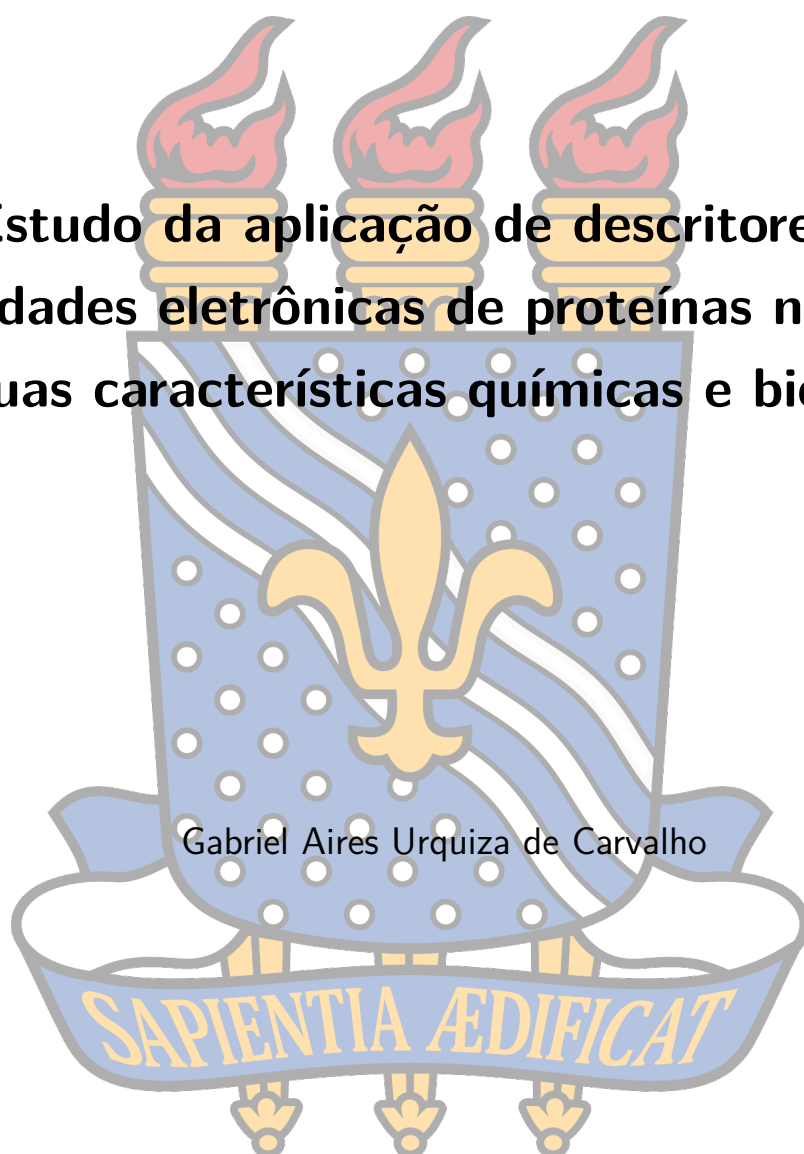


Universidade Federal da Paraíba
Centro de Ciências Exatas e da Natureza
Programa de Pós-Graduação em Química

Estudo da aplicação de descritores de densidades eletrônicas de proteínas no estudo de suas características químicas e biológicas.



Gabriel Aires Urquiza de Carvalho

João Pessoa - PB, Brasil

26 de Junho de 2017

Gabriel Aires Urquiza de Carvalho

**Estudo da aplicação de descritores de densidades
eletrônicas de proteínas no estudo de suas características
químicas e biológicas.**

Tese apresentada como requisito para obten-
ção do título de Doutor em Química pela
Universidade Federal da Paraíba.



Universidade Federal da Paraíba
Centro de Ciências Exatas e da Natureza
Programa de Pós-Graduação em Química

Orientador: Prof. Dr. Gerd Bruno da Rocha
Coorientador: Prof. Dr. Sávio Torres de Farias

João Pessoa - PB, Brasil

26 de Junho de 2017

C331e Carvalho, Gabriel Aires Urquiza de.
Estudo da aplicação de descritores de densidades
eletrônicas de proteínas no estudo de suas características
químicas e biológicas / Gabriel Aires Urquiza de Carvalho. -
João Pessoa, 2017.
120 f.: il. -

Orientador: Gerd Bruno de Rocha.
Coorientador: Sávio Torres de Farias.
Tese (Doutorado) - UFPB/CCEN

1. Química Analítica. 2. Invariantes Rotacionais.
3. Densidade Eletrônica. 4. Canterakis-Zernike. 5. Proteínas
Ancestrais. I. Título.

UFPB/BC

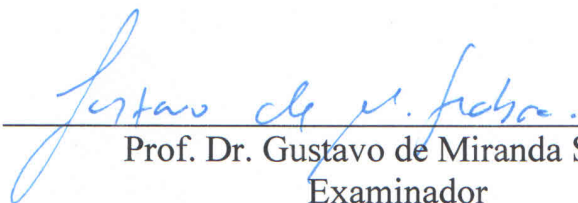
CDU: 543(043)

Estudo da aplicação de descritores de densidades eletrônicas de proteínas no estudo de suas características químicas e biológicas.

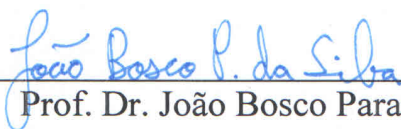
Tese de doutorado apresentada pelo aluno Gabriel Aires Urquiza de Carvalho e aprovada pela banca examinadora em 26 de junho de 2017.



Prof. Dr. Gerd Bruno da Rocha
Orientador/Presidente



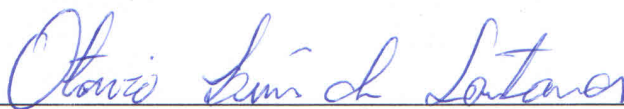
Prof. Dr. Gustavo de Miranda Seabra
Examinador



Prof. Dr. João Bosco Paraíso da Silva
Examinador



Prof. Dr. Wagner de Mendonça Faustino
Examinador



Prof. Dr. Otávio Luís de Santana
Examinador

*Este trabalho é dedicado às crianças adultas que,
quando pequenas, sonharam em se tornar cientistas.*

Agradecimentos

Um trabalho deste tamanho raramente vem ao mundo pela vontade e o esforço de uma única pessoa. Ao longo de mais de quatro anos de trabalho, muitas pessoas e instituições acabam participando, voluntaria ou involuntariamente do processo de desenvolvimento. Interações e contribuições que, no início, parecem pequenas, podem acabar tornando-se eventos nos quais depende todo o sucesso posterior do trabalho, em uma forma de efeito borboleta virtuoso. Portanto aqui neste pequeno segmento gostaria de abrir um espaço para agradecer a algumas pessoas, grupos e instituições que são colaboradores neste trabalho, de uma forma ou de outra.

Antes de qualquer um, eu gostaria de agradecer a minha família. Meu pai Francisco e minha mãe, Ana Maria em especial, que permitiram que eu tivesse a educação e as oportunidades todas alinhadas em meu favor, sempre que possível e que sem dúvida foram os maiores responsáveis por eu ter conseguido abraçar a oportunidade de me dedicar a este trabalho quando ela apareceu. Não é sempre na vida que temos simultaneamente a oportunidade e os meios de materializarmos um sonho e com certeza meus pais lançaram mão de tudo em seu alcance, e fizeram todos os sacrifícios necessários para permitir que esse fosse o meu caso. Agradeço, na verdade, duas vezes e meia a minha mãe, pois ela continuou segurando as pontas e fazendo todos os esforços para que eu pudesse continuar no meu caminho até sua conclusão. Embora ela não tenha depositado uma só letra sobre este papel, ela fez muitos sacrifícios para permitir que eu o fizesse.

Gostaria de agradecer, também, a todos os meus professores, não apenas aqueles da universidade, ou aqueles com diplomas e concursos. O universo é um lugar grande e quando honestamente tentamos aprender as coisas sobre o mundo natural com alguma profundidade, rapidamente nos damos conta de como semelhante aventura é vasta, complexa e provavelmente interminável. Para quem busca a ciência honestamente, dar-se conta da enormidade da tarefa deve ser uma experiência que nos faz sentir humildes e pequenos diante da vastidão da natureza. Por esta razão dedico um carinho muito especial a todos aqueles que, quando eu ainda era jovem neste mundo e não sabia de muitas coisas, pacientemente me tomaram pela mão e me mostraram esse gigantesco quebra-cabeça a ser montado. A tarefa é muito grande pare ser tomada de forma solitária de modo que todos aqueles que tiveram o cuidado de ir antes de mim e iluminar as trilhas mais comuns são pessoas amigas.

São muitos para serem enumerados, mas eu gostaria de citar em especial o Prof. Gutemberg, que foi meu prof. de Matemática no terceiro ano do Ensino Médio e o Prof. João Neto, meu prof. de Química da oitava série ao segundo ano. Se eu tivesse que apontar

onde se iniciou o processo acadêmico que me levaria até o doutorado, meu ensino médio e estes professores certamente estariam entre os momentos e personagens cruciais. Já na graduação, gostaria de agradecer à Prof. Dra. Bagnólia Araújo e ao Prof. Dr. Luis Fernando Marques, meu primeiro orientador, amigo e provavelmente a pessoa quem primeiro me ensinou alguma coisa sobre como ciência funciona na prática.

Ciência é uma atividade que não se encerra quando o expediente acaba. O que quer que se faça no laboratório nos acompanha para casa, interrompe o pensamento e dorme conosco. Trata-se de uma atividade que, se permitida, nos consome por completo. Embora esta não seja a receita para uma forma saudável de lidar com o trabalho, vez por outra todos que trabalham com ciência se vêem dedicando tempo demais ao trabalho e de menos a todo o resto. Gostaria de agradecer a todo mundo que estava em "todo o resto" nestes lapsos de excesso de dedicação. Toda a minha família, todos os meus amigos e amigas e todo mundo que gostaria de ter contado comigo para alguma coisa, mas eu simplesmente não estava lá, esgotado pelas exigências do trabalho.

Em especial gostaria de agradecer, pela paciência e pelo apoio, aos meus irmãos Emmanuel e Abraão, a minha mãe Ana Maria, novamente, e minha namorada Tereza Gracielle. Gostaria de agradecer a paciência e pedir desculpas por todas as vezes que meu trabalho sugou o melhor de mim e eles tiveram que se contentar com o que sobrou.

Fazer ciência é uma atividade recompensadora, mas também laboriosa e cara. Como uma única pessoa, pertencente a uma classe social não muito alta, jamais eu teria condições de obter por meus próprios meios e contatos o equipamento e o espaço necessário para desenvolver o trabalho. Sou completamente grato a CAPES e CNPq pelo financiamento, ao CENAPAD Unicamp e ao LMMRQ da UFPB pela disponibilidade de equipamento e tempo computacional sem os quais teria sido efetivamente impossível desenvolver o trabalho.

Ainda que uma boa parte do trabalho tenha sido efetuada, no dia-a-dia, por mim apenas, nunca me senti sozinho no laboratório. Em grande parte devo isso a pessoas que começaram como colegas de trabalho e tornaram-se alguns dos amigos mais próximos e leais com quem eu poderia contar. Obrigado a Italo Curvelo dos Anjos, Higo Lima, Jessé Azevedo, Felipe Nóbrega, Juliana Monteiro, a meu primo Júlio Daniel de Carvalho Maia e a todos que participaram das vivências e convivências do laboratório e, de forma geral, da minha vida científica, dividindo histórias e hipóteses. Sem vocês, a jornada teria sido bem solitária.

Finalmente, mas não menos importante, eu agradeço ao meu amigo e orientador Prof. Dr. Gerd Bruno da Rocha. Agradeço por ter acreditado no trabalho e em minha capacidade para desenvolvê-lo, por todos os puxões de orelha, por todos os momentos filosóficos e também pelos momentos pé-no-chão. Agradeço por ter selecionado a mim para este papel e tornado possível o desenrolar de todo esse arco, da concepção à conclusão do trabalho. Sem que ele abrisse a porta de sua sala para me orientar, provavelmente este

trabalho jamais teria visto a luz do dia. Muito obrigado, meu amigo.

"A química está nos elétrons."
Prof. Dr. Gerd Bruno da Rocha

Resumo

Proteínas são macromoléculas de aminoácidos cuja geometria final, semelhante a um complexo novo, permite que estas moléculas sejam bastante versáteis, utilizadas em uma vasta gama de papéis químicos nas células vivas. A geometria de equilíbrio das proteínas em suas condições normais são determinadas, em grande parte, por sua sequência primária, que é produto da evolução biológica. Neste trabalho foram utilizados os momentos de Canterakis-Zernike e suas Invariantes Rotacionais para estudar possíveis semelhanças entre as densidades eletrônicas de proteínas modernas e aquelas produzidas a partir de estruturas obtidas de suas sequências ancestrais retiradas da literatura. As tRNA-Aminoacil Sintetases foram selecionadas como conjunto de análise em virtude da riqueza de informação filogenética sobre estas moléculas na literatura, entre estas informações inclusive a presença de sequências de ancestrais obtidas pela técnica de Reconstrução de Sequências Ancestrais (ASR). As estruturas das sequências modernas e ancestrais foram obtidas utilizando-se o algoritmo de predição de estrutura terciária I-TASSER. A ferramenta BLAST foi utilizada para identificar sítios de dimerização nas sequências e o algoritmo ClusPro utilizado para realizar o ancoramento molecular entre as duas subunidades das proteínas que eram homodímeros. As proteínas foram submetidas a dinâmicas moleculares para que estruturas finais com estabilidade conformacional pudessem ser obtidas a partir daquelas produzidas pelo I-TASSER. As densidades eletrônicas foram obtidas destas estruturas finais através de cálculos single point semiempíricos utilizando o Hamiltoniano PM7 com o MOPAC2012. Os momentos de Canterakis-Zernike e suas Invariantes Rotacionais foram obtidas através de um novo algoritmo implementado neste trabalho para a projeção de densidades analíticas produzidas pelo Gaussian2009 e MOPAC2012 no espaço de Canterakis-Zernike, que encontra-se implementado no programa DAMQT. As densidades foram então agrupadas segundo suas invariantes rotacionais para observar se estes descritores seriam capazes de agrupar as densidades eletrônicas segundo a origem evolutiva das proteínas de onde elas foram calculadas. Concluiu-se que as densidades eletrônicas são capazes de agrupar as proteínas segundo seu estado de dimerização que, embora não sejam diretamente ligados ao passado evolutivo destas moléculas, é relacionado o suficiente para exigir investigações posteriores mais profundas e cuidadosas.

Palavras-chave: Química Quântica, Invariantes Rotacionais, Densidade Eletrônica, Canterakis-Zernike, Proteínas Ancestrais

Abstract

Proteins are macromolecules composed of aminoacids for which the final geometry, like that of a complex bundle, allows these molecules to be highly versatile and to occupy a wide array of chemical roles in living cells. The equilibrium geometry of proteins in their native conditions are determined, in great part, by their primary sequence which is itself a product of biological evolution. In this work the Canterakis-Zernike Moments and the Rotationally Invariant Fingerprints have been used to study possible similarity patterns between the electron densities of modern proteins and those obtained from the tertiary structures predicted from their ancestral sequences, obtained from the literature. tRNA-Aminoacyl Synthetases were chosen as the set of proteins to be analyzed due to the wealth of phylogenetic information about these molecules in the literature, including ancestral sequences obtained through the application of the Ancestral Sequence Reconstruction (ASR) technique. The structures for the modern and ancestral sequences were obtained through the tertiary structure prediction algorithm I-TASSER. Sequences were BLASTed in order to identify possible dimerization sites. The ClusPro tool was used to carry out the molecular docking between the two subunits of the homodimeric proteins. The I-TASSER/ClusPro structure of the proteins were then submitted for Molecular Dynamics runs for the obtention of final structures with conformational stability. The electron densities were obtained from these final structures by means of semiempirical single point calculations using the PM7 Hamiltonian implemented on MOPAC2012. The moments of Canterakis-Zernike and their rotationally invariante Fingerprints were obtained via a new algorithm developed in this work for the projection of the analytical densities computed by Gaussian2009 and MOPAC2012 in Canterakis-Zernike space, which is implemented in the DAMQT software. The densities were grouped according to their Fingerprints in order to study whether or not these descriptors could be used to group electron densities according to the evolutionary origins of the proteins from which they had been computed. It was concluded that the electron densities are capable of clustering the proteins according to their dimerization state which, although not directly linked to their evolutionary past, is related enough to warrant further and more detailed investigations.

Keywords: Quantum Chemistry, Rotational Invariants, Electron Density, Canterakis-Zernike.

Lista de ilustrações

Figura 1 – Esquema mostrando as características comuns aos 20 aminácidos proteínogênicos.	3
Figura 2 – Ângulos de Ramachandran mostrados em relação a um resíduo de aminoácido (em vermelho).	4
Figura 3 – Esquema ilustrando uma proteína assumindo sua Estrutura Terciária de forma autônoma a partir de uma conformação desnaturada.	5
Figura 4 – Exemplo ilustrativo de um cromossomo com quatro genes e cinco processos de mutação que poderiam, teoricamente, ocorrer durante um ciclo de reprodução. As mutações são: 1) Anulação de um gene; 2) Duplicação de um gene; 3) Deleção de um gene; 4) Enfraquecimento de um gene; 5) Troca de genes.	9
Figura 5 – Esquema ilustrativo indicando os processos de Anagênese e Cladogênese das espécies.	10
Figura 6 – Árvore filogenética produzida por Norm Pace baseado nos trabalhos de Carl Woese (WOESE; FOX, 1977), na descoberta do reino Archaea. . .	11
Figura 7 – Esquema ilustrativo mostrando como as características de um ancestral podem ser inferidas a partir das características do conjunto de espécies. . .	12
Figura 8 – Representação gráfica do referencial usado na expansão de produtos de orbitais atômicos em termos de Harmônicos Esféricos e fatores radiais . . .	18
Figura 9 – Amostra de um arquivo de saída do DAMQT.	24
Figura 10 – Fluxograma ilustrando os passos envolvidos no cálculo dos momentos de Canterakis-Zernike e seus invariantes rotacionais derivados.	24
Figura 11 – Amostra de um arquivo de saída do Cluto 2.1, com destaque em vermelho para as métricas dos grupos.	27
Figura 12 – Tipos de grupos baseados nos valores de similaridade interna (S) e desvio-padrão da similaridade interna (σ).	28
Figura 13 – Seção do arquivo de saída mostrando a solução aglomerativa. Métricas de qualidade de super-grupos estão destacadas em vermelho.	29
Figura 14 – Exemplo do resultado de um BLAST. A ênfase destaca a indicação de domínios associados com a estrutura quaternária da sequência submetida.	32
Figura 15 – Exemplo de uma proteína em uma caixa de simulação octaédrica truncada. A proteína está em preto, e a água está colorida por elemento: vermelho para oxigênio e branco para hidrogênio.	36
Figura 16 – Representação esquemática do método de análise 3SIEVE.	43

Figura 17 – Gráfico dos RMSD dos átomos da cadeia principal polipeptídica, em relação a primeira estrutura pelo tempo de trajetória simulado. As trajetórias mostradas são referentes às das aminoacil-tRNA Sintetases modernas que estabilizaram após dois ciclos de 10 ns.	45
Figura 18 – Gráficos de RMSD x Tempo das proteínas com trajetória de equilíbrio de 10 ns.	47
Figura 19 – Cadeias das proteínas ancestrais 19b e modernas 19a a serem utilizadas na obtenção das densidades.	49
Figura 20 – Cadeias das Lisil-tRNA Sintetases utilizadas na obtenção das densidades.	49
Figura 21 – Formato da densidade do benzeno para quantidades diferentes de funções da base de Canterakis-Zernike. As subfiguras mostram a) A reconstrução de uma densidade para as combinações de $K=L=10$, b) A reconstrução da densidade original com $K=L=20$ e c) A densidade GTO STO-3G original.	50
Figura 22 – Gráfico de $\sqrt{V(r^*)_0}\omega_0^0(r^*)$ para todos os valores inteiros de r^* entre 1 e 10 Å para a água com expansão $l_{max} = k_{max} = 15$ e a quadratura com 1024 pontos. A densidade original foi computada no Gaussian 2009, usando o método do Hartree-Fock restrito com base STO-3G no vácuo.	53
Figura 23 – Gráfico mostrando a relação entre o quociente dos raios e o número de funções de base da molécula. A linha vermelha mostra o ajuste exponencial.	54
Figura 24 – Algoritmo para a obtenção dos agrupamentos.	55
Figura 25 – Grupo com $S = 0,769$ e $\sigma_S = 0,034$	56
Figura 26 – Grupo com $S = 0,762$ e $\sigma_S = 0,047$	57
Figura 27 – Agrupamento de núcleos heterocíclicos para qualquer combinação de parâmetros.	58
Figura 28 – Mapa de contornos da densidade eletrônica no plano do anel heterocíclico do grupo dos heterociclos de cinco membros.	59
Figura 29 – Agrupamento de moléculas lineares sem insaturações. As caixas em cinza ilustram a estrutura de subgrupo sugerida.	59
Figura 30 – Super-grupo de moléculas lineares contendo insaturações. O delimitador preto mostra onde cada grupo pertencente a esse super-grupo começa e termina.	60
Figura 31 – Subconjunto das moléculas que mudam de grupo quando comparadas as densidades de <i>single point</i> e aquelas de geometria otimizada.	61
Figura 32 – Diferença entre a entalpia de formação de cada estrutura nativa e a linha de $-3s$ calculada segundo o 3SIEVE. A diferença é calculada para cada método em cada conjunto de estruturas.	66

Figura 33 – Gráficos de correlação entre a fração de contatos nativos do melhor decoy selecionado por cada método, na ausência da nativa, e a fração de contatos nativos do decoy com maior fração de contatos nativos em cada grupo.	68
Figura 34 – Imagem do arquivo de saída em que os grupos de proteínas são caracterizados segundo suas invariantes rotacionais.	70
Figura 35 – Bandas de Harmônicos Esféricos até $l = 9$. As partes vermelhas indicam valores negativos e as partes verdes indicam valores positivos.	96
Figura 36 – Tipos de Harmônicos Esféricos segundo suas propriedades de simetria.	97
Figura 37 – Sistema de coordenadas (θ, ϕ, ψ) utilizado para especificar uma rotação em torno de um eixo.	98

Lista de tabelas

Tabela 1 – Somas parciais na avaliação das funções de Canterakis-Zernike com $l = 15$, $k = 15$ e $r/r^* = \frac{1}{2} (1 + \cos \frac{633\pi}{4096})$ usando o Esquema de Horner.	23
Tabela 2 – Estruturas Quaternárias das proteínas modernas estudadas, com exceção das Lisil-tRNA Sintetases.	33
Tabela 3 – Estruturas Quaternárias das Lisil-tRNA Sintetases	34
Tabela 4 – Caracterização dos Conjuntos de <i>Decoys</i> utilizados na comparação dos Hamiltonianos semiempíricos.	41
Tabela 5 – Raio de giro médio e variância ao longo da trajetória de produção NPT de cada aminoacil-tRNA Sintetase do conjunto usado no trabalho, em Å.	48
Tabela 6 – Na coluna da esquerda tem-se o Z-escore da estrutura nativa, e na coluna da direita a diferença entre a entalpia de formação da nativa e a entalpia de formação do melhor <i>decoy</i> , expressa em kcal/mol. Os cálculos foram realizados antes da aplicação do 3SIEVE.	64
Tabela 7 – A diferença entre as entalpias de formação entre a estrutura nativa de cada conjunto de estruturas e suas linhas de $-3s$ correspondente, em kcal/mol. Estes valores foram calculados após a população passar pelo 3SIEVE.	67
Tabela 8 – Distribuição das densidades nos grupos em cada tipo de classificação considerado.	71
Tabela 9 – Quantidade de tRNA-aminoacil Sintetases de Classe 1 agrupadas pelo aminoácido substrato, indicados pelos seus códigos de uma letra.	71
Tabela 10 – Quantidade de tRNA-aminoacil Sintetases de Classe 2 agrupadas pelo aminoácido substrato, indicados pelos seus códigos de uma letra.	72

Lista de abreviaturas e siglas

3SIEVE	Expositor Iterativo de Semelhança Nativa até 3σ ;
ASR	Reconstrução de Sequências Ancestrais;
BLAST	Ferramenta de Busca Básica por Alinhamento Local;
COSMO	Modelo de Varredura Semelhante a Condutor;
DFT	Teoria do Funcional da Densidade;
DNA	Ácido Desoxirribonucleico;
EP	Enovelamento de Proteínas;
GTO	Orbital Tipo-Gaussiana;
LCAO	Combinação Linear de Orbitais Atômicos;
NCBI	Centro Nacional de Informação em Biotecnologia;
NPT	Sigla utilizada para referir-se a um <i>ensemble</i> termodinâmico contendo microestados acessíveis ao sistema mantendo a quantidade de partículas (N), pressão (P) e temperatura (T) constantes.
PCBM	Paradigma Central da Biologia Molecular;
PDB	Banco de Dados de Proteínas;
QTAIM	Teoria Quântica de Átomo em Moléculas;
RMSD	Raiz do Desvio Quadrático Médio;
RNA	Ácido ribonucleico;
SCF	Campo Auto-Consistente;
SCOP	Classificação Estrutural de Proteínas;
STO	Orbital Tipo-Slater;
STO-3G	Base de cálculo de química quântica <i>ab initio</i> em que cada STO é aproximado por um conjunto de três GTOs;
RHF	Hartree-Fock Restrito;

TIP3P	Modelo de molécula de água para simulações de dinâmica molecular em solvente explícito;
tRNA	RNA transportador;

Lista de símbolos

ρ	Densidade Eletrônica;
${}_k\rho_l^m$	Projeções da densidade eletrônica. Qualquer combinação de índices omissos ou explícitos gera diferentes projeções, todas válidas;
$\nabla\rho$	Gradiente da Densidade Eletrônica;
ψ_0	Função de onda do estado fundamental;
ρ_0	Densidade eletrônica gerada pelo estado fundamental;
E_0	Energia do estado fundamental;
$\tilde{\mathbf{P}}$	Conjunto das densidades eletrônicas reais;
$\mathbf{P}$	Conjunto das densidades eletrônicas escritas na base de Canterakis-Zernike;
Ψ	Conjunto das funções de onda;
$\mathbb{R}$	Conjunto dos números reais;
R_m^n	Polinômio radial 2D de Zernike;
χ	Função de base centrada em átomos;
${}_k\mathcal{F}_l^m$	Função genérica ortonormal na esfera unitária;
${}_k\tilde{\mathcal{F}}_l^m$	Função genérica ortonormal na esfera de raio r^* ;
r^*	Limite de Integração radial para a decomposição de Canterakis-Zernike;
${}_k\omega_l^m$	Momentos da expansão em termos das funções ${}_k\tilde{\mathcal{F}}_l^m$
${}_kZ_l^m$	Função não-normalizada de Canterakis-Zernike;
${}_k\tilde{Z}_l^m$	Função normalizada de Canterakis-Zernike;
Z_{kl}	Polinômio Radial de Canterakis-Zernike não-normalizado;
$\tilde{Z}_{kl}$	Polinômio Radial de Canterakis-Zernike normalizado;
$\mathcal{N}_l^m$	Constante de Normalização dos Harmônicos Esféricos;
Y_l^m	Harmônico Esférico não-normalizado;

$\tilde{Y}_l^m$	Harmônico Esférico normalizado;
δ_{kl}	Delta de Kronecker em relação aos índices k e l ;
$\mathcal{R}^l$	Matriz de rotação dos Harmônicos Esféricos de mesmo l ;
f_{lm}	Fatores radiais no referencial molecular;
$\hat{f}_{lm}(r)$	Fatores radiais no referencial padrão;
Å	Unidade de medida de comprimento Ångström;
$\bar{H}$	Entalpia de formação média de uma população;
q_{kl}^ν	Função radial parcial de Canterakis-Zernike;
N^{e^-}	Número de elétrons;
$V(r)$	Volume contido em uma esfera de raio r ;
$G(r)$	Função de distribuição radial de elétrons em torno do centro geométrico da molécula;
Ω	Conjunto de pontos na esfera de raio r^* ;
$d\Omega$	Intervalo de integração na esfera Ω ;
$SO(3)$	Grupo de todas as rotações em torno da origem em $\mathbb{R}^3$;
Q_l^m	Momentos de multipolo;
$\mathbf{F}$	Matriz das Invariantes Rotacionais;
F_{kl}	Elemento da matriz $\mathbf{F}$;
U	Potencial eletrostático de longo alcance;
$\mathcal{H}_2$	Função-utilidade implementada no programa de agrupamento Cluto2.1;
$\mathcal{I}_2$	Função-utilidade implementada no programa de agrupamento Cluto2.1.

