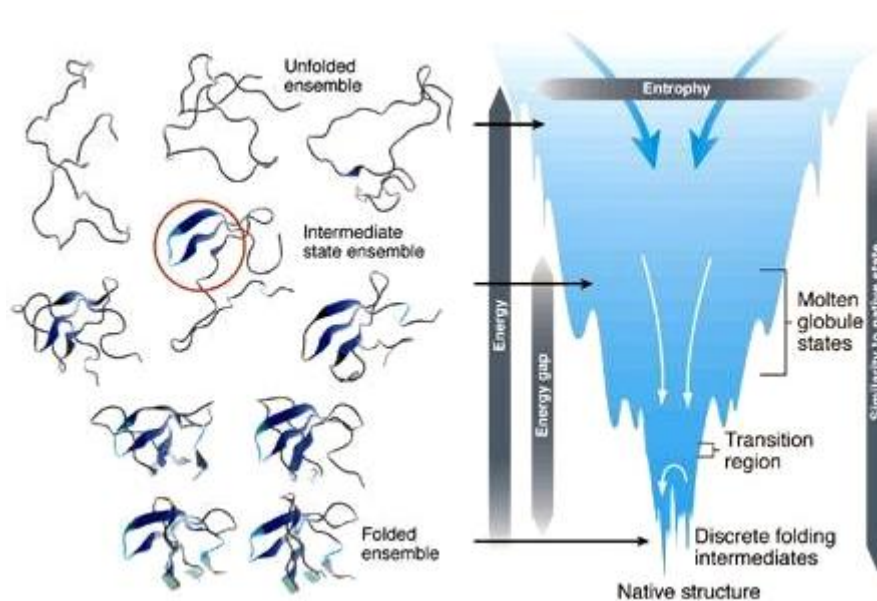


Figura 2- Representação da superfície de energia proposto por Leopold [8]

A superfície de energia é rugosa, devido aos mínimos locais que são como armadilhas para o enovelamento da proteína. Como a proteína deve se enovelar a tempo de exercer sua função, o funil deve possuir alta inclinação, de modo que seja o suficiente para vencer o armadilhamento e chegar ao estado nativo. Essa teoria resolveu o Paradoxo de Levinthal e, hoje, é amplamente utilizada no estudo de enovelamento de proteínas.

As proteínas não são estruturas aleatórias, portanto elas passaram pelo processo evolutivo para chegar a suas sequências e estruturas atuais.

1.3. Evolução das Proteínas

No estudo da evolução molecular, relacionam-se os processos evolutivos que resultaram em modificações no DNA, RNA e, conseqüentemente, nas proteínas ao longo do tempo. Ocorrendo modificações por mutações no DNA, serão codificadas proteínas com alterações em seus aminoácidos. Quando se trata de evolução proteica, deve-se falar de sequências homologas e similares. Homologia é um conceito qualitativo em biologia, estando relacionado à origem evolutiva de uma sequência. Já a similaridade é um conceito quantitativo que se relaciona à semelhança existente entre duas sequências. Algumas sequências homólogas apresentam um grande grau de similaridade, enquanto outras, devido aos processos evolutivos, apresentam baixo grau de similaridade.

O processo de evolução de proteínas é desafiador para a Biofísica e Biologia Molecular por ser um processo que aborda diversos fatores condicionantes. Ao longo do enovelamento de determinada proteína, seus aminoácidos interagem formando contatos que a direcionam ao estado nativo, ou seja, estado de mínima energia. Porém, a proteína não está apta a satisfazer todas as condições de minimização de energia, sendo isso chamado de frustração. Como a frustração não pode ser evitada, o estado nativo da proteína segue um padrão de mínima frustração [9,10]. A estabilidade de uma proteína nativa é um critério muito importante para sua funcionalidade [11], portanto, há evidências na literatura de que a evolução se limitou de forma a evitar sequências que proporcionam baixa estabilidade [12].

Durante o processo de evolução, algumas sequências proteicas se alteraram drasticamente, mantendo a mesma função, enquanto outras permaneceram fortemente conservadas. Algumas características contribuem para esse fato e, uma delas, é a hidrofobicidade média de cada proteína [1].

1.4. Hidrofobicidade

A hidrofobicidade é um dos fatores de maior influência no enovelamento e está relacionada às características físicas dos aminoácidos. Nos anos 50, Walter Kauzmann indicou a hidrofobicidade como fonte de estabilidade das proteínas [13].

Em ambientes aquosos polares, ocorrem diversas interações não covalentes entre diferentes domínios de uma molécula biológica, como interações eletrostáticas, ligações de hidrogênio ou interações de Van der Waals. A interação desfavorável entre a água e as moléculas não polares, como os resíduos de aminoácidos das proteínas, é muito relevante no âmbito molecular. A hidrofobicidade denota a repulsão entre moléculas apolares e a água, essa característica resulta da maior atração das moléculas de água umas pelas outras do que pelo soluto [14]. Ao inserir compostos apolares em água, surge uma interação desfavorável e ocorre uma variação negativa na entropia. Este efeito provém da capacidade da água de formar ligações de hidrogênio com moléculas de água vizinhas. Devido à alta eletronegatividade do oxigênio, a molécula de água é polar, possuindo cargas negativas próximas ao oxigênio e positivas perto dos átomos de hidrogênio. Desta forma, o oxigênio de uma molécula de água é fortemente atraído pelo hidrogênio de outra molécula, fazendo-se, então, a ligação entre duas moléculas de água mais energeticamente favorável do que a ligação com um soluto apolar. Portanto, nos sistemas que possuem água como solvente, as interações com grupos hidrofóbicos criam núcleos, aglomerados, e esta compactação é chamada de Efeito Hidrofóbico [42].

Os compostos apolares não possuem grupos para formar pontes de hidrogênio com a água, portanto forma-se uma cavidade hidrofóbica ao redor da molécula apolar, constituída por ligações de hidrogênio entre as moléculas de água, resultando em uma estrutura ordenada. Com o efeito hidrofóbico, a água expulsa o soluto apolar, fazendo-o se agregar em núcleos compactos em tais cavidades hidrofóbicas. Agregando grupos apolares é possível diminuir a área de contato com a água.

1.5. Hidrofobicidade no enovelamento de proteínas

Foi verificada uma relação entre a hidrofobicidade e o enovelamento de proteínas, sendo um dos fatores mais importantes para a determinação da topologia do funil de enovelamento. Notou-se que proteínas com alta hidrofobicidade média possuíam um enovelamento diferente das proteínas com um regime de baixa hidrofobicidade [1].

Os aminoácidos altamente hidrofóbicos interagem de forma desfavorável com a água, portanto, devido ao efeito hidrofóbico, eles são excluídos pelo solvente. Desta forma, em proteínas que estão em um regime de alta hidrofobicidade, seu enovelamento é dado em duas etapas. Inicialmente, seus aminoácidos hidrofóbicos são movidos para o centro da estrutura, formando um núcleo compacto e este estado é chamado de Molten Globule. Em seguida, o Molten Globule sofre um rearranjo que direciona lentamente o enovelamento para a estrutura nativa.

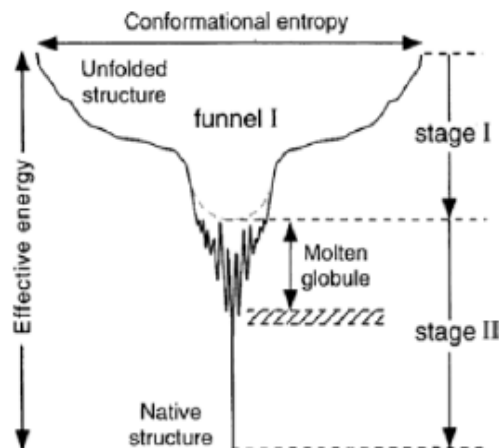


Figura 3- Funil de enovelamento representando o Molten Globule.

Proteínas em regime de baixa hidrofobicidade possuem um enovelamento que segue diretamente, de forma ordenada, para o estado nativo.

1.6. Motivação

Em um trabalho anterior [1], foram estudadas proteínas otimizadas, ou seja, sequências cujo estado nativo não possuía contatos frustrados entre monômeros e sequências mutadas cujo estado nativo possuía certo grau de frustração. O trabalho visava mostrar que as taxas de enovelamento dependem do grau de frustração e do regime de hidrofobicidade em que a proteína se encontra. Foi notado que a frustração alterava fortemente a temperatura de enovelamento de uma sequência de baixa hidrofobicidade, mostrando que esta não é robusta a mutações. Por outro lado, a frustração não alterou de forma considerável as sequências de alta hidrofobicidade. No caso cinético, as mutações afetaram o tempo de enovelamento de sequências de baixa hidrofobicidade, enquanto as sequências de alta hidrofobicidade se mostraram praticamente inalteradas.

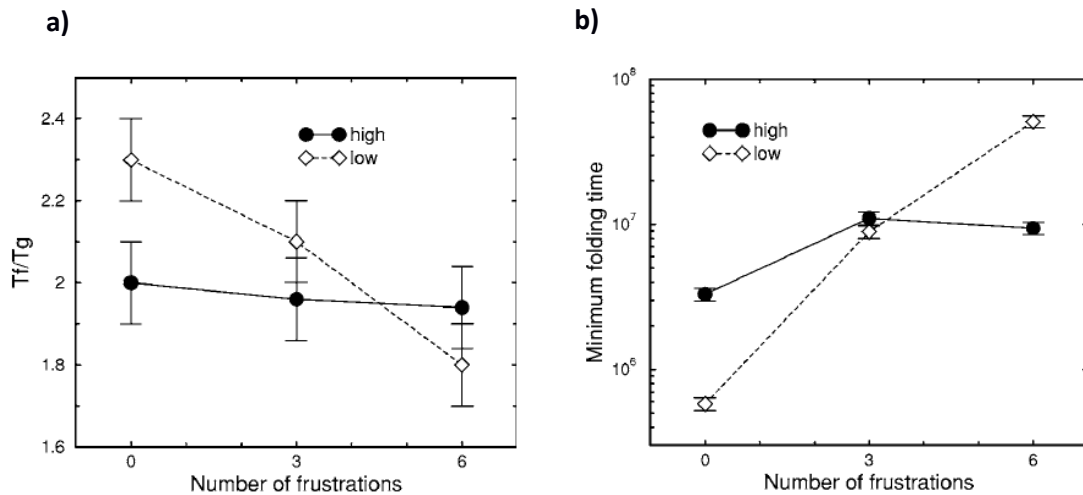


Figura 4 - a) Parâmetro de enovelamento T_f/T_g em função do número de contatos frustrados f na conformação nativa e da hidrofobicidade¹. b) Tempo mínimo de enovelamento em função da frustração f ¹.

A figura 4-a mostra que, ao aumentar o número de frustrações em uma proteína otimizada, o efeito é diferente dependendo do regime de hidrofobicidade que a proteína se encontra. Para a proteína de baixa hidrofobicidade, o aumento da frustração diminuiu a estabilidade do enovelamento, dada pela razão das temperaturas de enovelamento T_f e de armadilhamento (Temperatura de glass, T_g), ou seja, T_f/T_g . Na figura 4-b, é apresentada a alteração no tempo mínimo de enovelamento de acordo com o número de frustrações e com o regime de hidrofobicidade.

Baseando-se nestes resultados, o aluno André S. Maizel, em seu trabalho de conclusão de curso [16], realizou análises com quatro proteínas reais, de hidrofobicidade globais diferentes. Através de alinhamentos múltiplos com sequências de 41 espécies, notou-se que proteínas de baixa hidrofobicidade apresentavam uma grande conservação sequencial e proteínas de alta hidrofobicidade apresentavam maior variabilidade. Desta forma, fez necessária uma análise específica deste caso, buscando encontrar uma relação entre a hidrofobicidade média e a variabilidade sequencial, utilizando mais de uma escala de hidrofobicidade existente para os aminoácidos.

Este caso foi estudado no trabalho de doutorado do Ricardo H. T. Silva [17], utilizando quatro proteínas reais e seis escalas de hidrofobicidade. O trabalho visava encontrar a relação entre Hidrofobicidade e Dissimilaridade, utilizando 41 espécies de animais. Os resultados encontrados apresentaram barras de erro muito altas, sendo inconclusivos. A figura 5 mostra o gráfico obtido utilizando a Escala Consensus, desenvolvida por Eisenberg et al [24], para a Hidrofobicidade média e Dissimilaridade média.

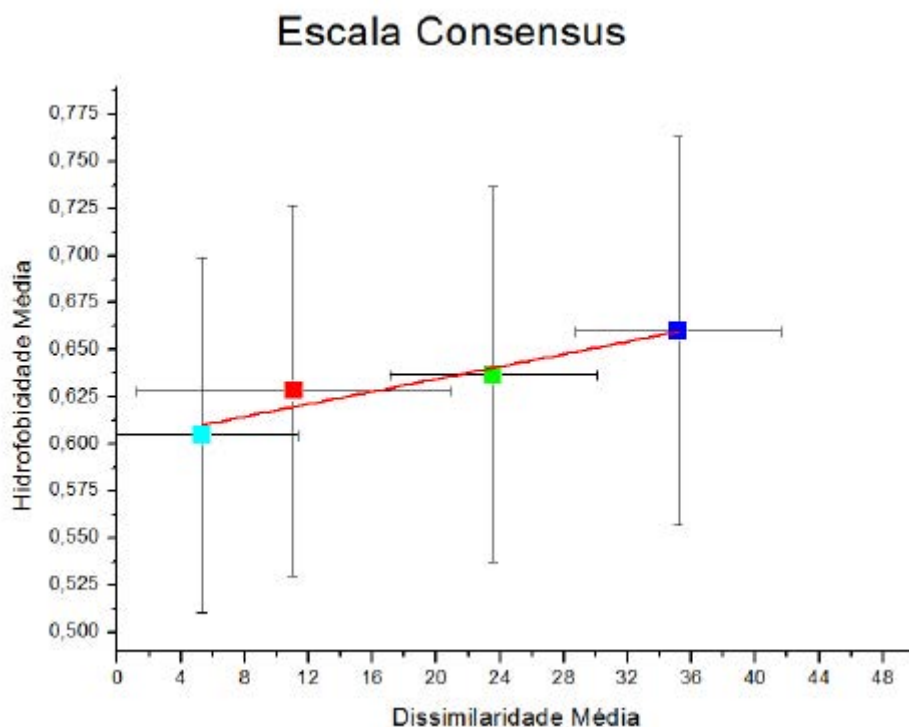


Figura 5 - Relação entre Hidrofobicidade e Dissimilaridade referentes à Escala de Eisenberg. Resultado referente à Tese de Doutorado de Ricardo H. T. Silva [17]. Legenda de cores: Ciano – Histona, Vermelho – Citocromo C, Verde – Mioglobina, Azul – Lisozima.

Resultados similares foram obtidos para todas as escalas utilizadas no trabalho. Nota-se que, a partir destes resultados, não é possível afirmar com exatidão que as duas

variáveis possuem uma relação linearmente crescente, pois o erro obtido permite outras configurações. Desta forma, motivou-se o desenvolvimento de uma nova metodologia, a fim de reduzir o erro nas análises e verificar correlação entre Hidrofobicidade e Dissimilaridade.

PARTE I

2. OBJETIVOS ESPECÍFICOS

Através do trabalho desenvolvido com proteínas otimizadas [1], foi visto que proteínas de alta hidrofobicidade são robustas às mutações. A partir desses resultados, foi ponderado qual seria o efeito deste fato no desenvolvimento evolutivo de uma proteína.

O objetivo é constatar a influência da hidrofobicidade na variabilidade sequencial das proteínas. Para este propósito são estudadas quatro proteínas reais de diferentes hidrofobicidades médias: Histona, Citocromo C, Mioglobina e Lisozima. Foram utilizadas as seis escalas de hidrofobicidade já trabalhadas ao longo da Tese de Doutorado citada [17]. Neste trabalho, as espécies foram divididas em grupos com distâncias evolutivas bem definidas. Esses são constituídos por animais pertencentes às classes dos peixes, aves e mamíferos. Os mamíferos estão representados por animais pertencentes apenas à ordem dos primatas, portanto os grupos foram estabelecidos como: Primatas, Peixes e Aves.

3. METODOLOGIA

3.1. Proteínas estudadas

Foram utilizadas quatro proteínas para o estudo: Lisozima, Citocromo C, Mioglobina e Histona H3. Tais proteínas foram escolhidas em trabalhos anteriores [16,17] devido à sua história evolutiva.

3.1.1. Lisozima

A Lisozima é uma pequena enzima responsável pela clivagem de carboidratos em bactérias e é utilizada como mecanismo de defesa. É encontrada nas lágrimas e no muco, possui um tamanho médio de 130 aminoácidos e é uma proteína com hidrofobicidade alta.

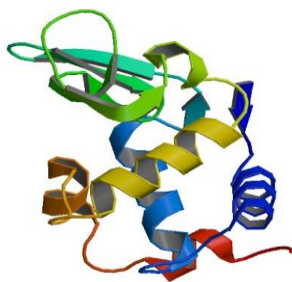


Figura 6- Estrutura de Raio-X da lisozima da clara do ovo de peru. Fonte: Protein Data Bank.

A lisozima é uma proteína que possui alta variabilidade sequencial ao longo da evolução.