Sumário

1	INTRODUÇÃO	1
2	OBJETIVOS	7
I	FUNDAMENTAÇÃO TEÓRICA	8
3	EVOLUÇÃO	9
4	EXPANSÃO GERAL DE UMA DENSIDADE MOLECULAR EM UMA ESFERA	13
II	METODOLOGIA	16
5	CÁLCULO DOS MOMENTOS DE CANTERAKIS-ZERNIKE	17
5.1	Expansão de Densidades Usando Técnicas de Translação	17
5.2	Funções de Canterakis-Zernike	20
5.3	Expansão da Densidade em Funções de Canterakis-Zernike	22
6	AGRUPAMENTOS UTILIZANDO AS INVARIANTES ROTACIONAIS	26
7	OBTENÇÃO DE ESTRUTURAS DE PROTEÍNAS	31
7.1	Seleção de Conjunto	31
7.2	Preparação de Estruturas	32
7.3	Dinâmica Molecular	35
7.4	Seleção do Hamiltoniano	37
7.4.1	Cálculos	39
7.4.2	Comparação dos Métodos	41
III	RESULTADOS	44
8	SELEÇÃO DAS ESTRUTURAS	45
8.1	Dinâmica Molecular	45
9	DECOMPOSIÇÃO DA DENSIDADE	50
9.1	Selecionando o Raio de Integração	51

10	AGRUPAMENTOS DAS MOLÉCULAS PEQUENAS USANDO AS INVARIANTES ROTACIONAIS	55
11	SELEÇÃO DO HAMILTONIANO PARA GERAÇÃO DE DENSIDADE ELETRÔNICA DE PROTEÍNAS	63
12	AGRUPAMENTO DE PROTEÍNAS	70
13	CONCLUSÃO	73
14	PERSPECTIVAS FUTURAS	75
	REFERÊNCIAS	77
	APÊNDICES	88

1 Introdução

A química teórica divide-se, basicamente, em dois grandes ramos de formalismo. De um lado, a complexidade computacional dos métodos de química quântica restringiu a aplicação destes métodos à descrição de sistemas pequenos e/ou de alta simetria. Grande esforço foi empregado para, partindo-se dos primeiros princípios físicos, chegar a descrições sólidas de propriedades e conceitos químicos. Esse esforço produziu métodos computacionais de química quântica para o cálculo de propriedades espectroscópicas (PALSHIN *et al.*, 2002; WELACK *et al.*, 2008; PUZZARINI; STANTON; GAUSS, 2010), o cálculo de reatividade e potencial químico (FURNESS *et al.*, 2016; GOTZ *et al.*, 2009; GEERLINGS; PROFT, 2008; GEERLINGS *et al.*, 2012; PINTER *et al.*, 2013), a detecção de ligações químicas (GRABOWSKI, 2012; JABLONSKI; PALUSIAK, 2012; NIKOLAIENKO; BULAVIN; HOVORUN, 2012; GRABOWSKI; UGALDE, 2010), a cinética e termodinâmica de reações químicas (DENG *et al.*, 2014; POLYAK; SEN; THIEL, 2013; SAKATA, 2000; GUO *et al.*, 2008; EMAMIAN, 2017), entre outros conceitos (YI *et al.*, 2017; SILVI, 2017).

No outro lado da modelagem encontram-se técnicas de mecânica molecular que permitiram o emprego da dinâmica molecular no estudo de sistemas maiores que aqueles tipicamente estudados pela mecânica quântica. Como o cálculo da energia em métodos de mecânica molecular é bem mais rápido, não apenas é possível estudar geometrias de equilíbrio e modos normais clássicos de moléculas maiores, mas também empregar técnicas de Dinâmica Molecular para estudar propriedades termodinâmicas de sistemas bastante heterogêneos como misturas de líquidos (WU *et al.*, 2005; RIVAS; REY, 2017; ALCALDE *et al.*, 2017; FERRARIO *et al.*, 1990), biomoléculas em solução (LUAN; ZHOU, 2017; KAGAMI *et al.*, 2017; KLEPEIS *et al.*, 2009; ADCOCK; MCCAMMON, 2006; KARPLUS; MCCAMMON, 2002) ou nanotubos (SEARS; BATRA, 2004; CHANG; GAO, 2003), para citar alguns exemplos.

O espaço entre os dois campos são ocupados por métodos e estratégias que visam utilizar química quântica para estudar sistemas grandes. Entre estas técnicas encontram-se os métodos semiempíricos de química quântica e técnicas de escalonamento linear. Já em 1996, J. P. P. Stewart publicou seu algoritmo MOZYME (STEWART, 1996), utilizado até hoje, e demonstrou a capacidade do método de calcular entalpia de formação de enzimas inteiras. O MOZYME foi seguido, também em 1996, pelo DivCon (DIXON; MERZ, 1996), o FMO (KITAURA *et al.*, 1999) em 1999 e, em 2004, pelo LocalSCF (ANIKIN *et al.*, 2004), até hoje um dos métodos de escalonamento linear mais bem sucedidos, do ponto de vista de desempenho computacional. Todas estas técnicas, de alguma forma, exploram o fato de que as matrizes usualmente envolvidas em cálculos de química quântica apresentam padrões de blocagem e esparsidade. Esta propriedade torna-se cada vez mais pronunciada

à medida que a escala do sistema aumenta, gerando bolsões de interações locais que podem ser particionados e resolvidos com relativa independência sem que se incorra em grande penalidade na exatidão das soluções. Técnicas de QM/MM são ainda mais antigas (WARSHEL; LEVITT, 1976; SENN; THIEL, 2009; BRUNK; ROTH LISBERGER, 2015).

A drástica expansão das capacidades computacionais, em particular o surgimento do ambiente de programação em unidades de processamento gráfico (GPUs) e seu hardware hiper-paralelo, possibilitou grandes contribuições para a interface entre os domínios da química quântica *ab initio* e da dinâmica/mecânica molecular. Hoje em dia existem implementações em GPUs para a propagação de trajetórias de Dinâmica Molecular (XU; WILLIAMSON; WALKER, 2010; ANDERSON; LORENZ; TRAVESSET, 2008; FRIEDRICH *et al.*, 2009), Cálculo de Integrais de Repulsão Eletrônica (BROWN *et al.*, 2010; UFIMTSEV; MARTINEZ, 2009; LUEHR; UFIMTSEV; MARTINEZ, 2011), e mesmo subrotinas centrais do próprio algoritmo do Campo Auto-Consistente (SCF) (UFIMTSEV; MARTINEZ, 2009; KINDRATENKO *et al.*, 2010), entre outras estratégias aplicáveis a problemas mais específicos.

Com a grande aceleração possibilitada por esse novo tipo de hardware, surgiram novas contribuições para o estudo dentro desta interface. Em particular, o TeraChem (LUEHR; JIN; MARTINEZ, 2015), um software de dinâmica molecular criado, desde sua concepção, para realizar Dinâmica Molecular *ab initio* utilizando o máximo do poder das GPUs, bem como a implementação, pelo nosso próprio grupo, de métodos de cálculo semiempírico do MOPAC nas GPUs (MAIA *et al.*, 2012).

Depois do *boom* dos GPUs, o tamanho dos sistemas tratáveis pela mecânica quântica cresceu vertiginosamente. Surgiram, naturalmente, implementações que visam portar confortavelmente conceitos e técnicas desenvolvidas para sistemas menores para estes novos sistemas ultra-grandes. A implementação e adaptação de técnicas computacionais já consagradas para esta nova fronteira de sistemas ainda é, até hoje, uma área ativa de pesquisa de grandes centros. Métodos de QTAIM (HERNANDEZ-ESPARZA *et al.*, 2014) e Coordenada Intrínseca de Reação (MAEDA SAND HARABUCHI *et al.*, 2014), por exemplo, possuem implementações voltadas para macromoléculas.

Em 2008, Fukushima, Sakurai e Wada publicaram um trabalho (FUKUSHIMA; WADA; M., 2008) em que demonstraram, entre outras coisas, que os orbitais moleculares de fronteiras de 112 enzimas pareciam estar localizados sobre o sítio ativo dessas proteínas e que suas propriedades de localização ou de-localização estavam correlacionadas com a especificidade das enzimas por seus respectivos substratos. Este trabalho é uma amostra do tipo de desenvolvimento que se seguiu à expansão da fronteira dos métodos de química quântica para incluir sistemas maiores. Expandir os métodos existentes para incluir um maior número de átomos faz emergir novas quantidades que são indefinidas para sistemas menores, como localização de orbitais moleculares estudada por estes autores.

Trabalhos como o de Fukushima e seus colaboradores mostram que o problema da descrição de sistemas grandes é, na verdade, multiescala, isto é, propriedades de interesse surgem ou desaparecem dependendo da escala em que a observação do sistema é realizada. A Metadinâmica (LAIO; GERVASIO, 2008; BARDUCCI; BUSSI; PARRINELLO, 2008) e a estratégia de *Coarse-Graining* (PANDEY; JACOBS; FARMER, 2017) podem ser enquadrados dentro das técnicas que levam em consideração a natureza multi-escala dos sistemas maiores. De forma geral, a comunidade parece ainda estar se expandindo para o espaço de intersecção entre química quântica e dinâmica molecular, tornado possível pelo surgimento da nova categoria de hardwares. A expansão da comunidade, por ser uma tarefa colaborativa laboriosa, não tem condições de acontecer na mesma rapidez com que o hardware se aprimora, de modo que a sugestão, teste e reformulação de descritores novos e antigos ainda é um dos temas dominantes nos círculos de modelagem multi-escala. O desenvolvimento de técnicas que permitam o tratamento de sistemas complexos desta forma rendeu o Prêmio Nobel de Química 2013 a Michael Levitt, Arieh Warschel e Martin Karplus (LEVITT; WARSCHHEL; KARPLUS, 2013). Entre os muitos sistemas que se beneficiaram destas técnicas, encontram-se as proteínas.

Proteínas são heteropolímeros formados por (em geral) até vinte e dois tipos diferentes de monômeros, os aminoácidos. Vinte dos vinte e dois aminoácidos proteinogênicos encontram-se codificados diretamente no código genético e são considerados "padrão". Os outros dois, a Selenocisteína e a Pirrolisina são inseridas na cadeia polipeptídica crescente através de alterações co-trauducionais na interpretação, pelo ribossomo, do códon UGA e UAG respectivamente, ambos interpretados normalmente pelo ribossomo como códons de parada. Estes dois aminoácidos proteinogênicos extra requerem a expansão dos tRNAs expressos normalmente na célula além de genes específicos para tRNA-Aminoacil Sintetases específicas.

Aminoácidos proteinogênicos são monômeros essencialmente compostos por quatro grupos ligantes conectados a um carbono assimétrico, referido como carbono- α : Um hidrogênio, um grupo amina, um grupo ácido e uma cadeia lateral R, específica de cada aminoácido. A estrutura básica é ilustrada na figura 1.

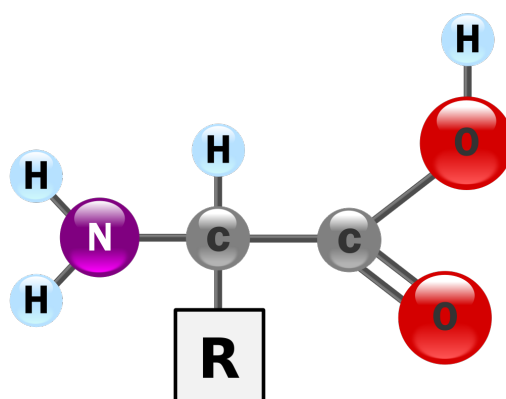


Figura 1 – Esquema mostrando as características comuns aos 20 aminoácidos proteínogênicos.

A sequência de aminoácidos de uma proteína é dita sua Estrutura Primária (ou Sequência Primária) e é a identidade mais básica que uma proteína pode ter. Entretanto, uma proteína não é definida apenas por sua sequência primária, porque proteínas não são fibras esticadas (como o kevlar). As ligações do carbono- α são ligações simples e podem girar em torno do próprio eixo, ao longo de ângulos de torção chamados Ângulos de Ramachandran. O ângulo ϕ é o ângulo de torção em torno da ligação entre o carbono- α e o nitrogênio do grupo amina e o ângulo ψ é o ângulo de torção em torno da ligação entre o carbono- α e o carbono da carbonila do grupo ácido. Existe ainda um ângulo ω entre o carbono da carbonila de um aminoácido e o nitrogênio do grupo amina de outro, os átomos que compõem a chamada ligação peptídica. Em princípio este seria um grau de liberdade torsional a mais, mas existe uma estrutura de ressonância que dá a essa ligação um caráter de ligação dupla, tornando-a mais rígida e proibindo a rotação em torno dela na mesma escala de temperatura das rotações ao longo dos ângulos ϕ e ψ .

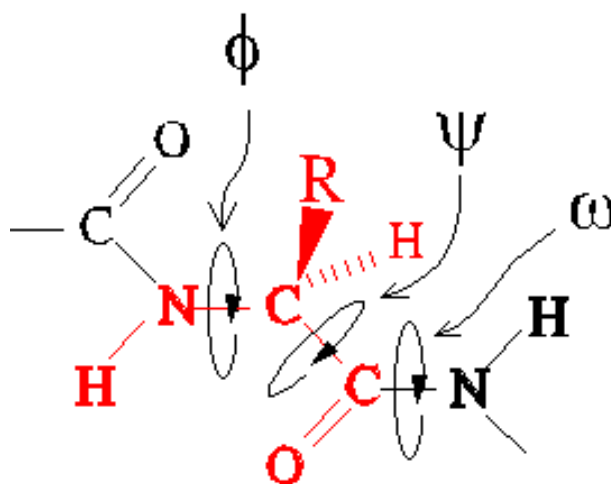


Figura 2 – Ângulos de Ramachandran mostrados em relação a um resíduo de aminoácido (em vermelho).

A liberdade torsional dos ângulos de Ramachandran significa que a cadeia principal das proteínas pode se torcer sobre ela mesma, um processo chamado enovelamento (o termo em inglês *folding* também é muito utilizado). Isto significa que a estrutura geométrica da proteína é, na verdade, um complicado novelo. Esta estrutura geométrica é específica de cada proteína e determina sua função biológica (DILL; MCCALLUM, 2012).

As geometrias na qual uma proteína é biologicamente ativa são chamadas geometrias, conformações ou estruturas nativas, pois é a forma como é encontrada na natureza (DILL; MCCALLUM, 2012; HARTL; HAYER-HARTL, 2009). O padrão geométrico tridimensional nativo de uma proteína é sua Estrutura Terciária. Toda estrutura terciária pode ser subdividida em blocos menores de construção chamados de Estruturas Secundárias, ligados por *loops* (regiões da cadeia de geometria um tanto amorfa, responsável em grande parte pela plasticidade da molécula). Algumas proteínas são compostas por mais de uma cadeia polipeptídica, o arranjo final de cada subunidade interagindo entre si por interações intermoleculares é a Estrutura Quaternária.

Christian Anfinsen foi um dos ganhadores do Prêmio Nobel de Química de 1972 por seus estudos demonstrando como uma amostra de Ribonuclease (uma enzima e, portanto, uma proteína) podia ser desnaturada e que, até certo ponto, readquiria sua atividade uma vez que o agente desnaturante fosse removido, sugerindo que a informação necessária para que uma proteína adquirisse sua configuração final estivesse contida totalmente na sequência primária (ANFENSEN, 1973b).

Anfinsen sugeriu que uma proteína é capaz de, partindo de uma configuração desnaturada qualquer, realizar mudanças conformacionais de forma autônoma e chegar a sua Estrutura Terciária (Figura 3) desde que o meio em que estivesse contida tivesse semelhanças com o meio onde é naturalmente encontrada, pelo menos no que diz respeito a pressão, temperatura e pH.

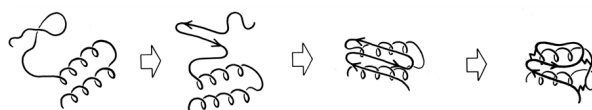


Figura 3 – Esquema ilustrando uma proteína assumindo sua Estrutura Terciária de forma autônoma a partir de uma conformação desnaturada.

Desde que cálculos de mecânica quântica tornaram-se possíveis, existe um interesse em desenvolver descritores quânticos para estas moléculas. Em 2012, Dokholyan e Yin propuseram um método (YIN; DOKHOLYAN, 2011) para utilizar descritores rotacionalmente invariantes calculados a partir de densidades eletrônicas, obtidas por crio-eletromicroscopia, e comparar moléculas segundo estes descritores, essencialmente comparando as duas densidades. A idéia geral do trabalho consistiu em propor um método de calcular os invariantes rotacionais de uma densidade eletrônica experimental e usá-los como "impressões digitais"

para encontrar, em um banco-de-dados, as moléculas cuja densidade são mais semelhantes à densidade de entrada, e assim usar tais estruturas para auxiliar na resolução da densidade de entrada.

O método é baseado em uma medida de distância e é capaz, em princípio, de agrupar as proteínas em grupos de estruturas próximas. Esta coleção de grupos e sub-grupos forma uma estrutura hierárquica que, em princípio, poderia ser utilizada para classificar proteínas segundo suas semelhanças. Embora o trabalho deles seja focado em semelhanças entre descritores da densidade eletrônica, apenas, o resultado de uma busca quase sempre retornou proteínas dentro da mesma classe funcional da proteína de entrada. Este resultado sugere que, de alguma forma, o agrupamento utilizando apenas as densidades eletrônicas resgatou alguma informação a respeito da função das proteínas agrupadas.

Na biologia evolutiva, não é incomum que agrupamentos de proteínas pela sua sequência também incidentalmente as agrupe segundo suas funções biológicas. Isto acontece com notável frequência porque, em geral, proteínas de mesma função possuem uma mesma ascendência evolutiva. Ao mesmo tempo, este tipo de agrupamento não é perfeito porque proteínas de mesma função, ocasionalmente, são fruto do que se chama de evolução convergente. Isto é, múltiplas linhagens de proteínas convergem para desempenhar uma função em particular (ALBERTS *et al.*, 2014).

A obtenção de uma linhagem evolutiva para uma estrutura protéica é possível. A linhagem trata-se de um conjunto de ancestrais cada vez mais remotos que seriam ancestrais comuns de uma coleção crescente de proteínas modernas. Para realizar a comparação utilizando os Invariantes Rotacionais empregados por Yin e Dokholyan neste trabalho foi necessário criar uma forma de obter as estruturas dos ancestrais e, em seguida, suas densidades eletrônicas.

O estudo combinado da ancestralidade e da densidade eletrônica de proteínas é um campo ainda muito novo, que carece de literatura. Apenas recentemente (VALLEAU *et al.*, 2017) é que um trabalho envolvendo a intersecção dos dois assuntos foi proposto para publicação.

Em vista disto, nesta tese foi decidido aplicar os métodos de Yin e Dokholyan em proteínas e seus ancestrais com o objetivo de investigar a ocorrência de algum tipo de padrão comum a todas as densidades de proteínas evolutivamente relacionadas. Uma vez que os métodos de Yin e Dokholyan foram desenvolvidos para a utilização com densidades eletrônicas experimentais, foi necessário implementar um método que pudesse ser utilizado para qualquer densidade eletrônica analítica obtida dos programas de química quântica.

2 Objetivos

Geral

Investigar a possibilidade de que as proteínas possam ser agrupadas segundo sua origem evolutiva através da densidade eletrônica obtida por métodos computacionais de Química Quântica.

Específicos

1. Obter linhagens de tRNA-Aminoacil Sintetases associadas;
2. Utilizar o método I-TASSER de predição de estrutura terciária de proteínas para obter estruturas iniciais das proteínas;
3. Utilizar dinâmica molecular nas estruturas iniciais de todas as proteínas;
4. Obter estruturas finais das trajetórias de dinâmica molecular;
5. Calcular a densidade eletrônica das estruturas finais;
6. Implementar a Projeção da Densidade no Espaço de Canterakis-Zernike;
7. Decompor as densidades eletrônicas das estruturas finais para obter suas invariantes rotacionais;
8. Selecionar um algoritmo de agrupamento;
9. Comparar as densidades obtidas e agrupá-las com base nas invariantes rotacionais;
10. Analisar o agrupamento das densidades eletrônicas de proteínas e determinar se existem padrões evolutivos.

Parte I

Fundamentação Teórica

3 Evolução

Segundo a teoria da evolução biológica pela seleção natural, as espécies que existem hoje são frutos de lentas mudanças graduais que ocorrem em virtude da variação natural entre indivíduos e sua progênie, e o fato que esta variação pode acabar gerando características mais vantajosas. (F; MILLER; SUZUKI, 2000).

Mutações e recombinações genéticas produzem variações na população que são redistribuídas através de fluxo genético (associado com a prática da migração de indivíduos), a deriva genética e o processo de seleção natural. Condições de visibilidade, pressão, clima, disponibilidade de alimento, território e a presença de predadores influenciam nas pressões evolutivas que uma população sofre oriundas da seleção natural, que estabelecem a eficácia com que indivíduos portadores de cada variedade de genes na população são capazes de sobreviver.

O que determina uma espécie é seu genoma. O genoma é o conjunto de todos os genes de uma espécie. Genomas são compostos de cromossomos, que são fitas contínuas de DNA, onde os genes se organizam em sequência. Como os genes se modificam ao longo do tempo, em virtude das pressões evolutivas, estas fitas vão se alterando através de processos de mutações como alguns exemplificados na Figura 4.

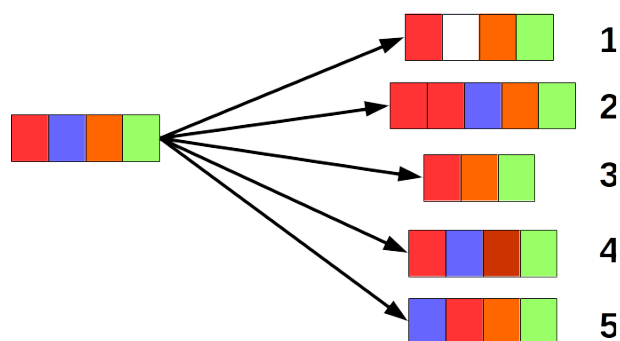


Figura 4 – Exemplo ilustrativo de um cromossomo com quatro genes e cinco processos de mutação que poderiam, teoricamente, ocorrer durante um ciclo de reprodução. As mutações são: 1) Anulação de um gene; 2) Duplicação de um gene; 3) Deleção de um gene; 4) Enfraquecimento de um gene; 5) Troca de genes.

Como as mutações desfavoráveis rapidamente desaparecem, juntamente com seus portadores, da população local, e as mutações vantajosas com o tempo se disseminam, rapidamente as versões mais eficientes dos cromossomos se difundem pela população. Com o tempo, as mudanças acumuladas nos genes alteram o genoma de uma espécie de forma a deixá-lo irreconhecível, uma espécie completamente diferente da espécie ancestral. Este processo de constante mudança pelo qual uma espécie ancestral eventualmente transforma-

se em uma espécie nova é chamado Anagênese (ALBERTS *et al.*, 2014; F; MILLER; SUZUKI, 2000).

Populações sujeitas às mesmas pressões evolutivas tendem a evoluir genomas semelhantes, que produzam genes que conferem características mais adaptadas ao *habitat*. Isto significa que populações de uma mesma espécie que sejam isoladas umas das outras por tempo suficiente sofrem pressões evolutivas diferentes, de modo que seus genomas em geral se modificam de formas distintas, gerando espécies diferentes. Este processo pelo qual uma espécie ancestral dá origem a dois ramos de uma linhagem que eventualmente evoluem em espécies distintas é chamado Cladogênese. Este processo, juntamente com a Anagênese estão ilustrados na Figura 5.

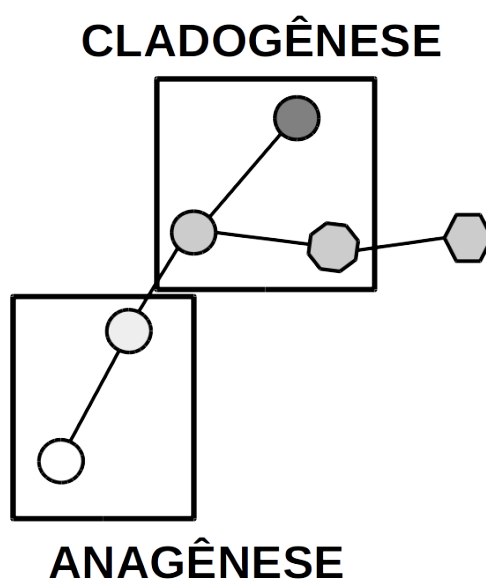


Figura 5 – Esquema ilustrativo indicando os processos de Anagênese e Cladogênese das espécies.

O processo de cópia do DNA pelas células vivas é mais eficiente em prevenir erros e amplificar mutações. Por esta razão, esse processo de convergência da população para variedades mais adaptadas de genes é mais lenta, mas não menos previsível. Com o tempo, é inevitável que as mudanças acumuladas no genoma pelas mutações e seleção natural alterem o genoma de uma espécie de forma a deixá-lo irreconhecível.

Não está claro se, ao seguirmos todos os eventos de Anagênese e Cladogênese, partindo das espécies modernas, em direção ao passado, encontraremos um único ancestral comum a todas as formas de vida. Esta hipótese foi proposta, de fato, mas não é a única. Existe uma série de hipóteses concorrentes sobre a natureza deste Ancestral Comum Universal (ACU) (THEOBALD, 2010; GLANSDORFF; YING; LABEDAN, 2008; WOESE, 1998; VALAS; BOURNE, 2011). Em grande parte, a natureza deste ACU (se este seria uma única espécie ou se é mais complicado que isto) deve-se à importância atribuída à transferência horizontal de material genético entre espécies coexistentes e à natureza

química do processo que deu origem à vida como a conhecemos (MCINERNEY; PISANI; OCONNELL, 2015; LANIER; WILLIAMS, 2017; FORTERRE, 2015; YAO; ODONNELL, 2016; KOONIN, 2014).

Sabe-se que o modelo mais simples da evolução como a transferência puramente vertical de material genético em direção ao futuro é incompleto, em particular quando se tenta inferir sobre o passado em escalas de tempo geológicas. Embora os mecanismos de transferência horizontal dos genes seja um mecanismo importante de difusão de variabilidade genética através dos ramos da árvore da vida (que neste ponto talvez fosse uma teia, mais que uma árvore), ainda é considerado que a transferência vertical de genes é o método dominante na propagação dos genes ao longo das gerações.

Reconstrução de Ancestrais

Embora os conceitos de Anagênese e Cladogênese tenham sido definidos, *a priori*, para espécies inteiras, eles podem ser utilizados também ao se analisar a história evolutiva de genes especificamente. Entre 1963 e 1965, Linus Pauling e Emile Zuckerkandl propuseram a utilização das sequências genéticas em microorganismos, ao invés de suas características morfológicas, para entender como eles se relacionavam entre si (ZUCKERKANDL; PAULING, 1965), uma ideia que acabou sendo o trabalho seminal do campo da filogenia molecular.

Em 1977, Carl Woese, utilizando as técnicas propostas por Pauling e Zuckerkandl, descobriu o reino Archea ao comparar a ancestralidade vertical de RNA ribossomal. A árvore que ele descobriu isola claramente os três grandes reinos e consegue classificar microorganismos de forma a eliminar inconsistências na classificação antiga (WOESE; FOX, 1977). Uma renderização da árvore obtida por Woese pode ser vista na Figura 6.

Como a ASR ignora a transferência horizontal de genes, quaisquer conjuntos de sequências terão um ancestral comum referente à raiz da árvore filogenética, mas quanto mais diferentes forem as sequências de entrada, mais difícil é o processo. Formalmente, dada uma matriz de substituições e um banco de dados de sequências as técnicas de ASR podem gerar todos os ancestrais intermediários para quaisquer par de sequências no banco-de-dados de entrada, mas nem todos eles serão significativos.

Embora a ASR não inclua a transferência horizontal em suas premissas para realizar a reconstrução das sequências ancestrais, ela é uma ferramenta amplamente utilizada no meio da filogenia e genética computacional na identificação de linhagens, com aplicações sólidas (HANSON-SMITH; KOLACZKOWSKI; T JOSEPH, 2010; ASHKENAZY *et al.*, 2012; AKANUMA *et al.*, 2013; ZINN *et al.*, 2015; BAR-ROGOVSKY *et al.*, 2015; WILSON *et al.*, 2015). Isto sugere que, mesmo ignorando a transferência horizontal de genes, a ASR consegue obter resultados bastante significativos.

Uma vez obtida as sequências ancestrais, foi possível utilizar técnicas de predição de estruturas terciárias para a obtenção de estruturas ancestrais, com as quais é possível realizar cálculos de química quântica. Procedendo desta forma, a ASR foi utilizada para estudar as propriedades fotoquímicas do complexo de Fenna-Mathew-Olson em um trabalho recente (VALLEAU *et al.*, 2017). Espera-se, portanto, que a utilização da ASR para estudar a ancestralidade das densidades eletrônicas seja uma metodologia apropriada.

4 Expansão Geral de Uma Densidade Molecular em uma Esfera

No esquema LCAO, a densidade eletrônica em torno de uma molécula em um sistema com M funções de base é expressa como:

$$\rho(\mathbf{r}) = \sum_{r=1}^M \sum_{s=1}^M \rho_{rs} \chi_r(\mathbf{r}) \chi_s(\mathbf{r}), \quad (4.1)$$

na qual $\mathbf{r}$ é um vetor em um sistema de coordenadas molecular, ρ_{rs} denota os elementos da matriz densidade de primeira ordem e as funções $\chi(\mathbf{r})$ são orbitais atômicos da forma expressa na Equação (4.2):

$$\chi_{lm}^n(\mathbf{r}) = R_n(r) \tilde{Y}_l^m(\theta, \phi), \quad (4.2)$$

onde a parte radial $R_n(r)$ depende do conjunto de base utilizado e os termos $\tilde{Y}_l^m(\theta, \phi)$ são os harmônicos esféricos normalizados. O formalismo é válido tanto para orbitais tipo Slater (STOs) quanto para orbitais tipo Gaussiano (GTOs). O somatório na Equação 4.1 ocorre para todas as combinações de pares das M funções de base.

O objetivo da decomposição pretendida nesta tese é produzir uma representação da densidade em que o sistema de coordenadas dos elétrons é totalmente molecular, ao invés de centrado nos átomos. Isto torna mais fácil a comparação de densidades de moléculas distintas, uma vez que coloca todas as densidades na mesma base e sistema de coordenadas. Para uma decomposição deste tipo, utiliza-se como base um conjunto de funções ${}_k\mathcal{F}_l^m$ ortonormais na esfera unitária (isto é $r^* = 1$) de tal modo que:

$$\langle {}_k\mathcal{F}_l^m | {}_{k'}\mathcal{F}_{l'}^{m'} \rangle_1 \equiv \int_0^1 \int_0^\pi \int_0^{2\pi} {}_k\mathcal{F}_l^m(\mathbf{r}) {}_{k'}\mathcal{F}_{l'}^{m'}(\mathbf{r}) r^2 \sin \theta \, d\theta \, d\phi \, dt = \delta_{kk'} \delta_{ll'} \delta_{mm'}. \quad (4.3)$$

O subscrito 1 no *braket* de integração indica que o procedimento é feito na esfera unitária. Para manter a ortonormalidade das funções ${}_k\mathcal{F}_l^m(\mathbf{r})$ dentro de uma esfera de raio arbitrário r^* , elas devem ser re-definidas como:

$${}_k\tilde{\mathcal{F}}_l^m(\mathbf{r}) \equiv \frac{1}{\sqrt{r^{*3}}} {}_k\mathcal{F}_l^m(\mathbf{r}/r^*). \quad (4.4)$$

Para assegurar-se de que as funções ${}_k\tilde{\mathcal{F}}_l^m(\mathbf{r})$ são de fato ortonormais na esfera de raio r^* , basta substituir a equação (4.4) na equação (4.3) e fazer a mudança de variável $t = r/r^*$ e $\mathbf{t} = (t, \theta, \phi)$

$$\begin{aligned}
\langle {}_k\tilde{\mathcal{F}}_l^m | {}_{k'}\tilde{\mathcal{F}}_{l'}^{m'} \rangle_{r^*} &\equiv \int_0^{r^*} \int_0^\pi \int_0^{2\pi} {}_k\tilde{\mathcal{F}}_l^m(\mathbf{t}) {}_{k'}\tilde{\mathcal{F}}_{l'}^{m'}(\mathbf{t}) t^2 \sin\theta \, dt \, d\theta \, d\phi \\
&= \delta_{kk'} \delta_{ll'} \delta_{mm'}.
\end{aligned} \tag{4.5}$$

Assim, a equação (4.5) é reduzida a equação (4.3). Agora, o subscrito r^* no *braket* indica que a integração é feita em uma esfera de raio r^* .

A expansão de uma densidade eletrônica $\rho(\mathbf{r})$ dentro de uma esfera de raio r^* em termos de funções ${}_k\tilde{\mathcal{F}}_l^m$ é dada por:

$$\rho(\mathbf{r}) = \sum_k^K \sum_l^L \sum_{m=-l}^l {}_k\omega_l^m {}_k\tilde{\mathcal{F}}_l^m(\mathbf{r}) \tag{4.6}$$

onde K e L são os valores máximos de k e l , índices da expansão. Na expansão infinita, a densidade de Canterakis-Zernike e a densidade LCAO são idênticas, mas para propósitos computacionais, é necessário truncar a expansão em algum ponto. K e L são parâmetros de entrada do cálculo.

Para obter expressões para os coeficientes ${}_k\omega_l^m$, multiplica-se a equação (4.6) por ${}_{k'}\tilde{\mathcal{F}}_{l'}^{m'}$ a esquerda e realiza-se a integração na esfera de raio r^* , levando em consideração as condições de ortonormalidade dadas na equação (4.5).

$$\begin{aligned}
\langle {}_{k'}\tilde{\mathcal{F}}_{l'}^{m'} | \rho \rangle_{r^*} &= \sum_k^K \sum_l^L \sum_{m=-l}^l {}_k\omega_l^m \langle {}_{k'}\tilde{\mathcal{F}}_{l'}^{m'} | {}_k\tilde{\mathcal{F}}_l^m \rangle_{r^*} \\
&= \sum_k^K \sum_l^L \sum_{m=-l}^l {}_k\omega_l^m \delta_{kk'} \delta_{ll'} \delta_{mm'} = {}_{k'}\omega_{l'}^{m'}
\end{aligned} \tag{4.7}$$

Substituindo a equação (4.7) na equação (4.1), e retirando os apóstrofes sobre os índices, uma vez que eles servem apenas como índices mudos para indicar que ${}_k\mathcal{F}_l^m \neq {}_{k'}\mathcal{F}_{l'}^{m'}$, obtém-se que os momentos das expansões de uma densidade eletrônica LCAO na base das funções ${}_k\mathcal{F}_l^m$ são dados por:

$$\begin{aligned}
{}_k\omega_l^m &= \langle {}_k\mathcal{F}_l^m | \rho \rangle_{r^*} \\
&= \sum_{r=1}^M \sum_{s=1}^M \rho_{rs} \langle {}_k\mathcal{F}_l^m | \chi_r \chi_s \rangle
\end{aligned} \tag{4.8}$$

Portanto é necessário resolver as integrais que aparecem em:

$$\langle {}_k\mathcal{F}_l^m | \rho \rangle_{r^*} = \sum_{r=1}^{M_A} \sum_{s=1}^{M_B} \rho_{rs} \int_0^{r^*} \int_0^\pi \int_0^{2\pi} {}_k\mathcal{F}_l^m(\mathbf{r}) \chi_r(\mathbf{r}_A) \chi_s(\mathbf{r}_B) r^2 \sin\theta \, dr \, d\theta \, d\phi, \tag{4.9}$$

onde introduz-se os índices A e B para denotar que a integral envolve funções centradas em pontos diferentes: as funções da expansão ${}_k\mathcal{F}_l^m$, centradas na origem O , e os orbitais χ_r que estarão centradas em um átomo A e os orbitais χ_s que estarão centradas em um átomo B . Nesta notação A pode ser igual a B , mas não necessariamente, e A e B são diferentes de O , o que dá origem a integrais de dois ou três centros, caras de se computar.

Parte II

Metodologia

5 Cálculo dos Momentos de Canterakis-Zernike

Como visto no referencial teórico, a projeção de uma densidade eletrônica analítica do tipo LCAO na base de Canterakis-Zernike dá origem a integrais complicadas de se resolver, mostradas na Equação (4.9). Aqui propõe-se um algoritmo para resolver o problema de forma a evitar o cálculo direto desta integral e obter os momentos de Zernike computando apenas integrais de uma variável.

5.1 Expansão de Densidades Usando Técnicas de Translação

Para evitar cálculos computacionalmente caros das integrais que aparecem na equação (4.9), um procedimento alternativo pode ser utilizado. Esta técnica se baseia na expansão de produtos de dois orbitais atômicos χ_{lm}^n em termos de fatores radiais f multiplicados por harmônicos esféricos Y_l^m centrados em um terceiro ponto, tomado aqui como sendo a origem do sistema de coordenadas, mostrado na Equação 5.1:

$$\chi_{L_A M_A}^{N_A}(\mathbf{r}_A) \chi_{L_B M_B}^{N_B}(\mathbf{r}_B) = \sum_{l=0}^{\infty} \sum_{m=-l}^l \mathcal{N}_l^m r^l Y_l^m(\theta, \phi) f_{lm}^{N_A L_A M_A N_B L_B M_B}(r). \quad (5.1)$$

Os fatores radiais $f_{lm}^{N_A L_A M_A N_B L_B M_B}(r)$ tornam-se computacionalmente mais simples em um referencial com uma orientação apropriada, com o centro A posto no eixo z e o centro B no plano xz , como mostrado na figura 8. Neste referencial os fatores radiais podem então ser calculados através de relações de recorrência (RICO *et al.*, 2005).

O fato de que este referencial particular em geral não coincide com o referencial molecular onde os átomos estão inseridos não é realmente um problema, uma vez que as rotações dos sólidos harmônicos $r^l Y_l^m(\theta, \phi)$ pode ser realizada de forma eficiente. Esta rotação transforma cada sólido harmônicos $r^l Y_l^m(\theta, \phi)$ em uma combinação linear de harmônicos esféricos regulares com o mesmo l e alguns (em geral todos) valores de m compatíveis. Isto implica que os fatores radiais rotacionados são apenas combinações lineares dos fatores radiais não-rotacionados. Assim, introduzindo uma matriz de rotações $\mathbf{R}^l$ que alterna os harmônicos esféricos normalizados de mesmo l entre ambos os referenciais, a equação (5.1) pode ser reescrita como:

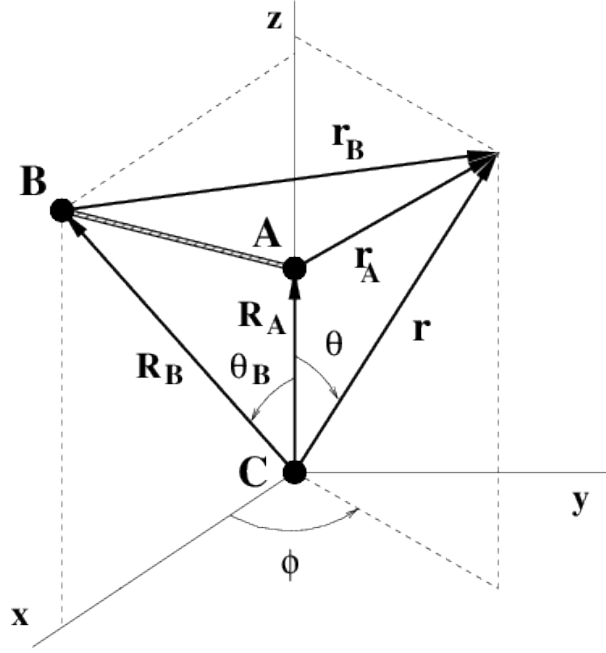


Figura 8 – Representação gráfica do referencial usado na expansão de produtos de orbitais atômicos em termos de Harmônicos Esféricos e fatores radiais

$$\begin{aligned}
 \chi_{L_A M_A}^{N_A}(\mathbf{r}_A) \chi_{L_B M_B}^{N_B}(\mathbf{r}_B) &= \sum_{l=0}^{\infty} \sum_{m=-l}^l [\mathcal{N}_l^m r^l Y_l^m(\theta, \phi)] f_{lm}^{N_A L_A M_A N_B L_B M_B}(r) \\
 &= \sum_{l=0}^{\infty} \sum_{m=-l}^l [\mathcal{N}_l^m r^l \hat{Y}_l^m(\theta, \phi)] \mathbf{R}^l \cdot (\mathbf{R}^l)^t \hat{f}_{lm}^{N_A L_A M_A N_B L_B M_B}(r), \quad (5.2)
 \end{aligned}$$

onde $\hat{Y}$ são os harmônicos esféricos no referencial orientado segundo a figura 8, e $\hat{f}$ são seus fatores radiais associados. Assim, obtém-se os fatores radiais associados a um produto de pares de STOs no referencial molecular, $f_{lm}(r)$, a partir daqueles computados segundo o referencial orientado, $\hat{f}_{lm}(r)$, como:

$$f_{lm}^{N_A L_A M_A N_B L_B M_B}(r) = \sum_{m'=-l}^l \mathbf{R}_{m'm}^l \hat{f}_{lm'}^{N_A L_A M_A N_B L_B M_B}(r) \quad (5.3)$$

Procedendo desta forma, para cada par de centros A e B , os fatores radiais para cada par de orbitais centrados em A e B são computados, multiplicados pela matriz densidade rotacionada e acumulados no referencial orientado. Uma vez que os fatores radiais tenham sido acumulados para todos os pares em A e B , os fatores radiais resultantes são rotacionados de volta para o referencial molecular e acumulados com os fatores oriundos dos outros pares de centros. Ao final do processo, a densidade eletrônica pode ser expressa como uma expansão de um centro no referencial molecular cujos termos são produtos de

Harmônicos Esféricos e fatores radiais $f_{lm}(r)$ de forma que:

$$\rho(\mathbf{r}) = \sum_{l=0}^{\infty} \sum_{m=-l}^l f_{lm}(r) [\mathcal{N}_l^m r^l Y_l^m(\theta, \phi)]. \quad (5.4)$$

Dessa forma, os coeficientes da expansão da densidade completa na esfera de raio r^* pode ser computada como:

$$\begin{aligned} {}_k\omega_l^m &= \langle {}_k\tilde{\mathcal{F}}_l^m | \rho \rangle_{r^*} \\ &= \sum_{l'=0}^{\infty} \sum_{m'=-l'}^{l'} \int_0^{r^*} \int_0^{\pi} \int_0^{2\pi} {}_k\tilde{\mathcal{F}}_l^m(\mathbf{r}) [\mathcal{N}_l^m r^{l'} Y_l^{m'}(\theta, \phi)] f_{l'm'}(r) r^2 \sin\theta \, d\theta \, d\phi. \end{aligned} \quad (5.5)$$

A integração na parte angular é trivial em virtude da ortonormalidade do conjunto dos harmônicos esféricos, resultando em:

$${}_k\omega_l^m = \int_0^{r^*} \tilde{\mathcal{R}}_{kl}(r) r^l f_{lm}(r) r^2 \, dr, \quad (5.6)$$

onde $\tilde{\mathcal{R}}_{kl}(r)$ é a parte radial normalizada da função ${}_k\tilde{\mathcal{F}}_l^m(\mathbf{r})$, que se relaciona com o fator radial da função ${}_k\mathcal{F}_l^m(\mathbf{r})$ como:

$$\tilde{\mathcal{R}}_{kl}(r) = \frac{1}{\sqrt{r^{*3}}} \mathcal{R}_{kl}(r/r^*). \quad (5.7)$$

A integral da parte radial pode ser resolvida numericamente utilizando uma regra de quadratura unidimensional, para as quais a exatidão da integração aproxima muito bem o cálculo direto de formas analíticas. O cálculo final deve ser:

$${}_k\omega_l^m = \sum_{i=1}^{\alpha} w_i^* \tilde{\mathcal{R}}_{kl}(r_i) r_i^{l+2} f_{lm}(r_i), \quad (5.8)$$

onde w_i^* e r_i são os pesos e as abscissas respectivamente da regra de quadratura de ordem α utilizada no intervalo $[0, r^*]$. Em geral, regras de quadratura são definidas no intervalo $[0, 1]$. Em tais casos, realizando uma mudança de variável $t = r/r^*$, a equação (5.8) pode ser reescrita como:

$${}_k\omega_l^m = r^{*l+3/2} \sum_{i=1}^{\alpha} w_i \mathcal{R}_{kl}(t_i) t_i^{l+2} f_{lm}(r^* \cdot t_i), \quad (5.9)$$

sendo w_i e t_i os pesos e abscissas da regra de quadratura de ordem α no intervalo $[0, 1]$, e a relação $w_i^* = r^* w_i$ está sendo levada em consideração.

5.2 Funções de Canterakis-Zernike

As funções de Zernike são funções polinomiais propostas por Fritz Zernike em 1934 para descrever propriedades de difração em microscópios de contraste de fase (ZERNIKE, 1934). Os polinômios de Zernike são funções ortogonais no disco unitário.

Escritos em coordenadas polares os polinômios de Zernike ímpares têm a forma $R_m^n(r)\sin(m\phi)$ e os pares $R_m^n(r)\cos(m\phi)$, onde a parte radial $R_m^n(r)$ é dada por:

$$R_m^n(r) = \sum_{k=0}^{\frac{n-m}{2}} (-1)^k \binom{n-k}{k} \binom{n-2k}{\frac{n-m}{2}-k} r^{n-2k} \quad (5.10)$$

para $n - m$ par e 0 para $n - m$ ímpar. Sendo um conjunto completo no disco unitário, quaisquer imagens 2D que possam ser mapeadas no disco podem ser decompostas em combinações lineares de polinômios de Zernike cujos coeficientes são os momentos de Zernike, que podem ser utilizados para reconstruir a função original de forma aproximada, retendo as características mais gerais e perdendo detalhes. Quanto mais termos forem utilizados na expansão, melhor a reconstrução. Portanto, tal combinação de funções se trata de uma série convergente. Adicionalmente, em seu artigo de 1934, Zernike mostrou ainda como os momentos podem ser utilizados para construir índices que são invariantes sob a rotação no plano, o que permite a comparação de imagens em qualquer orientação.

Em seu relatório técnico de 1999 (CANTERAKIS, 1999), Canterakis chamou atenção à invariância rotacional e à resistência a ruído que as funções apresentadas por Zernike apresentavam e demonstrou uma forma de obter um conjunto de polinômios na esfera¹ unitária que eram análogos aos polinômios de Zernike, mas em três dimensões, descritos nas coordenadas esféricas polares.

De forma análoga às funções de Zernike 2D, os momentos de Canterakis-Zernike podem ser utilizados para construir invariantes rotacionais que podem ser utilizados para comparar sólidos em 3D em qualquer orientação espacial. As funções de Canterakis-Zernike ${}_k\tilde{Z}_l^m(\mathbf{r})$ ortonormais na esfera de raio r^* são definidas como:

$${}_k\tilde{Z}_l^m(\mathbf{r}) = \tilde{Z}_{kl}(r)\tilde{Y}_l^m(\theta, \phi), \quad (5.11)$$

onde $\tilde{Y}_l^m$ são os harmônicos esféricos normalizados:

$$\tilde{Y}_l^m = \mathcal{N}_l^m Y_l^m(\theta, \phi) \quad (5.12)$$

¹ Tecnicamente, as funções de Canterakis-Zernike são os equivalentes das funções de Zernike na *bola* unitária, que seria uma esfera e todo o seu interior, em três dimensões. O termo *esfera*, na Matemática, é utilizado para se referir ao conjunto de pontos definidos por um azimute e uma latitude, sem uma coordenada radial. Nas ciências físicas, entretanto, o termo matemático raramente é utilizado, ficando o termo "esfera" ocupado com ambos os significados. Optou-se por utilizar a nomenclatura mais familiar aos físicos e químicos.

na qual $Y_l^m(\theta, \phi)$ são os harmônicos esféricos reais não-normalizados:

$$Y_l^m(\theta, \phi) = (-1)^m P_l^{|m|}(\cos \theta) \begin{cases} \cos m\phi & (m \geq 0) \\ \sin |m|\phi & (m < 0) \end{cases}, \quad (5.13)$$

onde os termos P_l^m são os Polinômios Associados de Legendre e $\mathcal{N}_l^m$ é uma constante de normalização, dada por:

$$\mathcal{N}_l^m = \sqrt{\frac{(2l+1)(l-|m|)!}{2\pi(1+\delta_{m0})(l+|m|)!}}. \quad (5.14)$$

As funções $\tilde{Z}_{kl}$ são os polinômios radiais de Canterakis-Zernike normalizados, mostrados na Equação (5.15), que são ortonormais na esfera de raio r^* . As funções Z_{kl} propostas por Canterakis são originariamente descritas na esfera unitária, mas é possível escalonar as funções para serem ortogonais em esferas de qualquer raio.

$$\tilde{Z}_{kl}(r) = \sqrt{\frac{3}{r^{*3}}} \sum_{\nu=0}^k q_{kl}^{\nu} \left(\frac{r}{r^*}\right)^{2\nu+l}, \quad (5.15)$$

onde os coeficientes q_{kl}^{ν} tem a forma:

$$q_{kl}^{\nu} = \frac{(-1)^k}{2^{2k}} \sqrt{\frac{2l+4k+3}{3}} \binom{2k}{k} (-1)^k \frac{\binom{k}{\nu} \binom{2(k+l+\nu)+1}{2k}}{\binom{k+l+\nu}{k}} t^{2\nu}. \quad (5.16)$$

Estas funções $\tilde{Z}_{kl}$ podem ser re-escritas de forma mais conveniente em termos de funções g_{kl} como:

$$\tilde{Z}_{kl}(r) = \frac{(-1)^k}{2^{2k}} \sqrt{\frac{2l+4k+3}{r^{*3}}} \binom{2k}{k} \left(\frac{r}{r^*}\right)^l g_{kl} \left(\frac{r}{r^*}\right), \quad (5.17)$$

cujas forma analítica é dada por:

$$g_{kl}(t) = \sum_{\nu=0}^k (-1)^{\nu} \frac{\binom{k}{\nu} \binom{2(k+l+\nu)+1}{2k}}{\binom{k+l+\nu}{k}} t^{2\nu}, \quad t = \frac{r}{r^*} \quad (5.18)$$

Em 2003 Novotni e Klein propuseram um algoritmo numérico para decompor um conjunto de dados volumétricos em um conjunto de Momentos de Canterakis-Zernike (NOVOTNI; KLEIN, 2003), e utilizaram dados volumétricos de formas sólidas para testar seu método. Em 2011, Dohkolyan e Yin aplicaram o algoritmo de Novotni e Klein em mapas de densidade eletrônica de proteínas (YIN; DOKHOLYAN, 2011) e utilizaram as invariantes rotacionais propostas como função de similaridade para procurar proteínas semelhantes em grandes bancos de dados.