3.1.2. Mioglobina

A Mioglobina é uma proteína encontrada nos músculos, tendo por função se ligar ao oxigênio. Possui cerca de 150 aminoácidos e um nível de hidrofobicidade médio-alta. Suas sequências possuem considerável variabilidade sequencial durante a evolução.

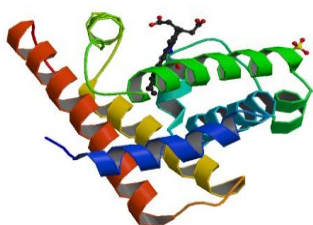


Figura 7- Estrutura e refinamento da oximioglobina ligada ao oxigênio. Fonte: Protein Data Bank

3.1.3. Citocromo C

O Citocromo C é uma proteína responsável pelo transporte de elétrons no processo de produção energética, sendo encontrado nas mitocôndrias. Possui um tamanho de cerca de 100 aminoácidos e uma hidrofobicidade médio-baixa. Suas sequências possuem uma maior conservação ao longo da evolução em comparação as proteínas anteriores.

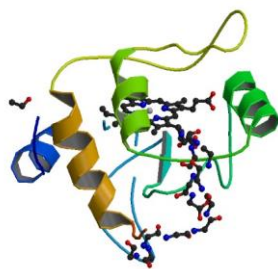


Figura 8- Estrutura do Citocromo C do Paracoccus denitrificans. Fonte: Protein Data Bank.

3.1.4. Histona H3

A Histona H3 é uma das proteínas envolvidas na formação do nucleossomo, auxiliando na compactação do DNA. Possui cerca de 95 aminoácidos e hidrofobicidade baixa. Suas sequências são altamente conservadas ao longo da evolução.

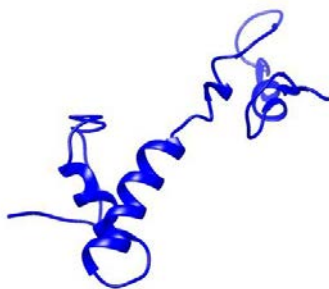


Figura 9- Histona H3 da galinha. Fonte: Protein Data Bank

3.2. Seleção das Sequências

As proteínas foram buscadas nos bancos de dados UNIPROT [18] e PIR [19], visando o maior número de sequências que poderiam ocorrer para diferentes organismos. Inicialmente, as buscas foram feitas para a Lisozima, e obteve-se cerca de quarenta mil sequências no banco de dados UNIPROT. O número de sequências depositadas no banco de dados foi atualizado e, provavelmente, hoje seria encontrada uma quantidade superior. Após obter as sequências, foram realizadas as análises dos dados e, ao notar redundância e discrepância de tamanhos, foi estipulado critérios de seleção:

- 1- As sequências seriam de vertebrados;
- 2- O tamanho não deveria exceder 35% do tamanho da sequência referente à proteína humana.

Utilizando estes critérios, foram encontradas duzentas e doze sequências de Lisozima para diferentes espécies. A partir das Lisozimas encontradas, buscou-se sequências para as outras três proteínas para os mesmos organismos. Foram encontradas quarenta e uma espécies diferentes com as quatro proteínas em comum

3.3. Escalas de Hidrofobicidade

Sabe-se que a hidrofobicidade é um tipo de interação de grande influencia no enovelamento das proteínas, mostrando a necessidade de quantificar as cadeias laterais dos aminoácidos por índices que mostram o nível de hidrofobicidade de cada um. Desta forma, muitos pesquisadores se envolveram em criar maneiras de classificar os aminoácidos, surgindo, então, escalas de hidrofobicidade.

Existem mais de cinquenta tipos de escalas de hidrofobicidade e podem ser encontrados no endereço eletrônico <http://www.expasy.ch/tools/protscale.html>. Essa variedade existe, pois a hidrofobicidade é uma grandeza de difícil quantificação, surgindo maneiras distintas de criar os parâmetros de classificação e suas respectivas escalas. Essas escalas são estabelecidas por métodos experimentais, que serão descritos a seguir.

3.3.1. Método do Particionamento

O método do particionamento é o mais utilizado para a determinação da hidrofobicidade de aminoácidos [44]. Consiste na utilização de solventes orgânicos que imitam o interior de uma proteína. A primeira escala foi proposta para nove aminoácidos, desenvolvida por Tanford e Nozaki [43] e utilizando etanol e dioxano.

Além de utilizar solventes em fase líquida, pode-se utilizar fases de vapor e micelares. Wolfenden et al. desenvolveu o método de utilização de fases de vapor, medindo a afinidade da cadeia lateral do aminoácido pela água [27]. Kyte & Doolittle [25] utilizaram um programa computacional para calcular a hidrofobicidade e hidrofobicidade de uma proteína, utilizando uma escala de hidropatia baseada em trabalhos encontrados na literatura, como os trabalhos de Tanford e Nozaki [43],

Wolfenden [27] e Chothia [29]. Roseman [28] construiu uma escala se baseando na partição de solutos entre água e solventes apolares.

3.3.2. Método da área acessível ao solvente

Este método utiliza modelagem molecular a fim de estimar a hidrofobicidade de cada aminoácido baseando-se no grau de sua exposição ao solvente [44]. Um programa possui a função de rodar uma molécula de solvente ao longo da proteína em questão e, assim, determinar a acessibilidade estática dos aminoácidos individuais à molécula do solvente. Lee e Richards desenvolveram a primeira escala utilizando este método [44]. Chothia comparou os resultados obtidos por Lee e Richards e os resultados de Tanford e Nozaki. Utilizando o mesmo método desenvolvido por Lee e Richards, Chothia calculou a área acessível para resíduos pertencentes a 12 proteínas.

3.3.3. Métodos cromatográficos

O método cromatográfico mais utilizado para o cálculo da hidrofobicidade de um soluto é a cromatografia de fase reversa (*Reverse Phase Liquid Chromatography – RPLC*). Isso ocorre, pois a fase estacionária apolar imita uma membrana biológica. Grande parte dos métodos utilizam peptídeos como solutos e realizam análise de retenção para relacionar a hidrofobicidade de um aminoácido à retenção do peptídeo [44].

3.3.4. Mutagênese de sítio dirigido

É um método bioquímico que utiliza a técnica do DNA recombinante, realizando substituições nas sequências de aminoácidos de proteínas que ocorrem naturalmente. Ao utilizar este método, é possível medir a estabilidade de uma proteína a partir da inserção de um aminoácido específico. Este método possui desvantagens como o alto custo, perturbações de outras interações e foi considerado como um método útil apenas para medir a estabilidade da proteína⁴⁵.

3.3.5. Método das Propriedades Físicas

Este método utiliza o cálculo de propriedades físicas como tensão superficial, temperaturas de transição e energia de solvatação, para estimar a hidrofobicidade dos aminoácidos [44].

Uma das escalas mais utilizadas, a escala Consensus, desenvolvida por Eisenberg et al [24]., foi desenvolvido um algoritmo que identifica as α - hélices envolvidas na interação das membranas das proteínas com bicamadas lipídicas e que as distingue de hélices em proteínas solúveis.

Segue abaixo uma tabela contendo os valores de hidrofobicidade, não normalizados, indicados por essa escala para cada aminoácido.

Tabela 1-Escala consensus de hidrofobicidade para cada resíduo

Aminoácido	Hidrofobicidade	Aminoácido	Hidrofobicidade
Ala	0,806	Leu	0,918
Arg	0	Lys	0,264
Asn	0,448	Met	0,81
Asp	0,417	Phe	0,951
Cys	0,721	Pro	0,677
Gln	0,43	Ser	0,601
Glu	0,458	Thr	0,634
Gly	0,77	Trp	0,854
His	0,545	Tyr	0,714
Ile	1	Val	0,923

Para uma comparação mais efetiva, foram escolhidas seis escalas de hidrofobicidade diferentes. Cada uma é construída a partir de procedimentos experimentais distintos. As escalas escolhidas foram:

- 1- Eisenberg et al [24] – Método das Propriedades Físicas;
- 2- Chothia [29] – Método da Área Acessível ao Solvente;
- 3- Kyte & Doolittle [25] – Método do Particionamento;
- 4- Miyazawa et al [26] – Método das Propriedades Físicas;
- 5- Roseman [28] – Método do Particionamento;
- 6- Wolfenden et al [27] – Método do Particionamento;

3.4. Divisão em Grupos

Após encontrar as espécies com as proteínas estudadas em comum, foi feita uma divisão desses grupos de acordo com a evolução, a fim de evitar uma comparação em larga escala. Proteínas de animais com proximidade filogenética, como, por exemplo, o humano e o chimpanzé, possuem alta similaridade sequencial. Emanuel Margoliash [21] e Emil Smith [20] mostraram, utilizando cerca de cem sequências de Citocromo C,

que a dissimilaridade entre as sequências é comparável a distancia evolutiva. Portanto, foi decidido dividir os animais em três grupos com diferenças evolutivas amplamente conhecidas e difundidas pela literatura, são eles os Primatas, as Aves e os Peixes. As informações relacionadas à evolução dos vertebrados e sua associação com trabalho se encontra no Apêndice B.

Dentre as espécies separadas de acordo com a disponibilidade das sequências para as quatro proteínas, os grupos escolhidos possuíam o maior número de espécies para o estudo, obtendo um total de quatro animais por grupo. A tabela a seguir mostra as espécies utilizadas neste trabalho e o código UNIPROT correspondente.

Tabela 2- Espécies analisadas e o código UNIPROT

Grupos	Organismo\Proteína	Histona	Citocromo C	Mioglobina	Lisozima
Primatas	Humano	P84243	P99999	P02144	P61626
	Chimpanze	H2QQ69	P99998	P02145	P61628
	Orangutango	Q5RCC9	Q5RFH4	P02148	P79239
	Gorila	G3RUH9	Q6WUX8	P02147	P79179
Aves	Galinha	P84229	P67881	P02197	P00698
	Mandarim	B5FZ57	B5FZX0	H0ZKN4	B5KFT9
	Peru	G1N7D3	P67882	G1NJB6	P00703
	Pato	P06902	P00020	Q7LZM2	P00705
Peixes	Truta	C1BEZ2	C1BFD3	D0FZZ1	P11941
	Salmão	B5DG71	B1A645	B9ENY2	B5XA65
	Baiacu	Q4S9G0	Q4SG99	Q701N9	P61944
	Peixe-Zebra	Q6PI20	Q6IQM2	Q6VN46	Q90YS5

3.5. Análise

Com a separação das espécies em grupos, foram realizadas as análises a fim de se obter a hidrofobicidade de cada sequência e a dissimilaridade.

3.5.1. Alinhamentos e o cálculo da dissimilaridade

O método mais utilizado para comparação de sequências de proteínas é através do alinhamento múltiplo de sequências (MSA), este se faz muito importante para a bioinformática.

O MSA é o alinhamento entre três sequências biológicas ou mais. É amplamente utilizado para detectar regiões de variabilidade e conservação em

sequências ou famílias proteicas, reconstrução filogenética, desenho de primers para PCR, entre outros. Dentre a variedade de métodos de MSA, o utilizado neste trabalho é o método do alinhamento progressivo, representado pela ferramenta ClustalW [22]. O ClustalW é uma versão recente do Clustal, em que o W representa *weight*, que está relacionado ao programa por utilizar pesos em seus parâmetros e nas sequências.

O método do alinhamento progressivo é dividido em três etapas, sendo que cada etapa gera informação para a próxima dar continuidade. O primeiro passo consiste em alinhar, por pares, as sequências de entrada, criando uma matriz de distâncias. Desta forma, o segundo passo consiste em criar uma árvore guia a partir da matriz obtida anteriormente. Essa árvore pode ser vista com as sequências representando os ramos e os alinhamentos representando os nós, de tal forma que o alinhamento completo represente sua raiz. O último passo é, a partir do primeiro nó, alinhar os nós filhos, de forma que todos os nós sejam adicionados ao alinhamento ao decorrer dos passos, até que todas as sequências sejam alinhadas.

A partir dessas definições, foi utilizado o ClustalW para a realização dos alinhamentos das sequências em estudo. Como matriz de substituição, foi utilizada a matriz BLOSUM 45. Matrizes BLOSUM (**B**locks of **A**mino Acid **S**ubstitution **M**atrix) são matrizes de substituição que pontuam alinhamentos entre sequências evolutivamente divergentes. Realizado o alinhamento com as sequências das espécies pertencentes aos três grupos, foi realizada uma análise de similaridade entre os aminoácidos de cada sequência, a partir do alinhamento, através do *Protein Variability Server* [23]. O resultado é obtido em porcentagem e, para a obtenção da dissimilaridade, neste trabalho foi admitida a relação:

$$\text{DISSIM}(i,j) = 1 - \text{SIM}(i,j) \quad (1)$$

Portanto, a dissimilaridade, $\text{DISSIM}(i,j)$, é obtida através da similaridade, $\text{SIM}(i,j)$. Este valor percentual foi convertido para valores entre 0 e 1.

3.5.2. Cálculo do RMSD (Root Mean Square Deviation)

Além do estudo sequencial, foi realizada uma análise estrutural. O cálculo do RMSD entre pares de estruturas de diferentes espécies foi realizado e comparado com a hidrofobicidade média do par. Para se calcular o RMSD é necessário obter a estrutura da proteína e, tais dados são encontrados no banco de dados PDB [30] (Protein Data Bank).

Ao obter as estruturas, utiliza-se um algoritmo chamado MUSTANG [31] (Multiple Structure Alignment Algorithm) para o cálculo do RMSD.

Tendo em vista que, de forma geral, as estruturas permanecem mais conservadas do que as sequências, foram realizados os cálculos do RMSD entre pares de grupos, cada elemento de um grupo com cada elemento de outro grupo, da forma: Primatas – Aves, Primatas – Peixes, Aves – Peixes. Após o cálculo por pares, foram tirados os valores médios e assim, as hidrofobicidades médias, para as seis escalas escolhidas, foram atribuídas. A tabela mostra as estruturas PDB utilizadas.

Tabela 3- Espécies analisadas e o código PDB

Grupos	Organismo\Proteína	Histona	Citocromo C	Mioglobina	Lisozima
Primatas	Humano	3AV2	1J3S	3RGK	133L
	Chimpanze	1KX5	2B4Z	1MBS	2NWD
	Orangutango	1KX5	2B4Z	1MBS	1JSF
	Gorila	1KX5	2B4Z	1MBS	1JSF
Aves	Galinha	2HIO	Cito_1	1MBS	193L
	Mandarim	1KX5	Cito_2	1H0Z	3LYO
	Peru	1KX5	Cito_3	1G1N	135L
	Pato	1KX5	1I54	1MBS	1GHL
Peixes	Truta	1KX5	Cito_4	1D0F	1BB6
	Salmão	1KX5	1CCR	1MYT	3LYO
	Baiacu	1KX5	Cito_5	1MYT	1GHL
	Peixe-Zebra	1KX5	1CCR	1MYT	2EQL

Algumas das estruturas ainda não foram resolvidas, portanto os códigos PDBs referentes ao Citocromo C com nomenclatura diferenciada (Cito_n), foram estruturas modeladas a partir da estrutura de uma espécie do próprio grupo que apresentava similaridade alta com a sequência da espécie que necessitava de uma estrutura.

4. RESULTADOS

Nesta seção serão apresentados os resultados obtidos neste trabalho. A primeira parte consiste em uma análise sequencial, em que são feitas as comparações entre a dissimilaridade e a hidrofobicidade. A segunda parte consiste em uma análise estrutural, mostrando a dissimilaridade através do RMSD e comparando, novamente, com a hidrofobicidade.

4.1. Análise Sequencial

A partir do processo descrito anteriormente, calculou-se a dissimilaridade média entre grupos e a hidrofobicidade média. A média, neste caso, é realizada por pares. As figuras a seguir mostram o gráfico da Hidrofobicidade pela Dissimilaridade para cada combinação entre grupos, utilizando a escala de Hidrofobicidade de Eisenberg. Os Gráficos referentes às outras escalas se encontram no apêndice A.