O método empregado (YIN; DOKHOLYAN, 2011) consiste em projetar a densidade eletrônica de muitos centros atômicos em um espaço de Canterakis-Zernike, que corresponde a um espaço de densidades de um centro. O resultado é uma versão aproximada da densidade normal, análoga à versão aproximada da figura 2D decomposta em termos das funções de Zernike originais, que contém características mais gerais da densidade e que pode ser utilizada para comparar duas densidade eletrônicas.

Os trabalhos de Yin e Dohkolyan abriram a possibilidade da utilização das funções de Canterakis-Zernike para o estudo da densidade eletrônica de moléculas, criando no processo uma medida de similaridade entre densidades bastante natural, uma tarefa que não é simples (LILIENFELD *et al.*, 2015; K, 2014; MORALES-BAYUELO; MATUTE; CABALLERO, 2015; ANTAL; WARBURTON; MEZEY, 2014). Entretanto, Yin e Dohkolyan utilizaram o método numérico de Novotni e Klein (NOVOTNI; KLEIN, 2003) para computar os momentos de suas densidades, que requer a criação e análise de uma grade 3D da densidade.

Para evitar a realização de uma integração numérica tripla, procurou-se, nesta tese, uma forma mais eficiente de calcular os momentos de Canterakis-Zernike de densidades eletrônicas analíticas, obtidas dos programas MOPAC (STEWART, 2009) e Gaussian2009 (FRISCH *et al.*, 2009), fazendo uso de determinadas simetrias dos Harmônicos Esféricos. Este método tem a vantagem de evitar o cálculo de integrais numéricas de duas ou três variáveis e se ater as de uma única variável.

Todas as densidades decompostas são escritas no método LCAO (do inglês, Combinação Linear de Orbitais Atômicos) como uma combinação linear de produtos de funções de base atômicas do tipo STO ou GTO, multiplicados por um elemento da matriz densidade. Isto possibilita a construção de algoritmos para tratar diretamente as densidades eletrônicas geradas por programas de cálculo de estrutura eletrônica.

5.3 Expansão da Densidade em Funções de Canterakis-Zernike

Substituindo na equação (5.9) o caso em que $\mathcal{R}_{kl}(r) = \tilde{Z}_{kl}$, tem-se as funções de Canterakis-Zernike, denotadas ${}_k\tilde{Z}_l^m$, que são ortonormais no intervalo $[0, r^*]$. Deste modo, a equação (5.9) se transforma na equação (5.19), onde os valores ${}_k\omega_l^m$ são chamados de momentos de Canterakis-Zernike.

$${}_k\omega_l^m = \langle {}_k\tilde{Z}_l^m | \rho \rangle_{r^*} = r^{*l+3/2} \sum_{i=1}^{\alpha} w_i \left[\tilde{Z}_{kl}(t_i) \right] t_i^{l+2} f_{lm}(r^* \cdot t_i) \quad (5.19)$$

A avaliação dos polinômios radiais $\tilde{Z}_{kl}(r)$ de Canterakis-Zernike seguindo a definição expressa na Equação 5.17 leva a problemas de instabilidade numérica para valores grandes de l e k . Como exemplo, a tabela 1 ilustra a avaliação do polinômio radial da função de Canterakis-Zernike dividido por $(r/r^*)^l$ usando o Esquema de Horner (CORMEN *et*

al., 2009) para o caso particular em que $k = 15$ e $l = 15$ e $r/r^* = \frac{1}{2} (1 + \cos \frac{633\pi}{4096}) \simeq 0,942219719359127$. O valor calculado com alta exatidão é: $-0,444\,008\,753\,2 \times 10^{-2}$.

Como é possível ver pela tabela, a perda em exatidão numérica devido a cancelamentos de termos pode eventualmente afetar, como neste caso em particular, todos os algarismos significativos em um cálculo de precisão dupla.

Tabela 1 – Somas parciais na avaliação das funções de Canterakis-Zernike com $l = 15$, $k = 15$ e $r/r^* = \frac{1}{2} (1 + \cos \frac{633\pi}{4096})$ usando o Esquema de Horner.

j	$q_{15\,15}^j$	$\sum_{\nu=0}^j q_{15\,15}^\nu (r/r^*)^{2\nu}$
0	$0,276\,016\,998\,8 \times 10^{12}$	$0,276\,016\,998\,8 \times 10^{12}$
1	$-0,277\,533\,575\,7 \times 10^{13}$	$-0,253\,029\,393\,8 \times 10^{13}$
2	$0,128\,788\,052\,6 \times 10^{14}$	$0,106\,324\,659\,7 \times 10^{14}$
3	$-0,365\,639\,643\,5 \times 10^{14}$	$-0,271\,246\,949\,8 \times 10^{14}$
4	$0,709\,771\,072\,6 \times 10^{14}$	$0,468\,963\,998\,1 \times 10^{14}$
5	$-0,997\,100\,085\,2 \times 10^{14}$	$-0,580\,764\,165\,0 \times 10^{14}$
6	$0,104\,633\,959\,6 \times 10^{15}$	$0,530\,749\,946\,9 \times 10^{14}$
7	$-0,834\,422\,715\,4 \times 10^{14}$	$-0,363\,234\,589\,3 \times 10^{14}$
8	$0,509\,322\,956\,2 \times 10^{14}$	$0,186\,851\,279\,1 \times 10^{14}$
9	$-0,237\,684\,046\,2 \times 10^{14}$	$-0,718\,015\,914\,4 \times 10^{13}$
10	$0,840\,034\,026\,3 \times 10^{13}$	$0,202\,595\,294\,2 \times 10^{13}$
12	$-0,220\,495\,487\,1 \times 10^{13}$	$-0,406\,358\,420\,8 \times 10^{12}$
13	$0,415\,426\,280\,0 \times 10^{12}$	$0,546\,702\,140\,3 \times 10^{11}$
14	$-0,529\,418\,106\,5 \times 10^{11}$	$-0,440\,679\,740\,9 \times 10^{10}$
15	$0,407\,244\,697\,3 \times 10^{10}$	$0,160\,189\,184\,7 \times 10^9$
16	$-0,142\,212\,434\,0 \times 10^9$	$-0,154\,018\,402\,1 \times 10^{-3}$
exato	—	$-0,444\,008\,753\,2 \times 10^{-2}$

Devido a estas instabilidades numéricas, formas recursivas das funções g_{kl} foram implementadas, cujas relações de recorrência encontram-se descritas na Equação (5.20) abaixo:

$$\begin{aligned}
 g_{k+1,l} &= \frac{4k+2l+3}{(2k+2l+3)(2k+1)} \left[\frac{4k(2k+2l+3)+4l(l+2)+3}{4k+2l+1} - (4k+2l+5)t^2 \right] g_{k,l} \\
 &- \frac{4k^2(4k+2l+5)(2k+2l+1)}{(4k+2l+1)(2k+2l+3)(2k-1)(2k+1)} g_{k-1,l}
 \end{aligned} \tag{5.20}$$

,

que podem ser iniciadas com $g_{-1,l} = 0$ e $g_{0,l} = 1$. Finalmente, para evitar a utilização de grandes números fatoriais na equação (5.17), utiliza-se a relação:

$$a_k \equiv (-1)^k \frac{1}{2^{2k}} \binom{2k}{k} = \frac{(-1)^k (k-1/2)!}{\sqrt{\pi} k!} \tag{5.21}$$

, a qual dá origem à uma recorrência simples: $a_{k+1} = -\frac{2k+1}{2k+2} a_k$, iniciada com $a_0 = 1$.

O algoritmo foi implementado pelo Professor Rafael López, da Universidade Autónoma de Madrid, no programa de análise de densidade eletrônica DAMQT 2.1.3 (KUMAR *et al.*, 2015), do qual é um dos autores. A opção de cálculo dos momentos de Canterakis-Zernike na interface gráfica do programa é ilustrada na Figura 9.

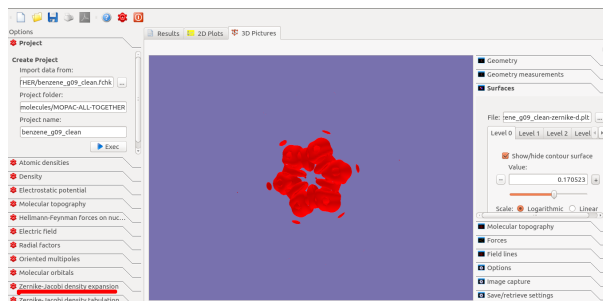


Figura 9 – Amostra de um arquivo de saída do DAMQT.

Apesar de estar implementado no DAMQT, que conta com uma interface gráfica para análises visuais mais finas, o programa de cálculo dos momentos e invariantes rotacionais possui uma forma separada *standalone* em linha de comando, mais apropriada para automatizar a decomposição de um grande número de densidades.

A interface entre os arquivos de saída tipo AUX do MOCAP2012 e os arquivos de entrada SGBSDEN do DAMQT para o cálculo das densidades foi desenvolvida no presente trabalho, mas as interfaces para o Gaussian2009 foram escritas pelo grupo que primeiro idealizou e programou o DAMQT em suas primeiras iterações. Um sumário dos passos envolvidos no cálculo dos momentos de Zernike e das invariantes rotacionais encontra-se ilustrado na Figura 10.

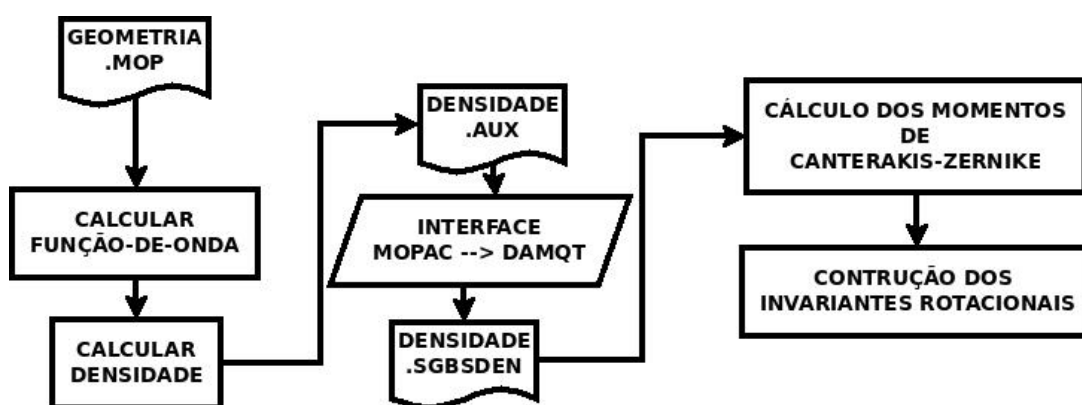


Figura 10 – Fluxograma ilustrando os passos envolvidos no cálculo dos momentos de Canterakis-Zernike e seus invariantes rotacionais derivados.

Os testes da implementação no DAMQT foram realizadas em um subconjunto de 110 sistemas moleculares, contendo entre 2 e 30 átomos, selecionados dentre as moléculas do conjunto de treino do método semiempírico RM1 (ROCHA *et al.*, 2006). Para cada uma

calculou-se a densidade eletrônica utilizando-se o Gaussian 2009 (FRISCH *et al.*, 2009). O método utilizado foi o RHF com a base STO-3G. Calculou-se a densidade eletrônica para as moléculas com ou sem otimização de geometrias e com ou sem solvente, utilizando o algoritmo SCRF para o Modelo do Contínuo Polarizável (PCM, no original em inglês). Dois solventes foram utilizados: água e tetracloreto de carbono.

Todas as combinações dos parâmetros acima foram calculadas, gerando um total de seis densidades para cada molécula. As invariantes rotacionais de cada densidade foram então calculadas com expansão $K = L = 20$ e 512 pontos de quadratura para a solução da equação (5.9). Esses dados foram utilizados para estimar o valor de r^* como função do tamanho da molécula de entrada, bem como testar a habilidade das invariantes rotacionais de agrupar as moléculas em classes de relevância química.

Para as proteínas, as densidades obtidas dos cálculos semiempíricos foram decompostas utilizando expansões $K = L = 5$ com quadraturas de 128 pontos.

6 Agrupamentos Utilizando as Invariantes Rotacionais

Os agrupamentos do conjunto de testes serviram para avaliar a capacidade das invariantes rotacionais em serem utilizadas para agrupar moléculas em classes com propriedades químicas semelhantes. Embora as invariantes já tenham sido utilizadas para comparar moléculas, inclusive com alto grau de acurácia, o agrupamento das moléculas não foi avaliado (YIN; DOKHOLYAN, 2011). Ter um bom método de agrupamento é necessário para que as moléculas sejam postas em classes semelhantes quando comparadas todas simultaneamente.

As densidades do conjunto de testes foram agrupadas segundo suas invariantes rotacionais com o programa de agrupamento Cluto 2.1 (ZHAO, 2003; ZHAO; KARYPIS; FAYYAD, 2005). Apenas as densidades com os mesmos parâmetros do cálculo da densidade eletrônica foram agrupadas. O agrupamento foi efetuado em duas fases, uma particional e outra aglomerativa. A fase particional produziu uma quantidade finita de grupos e a fase aglomerativa do procedimento aninhou os grupos de modo a produzir super-grupos e outros conjuntos de ordem mais alta.

Dentre os parâmetros de entrada, o método necessita de um critério de agrupamento, uma função de similaridade e uma função utilidade. O método de agrupamento escolhido foi o agrupamento direto, em que as moléculas são comparadas diretamente umas às outras uma por uma e agrupadas de modo a otimizar a função utilidade. As moléculas foram comparadas par-a-par seguindo uma função de similaridade. Foram consideradas duas funções de similaridade para objetos tipo vetor, denotado T_i . Os objetos a serem comparados, neste caso as densidades eletrônicas, são representadas como sequências de números que representam "características", as invariantes rotacionais. As duas funções de similaridade são a similaridade cossenóide, mostrada na Equação (6.1)

$$\cos(\mathbf{d}_i, \mathbf{d}_j) = \frac{\mathbf{d}_i^T \mathbf{d}_j}{\|\mathbf{d}_i\| \|\mathbf{d}_j\|}, \quad (6.1)$$

e o coeficiente de correlação, mostrada na Equação (6.2):

$$\text{corr}(\mathbf{d}_i, \mathbf{d}_j) = \frac{\text{Cov}(\mathbf{d}_i, \mathbf{d}_j)}{\sigma_i \sigma_j}. \quad (6.2)$$

As funções de utilidade são funções que quantificam a qualidade do arranjo total de grupos e o objetivo do algoritmo é otimizar a utilidade do agrupamento. Segundo as referências do *software* de agrupamento (ZHAO, 2003), que se encarregou de comparar

suas funções de forma bastante detalhada, as funções de melhor desempenho são a $\mathcal{H}_2$ e a $\mathcal{I}_2$, em combinação com as funções de similaridade, para obter os agrupamentos.

Como foi utilizada uma implementação de agrupamentos não-supervisionados, um dos parâmetros de entrada é o número de grupos no qual se deseja repartir o conjunto de entrada. Por isso foi necessário determinar uma forma de decidir qual o número de grupos a ser escolhido para realizar a análise. Para isto analisou-se as métricas de qualidade disponíveis, exemplificadas na Figura 11.

```

Solution -----
-----
19-way clustering: [H2=1.41e-02] [110 of 110]
-----
cid  Size  ISim  ISdev  ESim  ESdev  |
-----
  0     1 +1.000 +0.000 +0.355 +0.000 |
  1     3 +0.981 +0.002 +0.462 +0.020 |
  2     6 +0.837 +0.049 +0.375 +0.031 |
  3     6 +0.757 +0.035 +0.320 +0.056 |
  4     4 +0.966 +0.004 +0.548 +0.032 |
  5     9 +0.866 +0.026 +0.503 +0.055 |

```

Figura 11 – Amostra de um arquivo de saída do Cluto 2.1, com destaque em vermelho para as métricas dos grupos.

Na figura, o campo *ISim* indica a similaridade média entre os membros de um grupo, enquanto o campo *ESim* o quão próximos, em média, os membros do grupo estão dos elementos externos ao grupo. As colunas *ISdev* e *ESdev* mostram os desvios-padrão dos outros dois campos, respectivamente. Como o conjunto de dados utilizado é parte de um conjunto de parametrização do RM1 e, por sua natureza, contém moléculas de variada natureza química, supondo-se que os grupos em geral não representarão uma fração muito grande do conjunto total, os valores de similaridade externa tenderão a ter desvios-padrão bem maiores que os de similaridade interna. Os grupos, no arquivo de saída, são enumerados aproximadamente da maior similaridade interna para a menor, desempatando com a menor similaridade externa. As medidas de similaridade e desvio-padrão são apresentadas em escalas que vão de 0 a 1.

Baseado nos valores de similaridade interna e desvio-padrão, foi possível criar quatro classes de grupos, sumarizados na Figura 12. As classes são, da esquerda para a direita e de cima para baixo, na Figura 12: Fechado, Sobrecarregado, Artefato e Errático.

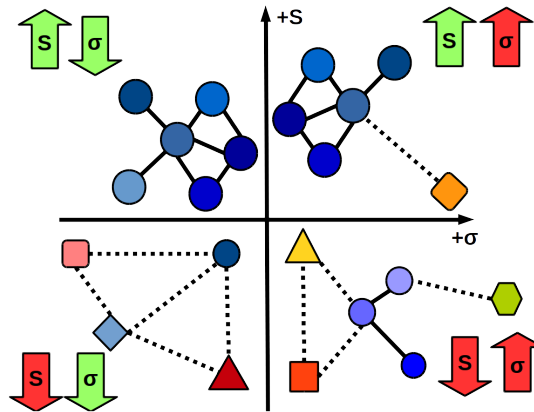


Figura 12 – Tipos de grupos baseados nos valores de similaridade interna (S) e desvio-padrão da similaridade interna (σ).

Grupos Fechados possuem um baixo desvio-padrão e alta similaridade, o que significa que, de modo geral, os elementos do grupo têm alta similaridade entre si e que, como a variação das similaridades em torno da média é muito baixa, não existem *outliers* evidentes. Nos grupos Fechados, não existe dúvida sobre a presença dos membros no grupo.

Grupos Sobrecarregados têm uma similaridade alta mas o desvio-padrão alto indica que pode haver divergências internas e que talvez a inclusão de alguns elementos no grupo seja um artefato, abaixando artificialmente a similaridade.

O grupo tipo Artefato é um grupo que aparece porque o número de grupos é um parâmetro de entrada portanto o algoritmo pode, em determinados momentos, ter que agrupar elementos que nada tem a ver simplesmente porque o número de grupos exigido nos parâmetros de entrada ainda não foi atingido. Isto leva o algoritmo de agrupamento a juntar elementos díspares, produzindo grupos espúrios.

O último tipo de grupo, Erráticos, indica que, apesar de a maioria dos elementos do grupo serem dissimilares, o fato de que existe um grande desvio padrão pode significar que alguns elementos sejam próximos uns dos outros o suficiente para que eles formem um subgrupo Fechado.

Para começar a analisar as classes dos grupos primeiro é preciso determinar o que é um valor "alto" de similaridade interna e de desvio-padrão. Para fazer isto, buscou-se exemplos de grupos de cada classe que tivessem diferentes valores de similaridade e desvio-padrão. Após a decisão sobre estes valores, os agrupamentos foram feitos de modo que os critérios de similaridade e desvio-padrão não fossem violados.

Os grupos até então obtidos são parte da fase particional do algoritmo de agrupamento. Na fase aglomerativa, o algoritmo tenta, respeitando as fronteiras dos grupos, criar uma árvore de similaridade entre cada elemento do conjunto. Isto cria uma estrutura que aninha grupos dentro de grupos para formar super-grupos, super-super-grupos e outras

estruturas de ordem mais alta e, no sentido interno, revelar os subgrupos contidos dentro de cada grupo. A Figura 13 mostra um exemplo de um arquivo de saída, destacando a seção das métricas de qualidade.

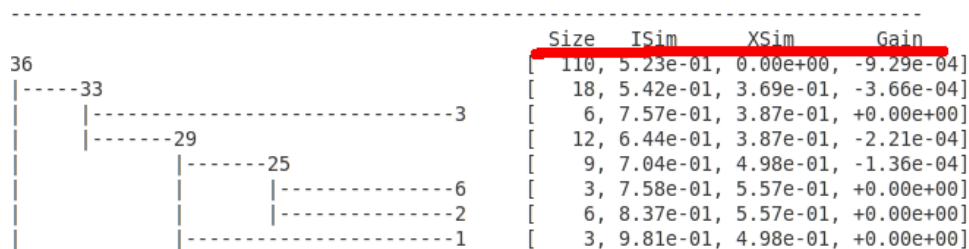


Figura 13 – Seção do arquivo de saída mostrando a solução aglomerativa. Métricas de qualidade de super-grupos estão destacadas em vermelho.

A figura mostra uma árvore de similaridade onde cada nó terminal é um grupo computado na fase particional e cada nó intermediário indica um super-grupo. Os super-grupos, assim como os grupos, têm uma medida de similaridade interna denotada *ISim* nos campos do arquivo de saída, mas também surgem duas novas métricas. A primeira é o *XSim*, que é como um *ESim*, mas envolvendo apenas os nós-filhos do mesmo nó-pai. Por exemplo, na Figura 13, o *XSim* do nó 6 é tomado apenas em relação ao nó 2, pois eles são os únicos nós-filhos do super-grupo 25. A segunda nova métrica é o *Gain*, que indica qual seria a mudança na função-utilidade caso os grupos que fazem parte do super-grupo fossem de fato fundidos. Como as duas funções-utilidades que foram utilizadas tornam-se ótimas em seus máximos, valores negativos de *Gain* indicam perda de utilidade, o que é esperado visto que o arranjo de saída é a solução ótima.

Apesar de todos os grupos estarem conectados uns aos outros em super-grupos cada vez maiores de modo que estejam todos inclusos em um super-grupo de topologia bastante complexa que inclui todos os elementos, isto na prática não acontece pois a maioria dos super-grupos de maior ordem possuem uma similaridade média baixa e desvio-padrão alto, indicando que sua existência é espúria e que se tratam de artefatos da fase aglomerativa do agrupamento. A topologia de grupos real inclui apenas aqueles super-grupos que possuem similaridade suficiente para serem relevantes, e portanto trata-se de uma árvore desconectada, sem raiz.

Nota-se que não existem métricas de qualidade para cada molécula individualmente de modo que os resultados envolvendo sub-grupos são mais difíceis de julgar através de um parâmetro numérico do que os resultados de super-grupos. Isto torna a confiança nos sub-grupos previstos inferior àquela posta no arranjo dos super-grupos.

Com isto, o método proposto para verificar se as invariantes rotacionais são capazes de tomar um conjunto de moléculas quimicamente distintas e classificá-las autonomamente em grupos de algum significado químico consiste em usar as quatro combinações de função

de utilidade e similaridade e verificar se todas produzem agrupamentos cujas classes incluam moléculas quimicamente semelhantes.

7 Obtenção de Estruturas de Proteínas

7.1 Seleção de Conjunto

A seleção do conjunto das proteínas com as quais realizou-se o estudo levou em consideração algumas limitações práticas. Primeiro, evitou-se considerar conjuntos de proteínas de membrana ou metalo-proteínas, visto que a montagem e simulação do sistema nas proteínas de membrana potencialmente tornariam o estudo dependente de simulações de difícil cômputo. No caso das metalo-proteínas, o motivo de evitar sistemas deste tipo é que tais simulações dependeriam de campos de força com parâmetros para metais de transição. A impossibilidade de ajustar a topologia do sistema durante a simulação para acomodar mudanças no número de coordenação dos metais, ou estados de oxidação e de spin torna o desenvolvimento de campos de força para este tipo de sistema uma tarefa difícil, de modo que nem todos os metais possuem parâmetros para simulações de dinâmica molecular. Por esta razão, optou-se por evitar a simulação de sistemas deste tipo.

O conjunto de sequências utilizado foi o dos ancestrais das tRNA-Aminoacil sintetases propostas por Farias e seus colaboradores (FARIAS; GUIMARAES, 2008; FARIAS, 2013). Tratam-se de enzimas de grande importância, sujeitas a grandes pressões evolutivas, que possuem apreciável representatividade nos bancos de dados de sequências, possibilitando um resgate com razoável certeza dos ancestrais das tRNA-Aminoacil Sintetases correspondentes aos aminoácidos principais.

A descendência evolutiva dessas enzimas é objeto de estudos que corroboram a hipótese de que elas se organizam em dois grandes grupos de descendência evolutiva aparentemente distintas (SMITH; HARTMAN, 2015; JAKO *et al.*, 2007; FARIAS; GUIMARAES, 2008; FARIAS, 2013), as tRNA-Aminoacyl Sintetases de Classe I e de Classe II. Isto significa dizer que, filogeneticamente, os ancestrais das enzimas da Classe I devem ser mais próximos uns dos outros que dos ancestrais das enzimas de Classe II.