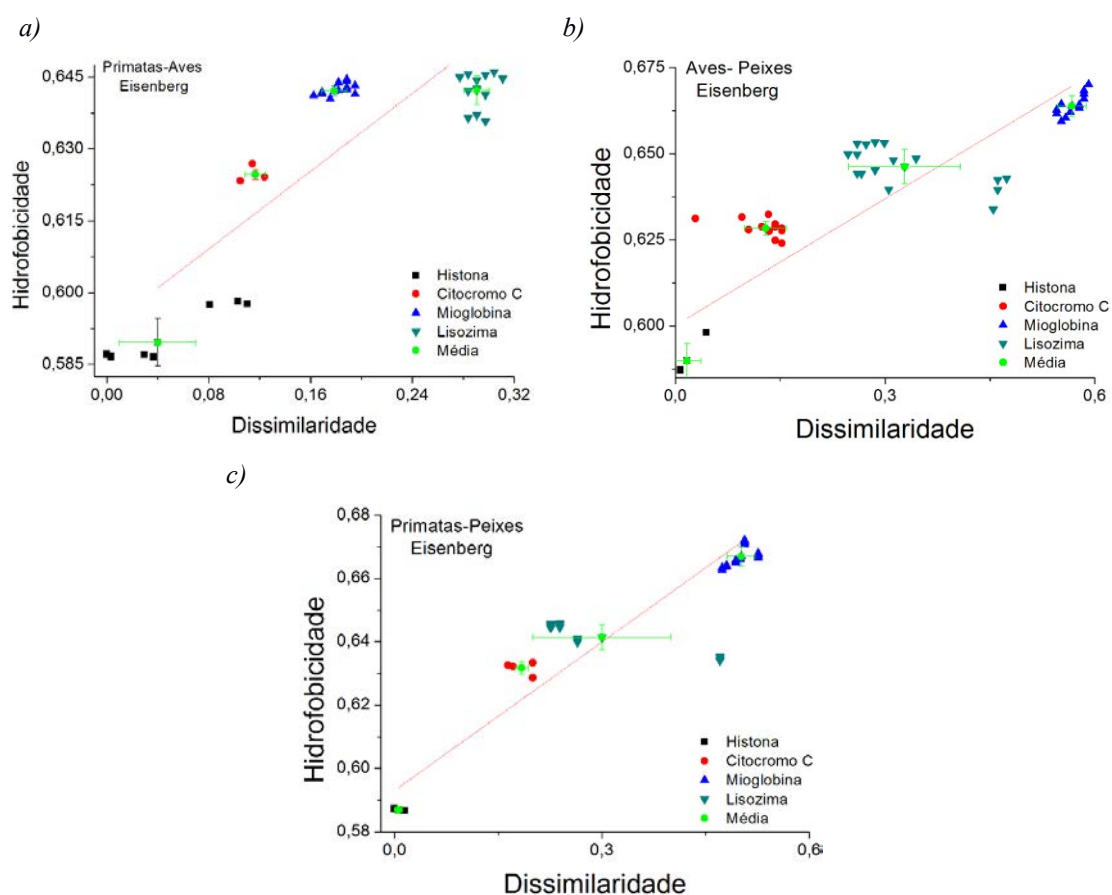


Figura 10-Relação entre Hidrofobicidade e Dissimilaridade. a) Grupos: Primatas e Aves. b) Grupos: Aves e Peixes. c) Grupos: Primatas e Peixes

Nos três gráficos, é possível notar uma relação linearmente crescente, de forma que, à medida que a hidrofobicidade média da proteína cresce a dissimilaridade entre as sequências também cresce. Cada ponto no gráfico refere-se a um par de espécies entre os grupos em questão e cada cor representa uma das proteínas estudadas. Pode-se notar que as proteínas se encontram clusterizadas, ou seja, é possível diferenciar regiões pertencentes a cada uma delas. A histona, por exemplo, se mostrou como uma proteína

de baixa hidrofobicidade e baixa dissimilaridade. A figura abaixo mostra o coeficiente angular (m) dos gráficos da Hidrofobicidade pela Dissimilaridade para cada um dos grupos e para cada escala. O eixo das coordenadas representa o coeficiente angular de cada um dos gráficos e o eixo das abscissas representa cada escala. Cada escala está representada por um número de acordo com a ordem mostrada no item 3.3 deste trabalho. É possível notar que os coeficientes são positivos para as duas escalas, indicando uma relação linearmente crescente. Embora seja possível notar a dependência linear entre essas grandezas e a clara clusterização das proteínas, sugerindo que a dissimilaridade sequencial aumenta com a hidrofobicidade global, é importante enfatizar que o coeficiente angular pode não ser muito preciso. Desta maneira, foram obtidos os valores de correlação entre as variáveis pelo método de correlação de Pearson [40].

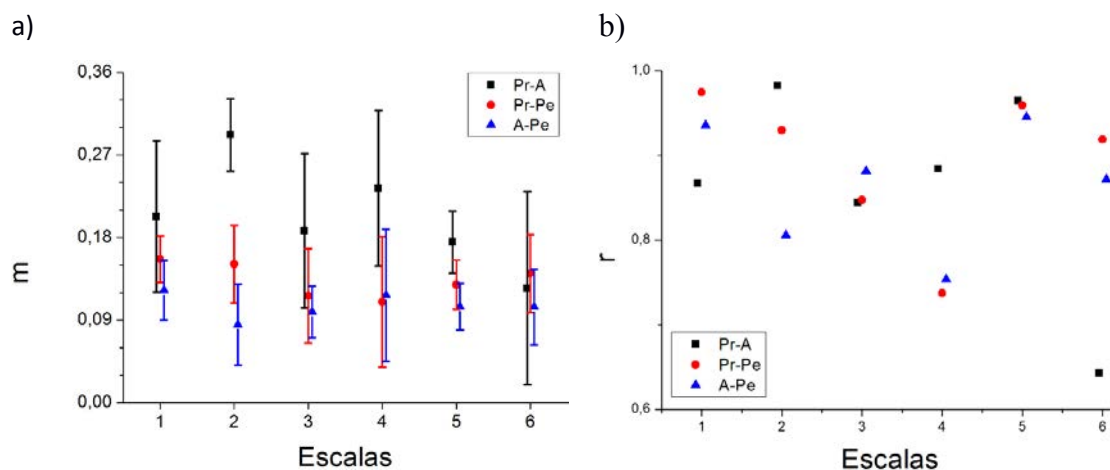


Figura 11- a) Coeficiente angular dos gráficos Hidrofobicidade x Dissimilaridade para cada escala de Hidrofobicidade. Pr – Primatas; Pe – Peixes; A- Aves. b) Coeficiente de correlação entre as variáveis Hidrofobicidade e Dissimilaridade

Todos os coeficientes de correlação obtidos se encontram em uma região acima de 0,5 e a maioria se encontra acima de 0,8. Coeficientes de correlação iguais a 0,5 sugerem uma correlação fraca e acima de 0,8 sugerem uma correlação muito forte. Assim, é possível notar que as variáveis em questão, de forma geral, apresentam uma correlação muito forte.

4.2. Análise Estrutural

Foi realizada uma análise análoga à anterior, porém utilizando as estruturas e os métodos mencionados na seção 3.4.2.. Calculou-se o RMSD das estruturas entre uma espécie de um grupo e uma espécie de outro grupo, portanto foi uma comparação intergrupos. Assim, foi realizada a média dos valores obtidos e, em seguida, atribuiu-se o valor médio das hidrofobicidades. A figura a seguir mostra o gráfico da Hidrofobicidade pelo RMSD referente à escala do Eisenberg. Os gráficos para as outras escalas se encontram no Apêndice A.

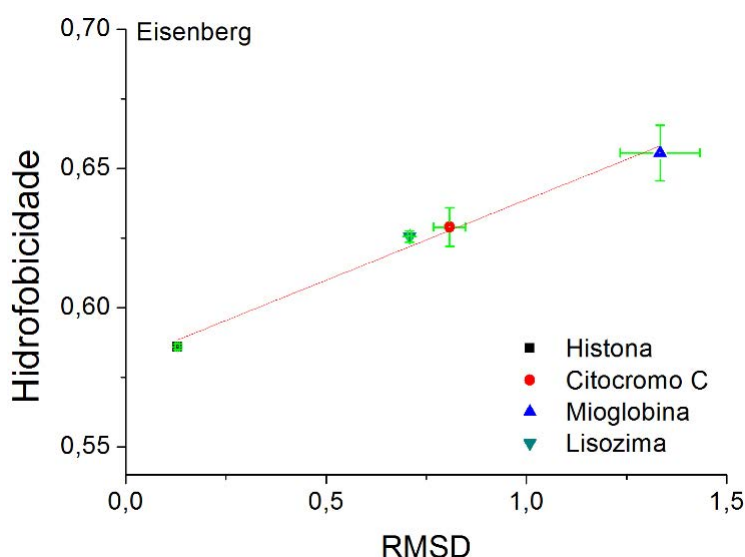


Figura 12- Relação entre Hidrofobicidade e RMSD em Angstrom.

Nesta etapa, não foram analisados os pares separadamente, como feito na análise sequencial, mas foi realizada uma média utilizando todos os grupos. Isso foi feito, pois a estrutura varia muito menos do que a sequência ao longo da evolução, portanto o efeito não é verificado ao se analisar separadamente.

Assim como na análise sequencial, podemos verificar uma dependência linearmente crescente entre as variáveis em questão. Para mostrar isso, também foi obtido o gráfico do coeficiente linear para cada uma das escalas de hidrofobicidade. A partir desse gráfico, é possível verificar que todos os coeficientes são positivos para todas as escalas, enfatizando a dependência linearmente crescente entre essas duas variáveis. Isso significa que as diferenças estruturais também aumentam de acordo com o aumento da hidrofobicidade. Mesmo que a estrutura seja forçada a permanecer praticamente inalterada, devido às restrições impostas pelos contatos entre os

aminoácidos específicos, as variações sequenciais alteram sutilmente a estrutura. Para complementar a análise, foi realizado o cálculo do coeficiente de correlação entre as variáveis.

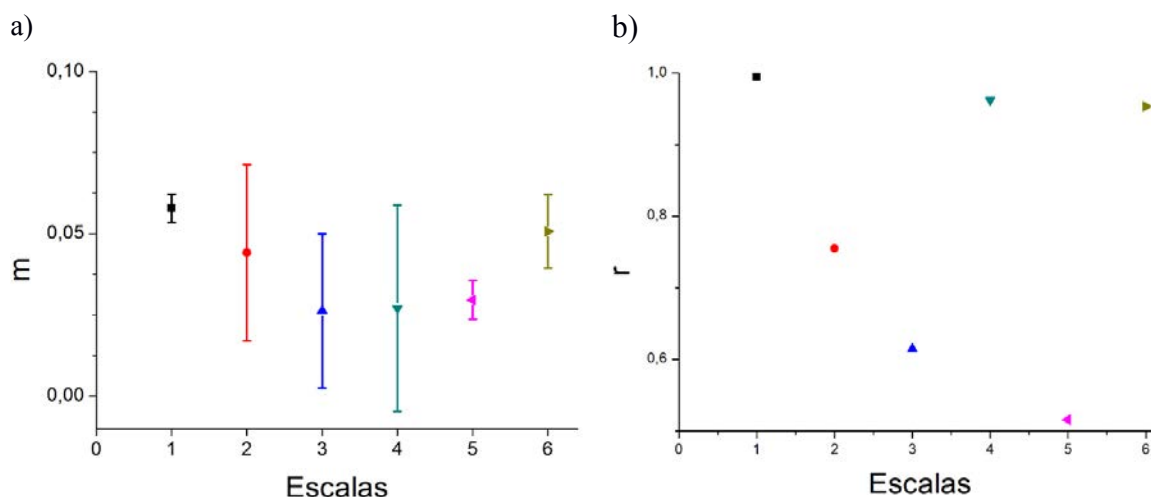


Figura 13- a) Coeficiente angular dos gráficos Hidrofobicidade x Dissimilaridade para cada escala de Hidrofobicidade. b) Coeficiente de correlação entre as variáveis Hidrofobicidade e RMSD(em Angstrom) para todas as escalas de hidrofobicidade.

Nas figuras acima, as escalas de hidrofobicidade estão representadas por números inteiros e a ordem consta no item 3.3 do trabalho. A partir dos resultados obtidos é possível concluir o trabalho de acordo com o esperado pela literatura utilizada.

5. CONCLUSÃO

Este trabalho tinha por objetivo analisar a relação entre a hidrofobicidade global de uma proteína e a variabilidade sequencial ao longo da evolução. Os gráficos obtidos nas análises sequenciais nos mostra a clusterização das proteínas, de forma que os pontos referentes a cada uma delas se encontram em uma região bem definida. Além deste comportamento, foram obtidos coeficientes angulares positivos para todos os gráficos, mostrando uma dependência crescente entre a hidrofobicidade e a dissimilaridade. Para comprovar a relação entre as duas variáveis, o cálculo da correlação foi realizado. Como mostrado nos resultados, todos os coeficientes de correlação se encontram acima de 0,5. No caso da análise sequencial, os coeficientes se encontram em uma região considerada como indicativo de correlação forte e muito

forte. Para a análise estrutural, foi possível observar comportamentos similares aos resultados da análise sequencial.

Os resultados obtidos estão de acordo com a literatura, constatando que a variabilidade sequencial possui uma correlação com a hidrofobicidade, exatamente como concluído no trabalho utilizando proteínas otimizadas [1]. À medida que se aumenta a hidrofobicidade global da proteína, a variabilidade de suas sequências aumenta ao longo da evolução. Com os resultados obtidos nesse trabalho e nos trabalhos utilizados como referência, pode-se dizer que as proteínas que apresentam um colapso no enovelamento, as proteínas de alta hidrofobicidade, são mais robustas às mutações, mantendo sua função e estabilidade. Por outro lado, as proteínas de baixa hidrofobicidade, mostram-se mais sensíveis às mutações, alterando seu enovelamento e comprometendo sua função.

PARTE II

Após concluir os objetivos iniciais, foram propostas novas metas. A primeira parte consistia em analisar a proteína de uma forma global, realizando a média das variáveis de interesse. A nova proposta exige a análise de cada aminoácido.

6. OBJETIVOS ESPECÍFICOS

O objetivo desta parte do trabalho é analisar de forma pontual a possível relação entre a co-evolução, a hidrofobicidade dos resíduos e uma grandeza denominada valores Φ . Essa grandeza quantifica a relevância de um aminoácido no estado de transição do enovelamento da proteína em questão. Para tal propósito, serão utilizados cálculos de acoplamento direto (*Direct Coupling Analysis*). As proteínas estudadas serão mesmas da primeira parte, que são Histona, Citocromo C, Mioglobina e Lisozima.

7. METODOLOGIA

7.1. *Direct Coupling Analysis*

É notável que as estruturas proteicas pouco variam ao longo da evolução, ao contrário de sua sequência. Assim, estudos [32-34] sugeriram que existem padrões de substituição entre resíduos de uma proteína e que estes restringem a variabilidade da estrutura. Essa restrição, conhecida como co-evolução entre resíduos, é descrita da seguinte maneira: Suponha um par de resíduos de uma proteína que formam um contato. Caso ocorra uma mutação em um dos resíduos, essa mutação será compensada no decorrer do tempo com uma substituição no outro resíduo, de forma a manter a interação [37].

Com o grande avanço em sequenciamento genético, hoje dispomos de dados para que possa ser feita uma análise evolutiva utilizando sequências de famílias de proteínas. Para analisar os resíduos que apresentam co-evolução, foram desenvolvidos diferentes métodos estatísticos para estudar correlação, porém algumas dificuldades foram encontradas. O maior obstáculo foi encontrar uma maneira que desvincule as correlações diretas e indiretas, ou seja, as análises iniciais obtiveram sinais de diferentes tipos de correlação. O objetivo era desenvolver um método que pudesse desvincular esses sinais e hoje temos o que é chamado de acoplamentos estatísticos diretos entre resíduos. Foi desenvolvida a Análise de Acoplamento Direto (*Direct Coupling*

Analysis), que se mostrou uma forma eficiente de predição de contatos nativos, sendo utilizada para predição estrutural.

7.2. Cálculo da Informação Direta

Inicialmente é necessário obter o alinhamento das sequências da família da proteína a ser estudada. O alinhamento, utilizando modelo oculto de Markov (*Hidden Markov Model*), é obtido através do banco de dados PFAM [36]. O acoplamento direto é calculado através do alinhamento pelo algoritmo descrito no trabalho de Morcos et al [37]. O algoritmo para Matlab pode ser obtido através do endereço eletrônico <http://dca.rice.edu/portal/dca>, assim como também está disponível para ser calculado online neste mesmo endereço.

Os métodos anteriores utilizados para se calcular a correlação entre dois aminoácidos utilizavam algoritmos de passagem de mensagens (*Message-Passing Algorithm*), porém ele demandava um alto custo computacional, não sendo possível utilizar sequências muito longas. Além disso, também eram feitos cálculos da Informação Mútua, que mede a probabilidade de duas variáveis aleatórias possuírem dependência. O cálculo da Informação Mútua não desvincula efeitos secundários dos cálculos, portanto, ao se calcular a relação entre dois aminoácidos que realizam um contato, o resultado é influenciado por outros aminoácidos que interagem com eles. O novo algoritmo visa desvincular efeitos diretos e indiretos, implementando um método de aproximação de campo médio, a fim de calcular a informação direta.

Para realizar o cálculo, é feita uma contagem de frequências $f_i(A)$ de um aminoácido A em uma das colunas do alinhamento e uma contagem da frequência do par $f_{ij}(A, B)$ do aminoácido A e B aparecerem simultaneamente na mesma posição nas colunas i e j. As frequências são definidas como:

$$f_{ij}(A, B) = \frac{1}{M} \sum_{a=1}^M \delta_{A, A_i^a} \delta_{B, A_j^a} \quad f_i(A) = \frac{1}{M} \sum_{a=1}^M \delta_{A, A_i^a} \quad (2)$$

Temos que $i, j = 1 \dots L$, e L é o comprimento da proteína, M é o número de proteínas no alinhamento e δ é o delta de Kronecker.