De forma geral, todas as enzimas específicas para a biossíntese de complexos de tRNA com um determinado aminoácido, independente da espécie, são majoritariamente de uma única classe (FARIAS; GUIMARAES, 2008). A exceção são as tRNA-Lisil Sintetases, que se encontram bem divididas entre ambas as classes (FARIAS; GUIMARAES, 2008).

Para este trabalho utilizou-se este conjunto de enzimas, devido a possibilidade de verificar se uma comparação das densidades eletrônicas é capaz de organizar as enzimas segundo sua filogenia esperada, função desempenhada ou apenas pela forma.

7.2 Preparação de Estruturas

O processo de resgates de ancestrais trabalha apenas com a obtenção de sequências primárias. Em seguida é necessário obter uma estrutura compatível com a sequência para que seja possível calcular as densidades eletrônicas destas molécula.

O problema de se chegar a uma estrutura tridimensional para uma proteína a partir de sua sequência primária não possui solução exata. Felizmente existem vários métodos heurísticos que são bons o suficiente para a predição de estruturas terciárias e quaternárias de proteínas em vários casos (ZHANG, 2008; DROZDETSKIY *et al.*, 2015; KARPLUS, 2009; SAEL *et al.*, 2008; BIASINI *et al.*, 2014; ZHANG; KOLINSKI; SKOLNICK, 2003; ZHANG; SKOLNICK, 2004; COMEAU *et al.*, 2004).

Os modelos tridimensionais foram gerados usando a ferramenta I-TASSER (WU; SKOLNIK; ZHANG, 2007). O I-TASSER trabalha com um índice de confiança nos modelos gerados chamado C-Score, que tipicamente varia de -5 a 2. Um C-Score mais alto significa uma confiança maior no modelo.

Após a geração dos modelos, é necessário obter os dímeros para as sequências cabíveis. As aminoacil-tRNA sintetases de Classe I são, em sua maioria, proteínas de uma única cadeia (aqui chamadas de "Singulares"), com uma fração delas se comportando como homodímeros. As de Classe II, por sua vez, são majoritariamente homodímeros (SMITH; HARTMAN, 2015; JAKO *et al.*, 2007; FARIAS; GUIMARAES, 2008; FARIAS, 2013).

Para saber se uma estrutura quaternária não trivial era necessária, foi feita uma busca no banco de dados de sequências no NCBI, utilizando a ferramenta BLAST. A ferramenta BLAST produz uma lista de sequências do banco de dados que estejam "próximas" de forma estatisticamente significativa da sequência buscada, mas, além disso, também identifica domínios conservados na sequência. Uma verificação no domínio pode facilmente dizer se a sequência apresenta ou não uma potencial interface de dimerização que seja conhecida. A Figura 14 mostra um exemplo da interface do BLAST para os domínios conservados e onde foram encontradas informações relevantes.

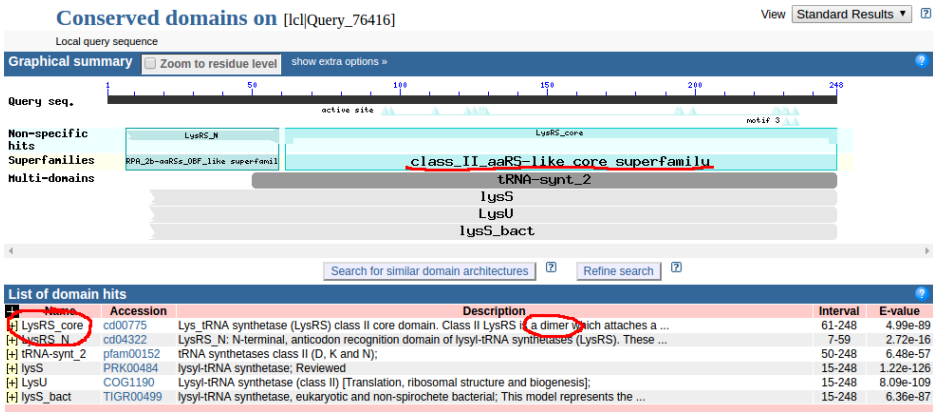


Figura 14 – Exemplo do resultado de um BLAST. A ênfase destaca a indicação de domínios associados com a estrutura quaternária da sequência submetida.

Este método circunspecto, baseado em extrapolações de informações conhecidas, foi utilizado porque um método geral de detecção de interfaces de interação deste tipo não é conhecido e a busca neste campo é bem intensa (GROMIHA; YUGANDHAR; JEMIMAH, 2017; COMEAU *et al.*, 2004; YAN *et al.*, 2017). Utilizando a verificação do BLAST para saber quais proteínas poderiam possuir um sítio de dimerização, atribuiu-se a estrutura quaternária a cada proteína utilizada no conjunto de dados.

As proteínas tRNA-Aminoacil Sintetases catalisam a síntese de complexos envolvendo um grupo aminoacil e um RNA transportador (tRNA) que, por sua vez, é capaz de interagir com o ribossomo durante o processo de biossíntese de proteínas. Assim, a enzima Cisteinil-tRNA Sintetase, por exemplo, catalisa a ligação de uma cisteína citoplasmática livre com seu RNA transportador. Neste trabalho utilizam-se os grupos aminoacil como forma de identificar cada enzima. Estas informações estão sumarizadas na Tabela 2 para as aminoacil-tRNA sintetases diferentes da Lisil-tRNA Sintetase.

Tabela 2 – Estruturas Quaternárias das proteínas modernas estudadas, com exceção das Lisil-tRNA Sintetases.

Grupo Aminoacil	Estrutura Quaternária	
	Modernas	Ancestrais
ARG (R)	Singular	Singular
CYS (C)	Singular	Singular
GLN (Q)	Singular	Singular
GLU (E)	Singular	Singular
ILE (I)	Singular	Singular
LEU (L)	Singular	Singular
MET (M)	Homodimérica	Homodimérica
TRP (W)	Homodimérica	Homodimérica
TYR (Y)	Homodimérica	Homodimérica
VAL (V)	Singular	Singular
ALA (A)	Homodimérica	Homodimérica
ASN (N)	Homodimérica	Homodimérica
ASP (D)	Homodimérica	Homodimérica
HIS (H)	Homodimérica	Homodimérica
PHE (F)	Homodimérica	Homodimérica
PRO (P)	Homodimérica	Homodimérica
SER (S)	Homodimérica	Homodimérica
THR (T)	Homodimérica	Homodimérica

Os resultados para as Lisil-tRNA Sintetases são apresentados separadamente na Tabela 3 porque, neste grupo de estruturas, existe uma parte que pertence à classe I e outra que pertence à classe II e, no entanto, todas são Lisil-tRNA Sintetases e têm um ancestral comum.

Tabela 3 – Estruturas Quaternárias das Lisil-tRNA Sintetases

Proteína	Estrutura Quaternária
Ancestral	Singular
B. subtilis	Homodimérica
T. burgdoferi	Singular
E. coli	Homodimérica
H. salinarum	Singular
T.Pallidum	Singular

Assim como no caso da construção de modelos tridimensionais a partir de sequências, o processo de se detectar interfaces de contato entre proteínas é feito de forma heurística. Saber se uma proteína possui uma superfície de contato não é um procedimento trivial, mas uma vez que haja motivo para crer que duas cadeias polipeptídicas interagem entre si, existem algoritmos para realizar a ancoragem de uma cadeia na outra. O ClusPro (COMEAU *et al.*, 2004) é uma ferramenta que faz isto e foi utilizado para a ancoragem dos homodímeros descobertos pelo BLAST.

A desvantagem de qualquer procedimento de ancoragem molecular é que o procedimento em geral ancora qualquer molécula em qualquer outra e gera um escore para a configuração final baseado em um número de parâmetros empíricos que dependem de cada algoritmo de ancoragem. A questão é que não há uma mensagem de erro ou um sinal claro de que as duas moléculas não interagem apreciavelmente nas temperaturas consideradas. Qualquer molécula pode ser ancorada em qualquer outra.

Para comparar com os ancestrais, é necessário obter as enzimas modernas. Isto pode ser feito escolhendo sequências dentre aquelas utilizadas para reconstruir os ancestrais. Como a média de sequências para cada conjunto de dados foi de cerca de 300 sequências, seria impraticável selecioná-las manualmente.

Para cada banco de dados de enzimas foram selecionadas cinco sequências do banco de dados de forma aleatória. Em cada grupo de cinco, todas foram postas através da ferramenta BLAST do Protein Databank (ROSE *et al.*, 2017) e o número de entradas com E-Value menor que 1×10^{-1} foram contadas. Das cinco, aquela com o maior número de entradas deste tipo no PDB BLAST foi selecionada para passar pelos algoritmos implementados no I-TASSER e no ClusPro (se necessário). Assim, ao final do processo, uma proteína moderna foi obtida para cada família de tRNA-Aminoacil Sintetases.

O motivo para as sequências modernas terem sido selecionadas desta forma tem a ver com a disponibilidade de estruturas tridimensionais que o algoritmo TASSER usa para inferir a configuração geométrica de uma sequência de entrada. Quanto maior a semelhança entre a sequência de entrada e as estruturas tridimensionais disponíveis, maior a confiança no modelo final predito.

Ao final, juntando os ancestrais e as modernas, obteve-se um conjunto com 44 proteínas: 10 proteínas modernas de classe I, 10 proteínas ancestrais de classe I, 9 proteínas modernas de classe II, 9 proteínas ancestrais de classe II, 5 proteínas modernas de Lisil-tRNA sintetases e 1 ancestral de Lisil-tRNA sintetases.

7.3 Dinâmica Molecular

Com os modelos finais criados, seguiu-se para a etapa de dinâmica molecular. Sabe-se que os sistemas químicos não se encontram congelados em um único microestado durante o curso de uma medição usual de suas geometrias, seja através de métodos espectroscópicos ou cristalográficos. Em geral, os resultados que se obtêm deste tipo de medição são uma média ponderada de vários microestados distintos entre os quais as várias moléculas que formam o sistema transitam durante o curso da medição, que ocorre em uma escala de tempo ordens de grandeza acima da escala de tempo envolvida nas transições entre os microestados.

É muito pouco provável que as estruturas geradas pelo I-TASSER representem essa média ponderada de microestados moleculares. O mais provável é que a estrutura produzida seja uma estrutura geométrica e energeticamente aceitável pelos padrões do algoritmo de predição de estruturas terciárias, mas que não necessariamente represente uma estrutura que seria obtida a partir de dados de uma medição em um *ensemble* de estruturas.

As estruturas obtidas pelo I-TASSER e os dímeros montados pelo ClusPro serviram como estruturas iniciais de onde foram propagadas trajetórias de dinâmica molecular em condições de simulação com pressão e temperatura constantes. Estruturas finais foram obtidas das trajetórias para servir de entrada para os cálculos de densidade eletrônica.

Todas as moléculas foram solvatadas em caixas de simulação octaédricas contendo moléculas de água utilizando o modelo TIP3P, que vem implementado nos campos de força do pacote AMBER14 (CASE *et al.*, 2014). As caixas de simulação foram montadas de modo que as arestas estivessem distantes da estrutura por, no mínimo, 5 Å, como mostrado na Figura 15.

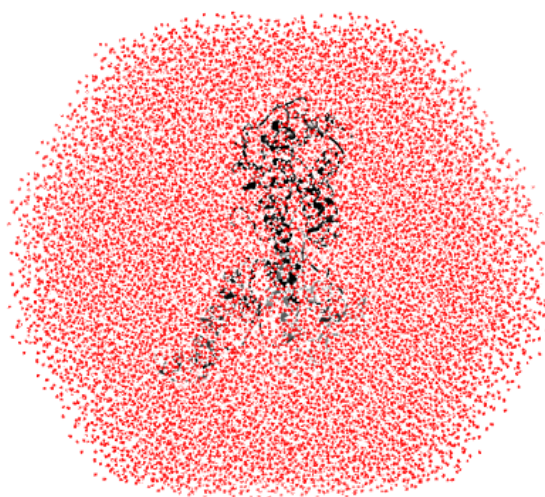


Figura 15 – Exemplo de uma proteína em uma caixa de simulação octaédrica truncada. A proteína está em preto, e a água está colorida por elemento: vermelho para oxigênio e branco para hidrogênio.

Contraíons foram adicionados para neutralizar a caixa de simulação e contrabalançar as cargas das cadeias laterais das proteínas. Isso é necessário para garantir a convergência das interações de longo alcance em sistemas com condições periódicas de contorno. Os hidrogênios foram adicionados pelo software tLeap, que é parte do pacote AMBER14.

O sistema foi submetido a 5000 passos de otimização de geometria dos hidrogênios, dos quais 500 utilizaram o método do máximo declive e o resto o dos gradientes conjugados. Em seguida, 7000 passos de otimização da geometria das moléculas de solvente e contraíons, também com 500 passos utilizando o método do máximo declive. Finalmente, 30.000 passos de otimização da geometria de todos os átomos do sistema simultaneamente, 600 dos quais utilizaram o método do máximo declive.

O sistema foi aquecido com uma dinâmica molecular curta no ensemble NVT de 250 mil passos utilizando o algoritmo SHAKE (RYCKAERT; CICCOTTI; C, 1977) para restringir o estiramento das ligações C—H, que no caso de proteínas normalmente é o grau de liberdade mais rápido do sistema, com estiramentos de período da ordem de uma oscilação por femtossegundo. A utilização do algoritmo SHAKE permitiu a seleção de um passo de dinâmica de 2,0 fs.

Uma vez aquecido, o sistema foi reiniciado a partir do ponto de restauração criado pela trajetória anterior e submetido a uma simulação no *ensemble* NPT. Cinco milhões de passos de 2,0 fs foram simulados, totalizando 10 ns de dinâmica molecular. Os arquivos de reinício foram atualizados a cada 5000 passos e os arquivos de saída e de trajetória foram atualizados a cada 500 passos de trajetória computada, totalizando 10 mil passos de trajetória armazenada por ciclo de estabilização conformacional.

Caso a estabilização não fosse atingida o sistema seria submetido a repetidos ciclos de 10 nanossegundos até que esta fosse atingida. Para todos os conjuntos de dados, o máximo de ciclos necessários foram 2, o que significa dizer que alguns sistemas estabilizaram em trajetórias de 10 ns e outros em trajetórias de 20 ns. Apenas cinco enzimas precisaram de dois ciclos de estabilização, todas as outras atingiram estabilidade conformacional adequada no primeiro ciclo de 10 ns.

A estruturas dos últimos 4000 passos de cada trajetória foram agrupados em *clusters* de estruturas com um critério de proximidade de 1,5 Å. O centróide do *cluster* de estruturas com maior população foi selecionado para ser a estrutura final da trajetória.

A estabilidade conformacional do sistema no segmento de 4000 passos selecionado para os agrupamentos finais foi verificada através da análise de gráficos de RMSD x tempo e pela flutuação no Raio de Giro ao longo deste segmento final da trajetória.

O raio de giro é definido em termos do seu quadrado segundo a equação 7.1:

$$R_g^2 = \frac{1}{N} \sum_{n=1}^N (\mathbf{r}_n - \bar{\mathbf{r}})^2, \quad (7.1)$$

onde N é o número de átomos utilizados no cálculo. No caso específico deste trabalho, estes são os átomos da cadeia principal das proteínas. O vetor $\mathbf{r}_n$ refere-se à posição cartesiana do n -ésimo átomo participante do somatório e o termo $\bar{\mathbf{r}}$ refere-se à posição do centróide de todos os átomos.

No caso da presente análise, $R_g = \sqrt{R_g^2}$ é calculado para todos os 4000 passos retirados para o agrupamento final e sua média e variância ao longo deste segmento da trajetória são calculados. A flutuação do raio de giro está associada à estabilidade conformacional do empacotamento da molécula (LOBANOV; BOGATYREVA; GALZITSKAYA, 2008; JANATI-FARD; REZA; MONHEMI, 2016; VANGA; SINGH; RAGHAVAN, 2015; TEE; RIPEN; MOHAMAD, 2016). Uma flutuação muito grande do raio de giro em torno de sua média implica que o polímero está constantemente se desviando de uma estrutura empacotada final, o que não é o esperado para uma proteína globular solúvel.

7.4 Seleção do Hamiltoniano

Os Hamiltonianos para os cálculos quânticos foram selecionados segundo suas capacidades de descrever adequadamente a estabilidade conformacional da cadeia polipeptídica. A descrição da estabilidade conformacional associada a cada estrutura de uma cadeia requer um modelo físico bastante exato, e para atingir esse grau de exatidão é necessário descrever corretamente as interações não-ligantes, que têm participação decisiva na termodinâmica da estabilidade das estruturas terciária e quaternária de proteínas, particularmente ligações de hidrogênio (ROSE *et al.*, 2006) e interações hidrofóbicas.

Testar métodos conforme sua habilidade de encontrar estruturas semelhantes à nativa é um sub-tópico associado ao problema da predição de estrutura terciária de proteínas, que é um obstáculo na biologia molecular a desafiar pesquisadores em química, biologia e física por mais de meio século (DILL; MCCALLUM, 2012; ROSE *et al.*, 2006; ADHIKARI; FREED; SOSNICK, 2011). Abordagens tradicionais tem tentado encontrar métodos apropriados para amostragens conformacionais e funções de energia efetiva que possam identificar corretamente conformações nativas.

Ainda que o vasto tamanho do espaço de busca e a complexidade da superfície de energia livre tornem especialmente complexa a amostragem extensiva e a otimização global do problema usando energias efetivas baseadas em modelos físicos, estas energias efetivas ainda são instrumentos importantes para o refinamentos de modelos de proteínas produzidos por métodos heurísticos, tais como Rosetta (BRADLEY; MISURA; BAKER, 2005; DAS *et al.*, 2007), TASSER (ZHANG; KOLINSKI; SKOLNICK, 2003; ZHANG; SKOLNICK, 2004; WU; SKOLNIK; ZHANG, 2007) e 3D-SHOTGUN (FISHER, 2003).

Estes métodos heurísticos citados mostraram ser capazes de produzir conjuntos de modelos que contém conformações relativamente acuradas e geometricamente próximas da nativa entre conformações incorretas, mas ainda assim geometricamente razoáveis, chamadas de *decoys*, palavra inglesa que significa "isca" ou "chamariz". Tais estruturas são assim chamadas por serem teoricamente capazes de enganar funções de energia efetiva pouco precisas. Estes conjuntos de *decoys* são criados utilizando-se uma variedade de algoritmos de amostragem (ZHANG; ZHANG, 2010; XU; ZHANG, 2012; SAMUDRALA; LEVITT, 2000; SELLERS *et al.*, 2008; WU; SKOLNIK; ZHANG, 2007).

Pelo fato da predição de estruturas de proteínas ser, em geral, baseada na hipótese termodinâmica de Anfinsen (ANFENSEN, 1973a), uma propriedade crucial de uma função de energia efetiva é sua habilidade de reconhecer quando uma conformação encontra-se em um mínimo de energia livre. Isto significa que um bom teste para uma função de energia efetiva é usá-la para filtrar uma grande quantidade de estruturas e separar aquelas que tenham conformações mais parecidas com a nativa. Essa tarefa é mais difícil se houverem muitos *decoys* no grupo de estruturas a ser filtrado pela função. Conjuntos de *decoys* foram desenvolvidos especificamente para este propósito pelos métodos supracitados e são usados amplamente (SHIROTA; ISHIDA; KINOSHITA, 2011; YASUDA *et al.*, 2011; CHATTERJEE; BHATTACHARYYA; VISHVESHWARA, 2012; ZHOU *et al.*, 2014).

Para um conjunto de conformações próximas, produzidas para terem geometrias razoáveis e com termos entrópicos igualmente favoráveis, a entalpia de formação em solução teria de ser capaz de definir qual a estrutura favorecida e semelhante à nativa. Nesse caso, a entalpia de formação em solução pode ser usada como função de energia efetiva para comparar os métodos em sua habilidade de descrever a estabilidade conformacional da proteína. Ao comparar os métodos segundo a geometria selecionada por eles de um

conjunto de *decoys* em comparação com a estrutura nativa conhecida dentro do conjunto, pode-se ter uma noção da exatidão do método.

Formalmente, a correta descrição das interações não-ligantes requer a recuperação de uma grande parte da correlação eletrônica, utilizando grandes conjuntos de base e métodos de química quântica com alto nível de teoria. Para evitar este tipo de cálculo custoso, correções *ad hoc* na energética das interações por dispersão foram propostas para DFT (GRIMME, 2004; GRIMME, 2006) e posteriormente para métodos semiempíricos (MCNAMARA; HILLIER, 2007; ANIKIN *et al.*, 2012). Correções similares foram publicadas para ligações de hidrogênio (KORTH *et al.*, 2010; KORTH, 2010; REZAC; HOBZA, 2012). Trabalhos posteriores relataram que a acurácia dos métodos semiempíricos que utilizam este tipo de correção é comparável em exatidão à de cálculos DFT corrigidos e até três vezes mais rápidos (YILMAZER; M, 2015).

Visto que os métodos semiempíricos possuem grande capacidade de previsão e dado seu ganho de velocidade sobre cálculos DFT e *ab initio* em geral, considerando a necessidade de se realizar os testes sobre um grande número de *decoys* e em seguida no conjunto em estudo, optou-se pelo uso de hamiltonianos semiempíricos.

As correções para os métodos semiempíricos já foram bastante testadas em sistemas biológicos (MIKULSKIS *et al.*, 2012; YILMAZER; KORTH, 2013; YILMAZER *et al.*, 2015a). Em particular, o método PM6 (REZAC *et al.*, 2009) e sua família de métodos derivados foram amplamente usados para modelar sistemas biológicos e outros sistemas complexos em várias ocasiões (BIKADI; HAZAI, 2009; FANFRLIK *et al.*, 2010; YILMAZER; KORTH, 2013; TEMELSO *et al.*, 2014; DUTRA *et al.*, 2013).

O PM6 também foi utilizado para modelar a geometria de proteínas com bom grau de sucesso (STEWART, 2009). Recentemente, foi relatado que o método PM6-DH+ (KORTH, 2010) (uma das correções do método PM6 para dispersão e ligações de hidrogênio) é uma alternativa rápida e acurada a um tratamento *ab initio* completo das interações entre proteínas e seus ligantes (YILMAZER *et al.*, 2015b).

Em vista disso, decidiu-se utilizar um hamiltoniano da família PM6 para obter as densidades eletrônicas a serem investigadas. PM6-DH+ (KORTH, 2010), PM6-D3 (GRIMME *et al.*, 2010) e PM7 (STEWART, 2013) foram selecionados para serem comparados com o PM6 (REZAC *et al.*, 2009) original em sua habilidade de descrever bem a energética das interações não-ligantes, centrais para a correta descrição da geometria de proteínas.

Todos estes métodos estão implementados no MOPAC2012 (STEWART, 2009; MAIA *et al.*, 2012). Embora o PM6-DH+ e o PM6-D3 sejam correções *ad hoc* para as interações de dispersão, o PM7 apresenta uma nova filosofia de correções, utilizando uma parametrização diferente que visa melhorar a descrição das interações não-ligantes *a priori*.

Visto que a estrutura tridimensional de uma proteína depende intimamente de uma descrição acurada das interações não-ligantes, é razoável supor que o método que mais acuradamente conseguir descrever a energética destas interações também será capaz de fornecer uma boa densidade eletrônica quando comparados com as densidades dos piores métodos, para uma mesma estrutura.

7.4.1 Cálculos

Para realizar a comparação dos métodos, adquiriu-se o pacote de conjuntos de *decoys* tipo II dos Laboratórios Zhang (ZHANG; ZHANG, 2010), gerados pelo algoritmo implementado na ferramenta I-TASSER (WU; SKOLNIK; ZHANG, 2007). O pacote contém 56 conjuntos, mas nem todos estavam em condições de uso para o MOPAC2012. Para alguns conjuntos do pacote, a estrutura nativa estava desfalcada de cadeias laterais inteiras para aminoácidos aparentemente aleatórios, ou ainda desfalcada de átomos na extremidade amino da cadeia, ou apresentavam outras irregularidades estruturais. Com a estrutura nativa em mal estado, tornou-se impossível utilizar estes conjuntos na comparação dos métodos, e eles tiveram que ser abandonados.

Para os 33 conjuntos restantes (Tabela 4), o software UCSF Chimera (PETTERSEN *et al.*, 2004) foi utilizado na adição de átomos de hidrogênio às estruturas nativas. Os estados de protonação dos resíduos de Histidina na cadeia foram determinados levando em consideração limitações estéricas e a possibilidade de formação de ligações de hidrogênio. Subsequentemente, as valências e as pontes de dissulfeto foram verificadas segundo as entradas tipo HEADER nos arquivos PDB das estruturas nativas. Para esta tarefa, utilizou-se o software SCIGRESS. Após a adição dos hidrogênios e verificação das valências, otimizações apenas das posições dos átomos de hidrogênio foram efetuadas com o MOPAC, para cada um dos quatro métodos selecionados, com limiar de convergência para a norma do gradiente de $1,0 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$.

Os métodos foram comparados segundo cálculos *single-point* de entalpia em solução aquosa. Ao todo, contando os quatro métodos através de todos os 33 conjuntos de *decoys*, foram efetuados 61280 cálculos semiempíricos utilizando o MOPAC2012 (STEWART, 2009; MAIA *et al.*, 2012) e o campo de solvente implícito COSMO (KLAMT; SCHUURMANN, 1993) com uma permissividade relativa de 78,0 e um raio efetivo para a sonda do solvente de $1,3 \text{ \AA}$.

Os cálculos foram efetuados utilizando a técnica de escalonamento linear MOZYME (STEWART, 1996), com raio de corte de $10 \text{ \AA}$ para a otimização dos hidrogênios nas estruturas nativas e $9 \text{ \AA}$ para os cálculos *single-point*. Após essa fase, qualquer *decoy* em qualquer conjunto que mostrasse grandes entalpias de formação positivas foram excluídos de seus respectivos conjuntos pois foi verificado individualmente para cada um desses casos que o SCF (do inglês, *Self-Consistent Field*). Método de resolução iterativas das equações

de Hartree-Fock) não convergiu. A Tabela 4 sumariza as características gerais dos conjuntos considerados para o trabalho.