É necessário fazer uma repesagem nas frequências, pois sequências biológicas apresentam vieses filogenéticos. Portanto, as frequências são definidas como:

$$f_i(A) = \frac{1}{\lambda + M_{eff}} \left(\frac{\lambda}{q} + \sum_{a=1}^M \frac{1}{m^a} \delta_{A, A_i^a} \right) \quad (3)$$

$$f_{ij}(A, B) = \frac{1}{\lambda + M_{eff}} \left(\frac{\lambda}{q^2} + \sum_{a=1}^M \frac{1}{m^a} \delta_{A, A_i^a} \delta_{B, A_j^a} \right) \quad (4)$$

Nas equações acima, temos um pseudo-contador λ , que estima probabilidades a partir de contagens em sequências biológicas [38]. M_{eff} é definido como $\sum_{a=1}^M 1/m^a$, $a = 1, \dots, M$, e significa o número efetivo de sequências independentes.

Para o cálculo da Informação Direta (DI), é definida a probabilidade:

$$P_{ij}^{(dir)}(A, B) = \frac{1}{Z_{ij}} \exp \left\{ e_{ij}(A, B) + \tilde{h}_i(A) + \tilde{h}_j(B) \right\} \quad (5)$$

Em que Z_{ij} é a função de partição, e_{ij} é a estimativa de acoplamentos por par, $h_i(A)$ e $h_j(B)$ são multiplicadores de Lagrange que possuem a interpretação de campos locais. Portanto, para se calcular DI:

$$DI_{ij} = \sum_{A, B=1}^q P_{ij}^{(dir)}(A, B) \ln \frac{P_{ij}^{(dir)}(A, B)}{f_i(A) f_j(B)} \quad (6)$$

Após o cálculo dos valores de Informação Direta, é necessário fazer a validação a partir de uma estrutura. A finalidade dessa validação é verificar quais são contatos verdadeiros da proteína. Assim, a partir de uma estrutura retirada do banco PDB é feito o mapeamento.

Após o mapeamento, é possível verificar quais são os contatos do PDB utilizado que possuem valores de Informação Direta. A figura a seguir mostra a utilização do método.

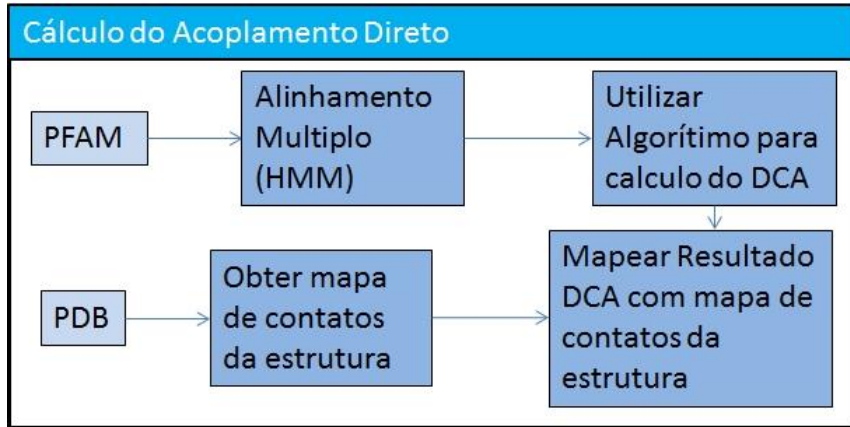


Figura 14 - Esquema do processo de utilização do método da Análise do Acoplamento Direto.

A figura acima descreve o método utilizado. O primeiro passo, o da inferência, consiste em buscar o alinhamento das sequências pertencentes à família da proteína, calcular o acoplamento direto através do algoritmo. Em seguida, na validação é feito o mapeamento utilizando uma estrutura obtida no banco de dados PDB, obtendo os contatos que possuem valores de acoplamento direto para uma estrutura específica. O método foi aplicado em todas as proteínas do estudo.

7.3. Valores Φ

Os valores Φ medem fornecem uma quantificação do grau de natividade de cada resíduo de uma cadeia no estado de transição.

O cálculo é feito medindo os efeitos na taxa de enovelamento que uma mutação em um ponto na cadeia da proteína provoca [35]. Os valores Φ são definidos da seguinte forma:

$$\Phi \equiv \frac{-RT \ln \frac{k_{mut}}{k_{wt}}}{\Delta\Delta G^0} \quad (7)$$

Tem-se que k_{mut} é a taxa de enovelamento da proteína com uma mutação inserida em sua cadeia e k_{wt} é a taxa de enovelamento da proteína sem mutações, K é a constante dos gases, T a temperatura absoluta e $\Delta\Delta G^0$ é a variação da energia entre a proteína mutante e nativa.

Utilizando a equação de Kramer

$$k = k_0 \exp(-\Delta G^\ddagger/RT) \quad (8)$$

onde k_0 depende da forma da barreira do sistema, podemos escrever os valores Φ como:

$$\Phi = \frac{\Delta\Delta G^\ddagger}{\Delta\Delta G^0} \quad (9)$$

e

$$\Delta\Delta G^\ddagger = \Delta G_{\text{mut}}^\ddagger - \Delta G_{\text{wt}}^\ddagger = -RT \ln k_{\text{mut}}/k_{\text{wt}} \quad (10)$$

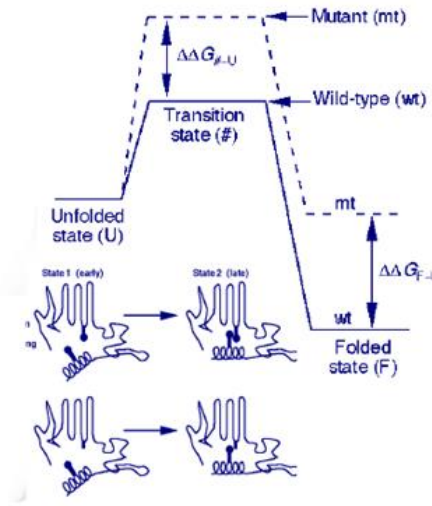


Figura 15- Variação da barreira de energia ao inserir uma mutação [39].

Portanto, os valores Φ podem ser vistos como a razão entre as variações de energia livre da barreira de enovelamento. Pode-se entender os valores Φ como uma taxa de variação de energia livre do conjunto de estados de transição sobre a mudança do conjunto do estado nativo. Se a barreira for de apenas alguns $K_B T$ e o deslocamento de sua posição ao longo de uma coordenada de reação sob mutação for pequena, é possível calcular os valores Φ através do método de perturbação na energia livre [35]:

$$\Phi = \frac{\Delta\Delta G^{\text{TS}} - \Delta\Delta G^{\text{U}}}{\Delta\Delta G^{\text{F}} - \Delta\Delta G^{\text{U}}} \quad (11)$$

Para realizar os cálculos dos valores Φ computacionalmente, é utilizada a seguinte aproximação:

$$\Phi = \frac{\Delta\Delta G^{TS-U}}{\Delta\Delta G^{N-U}} \approx \frac{P_{ij}^{TS} - P_{ij}^U}{P_{ij}^N - P_{ij}^U} \quad (12)$$

A quantificação é feita a partir da probabilidade de formação de contatos, onde P_{ij}^X é a probabilidade de formação do contato entre os resíduos i e j no estado X , que pode ser TS (Estado de transição), U (Desenovelado) e N (Nativo).

Os valores Φ variam entre 0 e 1, tais valores dizem respeito a participação do resíduo no processo de transição. Um valor próximo de 1 significa que a variação de energia livre entre a proteína mutante e nativa é praticamente a mesma no estado de transição e no estado nativo, ou seja, aquele contato envolvendo o resíduo mutado já estava formado no estado de transição. Desta forma, temos um resíduo com alta influencia no enovelamento. Por outro lado, valores próximos de 0 indicam que a variação de energia livre é a mesma no estado de transição de no estado desenovelado.

7.3.1. Cálculo dos valores Φ

Para calcular os valores Φ de uma determinada proteína, inicialmente são realizadas simulações de Dinâmica Molecular. Escolher um modelo para simular as proteínas pode ser complexo, pois descrever todas as interações de um sistema de forma completa e realística envolve complicações computacionais. É possível utilizar modelos minimalistas, que possui quantidade razoável de detalhes sem perder as informações necessárias para o estudo e poupando tempo computacional.

Neste trabalho, é utilizado um modelo minimalista baseado em estrutura, onde temos uma simplificação nos componentes do sistema e no potencial de interação entre eles [46-51]. O modelo C- α aproxima cada aminoácido por apenas um átomo, seus respectivos carbonos α , mantendo seu raio de volume de exclusão. Outra aproximação característica do modelo é o potencial de interação entre os carbonos α . Os parâmetros do potencial do modelo baseado em estrutura são obtidos a partir da estrutura nativa da proteína, tendo como mínimo de energia global correspondente à conformação nativa da proteína contida no Protein Data Bank (PDB). O potencial para uma conformação Γ , a

partir da conformação nativa Γ_0 é dado a partir da equação a seguir:

$$\begin{aligned}
V(\Gamma, \Gamma_0) = & \sum_{bonds} \epsilon_r (r - r_0)^2 \\
& + \sum_{angles} \epsilon_\theta (\theta - \theta_0)^2 \\
& + \sum_{backbone} \epsilon_\phi \left\{ [1 - \cos(\Phi - \Phi_0)] + \frac{1}{2} [1 - \cos(3(\Phi - \Phi_0))] \right\} \\
& + \sum_{contacts} \epsilon_c \left[5 \left(\frac{d_{ij}}{r_{ij}} \right)^{12} - 6 \left(\frac{d_{ij}}{r_{ij}} \right)^{10} \right] + \sum_{non-contacts} \epsilon_{NC} \left(\frac{\sigma_{NC}}{r_{ij}} \right)^{12}
\end{aligned} \tag{13}$$

O primeiro termo da equação nos dá informação a respeito da ligação entre dois carbonos α adjacentes, em que r_0 representa a distância entre dois carbonos α na estrutura nativa que estão ligados entre si. O segundo termo nos dá um potencial caracterizado pela interação entre três átomos consecutivos, em que θ_0 é o ângulo entre eles. Da mesma forma, Φ_0 é o ângulo diedral entre quatro carbonos consecutivos, fazendo com que o terceiro termo quantifique a torção na cadeia. No termo de contatos d_{ij} é obtido a partir do mapa de contatos da estrutura nativa e o termo é um potencial tipo Lennard-Jones com potencias 12-10 ao invés de 6-12. O ultimo termo caracteriza um potencial repulsivo a fim de estabelecer a distância máxima de aproximação entre os carbonos, de forma que σ_{NC} estabelece o volume aproximado que um carbono ocupa.

Após realizar a simulação da proteína, utiliza-se o algoritmo WHAM (weighted histogram analysis method) [52], que possibilita obter grandezas termodinâmicas importantes como o calor específico e a energia livre. A partir do gráfico do calor específico em função da temperatura, é possível determinar a temperatura de folding, que se encontra no ponto de máximo. Desta forma é possível obter as informações necessárias do estado de transição das proteínas e, por fim, calcular os valores Φ .

8. RESULTADOS

O cálculo do acoplamento direto foi realizado e, em seguida, feito o mapeamento da estrutura das proteínas utilizando o arquivo PDB. Desta forma, foi possível obter os contatos que possuem valores de acoplamento direto e, aqueles que não foram mapeados tiveram um valor nulo atribuído. Segundo a proposta do trabalho, o próximo passo é tentar observar uma correlação entre as três variáveis em questão: Acoplamento

Direto, Hidrofobicidade e Valores Φ .

Para completar o objetivo, foram obtidos os valores da hidrofobicidade de cada aminoácido pertencente às cadeias das proteínas, seguindo o mesmo procedimento da primeira etapa do trabalho. Nesta etapa, foi utilizada uma única escala, a princípio, que é a escala Consensus (Eisenberg et al [24]). Ao obter o valor correspondente de cada aminoácido, o valor da hidrofobicidade dos contatos é calculado utilizando a média da hidrofobicidade dos aminoácidos envolvidos. Assim, temos os valores das três grandezas para os contatos que apresentam acoplamento direto.

Inicialmente, foi realizado o gráfico 3D das três grandezas. A figura a seguir apresenta o gráfico da Lisozima.

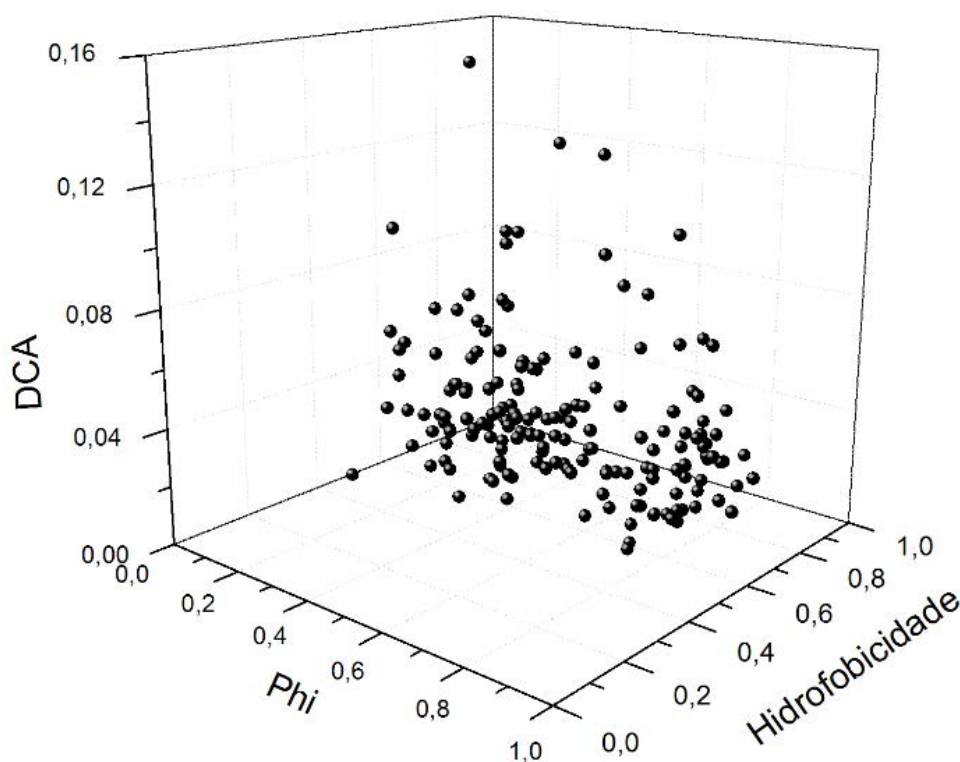


Figura 16 - Gráfico Acoplamento Direto (DCA) x Phi x Hidrofobicidade.

O gráfico foi realizado para todas as proteínas, evidenciando um padrão similar em todos eles. Tal gráfico se mostra inconclusivo, apresentando uma grande quantidade de pontos em uma região de baixo DCA e alta hidrofobicidade. A média dos valores Φ se encontra em 0,498 e do DCA em 0,0336. Para melhorar a visibilidade dos dados, foram realizados os gráficos separadamente entre Hidrofobicidade x DCA e Valores Φ x DCA. Seguem os gráficos obtidos para a lisozima.

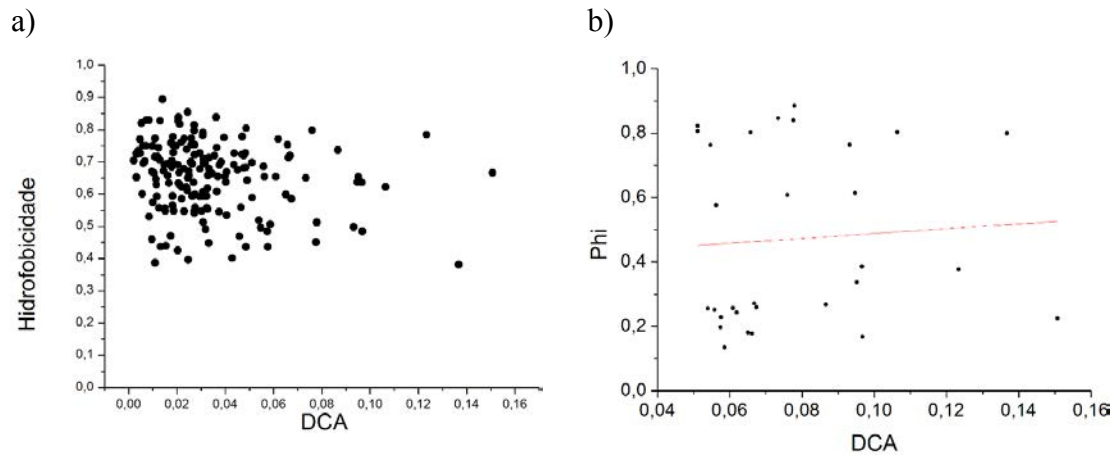


Figura 17 - Gráficos referentes à Proteína Lisozima. a) Hidrofobicidade x DCA. b) Valores phi x DCA.

Sabe-se que a lisozima é uma proteína de alta hidrofobicidade e a histona uma proteína de baixa hidrofobicidade. Para efeito de comparação, a fim de visualizar os dois extremos de hidrofobicidade, seguem os gráficos para a histona.

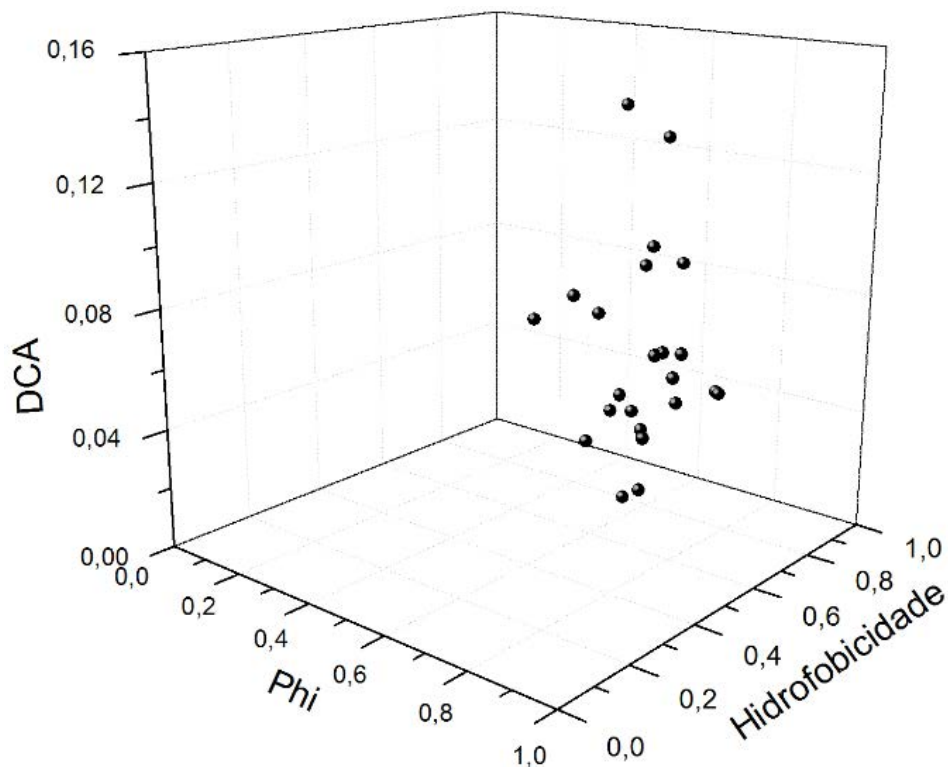


Figura 18 - Gráfico Acoplamento Direto (DCA) x Phi x Hidrofobicidade

Para a histona, foi encontrada uma quantidade menor de contatos com valores de Acoplamento direto. Como mostra o gráfico, a quantidade de pontos é muito menor e a

região em que estão concentrados possui uma hidrofobicidade média e valores Φ que se alternam entre 0,4 e aproximadamente 0,8. A seguir são apresentados os gráficos Hidrofobicidade X DCA e Hidrofobicidade X Valores Φ da Histona.

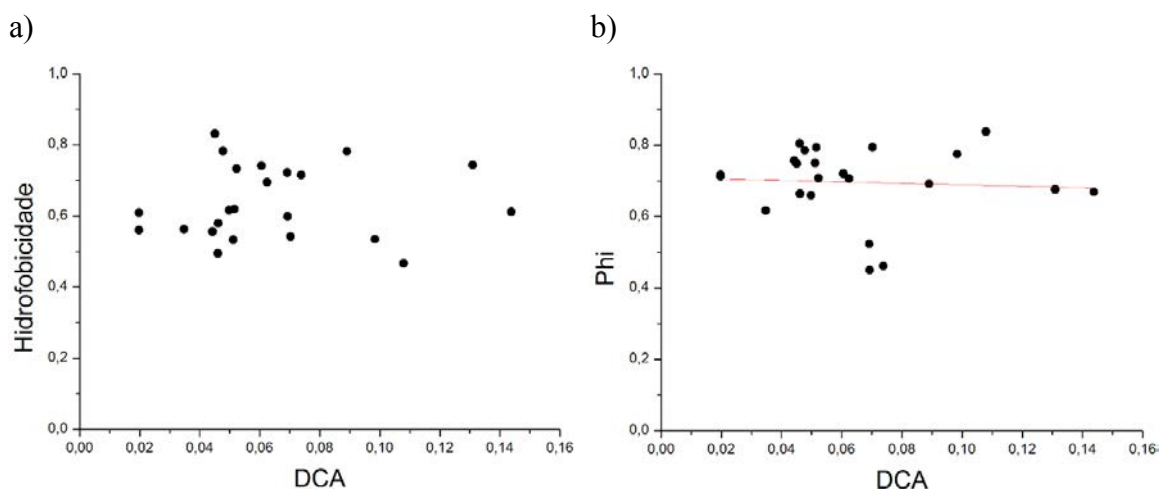


Figura 19 - Gráficos referentes à Proteína Histona. a) Hidrofobicidade x DCA. b) Valores Φ x DCA.

A média dos valores Φ da Histona se encontra em 0,697 e do DCA 0,064. Aparentemente, esses resultados sugerem que as proteínas de baixa hidrofobicidade possuem altos valores Φ e um DCA médio alto, enquanto as proteínas de alta hidrofobicidade apresentam um padrão oposto. Não é possível concluir isso com os dados obtidos, pois as proteínas Mioglobina e Citocromo C apresentaram um padrão diferente do esperado. A média dos valores Φ da Mioglobina, proteína de alta hidrofobicidade, se encontra em 0,527 e seu DCA em 0,039. Para o Citocromo C, proteína de baixa hidrofobicidade, os valores Φ estão em 0,387 e o DCA em 0,037. Os gráficos para as duas proteínas se encontram no Apêndice A.

Além desses gráficos, foi calculado o coeficiente de correlação entre os gráficos de Hidrofobicidade x DCA e Φ x DCA para cada proteína. Os valores obtidos não mostraram existir alguma dependência entre as variáveis em questão.

9. CONCLUSÃO

A primeira parte do trabalho mostrou haver uma dependência entre a hidrofobicidade e a dissimilaridade entre as sequências de proteínas, mostrando que proteínas com baixa hidrofobicidade média se mantiveram mais conservadas ao longo da evolução. A partir desses resultados, foi decidido investigar as proteínas de forma pontual, observando seus aminoácidos. O valor Φ revela a importância de um

aminoácido ou contato no estado de transição do enovelamento e o valor de acoplamento direto fornece a informação da co-evolução entre dois aminoácidos que realizam contato. Portanto, os valores Φ , valores de acoplamento direto e a hidrofobicidade de um contato poderia fornecer alguma informação, podendo haver alguma dependência entre essas variáveis.

Ao analisar as três variáveis juntas, como mostra o gráfico DCA contra Hidrofobicidade contra Valor Φ , não é possível notar alguma dependência clara. Assim, foi feita uma análise entre duas variáveis, como os gráficos Hidrofobicidade contra DCA e Hidrofobicidade contra Valores Φ , que mostraram pontos onde não é possível identificar um padrão.

Os resultados obtidos até então não forneceram informações o suficiente para concluir alguma dependência entre as variáveis. Os dados ainda podem ser explorados, e as análises podem continuar utilizando mais estruturas para realizar os cálculos de DCA e valores Φ , a fim de fazer uma comparação com um maior número de dados.

10. PERSPECTIVAS

Foram levantadas as hipóteses a respeito do possível comportamento da associação entre essas três variáveis. Até o momento, não foram obtidas correlações entre as três variáveis, fazendo com que haja a necessidade de analisar o caso e, eventualmente, conferir os dados e planejar outra metodologia.

A primeira análise levou em consideração apenas uma das escalas de hidrofobicidade e utilizou-se apenas uma estrutura para o cálculo do acoplamento direto. A fim de continuar o trabalho, é possível ampliar o número de estruturas a serem analisadas e fazer a mesma análise para todas as escalas de hidrofobicidade já utilizadas anteriormente. Além dos métodos utilizados, é possível aplicar métodos de clusterização. Ao realizar uma análise com todos os dados juntos, pode haver regiões clusterizadas, que permitem definir características das três variáveis e estudar sua possível relação.

Ao ampliar as análises e revisar os resultados é possível que se estabeleça uma conclusão para este assunto. Havendo um não uma correlação entre as variáveis, o atual estudo fornecerá uma fonte de referência para estudos posteriores, auxiliando o desenvolvimento desta área do conhecimento.

11. REFERÊNCIAS

- 1- Oliveira, L C; Silva, R T H; Leite, V B P; Chahine, J. - “Frustration and Hydrophobicity Interplay in Protein Folding and Protein Evolution”. The Journal of Chemical Physics 125, no 8.
- 2 - Nelson, D L.; Lehninger, CM. – “Princípios de Bioquímica”. 3ed. São Paulo: Sarvier, 2002.
- 3 – Leopold, PE; Montal, M e Onuchic, JN. – “Protein folding funnels: a kinetic approach to the sequence-structure relationship”. Proc Natl Acad Sci U S A, 1992.
- 4 - Anfinsen, CB - “Principles that govern the folding of proteins chains.” Science, 181:223–230, 1973.
- 5 – Levinthal, C. – “Are there pathways for protein folding?” J. Chem. Phys., 65:44–45, 1968.
- 6 - Onuchic JN, Nymeyer H, Garcia AE, Chahine J, Socci ND. – “The energy landscape theory of protein folding: insights into folding mechanisms and scenarios”. Department of Physics, University of California at San Diego, La Jolla 92093-0319, USA, Adv Protein Chem. 53:87-152, 2000.
- 7 - Onuchic JN, Luthey-Schulten Z, Wolynes PG. – “Theory of protein folding: the energy landscape perspective”. Annu Rev Phys Chem.48:545-600, 1997.
- 8 – Leopold, PE; Montal, M.; Onuchic, JN. – “Protein folding funnels - A kinetic approach to the sequence structure relationship”. Proc. Natl. Acad. Sci. USA, 18:8721–8725, 1992.
- 9 - Ferreira, DU; Hegler, JA; Komives, EA e Wolynes, PG . - “Localizing frustration in native proteins and protein assemblies”, Proceedings of the National Academy of Sciences, vol. 104, no. 50, p. 19819 -19824, 2007.

- 10 - Goldstein, RA; Luthey-Schulten, ZA e Wolynes, PG. - “Optimal protein-folding codes from spin-glass theory”, *Proceedings of the National Academy of Sciences*, vol. 89, no. 11, p. 4918 -4922, 1992.
- 11 - Bloom JD; Silberg JJ; Wilke CO; Drummond DA; Adami C e Arnold FH. - “Thermodynamic prediction of protein neutrality. *Proc*”. *Natl Acad. Sci. USA* 102, 606–611, 2005.
- 12 - Gong LI, Suchard MA e Bloom JD. – “Stability-mediated epistasis constrains the evolution of an influenza protein”. *eLife* 2, e00631, 2013.
- 13 – Kauzmann,W. – “Some factors in the interpretation of protein denaturation”. - *Advances in Protein Chemistry*, Volume 14, Pages 1–63, 1959.
- 14 - Pratt, LR. - “Molecular theory of hydrophobic effects: ‘She is too mean to have her name repeated.’*”. *Annual Review of Physical Chemistry* 53, no 1, 409–436, 2002.
- 15 – Huang, K. – “Lectures on Statistical Mechanics and Protein Folding”; World Scientific Publishing Co., Cingapura, pg.101.
- 16 – Maizel, AS. – “Papel da Hidrofobicidade na Variabilidade Sequencial de Proteínas”. Monografia (trabalho de conclusão de curso) – Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, São Paulo. 2011.
- 17 - Silva, RHT. – “Análise da Hidrofobicidade na evolução de proteínas”. Tese de Doutorado em Biofísica molecular - Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, São Paulo. 2009.
- 18 - The UniProt Consortium. “Activities at the Universal Protein Resource (UniProt)”. *Nucleic Acids Research* 42, 2013.

- 19 - Wu, Cathy H., Lai-Su L. Yeh, Hongzhan Huang, Leslie Arminski, Jorge Castro-Alvear, Yongxing Chen, Zhangzhi Hu, et al. "The Protein Information Resource". *Nucleic Acids Research* 31, no 1, 2013.
- 20 - Heller, J e Smith, EL. - "The amino acid sequence of cytochrome c of *Neurospora crassa*." *Proceedings of the National Academy of Sciences of the United States of America* 54, no 6, 1621–1625, 1965.
- 21 - Margoliash, E. - "Primary Structure and evolution of cytochrome c". *Proceedings of the National Academy of Sciences of the United States of America* 50, no 4, 672–679, 1963.
- 22 – McWilliam, H.; Li, W.; Uludag, M.; Squizzato, S.; Park, YM.; Buso, N.; Cowley, AP.; Lopez, R. - "Analysis Tool Web Services from the EMBL-EBI". *Nucleic Acids Research* 41, no Web Server issue: W597–600, 2013.
- 23 - Garcia-Boronat, M; Diez-Rivero, CM; Reinherz EL e Reche PA. – "PVS: a web server for protein sequence variability analysis tuned to facilitate conserved epitope discovery". *Nucleic Acids Res.*, 36:W35-41, 2008.
- 24 – Eisenberg, D; Weiss RM, Terwilliger TC. - The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature*. 1982.
- 25 - Kyte, J e Doolittle, RF. – "A simple method for displaying the hydropathic character of a protein". *J. Mol. Biol.* 157:105-132, 1982.
- 26 – Miyazawa, S e Jernigan, RL. - "Estimation of Effective Interresidue Contact Energies from Protein Crystal-Structures-Quasi-Chemical Approximation". *J Macromolecules* , 18:534-552, 1985.

- 27 –Wolfenden, RV; Andersson, L; Cullis, PM e Southgate CCF. – “Affinities of amino-acid side-chains for solvent water. *Biochemistry*”, 20(4): 849-855, 1981.
- 28 –Roseman, MA. – “Hydrophilicity of polar amino acid side-chains is markedly reduced by flanking peptide bonds”. *J. Mol. Biol.* 200:513-522, 1988.
- 29 - Chothia C. – “The nature of the accessible and buried surfaces in proteins”. *J. Mol. Biol.* 105:1-14, 1976.
- 30 - Berman, HM; Westbrook, J; Feng, Z; Gilliland, G; Bhat, TN; Weissig, H; Shindyalov, IN e Bourne, PE. – “The Protein Data Bank *Nucleic Acids Research*”. 28: 235-242, 2000.
- 31 - Konagurthu, Arun S, James C Whisstock, Peter J Stuckey, e Arthur M Lesk. - “MUSTANG: A Multiple Structural Alignment Algorithm”. *Proteins* 64, no 3, 559–574, 2006.
- 32 - Altschuh D, Lesk A, Bloomer A, Klug A . – “Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus”. *J Mol Biol* 193:693–707, 1987.
- 33 - Göbel U, Sander C, Schneider R, Valencia A. – “Correlated mutations and residue contacts in proteins”. *Proteins Struct Funct Genet* 18:309–317, 1994.
- 34 – Neher, E. – “How frequent are correlated changes in families of protein sequences?” *Proc Natl Acad Sci USA* 91:98–102, 1994.
- 35 - Clementi, C.; Nyemeyer H e Onuchic, J. N. – “Topological and energetic factors: What determines the structural details of the transition state ensemble and Em-route intermediates for protein folding? An investigation for Small Globular Proteins”. *Journal of Molecular Biology*, 298, 937-953, 2000.

- 36 - Finn, R.D. ; Bateman, A.; Clements, J.; Coghill, P.; Eberhardt, R.Y. ; Eddy, S.R.; Heger, A.; Hetherington, K.; Holm, L.; Mistry, J. ; Sonnhammer, E.L.L.; Tate, J. ; Punta, M. – “The Pfam protein families database”. *Nucleic Acids Research . Database Issue* 42:D222-D230, 2014.
- 37 - Morcos, F; Pagnani, A; Lunt, B; Bertolino, A; Marks, DS; Sander, C; Zecchina, R; Onuchic, J.N; Hwa, T e Weigt, M. - “Direct-coupling analysis of residue coevolution captures native contacts across many protein families,” *Proc Natl Acad Sci U S A*, vol. 108, no. 49, pp. E1293–301, 2011.
- 38 - Durbin, R; Eddy, S; Krogh, A e Mitchison, G. – “Biological sequence analysis: Probabilistic models of proteins and nucleic acids”. Cambridge Univ Pr, 1998.
- 39 - Fersht, A. R. – “Characterizing the transition states in protein folding: an essential step in the puzzle”. *Curr. Opin. Struct. Biol.* 5, 79-84, 1994.
- 40 - Wang, J – “Pearson Correlation Coefficient”. *Encyclopedia of Systems Biology*, 2013, Editora Springer.
- 41 - Morcos, F; Hwab, T; Onuchic, JN e Weigt, M – “Direct Coupling Analysis for protein contact prediction *Methods in Molecular Biology, Protein Structure Prediction 3rd Edition*”.
- 42 - Southall , NT; Dill, K A e Haymet , ADJ. - A View of the Hydrophobic Effect. *J. Phys. Chem. B*, 2002, 106 (3), pp 521–533.
- 43 – Nozaki, Y.; Tanford, C. – “The Solubility of Amino Acids and Two Glycine Peptides in Aqueous Ethanol and Dioxane Solutions. Establishment of a Hydrophobicity Scale” *J. Biol. Chem.* 246, 2211–2217, 1971.
- 44 – Biswas, K.M.; DeVido, D.R. e Dorsey, J.G. - Evaluation of methods for measuring amino acid hydrophobicities and interactions. *Journal of Chromatography A*, 2003.