Tabela 4 – Caracterização dos Conjuntos de *Decoys* utilizados na comparação dos Hamiltonianos semiempíricos.

PDB	Descrição	Tipo	Nº Estruturas
1ABV	Amino-término da subunidade δ da F1F0-ATP Sintetase	RMN	528
1AF7	Receptor Quimiotático CheR	Raio-X	528
1AH9	Fator de Inicialização Traducional IF1	RMN	511
1AOY	Amino-término do repressor ArgR	RMN	530
1B4B	Domínio de oligomerização do repressor ArgR	Raio-X	461
1BM8	Domínio ligante de DNA Mbp 1	Raio-X	330
1CEW	Cystatina da clara do ovo de frango	Raio-X	453
1CQK	Domínio anti-corpo CH3	Raio-X	285
1DCJ	Proteína YHHP de <i>Escherichia coli</i>	RMN	526
1EGX	Domínio EVH1 da da VASP Humana	RMN	353
1FAD	Domínio apoptótico FADD	RMN	515
1FO5	Forma reduzida de Dissulfeto Redutase	RMN	342
1G1C	Domínio I1 da titina	Raio-X	308
1GJX	Domínio lipoil da P64K	RMN	526
1GPT	Tionina Gama 1-H	RMN	470
1ITP	Inibidor 1 da Proteínase A	RMN	527
1KJS	Carbóxi-término da C5a	RMN	549
1KVI	Forma reduzida da ATP7A	RMN	551
1MLA	Proteína carreadora de Acil, transacilase	Raio-X	336
1N0U	Fetor de alongamento 2 de levedura	Raio-X	302
1NE3	Proteína ribossomal S28E	RMN	566
1OF9	Toxina formadora de poros da <i>E. histolytica</i>	RMN	508
1R69	Domínio amino-termina do Repressor do Fago 434	Raio-X	292
1SHF	domínio SH3 da proteína Fyn humana	Raio-X	537
1SRO	Domínio ligante de RNA S1	RMN	516
1TFI	Domínio ligante de ácidos nucleicos TFIIS	RMN	340
1TIF	Amino-término do fator translacional IF3	Raio-X	543
1TIG	Carbóxi-término do fator translacional IF3	Raio-X	565
1VCC	Fragmento amino-terminal da DNA Topoisomera 1	Raio-X	552
256B	Citocromo b562	Raio-X	507
2CR7	Domínio PAH do repressor transcricional	RMN	541
2F3N	Domínio SAM da Shank nativa	Raio-X	486
2PCY	Aplastocianina poplar	Raio-X	436

7.4.2 Comparação dos Métodos

Sabe-se há bastante tempo que o *ensemble* nativo de uma proteína é composto por estruturas que estão geometricamente distantes umas das outras numa faixa de até 2 Å de desvio quadrático médio (BROOKS; KARPLUS; PETTITT, 1988). Também é conhecido que a entropia do estado nativo pode ser aproximada pela contribuição de um pequeno

número de conformações do *ensemble* nativo, em relação ao vasto número de conformações não-enoveladas (KARPLUS; ICHIYE; PETTITT, 1987).

Tudo isto significa que o estado nativo deve estar localizado em um poço muito fundo na hipersuperfície de energia livre, o que implica que deve haver uma grande diferença de energia livre entre a estrutura nativa e qualquer estrutura não-nativa mediana.

Para uma população de conformações geometricamente próximas e razoáveis (como os *decoys*), nas quais a componente entrópica não seria tão diferente de estrutura para estrutura, a entalpia de formação teria que ser capaz de determinar qual a estrutura nativa dentre aquelas que satisfazem as limitações geométricas. Desta forma, a nativa também deveria encontrar-se em um poço fundo na hipersuperfície de entalpia, que localmente seria a componente mais variável da energia livre.

Assumindo-se que um *ensemble* aleatório de conformações *decoy* teriam uma distribuição Gaussiana de entalpias de formação, todas as conformações deveriam pertencer a uma população de estruturas com propriedades geométricas e energéticas semelhantes. A nativa, no entanto, deveria estar em um poço de entalpia profundo e fora da população de *decoys*, mesmo que estes fossem geometricamente razoáveis e conformações válidas.

Todo processo de amostragem produz sistematicamente valores que ocasionalmente encontram-se fora da população, chamados de *outliers* ou aberrantes. Adicionalmente, estes valores impactariam fortemente na média porque frequentemente possuem valores extremos. Assim, o primeiro passo para fazer a análise foi remover os *outliers* acima da entalpia média de forma sistemática. Para fazer isto, calculou-se a média da população $\bar{H}$ e o desvio-padrão s , e então os aberrantes com entalpias além do limite de dois desvios-padrão acima da média foram removidos. Então $\bar{H}$ e s foram recalculados e o processo foi repetido até que não houvessem mais *outliers* acima da média da população.

Depois de purificar a população dos aberrantes acima da média, verificou-se se a nativa encontrava-se abaixo da linha de três desvios-padrão abaixo da média desta nova população. O algoritmo completo é mostrado na figura 16. O processo foi chamado de 3SIEVE (*3Sigma Iterative Enthalpy natVe-likeness Exposer*) (URQUIZA-CARVALHO; FRAGOSO; ROCHA, 2016).

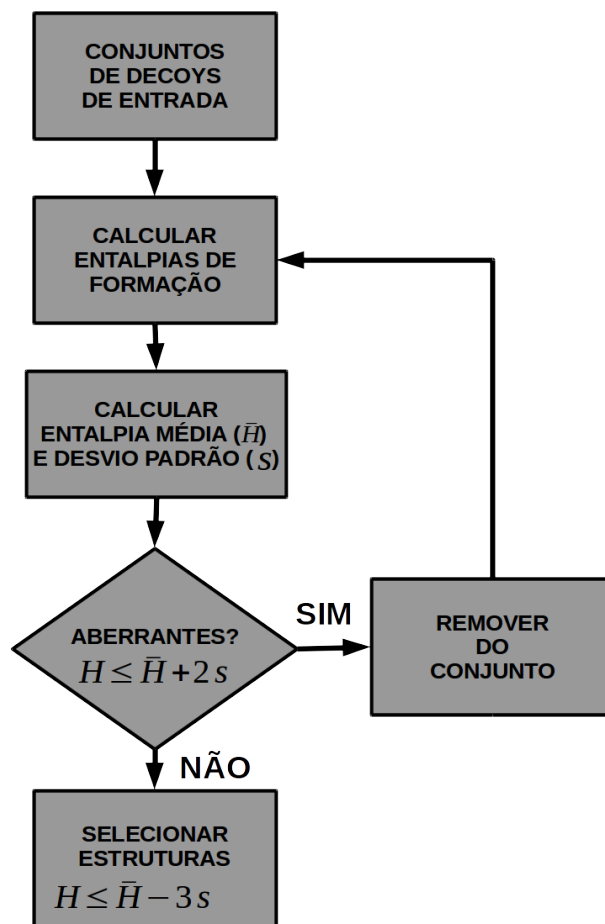


Figura 16 – Representação esquemática do método de análise 3SIEVE.

O procedimento descrito na Figura 16 auxilia a identificação dos métodos capazes de reconhecer a estrutura nativa em meio ao conjunto de *decoys*, como uma forma de quantificar a qualidade da descrição da superfície de energia potencial de cada método.

Após a seleção do hamiltoniano, os cálculos para as estruturas das tRNA-Aminoacil Sintetases modernas e ancestrais foi realizado utilizando os mesmos parâmetros utilizados no cálculo dos *decoys*. Após selecionado o hamiltoniano e geradas as estruturas finais, utilizou-se o programa MOPAC2012 para gerar os arquivos de matriz densidade. As estruturas finais foram submetidas a um cálculo de entalpia de formação no campo de reação COSMO (KLAMT; SCHUURMANN, 1993), utilizando o método de escalonamento linear MOZYME (STEWART, 1996).

Parte III

Resultados

8 Seleção das Estruturas

8.1 Dinâmica Molecular

As trajetórias foram avaliadas segundo a estabilidade conformacional das cadeias polipeptídicas simuladas. As proteínas foram submetidas a ciclos de simulação NPT de 10 ns e tiveram sua estabilidade conformacional verificada após cada ciclo desses através de medidas da Raiz do Desvio Quadrático Médio (RMSD) de cada passo gravado no arquivo de trajetória e o primeiro passo. Isso dá uma medida do quão geometricamente distinto cada passo é dos outros. Apenas cinco proteínas necessitaram passar por dois ciclos de estabilização e seus gráficos de RMSD x Tempo estão mostrados na Figura 17.

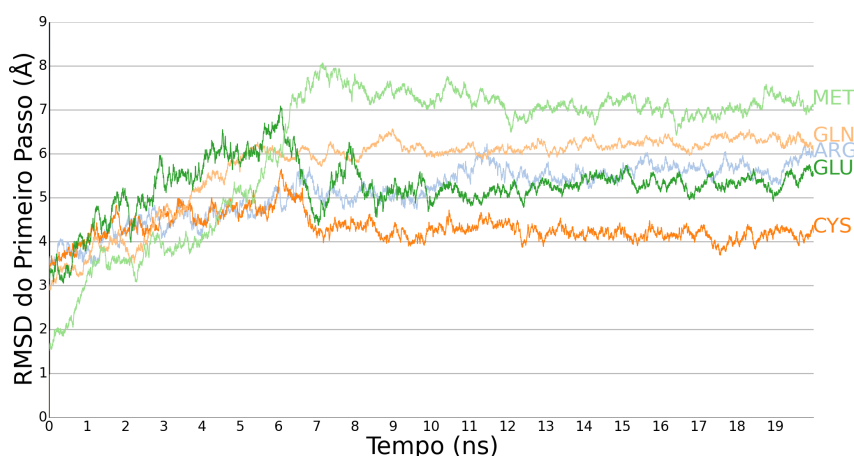


Figura 17 – Gráfico dos RMSD dos átomos da cadeia principal polipeptídica, em relação a primeira estrutura pelo tempo de trajetória simulado. As trajetórias mostradas são referentes às das aminoacil-tRNA Sintetases modernas que estabilizaram após dois ciclos de 10 ns.

Percebe-se como o gráfico atinge um platô após os 10 ns, onde o primeiro ciclo de simulação deveria ter terminado. Essas cinco proteínas são todas modernas e todas de Classe I. Este platô é uma indicação que a estabilidade conformacional foi atingido e que o RMSD entre as cadeias principais das conformações da trajetória e da conformação de partida atingiu um valor relativamente constante. Os gráficos de RMSD x Tempo das demais proteínas estão mostrados na Figura 18.

Nem todas as proteínas tiveram trajetórias estáveis e não puderam ser agrupadas, em particular, o ancestral de Glicil-TRNA aminoacil sintetase não estabilizou e apresentou instabilidades na simulação que prejudicaram os cálculos de dinâmica molecular.

Atribuímos isto ao baixo C-Score da geração das cadeias do I-TASSER. Como trata-se de uma proteína homodimérica, problemas no modelo da cadeia polipeptídica

poderiam prejudicar a interface entre as duas subunidades, gerando um conjunto que não interage, e que, potencialmente, se difundem para longe um do outro durante a simulação, podendo ultrapassar os limites da caixa de simulação, que é o que acontece. Não se sabe ao certo se o problema é a reconstrução do ancestral ou o *protein threading* para gerar a estrutura tridimensional final.

Como o ancestral da glicil-tRNA sintetase foi retirado das análises, decidiu-se também não incluir a versão moderna desta enzima na análise final. Com isso, o conjunto de 44 sequências passou para um conjunto de 42 estruturas finais para a obtenção da densidade eletrônica.

Na Figura 18 pode-se perceber que os platôs se formam bem dentro da margem de 10 ns, indicando estabilização conformacional apropriada. Uma observação mais cuidadosa revela que não apenas o gráfico atingiu um platô, como o desvio no platô é menor que 1,0 Å em relação ao RMSD médio.

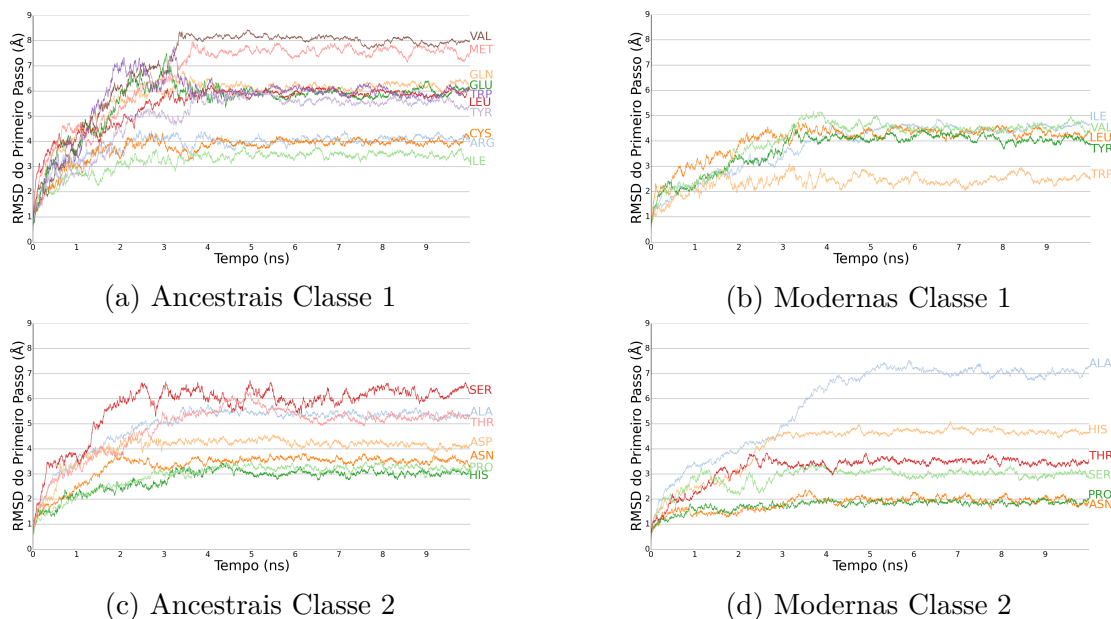


Figura 18 – Gráficos de RMSD x Tempo das proteínas com trajetória de equilíbrio de 10 ns.

Ao contrário do RMSD, o Raio de Giro reduz todas as complexidades conformacionais em um único número e não deve variar de forma abrupta ou expressiva ao longo da simulação, a menos que alguma anomalia conformacional de larga escala se apresente. Com isto em mente e para confirmar a estabilidade conformacional das proteínas, tomou-se o Raio de Giro de todas as proteínas nas simulações NPT e calculou-se o valor médio e a variância do Raio de Giro ao longo da trajetória. Estes resultados estão sumarizados na Tabela 5.

Tabela 5 – Raio de giro médio e variância ao longo da trajetória de produção NPT de cada aminoacil-tRNA Sintetase do conjunto usado no trabalho, em Å.

Proteína	Ancestrais		Modernas	
	Média	Variância	Média	Variância
ARG	23,5555	0,0427	26,0299	0,3074
CYS	22,9345	0,0642	27,6118	0,0245
GLN	49,9247	0,2379	43,5633	0,0833
GLU	21,7624	0,1134	32,6561	0,1198
ILE	29,9416	0,0723	39,8108	0,1793
LEU	25,6617	0,3909	32,5808	0,0878
MET	36,9083	0,3216	40,4784	0,3207
TRP	22,8601	0,4591	33,2740	0,0659
TYR	23,8938	0,0697	41,9653	0,0866
VAL	28,9915	0,0432	38,1827	0,1767
ALA	42,4359	0,0939	44,5658	0,0383
ASN	27,4770	0,0265	29,8644	0,0218
ASP	19,5322	0,0326	31,4127	0,0213
HIS	23,6247	0,0505	32,6198	0,0724
PHE	40,3301	0,0988	38,1822	0,0455
PRO	25,5974	0,0106	30,7596	0,0074
SER	30,5489	0,1141	36,0270	0,0956
THR	41,1952	0,1296	41,8934	0,1829

Como é possível ver na tabela, a variância no Raio de Giro de nenhuma das proteínas foi maior que 0,5 Å, o que mostra que ao longo de toda a trajetória as proteínas se mantiveram igualmente compactas. Isto, somado aos dados de RMSD x Tempo dá uma boa confirmação de estabilidade conformacional, uma pré-condição necessária de ser atingida antes de se realizar o agrupamento e conseguir estruturas finais.

Como o trabalho tem o interesse de estudar elementos evolutivos nas densidades eletrônicas das proteínas, e genes codificam cadeias polipeptídicas únicas (a dimerização ocorre após a Tradução), tomou-se apenas uma das cadeias de cada homodímero estudado para realizar o cálculo das densidades eletrônicas. Esta informação está sumarizada na Figura 19.

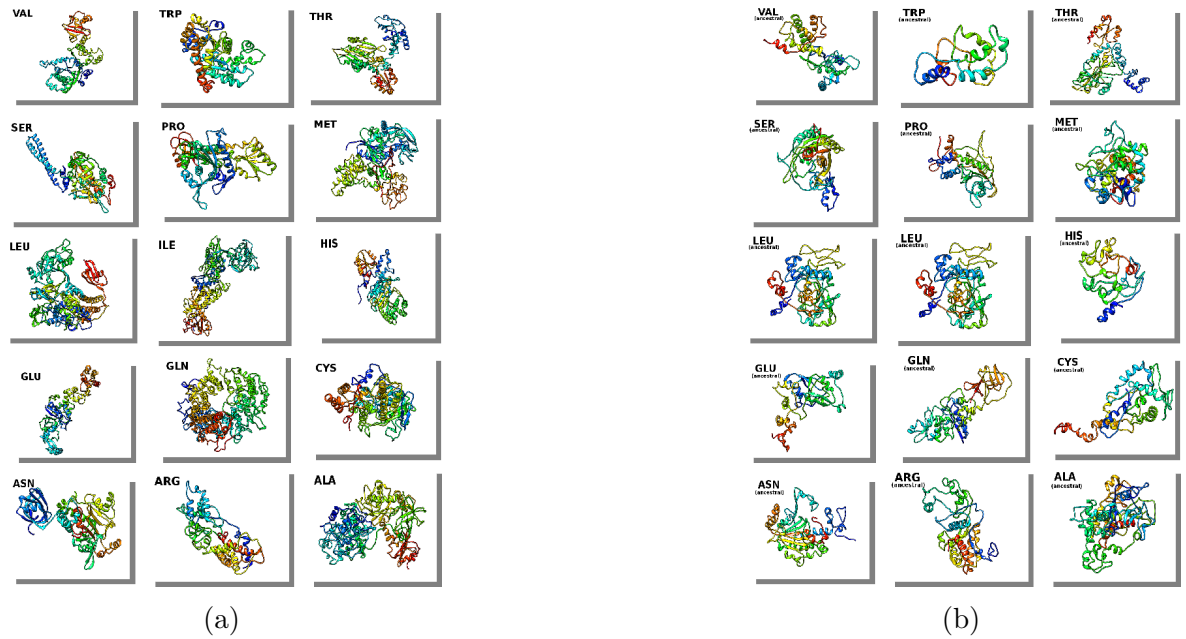


Figura 19 – Cadeias das proteínas ancestrais 19b e modernas 19a a serem utilizadas na obtenção das densidades.

As cadeias das Lisil-tRNA Sintetases modernas são relatadas separadamente, juntamente com seu ancestral, na Figura 20.

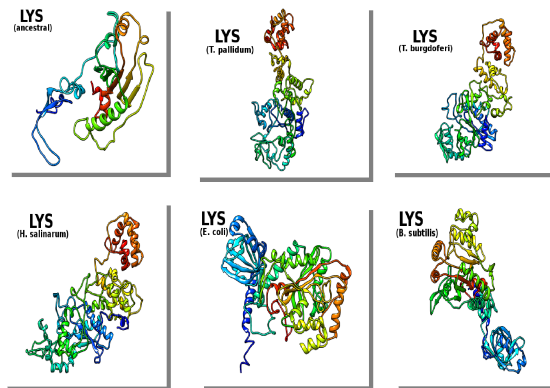


Figura 20 – Cadeias das Lisil-tRNA Sintetases utilizadas na obtenção das densidades.

As estruturas obtidas do agrupamento da dinâmica molecular foram as estruturas de onde se obtiveram as densidades eletrônicas para a decomposição na base de Canterakis-Zernike utilizando o programa DAMQT 2013, como descrito na metodologia.

9 Decomposição da Densidade

Com os momentos de Canterakis-Zernike calculados, é possível reconstruir a densidade eletrônica, mas mesmo nas melhores condições a reconstrução nunca será perfeita. A densidade de Canterakis-Zernike é apenas uma aproximação da densidade LCAO, utilizada aqui nas comparações entre as densidades por ter a propriedade de produzir os rotacionais invariantes. Na Figura 21 pode-se ver que, apesar de manter algumas características esperadas para a densidade da molécula, a densidade reconstruída não é exatamente igual a densidade original.

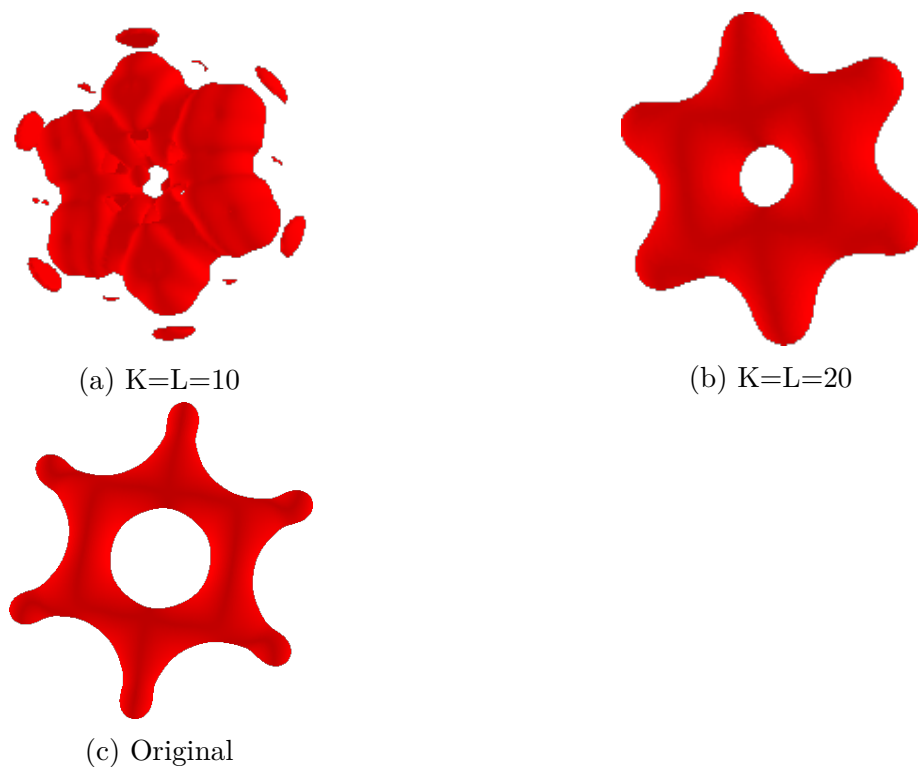


Figura 21 – Formato da densidade do benzeno para quantidades diferentes de funções da base de Canterakis-Zernike. As subfiguras mostram a) A reconstrução de uma densidade para as combinações de $K=L=10$, b) A reconstrução da densidade original com $K=L=20$ e c) A densidade GTO STO-3G original.

Além do tamanho da expansão da densidade, o outro parâmetro que controla a decomposição é o raio de integração r^* . Enquanto alguns termos podem ser adicionados ou retirados da expansão sem alterar outros termos, mudanças no raio de integração modificam todos os momentos da expansão, portanto é necessário ter uma forma padronizada de selecionar um raio de integração que leve em consideração o tamanho do sistema.

9.1 Seleccionando o Raio de Integração

Felizmente existem certas propriedades matemáticas da base de Canterakis-Zernike que podem ser exploradas para que seja possível chegar a uma forma razoável de estimar um bom raio de integração. Sejam as funções g_{kl} descritas na Equação 5.18. Das propriedades dos coeficientes binomiais, temos: $\binom{n}{0} = \binom{n}{n} = 1 \forall n \geq 0$. Com isto, tem-se que $g_{0l} = 1 \forall l \in \mathbb{Z}$ e $l \leq L$, o que leva à conclusão:

$$\tilde{Z}_{0l}(r) = \sqrt{\frac{2l+3}{r^{*3}}} \left(\frac{r}{r^*}\right)^l = \sqrt{\frac{2l+3}{r^{*2l+3}}} r^l. \quad (9.1)$$

Para $l = 0$ tem-se que: $\tilde{Z}_{00}(r) = \sqrt{\frac{3}{r^{*3}}}$, o que leva a concluir que: $\sqrt{\frac{r^{*3}}{3}} \tilde{Z}_{00} = 1$. Os polinômios radiais formam um conjunto ortogonal de função, daí segue:

$$\int_0^{r^*} \tilde{Z}_{kl}(r) \tilde{Z}_{k'l'}(r) r^2 dr = \delta_{kk'} f_{k,l,l}(r), \quad (9.2)$$

,

onde δ_{ab} é o delta de Kronecker e a função $f_{k,l,l}(r)$ representa a integral diferente de 0 para o caso $k' = k$. Por construção sabe-se que $\sqrt{\frac{r^{*3}}{3}} \tilde{Z}_{00} = 1$. Assim, para calcular a integral de qualquer polinômio radial na esfera unitária tem-se que:

$$\int_0^{r^*} \tilde{Z}_{kl}(r) r^2 dr = \int_0^{r^*} \tilde{Z}_{kl}(r) 1 r^2 dr = \sqrt{\frac{r^{*3}}{3}} \int_0^{r^*} \tilde{Z}_{kl}(r) \tilde{Z}_{00}(r) r^2 dr = \sqrt{\frac{r^{*3}}{3}} \frac{\sqrt{6l+9}}{l+3} \delta_{k0}, \quad (9.3)$$

porque, para esse caso, a função $f_{0,0,l}(r) = \sqrt{\frac{r^{*3}}{3}} \frac{\sqrt{6l+9}}{l+3}$.