- 45 - Taketomi, H.; Ueda, Y.; Go, N. Studies on protein folding, unfolding and fluctuations by computer simulation. 1. Effect of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Protein Res.* 1975, 7, 445-459.
- 46 - Go, N.; Scheraga, H.A. On the use of classical statistical mechanics in the treatment of polymer chain conformations. *Macromolecules* 1976, 9, 535-542.
- 47 - Ueda, Y.; Go, N. Theory of large-amplitude conformational fluctuations in native globular proteins: Independent fluctuating site model. *Int. J. Pept. Protein Res.* 1976, 8, 551-558.
- 48 - Ueda, Y.; Taketomi, H.; Go, N. Studies on protein folding, unfolding and fluctuations by computer simulation. 2. A three-dimensional lattice model of lysozyme. *Biopolymers* 1978, 17, 1531-1548.
- 49 - Go, N.; Taketomi, H. Studies on protein folding, unfolding and fluctuations by computer simulation. 3. Effect of short-range interactions. *Int. J. Pept. Protein Res.* 1979, 13, 235-252.
- 50 - Go, N.; Taketomi, H. Studies on protein folding, unfolding and fluctuations by computer simulation. 4. Hydrophobic interactions. *Int. J. Pept. Protein Res.* 1979, 13, 447-461.
- 51 - Kumar, S.; Bouzida, S.K.; Swendsen, R.H.; Kollman, P.A.; Rosenberg, J.M.: The Weighted Histogram Analysis Method for Free-Energy Calculations on biomolecules. I. The Method. *J. Comp. Chem.*, 13:1011–1021, 1992.
- 52 - Baillie, J.E.M., Hilton-Taylor, C. and Stuart, S.N. (Editors) 2004. 2004 IUCN Red List of Threatened Species. “A Global Species Assessment”. IUCN, Gland, Switzerland and Cambridge, UK.
- 53 – Pough, F.H.; Janis, C.M.; Heiser, J.B. “A vida dos vertebrados”. São Paulo: Atheneu. 3ª ed., 2003.
- 54 - Flicek, P., I. Ahmed, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, et al. “Ensembl 2013”. *Nucleic Acids Research* 41, nº D1 (30 de novembro de 2012): D48–D55. doi:10.1093/nar/gks1236.

APÊNDICE A

Este apêndice contém os resultados omitidos no texto

PARTE I

Resultados referentes à primeira parte do trabalho. Contem as análises sequenciais e estruturais.

A.1. Análise Sequencial

Escala Chothia

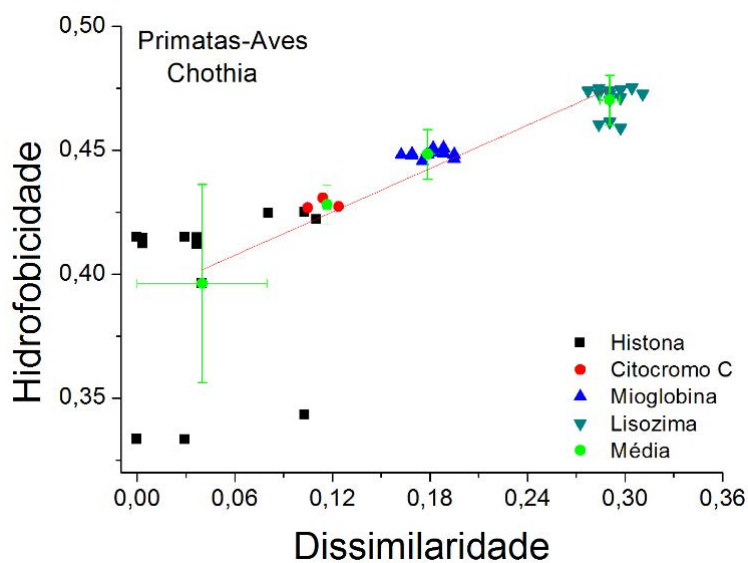


Figura 20- Relação entre Hidrofobicidade e Dissimilaridade para os grupos dos Primatas e das Aves. Escala Chothia.

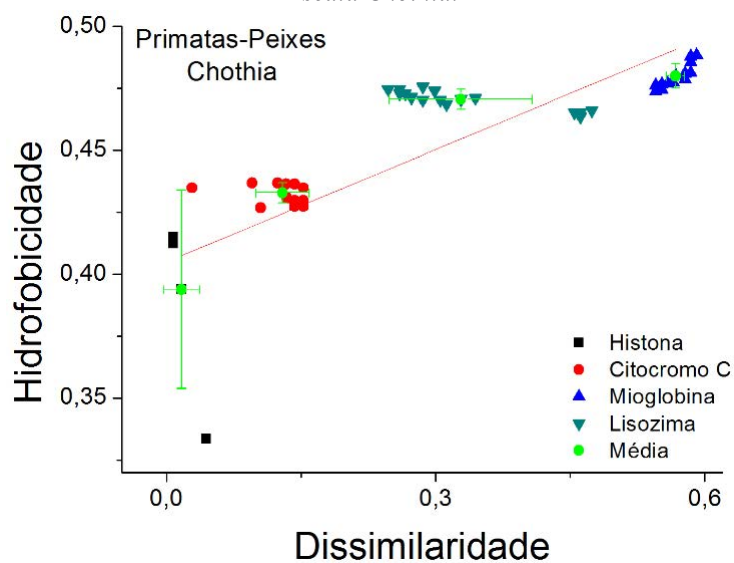


Figura 21- Relação entre Hidrofobicidade e Dissimilaridade para os grupos dos Primatas e dos Peixes. Escala Chothia.

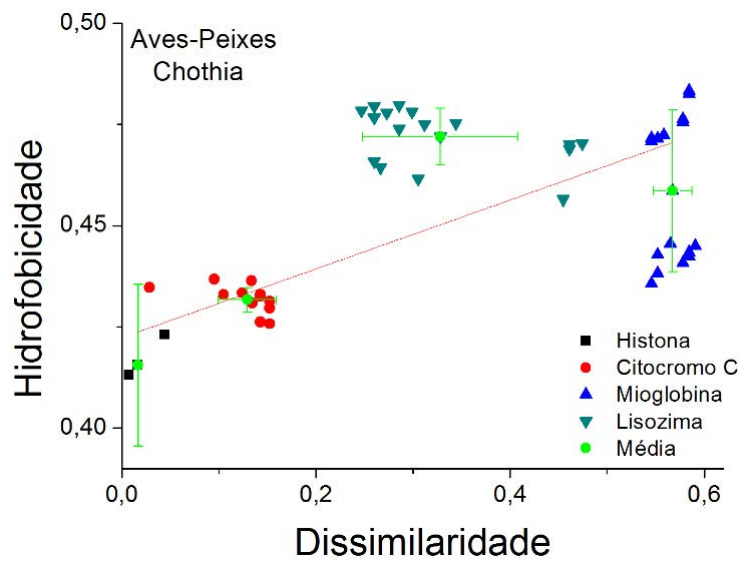


Figura 22- Relação entre Hidrofobicidade e Dissimilaridade para os grupos dos Primatas e dos Peixes. Escala Chothia.

Escala Wolfenden

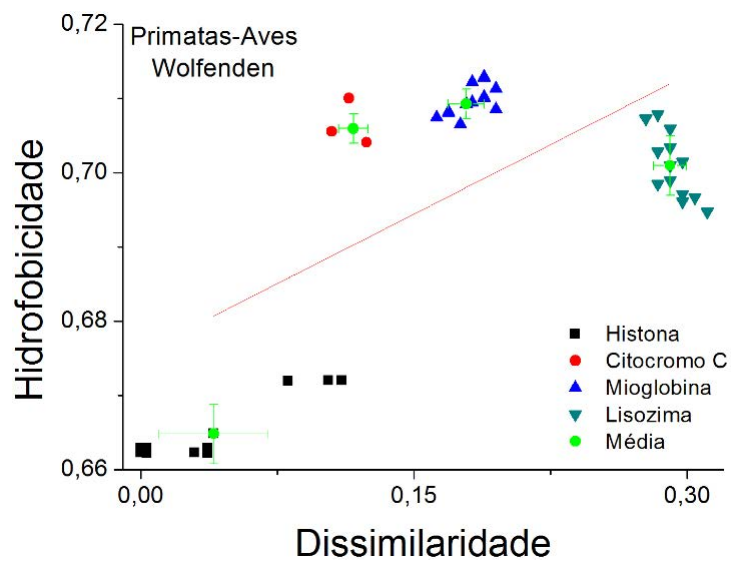


Figura 23- Relação entre Hidrofobicidade e Dissimilaridade para os grupos dos Primatas e das Aves. Escala Wolfenden.

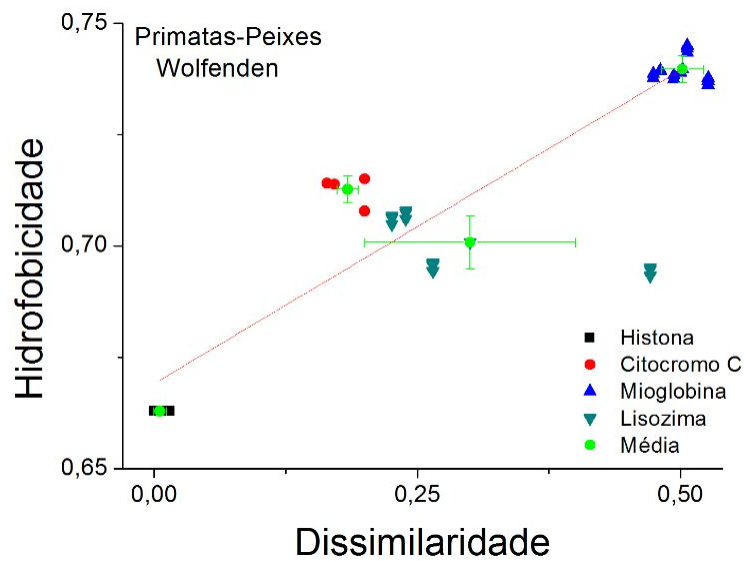


Figura 24 - Relação entre Hidrofobicidade e Dissimilaridade para os grupos dos Primatas e dos Peixes. Escala Wolfenden.

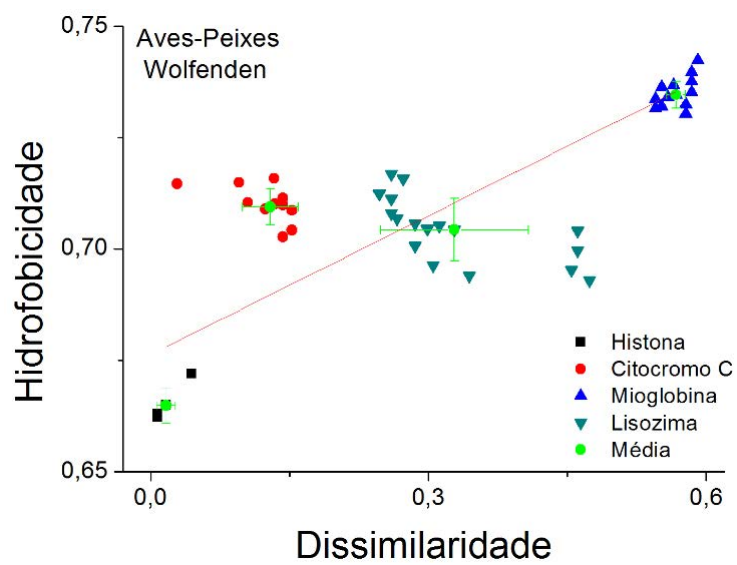


Figura 25- Relação entre Hidrofobicidade e Dissimilaridade para os grupos das Aves e dos Peixes. Escala Wolfenden.

Escala Miyazawa

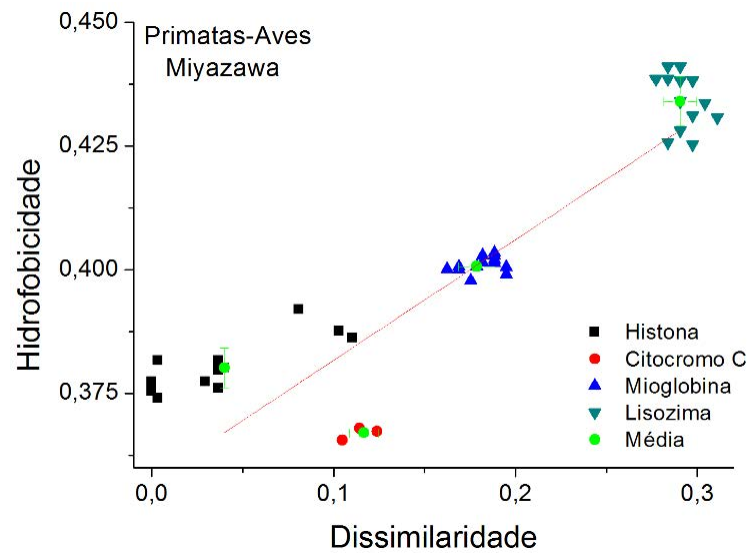


Figura 26- Relação entre Hidrofobicidade e Dissimilaridade para os grupos dos Primatas e das Aves. Escala Miyazawa.

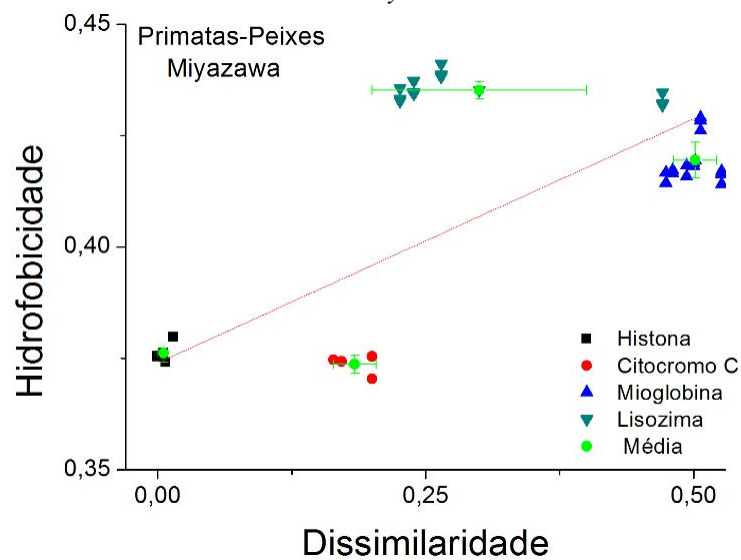


Figura 27- Relação entre Hidrofobicidade e Dissimilaridade para os grupos dos Primatas e dos Peixes. Escala Miyazawa.

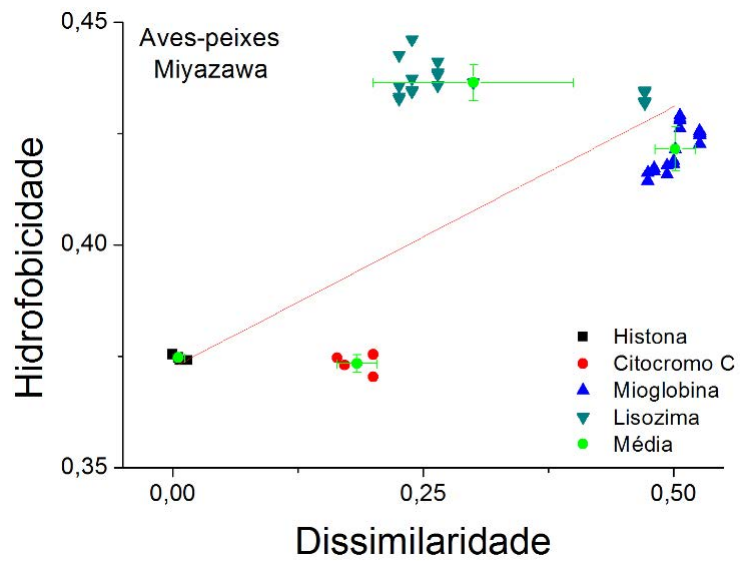


Figura 28 - Relação entre Hidrofobicidade e Dissimilaridade para os grupos das Aves e dos Peixes.
Escala Miyazawa.

Escala Roseman

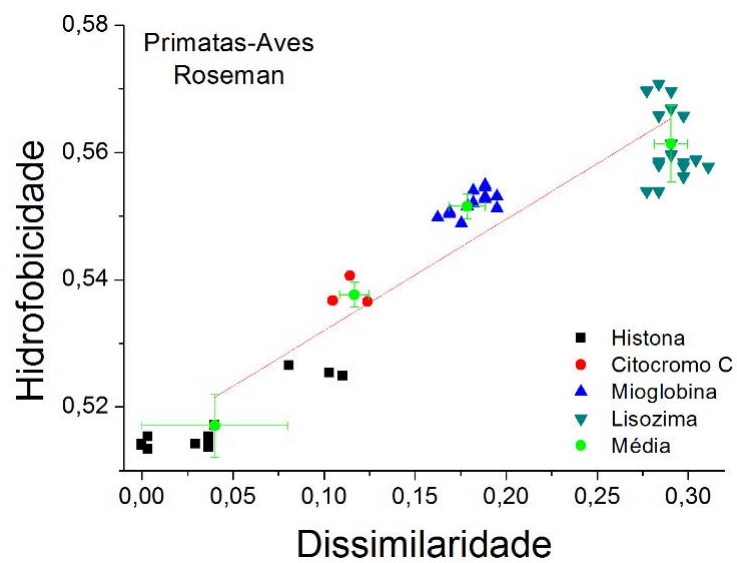


Figura 29- Relação entre Hidrofobicidade e Dissimilaridade para os grupos dos Primatas e das Aves.
Escala Roseman.

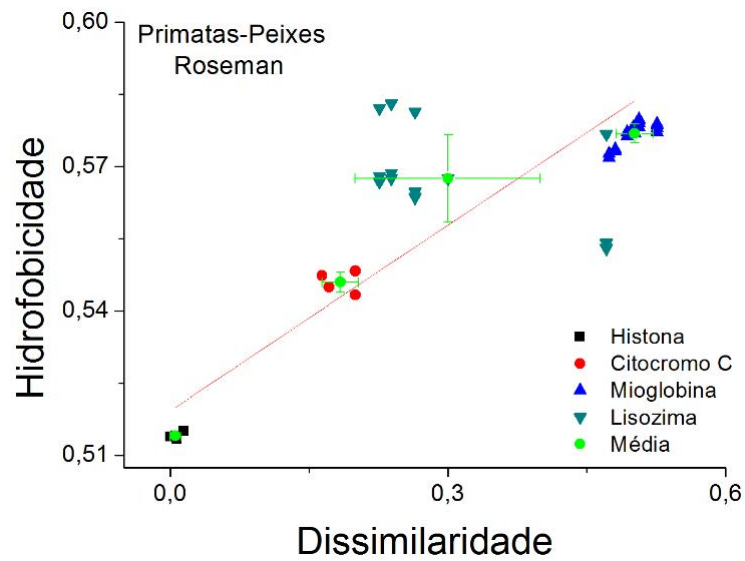


Figura 30 - Relação entre Hidrofobicidade e Dissimilaridade para os grupos dos Primatas e dos Peixes. Escala Roseman.

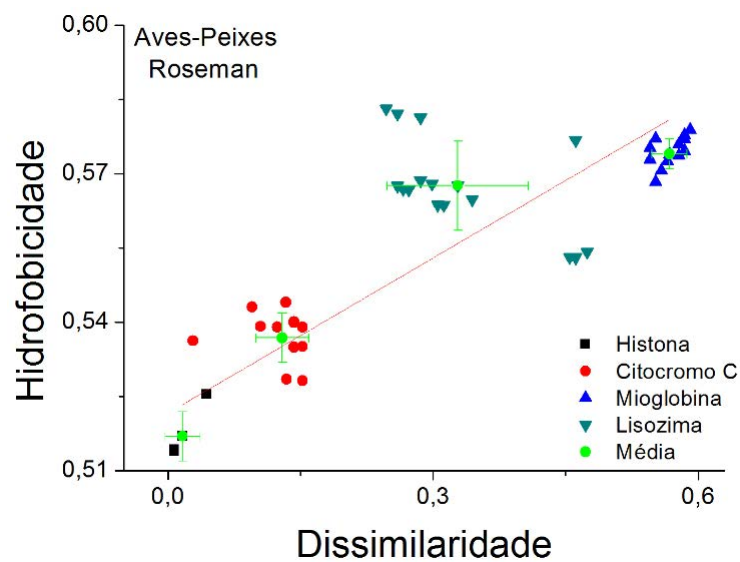
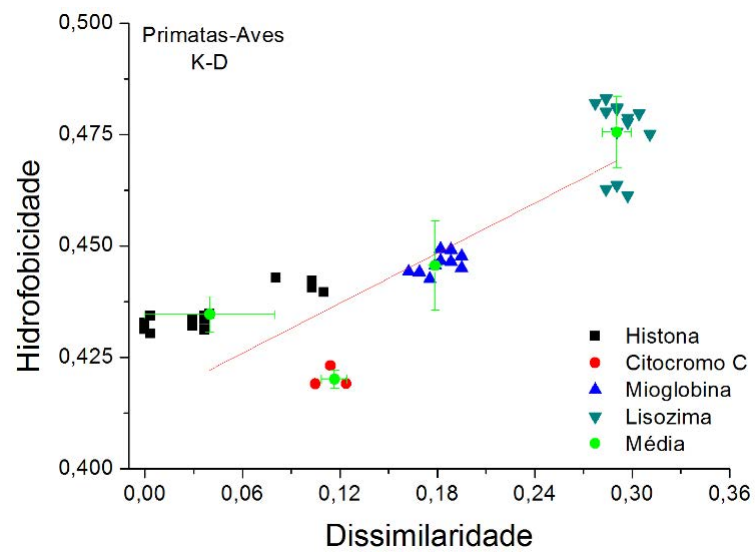
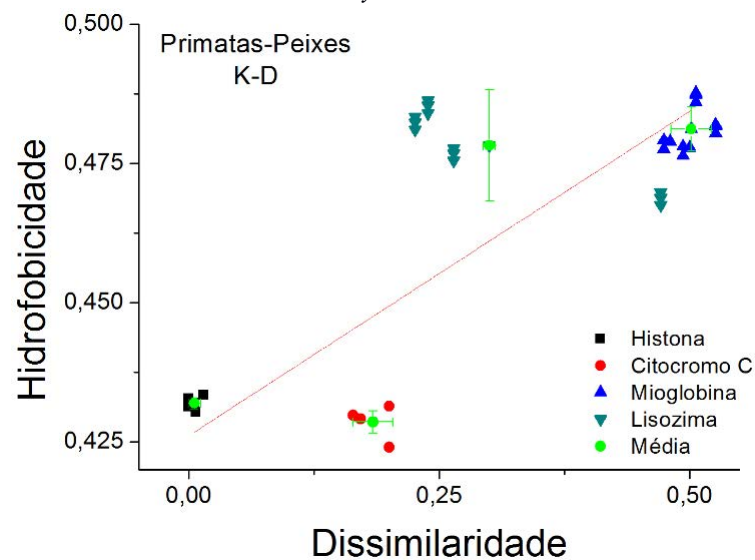


Figura 31- Relação entre Hidrofobicidade e Dissimilaridade para os grupos das Aves e dos Peixes. Escala Roseman.

Escala Kyte-doolittle



*Figura 32- Relação entre Hidrofobicidade e Dissimilaridade para os grupos dos Primatas e das Aves.
Escala Kyte-Doolittle.*



*Figura 33- Relação entre Hidrofobicidade e Dissimilaridade para os grupos dos Primatas e dos Peixes.
Escala Kyte-Doolittle.*

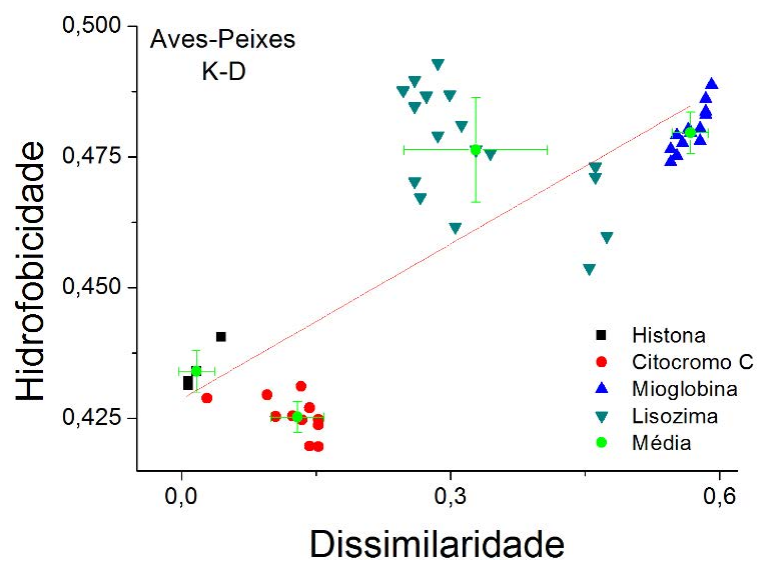


Figura 34 - Relação entre Hidrofobicidade e Dissimilaridade para os grupos das Aves e dos Peixes. Escala Kyte-Doolittle.

A.2. Análise Estrutural

Escala Cothia

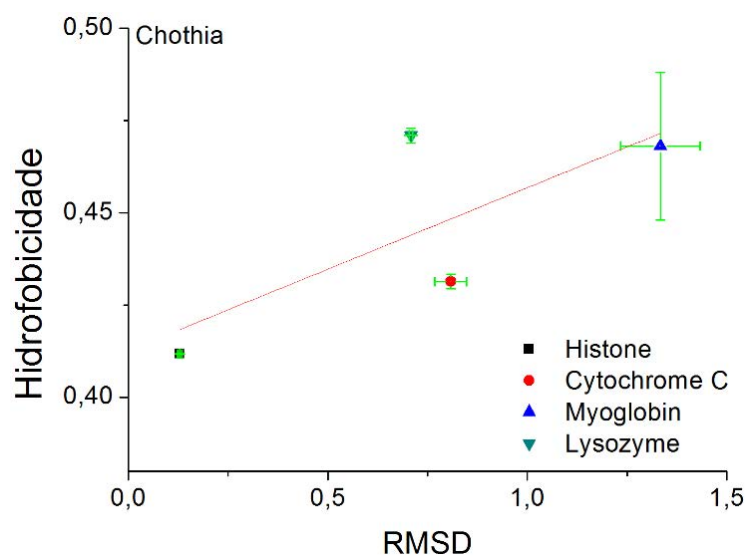


Figura 35- Relação entre Hidrofobicidade e RMSD para todos os grupos. Escala Chothia.

Escala Wolfenden

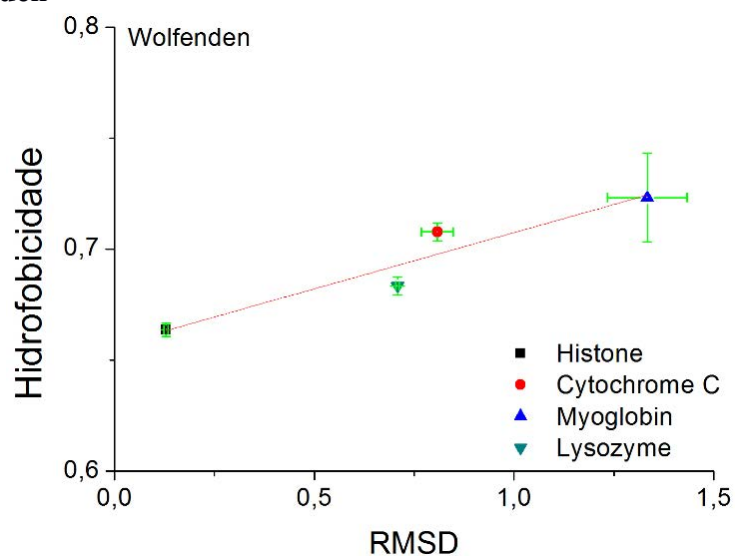


Figura 36- Relação entre Hidrofobicidade e RMSD para todos os grupos. Escala Wolfenden.

Escala Miyazawa

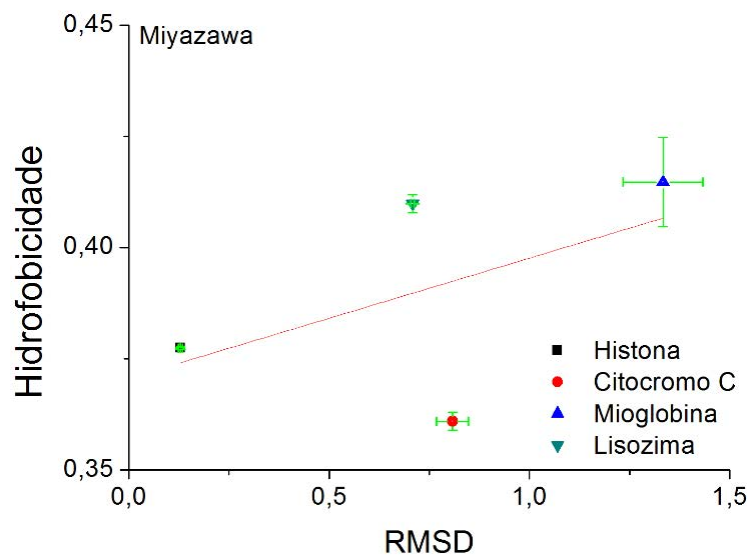


Figura 37- Relação entre Hidrofobicidade e RMSD para todos os grupos. Escala Miyazawa.

Escala Roseman

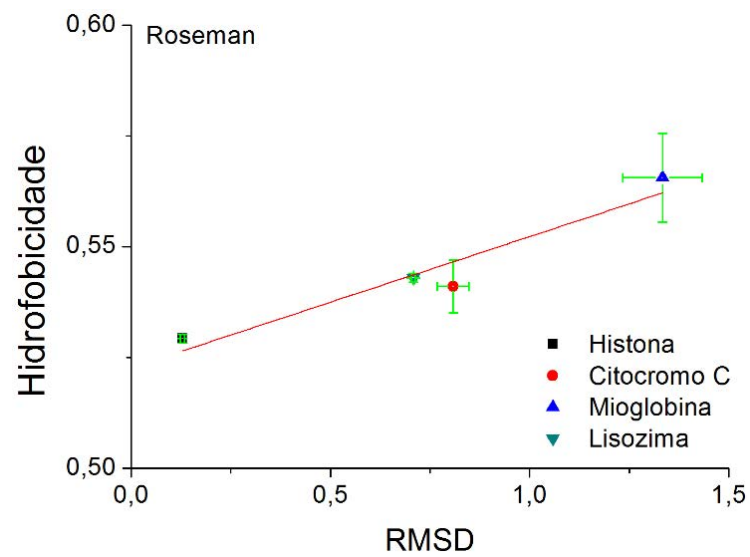


Figura 38 - Relação entre Hidrofobicidade e RMSD para todos os grupos. Escala Rose man.

Escala Kyte-Doolittle

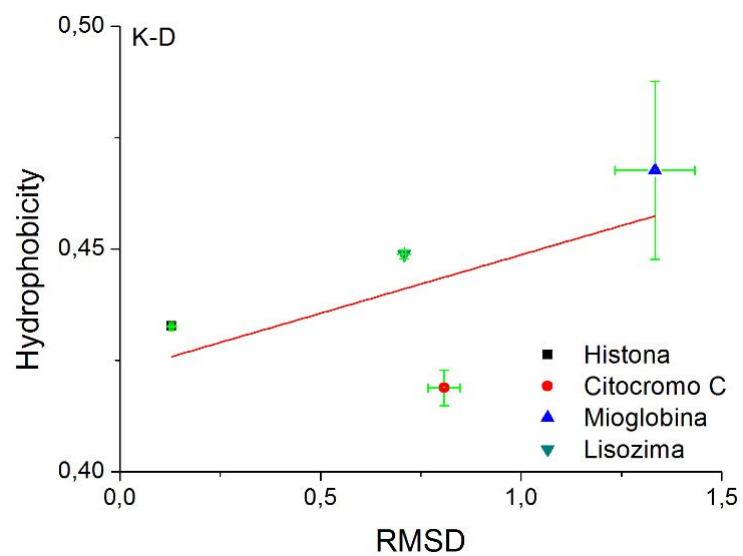


Figura 39- Relação entre Hidrofobicidade e RMSD para todos os grupos. Escala KD.

PARTE II

Resultados referentes à segunda parte do trabalho.