Os harmônicos esféricos respeitam uma relação de ortonormalidade:

$$\int_0^{2\pi} \int_0^\pi \tilde{Y}_l^m(\theta, \phi) \tilde{Y}_{l'}^{m'}(\theta, \phi) \sin \theta d\theta d\phi = \delta_{ll'} \delta_{mm'}. \quad (9.4)$$

O mesmo raciocínio pode ser aplicado quando tem-se que $Y_0^0 = 1$ e $N_0^0 = \sqrt{\frac{1}{4\pi}}$, que implica em $\tilde{Y}_0^0 = \sqrt{\frac{1}{4\pi}}$ e $\sqrt{4\pi} \tilde{Y}_0^0 = 1$. Logo, a integral de qualquer harmônico esférico ortonormal Y_l^m na superfície da esfera unitária será 0 se $Y_l^m \neq Y_0^0$, ou $\sqrt{4\pi}$, para o caso contrário.

Pode-se ver, portanto, que enquanto $k > 0$ a integral radial será 0, e se $l = 0$, então a integral radial será $\sqrt{\frac{r^{*3}}{3}}$. Como a integral angular só será diferente de 0 quando $l = 0$ e $m = 0$, caso em que será $\sqrt{4\pi}$, chega-se a conclusão que a única função de base que

possui uma integral diferente de 0 é a função ${}_0\tilde{Z}_0^0$. Dessa forma, a integração da densidade completa em toda a esfera de integração produz:

$$\int_{\Omega} \rho(\mathbf{r}) d\Omega = \int_0^{2\pi} \int_0^{\pi} \int_0^{r^*} \rho(\mathbf{r}) r^2 \sin\theta dr d\theta d\phi = \sqrt{\frac{r^{*3}}{3}} \sqrt{4\pi} {}_0\omega_0^0 = \sqrt{\frac{4\pi r^{*3}}{3}} {}_0\omega_0^0. \quad (9.5)$$

Sabe-se que durante o processo de decomposição da densidade em funções de Canterakis-Zernike é necessário truncar o espaço tridimensional e decompor a densidade em um subconjunto de $\mathbb{R}^3$ (CANTERAKIS, 1999). Este subconjunto é uma esfera de raio finito r^* . Isto significa que alguma densidade será deixada de fora da esfera de integração, algo que acontecerá para integração em qualquer subconjunto finito do espaço, uma vez que a integral da densidade eletrônica só será igual ao número total de elétrons do sistema se a integração acontecer em todo o espaço. Para casos finitos, pode-se aumentar a fração da densidade eletrônica considerada na computação à medida que a partição do espaço se torna mais completa, o que significa aumentar r^* .

Uma consequência do fato de se efetuar uma integral em uma esfera de raio arbitrário é que a integração em r sempre será uma função de r^* , a menos que $r^* \rightarrow \infty$, quando a integração tenderá ao resultado exato, e igual ao número total de elétrons, uma constante independente de qualquer variável. A medida que r^* cresce, mais e mais do espaço é levado em consideração na integração e a integral se aproxima assintoticamente de N_{e-} , o número total de elétrons do sistema. Assim, tem-se que:

$$\lim_{r^* \rightarrow \infty} \sqrt{V(r^*)} {}_0\omega_0^0 = N_{e-}. \quad (9.6)$$

Das propriedades dos limites:

$$\lim_{r^* \rightarrow \infty} \sqrt{V(r^*)} {}_0\omega_0^0 = \lim_{r^* \rightarrow \infty} \sqrt{V(r^*)} \lim_{r^* \rightarrow \infty} {}_0\omega_0^0 = N_{e-}. \quad (9.7)$$

Pela forma analítica de $\sqrt{V(r^*)}$, vista na Equação 9.5, observa-se que $\lim_{r^* \rightarrow \infty} \sqrt{V(r^*)} = \infty$, de modo que a função diverge a medida que r^* aumenta. Para que o limite do produto seja convergente, é necessário que $\lim_{r^* \rightarrow \infty} {}_0\omega_0^0 = 0$, na mesma velocidade que $\lim_{r^* \rightarrow \infty} \sqrt{V(r^*)} = \infty$.

A finitude do limite do produto será atingida apenas quando a taxa de crescimento de cada função se equilibrar exatamente a medida que r^* cresce. Escrevendo $\sqrt{V} = cr^{*\frac{3}{2}}$, fica evidente que se trata de uma função polinomial. A taxa de crescimento de uma função polinomial aumenta à medida que o grau do polinômio aumenta, o que significa que

${}_0\omega_0^0(r^*) = Ar^{*-3/2}$, e a equação anterior se torna:

$$\lim_{r^* \rightarrow \infty} \sqrt{V(r^*)} {}_0\omega_0^0(r^*) = \sqrt{\frac{4\pi}{3}} \lim_{r^* \rightarrow \infty} Ar^{*3/2} r^{*-3/2} = \sqrt{\frac{4\pi}{3}} \lim_{r^* \rightarrow \infty} Ar^{*0} = A\sqrt{\frac{4\pi}{3}} = A\sqrt{V(1)}. \quad (9.8)$$

Uma vez que $\rho(\mathbf{r}) \geq 0 \ \forall \ \mathbf{r} \in \Omega$ então à medida que mais e mais densidade é levada em consideração, nunca haverá uma subtração na quantidade total de densidade já acumulada. Isto é o mesmo que dizer que $\sqrt{V(r^*)} {}_0\omega_0^0(r^*)$ será monotonicamente crescente à medida que r^* se aproxima do infinito. Como $\sqrt{V(0)} {}_0\omega_0^0(0) = 0$, para qualquer valor real $\epsilon \in [0, 1]$, tem-se que $\exists r^*_0 \mid \sqrt{V(r^*_0)} {}_0\omega_0^0(r^*_0) = \epsilon \cdot N_{e-}$. O valor de ϵ é a fração da densidade eletrônica contida na esfera de integração. Com isto espera-se um comportamento assintótico de $\sqrt{V(r^*)} {}_0\omega_0^0(r^*)$ à medida que r^* aproxima-se do limite.

Para verificar que o algoritmo de integração reproduz este comportamento esperado, calculou-se a densidade eletrônica da água no vácuo com o software Gaussian 2009 (FRISCH *et al.*, 2009), usando o método de Hartree-Fock restrito com base STO-3G e uma decomposição de $L = K = 15$ com 1024 pontos de quadratura para cada valor inteiro de r^* entre 1 e 10 Å. O resultado deste cálculo é mostrado na figura 22.

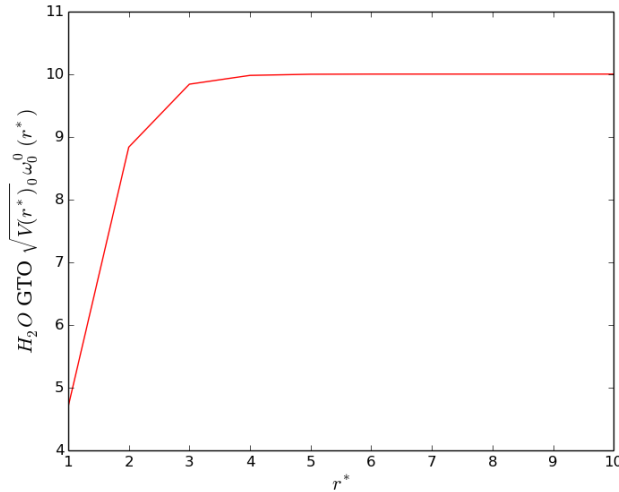


Figura 22 – Gráfico de $\sqrt{V(r^*)} {}_0\omega_0^0(r^*)$ para todos os valores inteiros de r^* entre 1 e 10 Å para a água com expansão $l_{max} = k_{max} = 15$ e a quadratura com 1024 pontos. A densidade original foi computada no Gaussian 2009, usando o método do Hartree-Fock restrito com base STO-3G no vácuo.

Esta propriedade das densidades eletrônicas expressas na base de Canterakis-Zernike foi utilizada para se determinar um valor apropriado para o raio da esfera de integração r^* . Se r^* for muito grande, tempo de computação será consumido sem contribuir para a exatidão do cálculo, e se for muito curto, densidade demais será deixada de fora da esfera unitária, prejudicando a obtenção de descritores da forma da molécula.

As 660 densidades eletrônicas obtidas dos cálculos de densidade eletrônica com as 110 moléculas foram sucessivamente decompostas até que a diferença da quantidade $\sqrt{V(r^*)_0\omega_0^0(r^*)}$ entre dois passos sucessivos fosse inferior a 0,0001, o r^* correspondente ao novo passo seria o raio de integração final r^*_0 .

Esse número r^*_0 foi associado a um valor r_{max} , que representa a distância entre o centro geométrico da molécula e seu átomo mais distante. Calcular r_{max} é simples até para moléculas muito grandes, ao passo que determinar r^*_0 através de repetidas integrações da forma realizada para as moléculas menores é muito custoso. Por esta razão, estas duas quantidades foram relacionadas com o número de funções de base da molécula e esta informação foi utilizada para realizar uma regressão de decaimento exponencial, cujos parâmetros estão mostrados na Equação (9.9).

$$q = Ae^{N_{bases}/t} + q_0, \quad (9.9)$$

que relaciona o número de funções de base N_{bases} ao quociente $q = r^*/r_{max}$ com os parâmetros $A = 3,79 \pm 0,14$, $t = 14,41 \pm 0,66$, e $q_0 = 1,61 \pm 0,02$. O ajuste atingiu a qualidade de $R^2 = 0,803\,944\,733\,992\,954$ e o gráfico mostrado na figura 23.

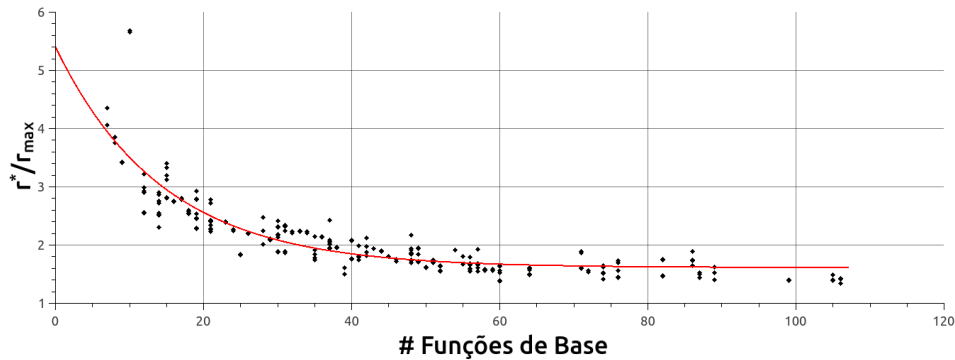


Figura 23 – Gráfico mostrando a relação entre o quociente dos raios e o número de funções de base da molécula. A linha vermelha mostra o ajuste exponencial.

Com isso, torna-se possível estimar o raio de integração sem que seja necessário realizar repetidas integrações. Basta ter uma estimativa da quantidade de funções de base utilizadas para obter o quociente e, juntamente com o raio r_{max} , de fácil obtenção a partir da geometria, é possível estimar o valor de r^*_0 para moléculas grandes.

10 Agrupamentos das Moléculas Pequenas Usando as Invariantes Rotacionais

Após os cálculos de densidade eletrônica com as 110 moléculas do conjunto de treino do RM1, as densidades foram obtidas e agrupadas com o objetivo de verificar se o agrupamento segundo as Invariantes Rotacionais produziria grupos com significado químico.

O primeiro passo do agrupamento envolveu quantificar o quanto seria uma similaridade interna (S) e um desvio-padrão de similaridade interna (σ_S) altos. Primeiro, estabeleceu-se valores de corte para σ_S e S correspondendo ambos a 0,5, que seria a metade da escala.

Em seguida, realizaram-se agrupamentos sucessivos, iniciando com $n = 2$ e incrementando n em uma unidade a cada ciclo. Em cada ciclo, verificou-se se a quantidade de interesse do ciclo (S ou σ_S) violava o raio de corte (para mais no caso de σ_S e para menos no caso de S). Em caso negativo, esta seria a condição de parada do procedimento. Eventualmente, chega-se a n tal que nenhum grupo viola os valores de corte. O algoritmo está sumarizado na Figura 24.

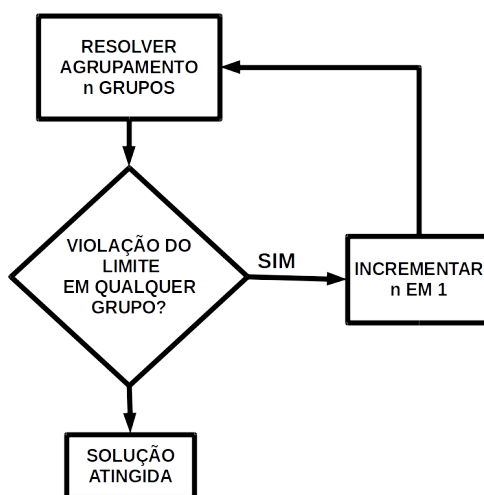


Figura 24 – Algoritmo para a obtenção dos agrupamentos.

Deu-se início a análise verificando os grupos que surgem para diferentes valores de S . O foco foi mais na similaridade interna porque ao contrário do desvio-padrão, esta propriedade também é uma característica que foi calculada para super-grupos, o que significa que é possível avaliar, também, quais são os super-grupos com similaridade razoável, e não apenas os grupos.

Ao analisar os agrupamentos que surgem quando se estabelece o valor de corte

$S \geq 0,5$ imediatamente se percebe que trata-se de um valor muito alto. O conjunto inteiro foi dividido em apenas dois grupos, cada um com $S \approx 0,55$. Dessa forma qualquer desvio-similaridade abaixo de 0,55 é baixa demais e inclui moléculas muito diversas entre si. Sabe-se que as moléculas são diversas entre si porque foram utilizadas no processo de parametrização do método semiempírico RM1 (ROCHA *et al.*, 2006), caso em que uma base de dados composta de moléculas diversas, contendo diferentes tipos de ligações químicas, é preferido. Isso significa que 0,55 ainda é um desvio-padrão que inclui muitas moléculas diversas entre si e que os dois grupos obtidos no processo são claramente Grupos Erráticos, que podem conter grupos Fechados.

Sabendo que 0,55 parece ser uma similaridade limite, abaixo da qual não faz sentido falar sobre grupos, valores limites diferentes foram testados para verificar que tipo de grupos surgiriam. Começou-se com 0,6. Com este valor, foi possível obter amostras de grupos com maior variedade de combinações de S e σ_S .

Começando por cálculos de *single point*, utilizando o coeficiente de correlação e a função $\mathcal{I}_2$, vendo-se na Figura 25 um grupo com $S = 0,769$ e $\sigma_S = 0,034$.

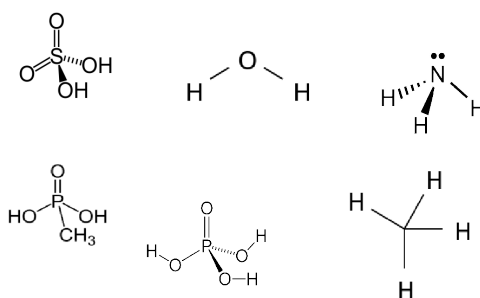


Figura 25 – Grupo com $S = 0,769$ e $\sigma_S = 0,034$.

No exemplo da Figura 25 é interessante observar que o agrupamento foi feito levando em consideração não a geometria molecular, mas a forma da densidade, inclusive os elétrons não-ligantes, como no caso da água e amônia.

Com $S = 0,762$, o segundo grupo, mostrado na Figura 26, também é um caso de grande similaridade. Entretanto, ao verificar visualmente o resultado, precebe-se que se trata de um conjunto bastante heterogêneo, possivelmente um grupo Sobrecarregado, composto, na verdade, de subgrupos que, em virtude de suas altas similaridades internas, compensam as disparidades entre o resto dos elementos.

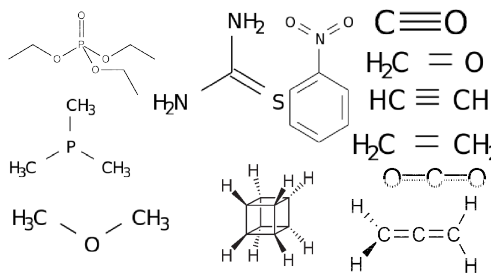


Figura 26 – Grupo com $S = 0,762$ e $\sigma_S = 0,047$.

Apesar da alta similaridade, parece que $\sigma_S \approx 0,05$ é demais para um conjunto deste tamanho. O resto dos grupos são todos de similaridade abaixo de 0,7 e são agrupamentos ainda piores e maiores que estes dois apresentados, para todos os casos. Baixando a exigência de desvio-padrão para $\sigma_S = 0,04$, a mudança é dramática. Grupos grandes se esfacelam rapidamente em dezenas de grupos menores, todos eles, aparentemente, Fechados.

As moléculas que participam do resto dos grupos mudam bastante de grupo entre as funções e as medidas de similaridade. De modo geral as variações sempre apresentam-se como conjuntos contendo moléculas cuja similaridade química pode ser notada.

De modo geral, pode-se dizer que tanto o coeficiente de correlação quanto a distância cossenóide se saíram igualmente bem em agrupar os elementos do conjunto de moléculas, que coincide com o que existe na literatura (SAEL *et al.*, 2008) para as invariantes rotacionais calculadas segundo o algoritmo de Cubatura (NOVOTNI; KLEIN, 2003).

Como tanto o cosseno quanto o coeficiente de correlação são igualmente bons, restringiu-se as próximas discussões sobre efeitos de otimização e solvatação para combinação de distância cossenóide com a função de similaridade $\mathcal{H}_2$, por motivos de eficiência, já que o cálculo da distância cossenóide é mais rápido.

Muitos arranjos de grupos e super-grupos específicos se conservam através de todas as combinações de parâmetros de computação, mesmo a estrutura de subgrupos que o algoritmo aglomerativo tenta inferir. Estes casos são notáveis porque não é regra que as topologias fiquem constantes ao longo das variações de parâmetros. O efeito do solvente não é apreciavelmente detectado de modo que todos os agrupamentos de *single point* variando apenas o solvente acabaram sendo idênticos.

Um exemplo é o grupo que inclui o Imidazol, o Pirrol e o Furano. A topologia de similaridade destes três compostos cíclicos de cinco membros com heteroátomos na cadeia principal tem a mesma topologia em todos os parâmetros

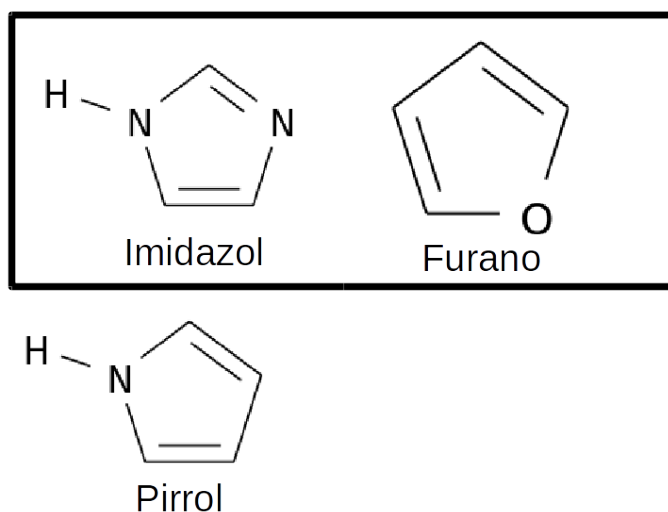


Figura 27 – Agrupamento de núcleos heterocíclicos para qualquer combinação de parâmetros.

Percebe-se que o Imidazol e o Pirrol não são mais próximos um do outro que o Imidazol do Furano. Uma investigação da densidade eletrônica no plano do anel, mostrada na Figura 28, nos revela porque o agrupamento ocorre desta forma.

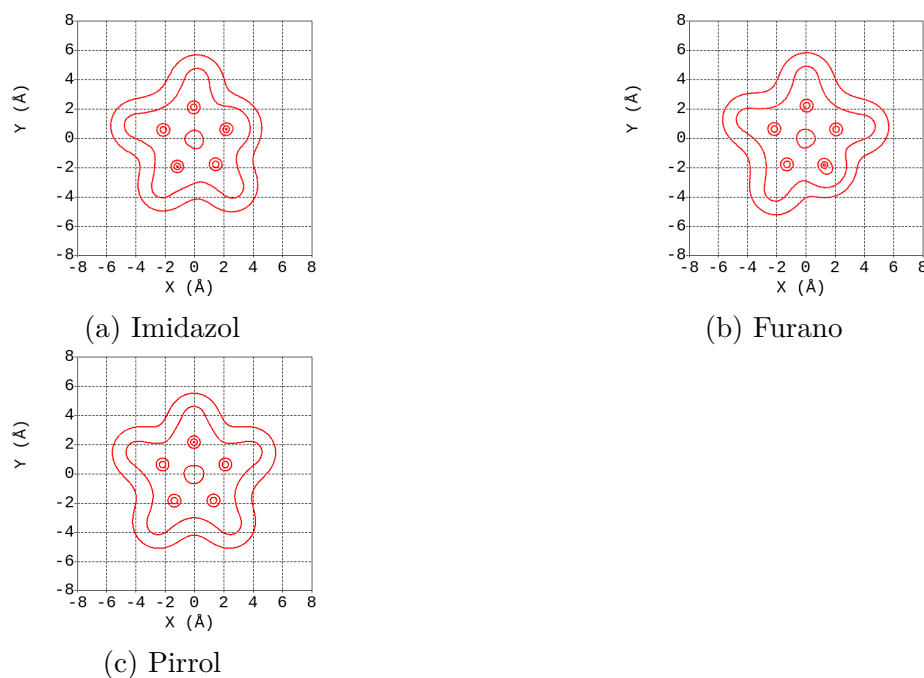


Figura 28 – Mapa de contornos da densidade eletrônica no plano do anel heterocíclico do grupo dos heterociclos de cinco membros.

Segundo a comparação utilizando as invariantes rotacionais, a presença de um heteroátomo cujos elétrons não-ligantes participam do sistema π do anel é a chave da semelhança. Uma possível explicação, ao olhar os mapas de contorno das densidades, é supor que o nitrogênio piridínico do imidazol se comporta mais como o oxigênio do furano. A presença de insaturações muda sobremaneira a forma da densidade.

A Figura 29 mostra um grupo de moléculas lineares com parâmetros e similaridade interna acima de 0,9.

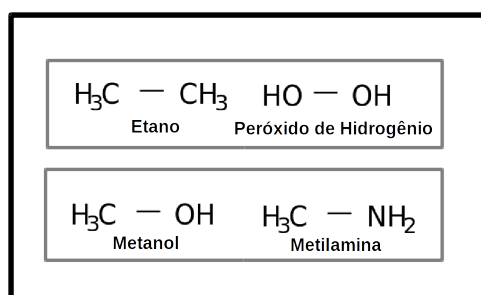


Figura 29 – Agrupamento de moléculas lineares sem insaturações. As caixas em cinza ilustram a estrutura de subgrupo sugerida.

A Figura 29 também ilustra a separação dos subgrupos sugerida pelo algoritmo de aglomeração. Normalmente essa estrutura interna não possui critérios de qualidade que possam ser comunicados ao usuário para serem avaliados, mas nesse caso esse arranjo de separação é constante para qualquer combinação de parâmetros. Observando atentamente,

as moléculas estão arranjadas, basicamente, pela polaridade. Moléculas polares permaneceram em um sub-grupo distinto das moléculas apolares. Apesar de serem formados por moléculas lineares, este grupo é diferente do grupo de moléculas lineares mostrado na Figura 30.

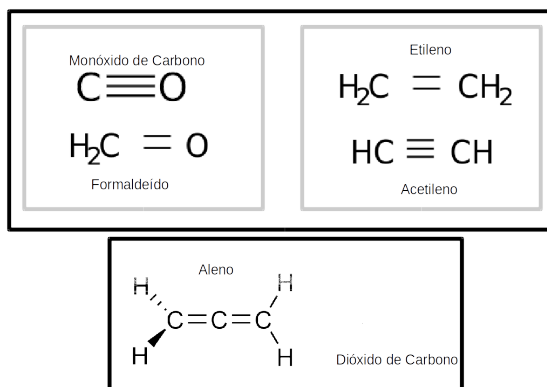


Figura 30 – Super-grupo de moléculas lineares contendo insaturações. O delimitador preto mostra onde cada grupo pertencente a esse super-grupo começa e termina.

Neste caso, tem-se um super-grupo onde as moléculas estão separadas em grupos de moléculas que contém dois ou três centros ao longo do eixo de simetria. As de dois centros se encontram adicionalmente subdivididas segundo polaridade. Pode-se supor que a divisão de acordo com a presença de insaturações é mais severa que a divisão segundo a polaridade.

Dentre as moléculas que mudam de grupo quando sofrem otimização de geometria, o efeito do solvente também é um grande fator de diferenças. Ao todo 46 moléculas mudaram de grupo comparando-se os agrupamentos feitos com densidades de *single point* em fase gasosa e aqueles feitos com as densidades oriundas de uma otimização de geometria nas mesmas condições. Apenas 2 grupos permaneceram inalterados, que são os grupos contidos no super-grupo da Figura 30 e 6 grupos permaneceram como estavam mas sofreram adição de outras moléculas de outros grupos que desapareceram.

A maior parte de todas as moléculas que mudam de grupo com a otimização de geometria são moléculas alifáticas com grande número de graus de liberdade torsionais, e em geral são polares. Algumas destas moléculas estão apresentadas na Figura 31.

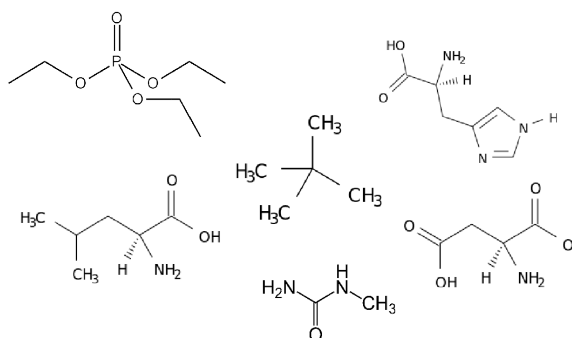


Figura 31 – Subconjunto das moléculas que mudam de grupo quando comparadas as densidades de *single point* e aquelas de geometria otimizada.