A.3. Gráficos Citocromo C

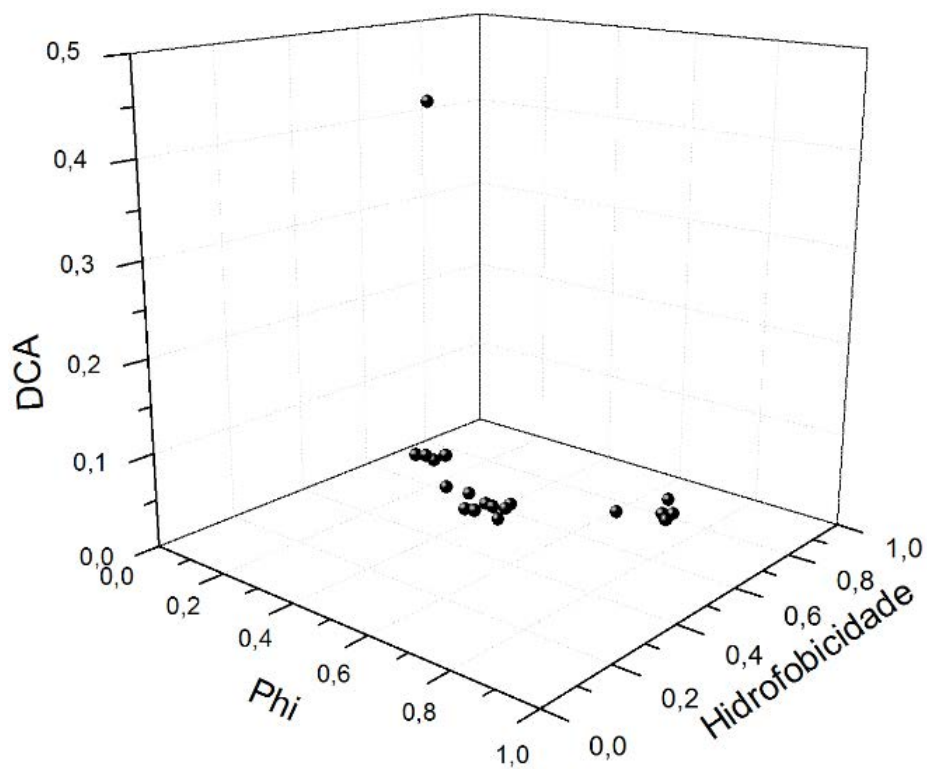


Figura 40 - Gráfico Acoplamento Direto (DCA) x Phi x Hidrofobicidade

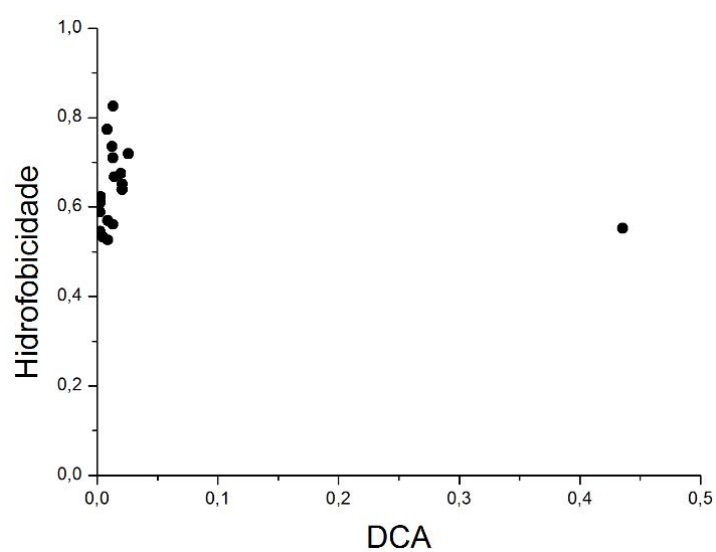


Figura 41- Gráfico Hidrofobicidade x DCA referente ao Citocromo C

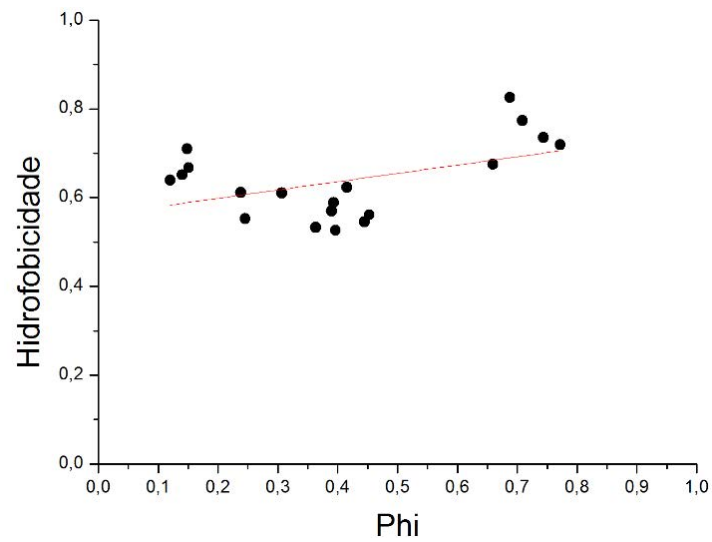


Figura 42 - Gráfico Hidrofobicidade x Phi referente ao Citocromo C

A.4. Gráficos Mioglobina

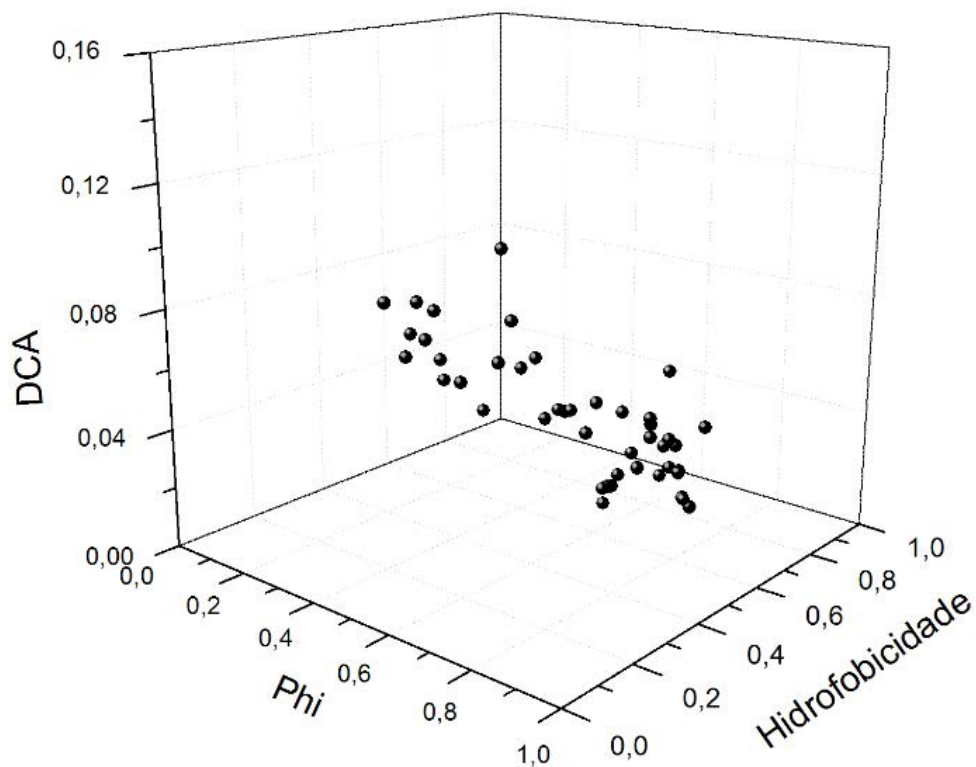


Figura 43- Gráfico Acoplamento Direto (DCA) x Phi x Hidrofobicidade

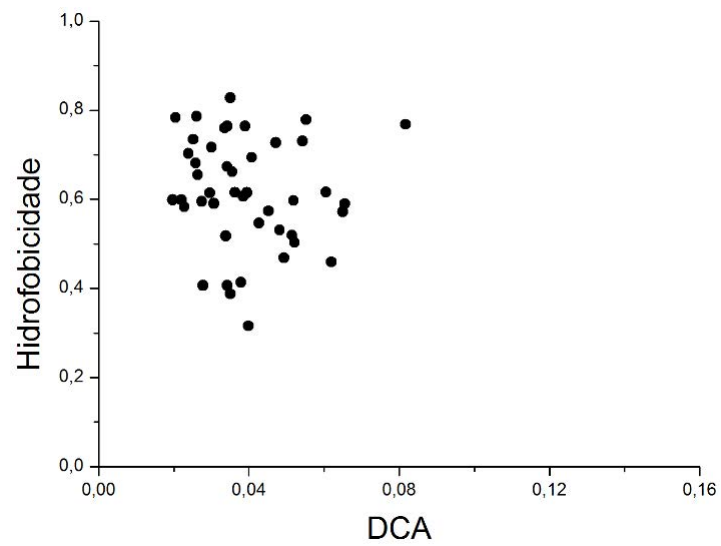


Figura 44- Gráfico Hidrofobicidade x DCA referente a Mioglobina

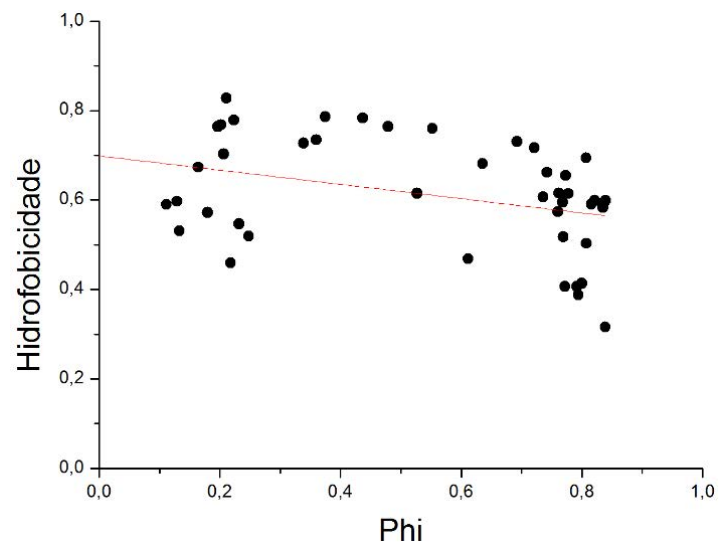


Figura 45- Gráfico Hidrofobicidade x Phi referente a Mioglobina

APÊNDICE B

Filogenética dos Vertebrados

Os animais do filo chordata que possuem coluna vertebral pertencem ao subfilo vertebrata. A palavra “vertebrado” vem das vértebras que são arranjadas em série e, assim, formam a coluna vertebral. Os vertebrados são muito diversificados, possuindo cerca de 64.000 espécies descritas até hoje⁵², representando a maior parte do filo chordata.

Carl von Linné, um naturalista sueco, desenvolveu a nomenclatura binomial, que estabelecia um conjunto de normas para atribuir nomes científicos aos animais. O método designa um nome formado por duas palavras (binomial) para cada espécie. Linné também contribuiu para o desenvolvimento do sistema natural de classificação. Desta forma, foi estabelecido que a espécie é o nível básico da classificação biológica. As espécies, por sua vez, pertenciam a um gênero e este era definido baseado em caracteres específicos, geralmente anatômicos. Os gêneros se agrupam em famílias, famílias em ordens, ordens em classes e as classes em filos. Este método parecia eficiente na época em que foi desenvolvido, pois os animais eram agrupados de acordo com suas similaridades anatômicas. Com o avanço dos conhecimentos biológicos, hoje se sabe que as características biológicas geneticamente determinadas que expressam as similaridades entre os organismos⁵³.

Todos os métodos antigos de classificação de organismos são baseados nas similaridades e diferenças entre espécies, não levando em consideração que algumas delas são mais relevantes que outras. Como exemplo é possível citar que todos os vertebrados possuem coluna vertebral, porém nem todos possuem glândulas mamárias. Dizer que duas espécies possuem glândulas mamárias fornece mais informações a respeito de seu parentesco do que apenas dizer que possuem coluna vertebral. A partir desse fato, Willi Hennig desenvolveu um método que determina as relações evolutivas entre as espécies e este é conhecido como sistemática filogenética ou cladística. Com a contribuição de Hennig, os grupos de animais começaram a ser identificados utilizando como base os caracteres derivados. Ao se tratar de caracteres derivados, a palavra “derivado” possui como significado “diferente do ancestral”.

A sistemática filogenética nos permite dizer que organismos agrupados compartilham um ancestral comum, que é responsável por suas similaridades. À medida que um organismo se distancia de outro ao longo da evolução, são originados novos clados. A partir das informações disponíveis a respeito dos vertebrados, segue a relação filogenética atual.

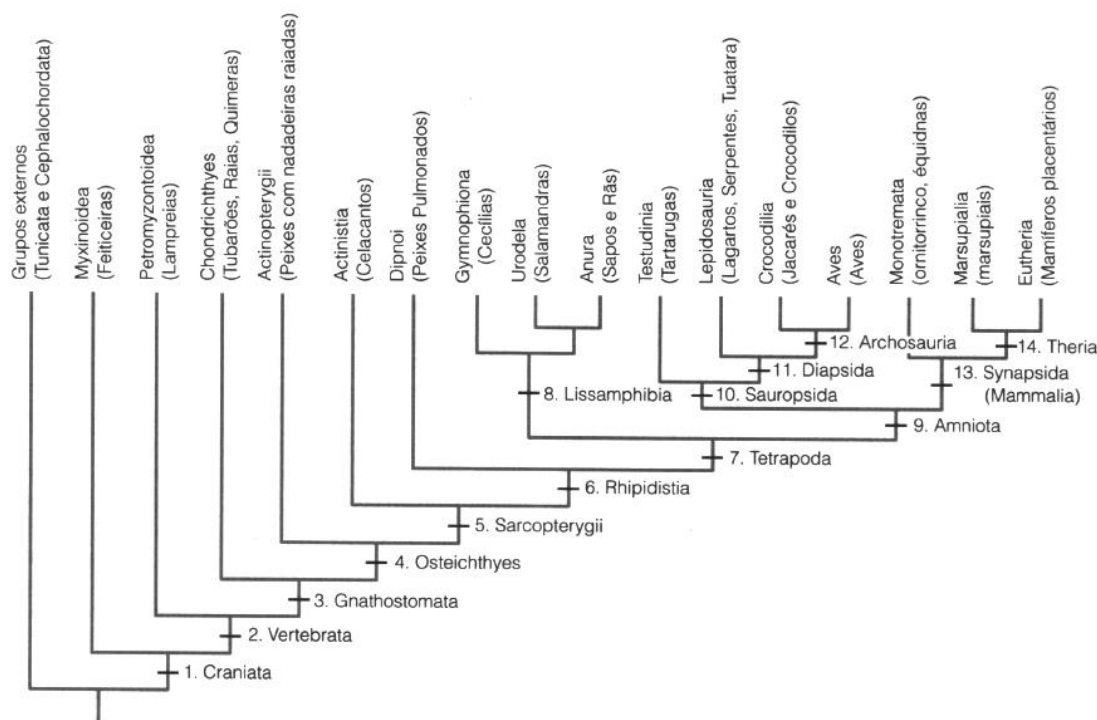


Figura 46 - Relações Filogenéticas entre os principais grupos de vertebrados vivos⁵³

Portanto, com base nas relações filogenéticas é possível assumir que mamíferos placentários possuem uma longa distância filogenética de peixes com nadadeiras raiadas.

Este trabalho utilizou as informações a respeito das relações filogenéticas para desenvolver a metodologia adequada. A metodologia consiste em dividir as espécies em grupos de acordo com sua distância evolutiva e realizar a comparação entre pares de espécies pertencentes a grupos distintos. De acordo com os conhecimentos em evolução, espécies com proximidade filogenética, ou seja, com um ancestral comum recente, como Chimpanzé e Humanos, são muito similares. Assim, ao se comparar as diferenças entre espécies com alta proximidade filogenética e espécies com baixa proximidade, é gerado um alto desvio com relação a média.

Utilizando as espécies estudadas, foi construída uma árvore filogenética utilizando o DNA mitocondrial dos organismos disponíveis no banco de dados Ensembl⁵⁴. Nem todas as espécies deste trabalho possuem tais informações nos bancos

de dados. A figura a seguir mostra o cladograma obtido com nove espécies utilizadas no trabalho. É possível comparar com o cladograma dos vertebrados, de modo a evidenciar sua consistência.

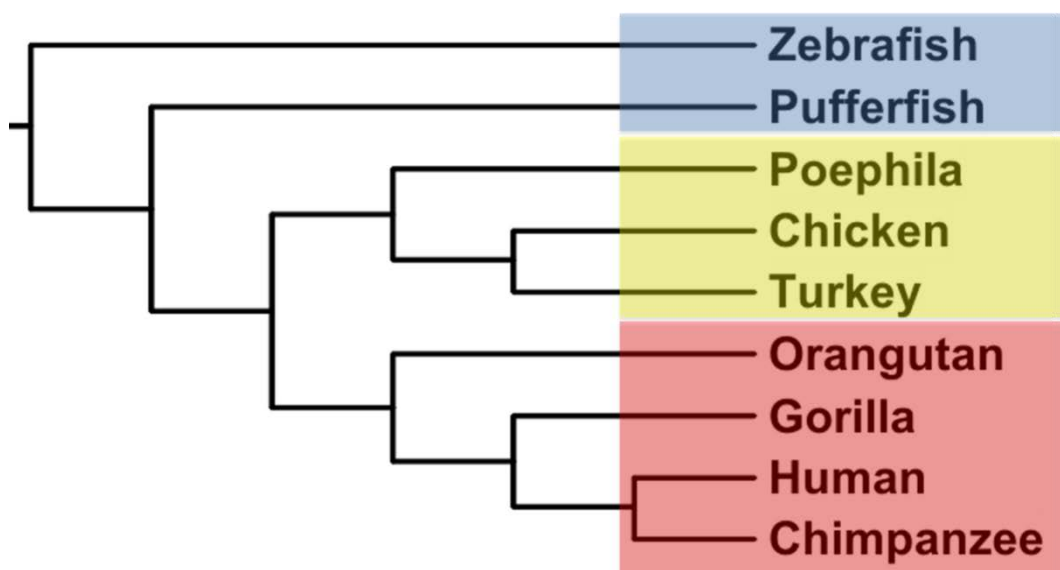


Figura 47 - Relação filogenética entre as espécies envolvidas no trabalho. Construída utilizando o software Winclada com Ri=63 e Ci=76. As cores indicam os grupos: Azul – Peixes, Amarelo-Aves, Vermelho-Primatas.

É possível notar a separação entre os grupos dos Peixes, Aves e Primatas. Essa é uma árvore simplificada, mostrando apenas a relação entre as espécies utilizadas no trabalho, de forma que o tempo de evolução e as outras espécies que se desenvolveram ao longo dos anos não estão evidenciados. Dessa forma, assumir que os grupos definidos no trabalho possuem distâncias evolutivas diferentes, de modo que as espécies contidas em cada um deles possuem diferenças geneticamente determinadas, está de acordo com a literatura.