Com isso pode-se argumentar que os agrupamentos funcionam melhor quando utilizados em moléculas cujas variações nos graus de liberdade torsionais não representam grande variação na geometria e, portanto, na forma da densidade eletrônica. Para cadeias alifáticas, a medida que o número de graus de liberdade torsionais aumenta, o espaço conformacional da molécula passa a conter estruturas muito distintas, que possuem densidades eletrônicas com características diferentes.

Comparações entre moléculas alifáticas ainda é possível, mas, neste caso, precauções precisam ser tomadas para a análise do problema. Pode ser argumentado, por exemplo, que nem todas as conformações são preferidas e que uma comparação racional entre duas cadeias alifáticas deveria tomar em conta apenas este subconjunto das conformações, e comparar as moléculas apenas no contexto destas conformações preferidas, ao invés de levar todo o espaço conformacional de cada molécula em questão.

Para utilizar-se esta forma de comparação, é necessário ser capaz de particionar o espaço conformacional entre as conformações que são preferidas e as que não são. Para obter esta informação, é preciso que a função de partição de cada molécula seja computada.

Visto que a seleção do *ensemble* no qual simular as moléculas é um parâmetro livre da simulação, requer-se que este seja especificado no momento da simulação. Modificar os parâmetros da simulação modifica o *ensemble* amostrado, uma vez que a molécula apresentará diferentes preferências para cada conformação em diferentes condições de solvatação, temperatura, pressão e concentração, devido a vínculos termodinâmicos.

A liberdade de selecionar os parâmetros da simulação significa que, para um caso geral, não existe um conjunto obviamente inambíguo de conformações preferidas para serem comparadas. As moléculas podem ser comparadas através dos métodos descritos aqui em diferentes *ensembles* e esta comparação conterà informações sobre as diferenças químicas entre as duas cadeias alifáticas entre cada molécula e cuidado deve ser tomado quando comparando duas moléculas simuladas em ambientes diferentes.

Além do problema de definir as condições preferidas de simulação, existe ainda o problema da exatidão. O conjunto de conformações é obtido essencialmente agrupando-se as conformações atravessadas pela simulação. O algoritmo da simulação é feito de tal forma que no limite da simulação de tempo infinito, a função de partição aproximada converge para a exata, seguindo o princípio ergódico. Isto significa que, para um tempo finito, conformações com preferência muito baixa no *ensemble* da simulação podem não ser visitadas sequer uma vez durante a simulação inteira. Quando mais de uma conformação têm níveis similares de preferência, decidir qual delas é preferível pode não ser um problema trivial.

Em resumo, a comparação de moléculas alifáticas é possível, mas muito mais complexa. A grande quantidade de variações que se pode analisar de tais moléculas torna quase impossível qualquer análise significativa. Entretanto, à medida que os graus de liberdade são reduzidos, pode-se melhorar o processo de comparação. As classes de sistemas alifáticos para as quais uma comparação simples de agrupamento como a realizada aqui pode ser feita com maior possibilidade de sucesso são aquelas que possuem as seguintes características:

- Existe, para a molécula, condições de simulação preferidas e bem definidas, permitindo que a seleção do *ensemble* estatístico no qual simular o sistema não seja arbitrária;
- Existe uma conformação preferida dominante que a molécula irá assumir em suas condições de simulação preferidas, o que significa que a função de partição aproximada será menos provável de ser imprecisa para níveis razoáveis;
- O espaço torsional em volta da conformação dominante deve conter apenas conformações similares, de modo que haja confiança que os graus torsionais, por mais numerosos que sejam, não estão sujeitos a variações expressivas nas condições preferidas de simulação.

As classes de cadeias alifáticas que satisfazem estes três requisitos são aquelas com conformações bem definidas o suficiente para serem utilizadas em agrupamentos e comparações similares às realizadas anteriormente. Cadeias polipeptídicas, de forma notável, satisfazem todas essas características.

11 Seleção do Hamiltoniano Para Geração de Densidade Eletrônica de Proteínas

Ao avaliar a qualidade dos hamiltonianos em sua capacidade de funcionar como função-escore, os métodos foram testados de duas formas: (1) suas capacidades de detectar a nativa no conjunto e (2) a qualidade geral dos modelos selecionados na ausência da estrutura nativa. Esta parte da tese deu origem a um trabalho publicado em conjunto com o Prof. Dr. Wallace Duarte Fragoso da Universidade Federal da Paraíba (URQUIZA-CARVALHO; FRAGOSO; ROCHA, 2016).

Para analisar a habilidade geral do método de apontar a nativa foi utilizado o Z-escore, mostrado na equação 11.1. Nesta equação, $\bar{H}$ e s são a entalpia média e desvio padrão da população de estruturas respectivamente; H é a entalpia de uma conformação em particular do conjunto de modo que cada estrutura possui um Z-Score em relação à população.

$$Z(H) = \frac{H - \bar{H}}{s} \quad (11.1)$$

Na tabela encontram-se os valores de $Z(H_{nativa})$ para todos os conjuntos e métodos para ver quão distante a entalpia da nativa está da entalpia do resto do conjunto. Esta é uma medida bastante geral para uma função escore porque à medida que a nativa se distancia da estrutura média, torna-se cada vez mais provável que ela se torne a estrutura de menor ou maior entalpia. Estes dados estão sumarizados na Tabela 6.

Tabela 6 – Na coluna da esquerda tem-se o Z-escore da estrutura nativa, e na coluna da direita a diferença entre a entalpia de formação da nativa e a entalpia de formação do melhor *decoy*, expressa em kcal/mol. Os cálculos foram realizados antes da aplicação do 3SIEVE.

PDB	Z-escore				$\Delta\Delta H_f$			
	PM7	PM6	PM6-DH+	PM6-D3	PM7	PM6	PM6-DH+	PM6-D3
1ABV	-3,46	-2,60	-2,49	-2,42	-870,48	-536,73	-698,24	-1284,52
1AF7	-2,89	-3,52	-2,41	-2,80	-861,99	-602,77	-692,42	-1145,81
1AH9	-3,75	-3,90	-3,31	-3,34	-468,40	-279,93	-415,37	-681,48
1AOY	-7,12	-5,53	-4,91	-4,77	-743,22	-576,77	-647,31	-1127,31
1B4B	-3,30	-0,06	-0,06	-5,32	-956,17	-689,34	-805,49	-1850,38
1BM8	-3,61	-6,60	-4,26	-5,38	-1590,28	-958,35	-1322,55	-1819,25
1CEW	-3,35	-0,11	-0,11	-1,84	-250,27	10,42	5,57	-756,23
1CQK	-2,59	-3,62	-3,09	-2,98	-1415,23	-1188,58	-1286,27	-1617,10
1DCJ	-7,82	-3,37	-1,87	-2,70	-388,45	31,73	-157,13	-1147,89
1EGX	-3,98	-5,07	-4,09	-4,66	-1131,18	-751,22	-928,31	-1864,95
1FAD	-13,56	-2,78	-2,15	-2,14	-503,56	-236,91	-291,80	-751,08
1FO5	-4,28	-4,67	-4,38	-4,71	-804,76	-711,80	-769,46	-1360,42
1G1C	-1,90	-5,38	-3,80	-4,47	-1316,76	-1047,60	-1182,52	-1979,27
1GJX	-4,80	-4,50	-3,78	-4,34	-608,52	-411,19	-510,89	-1165,93
1GPT	-3,00	-3,80	-3,11	-3,61	-457,99	-308,44	-404,03	-1016,37
1ITP	-4,53	-3,88	-3,23	-3,39	-575,25	-376,21	-471,57	-794,16
1KJS	-4,75	-1,85	-0,86	-1,33	35,31	274,23	186,85	-549,92
1KVI	-8,88	-2,90	-2,32	-2,41	-316,05	-168,74	-185,43	-574,59
1MLA	-1,63	-4,19	-2,95	-3,26	-810,77	-534,78	-700,02	-1160,92
1N0U	-3,36	-5,10	-4,54	-4,59	-861,19	-704,79	-767,21	-1336,60
1NE3	-4,65	-4,45	-3,96	-4,25	-608,33	-471,94	-536,50	-802,74
1OF9	-8,47	-2,66	-1,76	-1,99	-361,51	-151,56	-211,93	-478,35
1R69	-1,49	-4,75	-3,98	-4,36	-664,23	-462,91	-562,06	-1315,62
1SHF	-2,07	-3,47	-2,65	-2,86	-606,21	-365,22	-512,21	-876,87
1SRO	-5,57	-3,13	-1,95	-2,04	-341,72	3,92	-131,90	-498,17
1TFI	-2,76	-2,63	-1,82	-2,07	-225,25	13,08	-117,50	-412,00
1TIF	-7,96	-4,28	-3,64	-3,87	-770,25	-577,23	-672,78	-1237,33
1TIG	-3,19	-5,85	-4,79	-5,46	-1154,10	-778,83	-961,78	-1326,70
1VCC	-5,00	-3,74	-2,71	-2,94	-1096,28	-878,49	-988,63	-1453,25
256B	-7,11	-4,44	-3,72	-4,06	-1028,09	-738,99	-865,25	-1457,73
2CR7	-2,49	-3,43	-2,93	-2,79	-688,88	-608,87	-611,30	-1073,85
2F3N	-3,33	-3,43	-2,76	-3,12	-671,03	-450,32	-523,52	-921,50
2PCY	-4,07	-5,68	-5,04	-5,64	-1149,54	-902,25	-1061,76	-1612,31
Média	-4,57	-3,80	-3,01	-3,51	-735,17	-489,01	-600,02	-1134,87
σ	2,57	1,44	1,27	1,19	367,83	341,60	366,90	431,42

Como mostrado na Tabela 6, o método PM7 e PM6-D3 saíram-se melhor, de forma geral, que os demais. Em média, o PM7 posiciona a estrutura nativa cerca de 4,5 desvios-padrão abaixo da média da população, tornando-a uma estrutura bastante distinta do resto do conjunto, uma característica que se esperaria observar na superfície de energia da proteína. Observa-se que as nativas referentes aos códigos 1CEW, 1KJS e 1OF9 possuem um Z-Score distintamente mais baixo que as demais, para o PM7. Supondo-se que estas estruturas estão artificialmente abaixando a média do Z-Score, calculamos que sem estes códigos, a média do PM7 seria cerca de 3,89, ainda assim mais alto que o PM6-D3.

Alguns casos problemáticos foram observados, como os conjuntos das proteínas de código PDB 1G1C, 1MLA e 1R69, nos quais a estrutura nativa é colocada a menos de 2 desvios-padrão abaixo da média. Os métodos PM6 e PM6-DH+ saíram-se pior que o PM7, em média, com casos em que os métodos colocaram a nativa dentro da população, como exemplificado pelos conjuntos com código 1B4B, 1KJS e 1CEW. Nestes conjuntos, ambos os métodos colocaram a nativa a menos de um desvio-padrão abaixo da entalpia média. PM6-D3 não foi tão bom quanto o PM7 mas melhor que os outros dois, pois não colocou a nativa abaixo de um desvio-padrão da média em nenhum conjunto, sendo comparável ao PM7 neste aspecto.

Ainda na Tabela 6, pode-se observar a diferença entre as entalpias da estrutura nativa e a entalpia da estrutura *decoy* de entalpia mais baixa. É possível distinguir os conjuntos para os quais os métodos falharam em ranquear a nativa, uma vez que para tais conjuntos o valor desta diferença é positivo. O método PM7 falhou em ranquear corretamente a nativa apenas para o código 1KJS, como todos os outros métodos. Entretanto, para o PM7 um único *decoy* teve sua entalpia calculada menor que a nativa, enquanto para o PM6, 101 estruturas foram ranqueadas como melhores que a nativa, contra 51 do PM5-DH+ e 22 do PM6-D3. Além da 1KJS, o método PM6 falhou em ranquear a estrutura nativa para os conjuntos 1CEW, 1TFI e 1SRO, enquanto o PM6-DH+ só falhou no conjunto 1CEW, assim como o PM6-D3, mas por uma margem menor. O PM6-D3 não atribuiu a nenhum *decoy* uma energia mais baixa que suas respectivas nativas. Embora o PM6-D3 possua, em média, valores de $\Delta\Delta H_f$ mais baixo, estes valores podem variar segundo a parametrização.

Embora o PM6-D3 possua, em média, valores de $\Delta\Delta H_f$ mais baixo, estes valores podem variar segundo a parametrização. Uma forma de testar a eficácia do método em distinguir a nativa levando em consideração todas as estruturas em um conjunto de decoys é através do 3SIEVE.

Os métodos foram comparados de acordo com a diferença nos calores de formação atribuídos à estrutura nativa em um conjunto e a linha que se encontra 3 desvios-padrão abaixo da entalpia média do conjunto após a aplicação do 3SIEVE na população, que é a linha de corte para o 3SIEVE. Estes dados estão mostrados na Tabela 7 e sumarizados na

Figura 32.

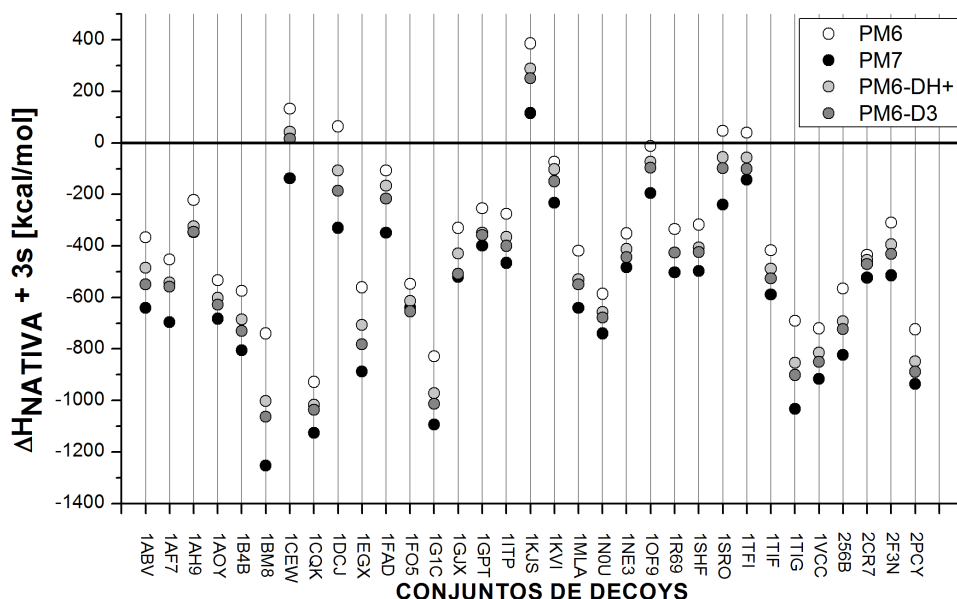


Figura 32 – Diferença entre a entalpia de formação de cada estrutura nativa e a linha de $-3s$ calculada segundo o 3SIEVE. A diferença é calculada para cada método em cada conjunto de estruturas.

Na Figura 32, os pontos que estão acima da linha representam estruturas nativas que o 3SIEVE considerou como parte da população de estruturas. De acordo com o 3SIEVE, alguns métodos não atribuíram uma entalpia de formação baixa o suficiente às estruturas nativas para inequivocamente diferenciá-las de uma estrutura que possui uma entalpia baixa por uma coincidência estatística de termos uma amostra aleatória de estruturas. As estruturas que estão acima da linha também incluem estruturas nativas mal ranqueadas, como aquelas vistas na Tabela 6.

Tabela 7 – A diferença entre as entalpias de formação entre a estrutura nativa de cada conjunto de estruturas e suas linhas de $-3s$ correspondente, em kcal/mol. Estes valores foram calculados após a população passar pelo 3SIEVE.

Set index	PDB	PM7	PM6	PM6-DH+	PM6-D3
1	1ABV	-640,22	-367,10	-485,42	-549,40
2	1AF7	-696,29	-452,45	-542,02	-558,10
3	1AH9	-347,40	-221,64	-324,56	-345,20
4	1AOY	-682,83	-533,54	-602,15	-628,18
5	1B4B	-805,92	-574,63	-685,57	729,69
6	1BM8	-1252,85	-740,33	-1002,91	-1063,74
7	1CEW	-136,92	132,85	41,95	16,00
8	1CQK	-1126,25	-928,06	-1017,75	-1036,97
9	1DCJ	-330,52	63,86	-107,39	-185,92
10	1EGX	-888,23	-561,15	-707,42	-781,92
11	1FAD	-349,34	-107,13	-165,85	-216,85
12	1FO5	-638,82	-548,11	-613,38	-654,20
13	1G1C	-1093,43	-828,36	-971,83	-1013,32
14	1GJX	-521,26	-330,04	-429,44	-507,72
15	1GPT	-398,76	-254,49	-348,38	-358,55
16	1ITP	-466,09	-275,57	-364,59	-399,67
17	1KJS	115,74	386,01	288,40	251,22
18	1KVI	-233,03	-72,77	-103,04	-149,53
19	1MLA	-640,99	-418,98	-529,86	-549,64
20	1N0U	-740,10	-586,03	-657,36	-679,04
21	1NE3	-483,07	-351,66	-412,52	-444,47
22	1OF9	-194,97	-12,46	-72,92	-96,18
23	1R69	-503,67	-335,35	-425,57	-426,41
24	1SHF	-498,12	-317,84	-405,69	-423,74
25	1SRO	-240,00	46,28	-56,07	-98,19
26	1TFI	-142,56	39,50	-57,55	-100,90
27	1TIF	-588,78	-416,67	-489,11	-525,62
28	1TIG	-1033,25	-690,70	-854,00	-901,25
29	1VCC	-916,64	-720,58	-814,29	-850,23
30	256B	-824,14	-565,88	-692,05	-722,60
31	2CR7	-523,17	-435,28	-455,54	-470,61
32	2F3N	-514,76	-309,58	-394,87	-431,01
33	2PCY	-936,04	-724,18	-848,78	-888,52

Para os conjuntos nos quais a conformação nativa foi ranqueada em posição pior que um ou mais *decoys*, nenhum destes foi detectado como tendo uma entalpia de formação baixa o suficiente para ser considerado uma estrutura nativa pelo 3SIEVE, mesmo após remover a conformação nativa do conjunto. Percebemos que o $\Delta\Delta H_f$ negativo referente ao método PM6-D3 não significa que o método foi capaz de distinguir de fato a estrutura nativa na população de estruturas *decoy* segundo o critério do 3SIEVE.

A detecção da estrutura nativa é uma característica interessante para uma função-

escore ter. Entretanto, durante um procedimento de amostragem, o algoritmo de amostragem muito raramente irá encontrar a estrutura nativa especificamente. Em uma abordagem mais realista, uma característica adicional para uma boa função escore seria a capacidade de discriminar estruturas geometricamente boas daquelas geometricamente ruins.

A avaliação geométrica da qualidade de um *decoy* pode ser baseada na fração de contatos nativos $Q(X)$, considerada um importante descritor geométrico da estrutura de proteínas (BEST; HUMMER; EATON, 2013; HENRY; BEST; EATON, 2013). O cálculo deste descritor é feito segundo a Equação (11.2):

$$Q(X) = \frac{1}{|S|} \sum_{(i,j) \in S} \frac{1}{1 + \exp[\beta(r_{ij}(X) - \lambda r_{ij}^0)]}, \quad (11.2)$$

onde X é a conformação. S é o conjunto de todos os pares de átomos pesados (i, j) pertencentes aos resíduos R_i e R_j de tal modo que estes dois resíduos estejam a mais de 3 resíduos de distância um do outro na cadeia e a distância entre os átomos i e j na conformação nativa seja menos que $r_{ij}^0 = 4,5 \text{ \AA}$. A distância entre os átomos i e j na conformação X é denotada por $r_{ij}(X)$. $|S|$ denota o número de elementos no conjunto S e β e λ são parâmetros numéricos com valores 5 e 1,8 respectivamente (BEST; HUMMER; EATON, 2013).

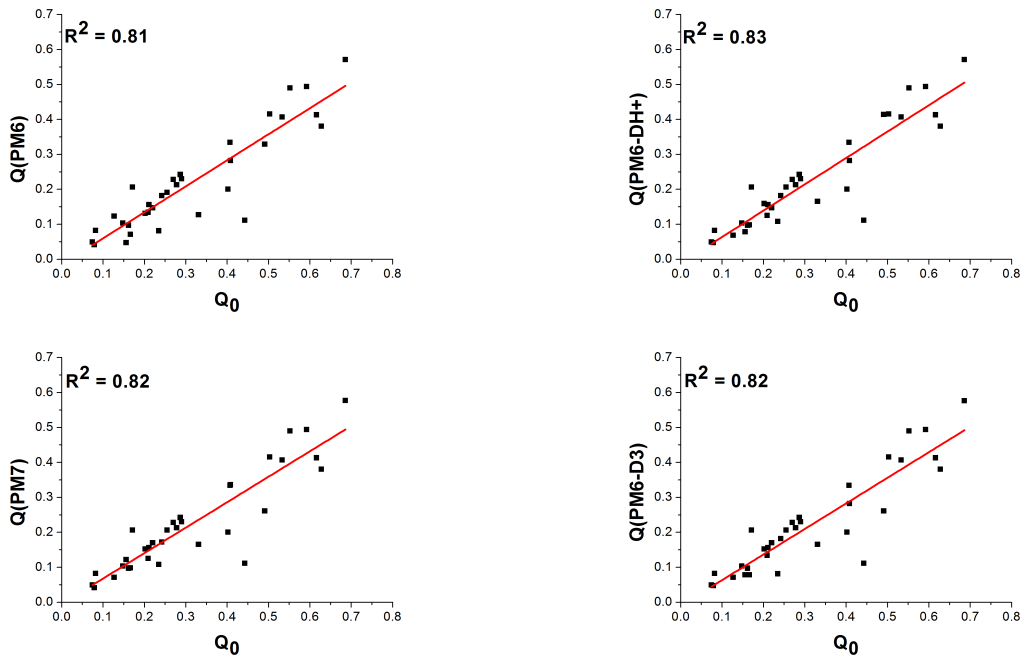


Figura 33 – Gráficos de correlação entre a fração de contatos nativos do melhor decoy selecionado por cada método, na ausência da nativa, e a fração de contatos nativos do decoy com maior fração de contatos nativos em cada grupo.

Utilizou-se a notação encurtada $Q(\text{método})$ para referi-se à fração de contatos nativo do *decoy* de menor entalpia de acordo com um método. Adicionalmente, em cada

conjunto de estruturas deve haver um *decoy* que possui uma fração de contatos nativos maior que a de qualquer outro no conjunto. O valor desta fração é o Q_0 do conjunto. Os gráficos na Figura 33 mostram a correlação entre Q_0 e $Q(\text{PM7})$, $Q(\text{PM6})$, $Q(\text{PM6-DH+})$ and $Q(\text{PM6-D3})$ para cada conjunto.

Os coeficientes de determinação para todos os três métodos são muito similares e todos eles foram levemente maiores que 0,8. Isto é um indicativo que os melhores *decoys* em cada conjunto têm aproximadamente a mesma fração de contatos nativos que aqueles geometricamente mais semelhantes a estrutura nativa, o que sugere que todos os três métodos conseguem encontrar boas conformações em um conjunto de estruturas.

Como todos os métodos se saíram de forma geral igualmente bem na avaliação geométrica, para os propósitos deste trabalho, decidiu-se utilizar o PM7 como Hamiltoniano para obter as densidades porque ele descreveu melhor as características energéticas dos conjuntos avaliados.

12 Agrupamento de Proteínas

As densidades eletrônicas de proteínas, obtidas do cálculo semiempírico utilizando o hamiltoniano PM7, foram decompostas com expansões $K = L = 5$.

As proteínas foram classificadas segundo sua classe evolutiva (classe I ou classe II), seu estado de dimerização (dímero ou singular) e sua modernidade (ancestral ou moderna). Em seguida, elas foram agrupadas utilizando todas as combinações das funções-utilidade $\mathcal{H}_2$ e $\mathcal{I}_2$ com as funções-similaridade de distância cossenóide e coeficiente de correlação. No entanto todas as combinações de parâmetros de agrupamentos produziram a mesma solução. Portanto, daqui em diante serão mostrados e discutidos apenas as soluções referentes à função utilidade $\mathcal{H}_2$ e função-similaridade de distância cossenóide. Ao todo 42 proteínas foram utilizadas no agrupamento.

Primeiro o conjunto de dados completo foi agrupado em dois grupos. Observou-se que os grupos são Fechados, com $\text{ISim} > 0,830$ e $\text{ISdev} < 0.03$ para ambos os grupos. A divisão de invariantes para cada grupo é bastante distinta, com mais de 70% da similaridade interna do grupo 1 dividida quase que igualmente entre os fingerprints $F_{0,0}$ e $F_{0,1}$ e mais de 60% da similaridade interna do grupo 0 concentrada totalmente no *fingerprint* $F_{0,1}$, mostrado na Figura 34

```
-----
2-way clustering solution - Descriptive & Discriminating Features...
-----
Cluster 0, Size: 16, ISim: 0.869, ESIm: 0.785
  Descriptive: (0,1) 64.7%, (0,0) 17.0%, (4,0) 2.1%, (0,5) 1.9%, (3,0) 1.9%
  Discriminating: (0,1) 24.7%, (0,4) 20.3%, (0,0) 18.2%, (1,2) 9.5%, (0,5) 5.4%

Cluster 1, Size: 23, ISim: 0.833, ESIm: 0.785
  Descriptive: (0,1) 37.1%, (0,0) 33.7%, (0,5) 5.2%, (0,4) 3.2%, (5,0) 2.5%
  Discriminating: (0,1) 24.7%, (0,4) 20.3%, (0,0) 18.2%, (1,2) 9.5%, (0,5) 5.4%
-----
```

Figura 34 – Imagem do arquivo de saída em que os grupos de proteínas são caracterizados segundo suas invariantes rotacionais.

O fingerprint $F_{0,0}$ é numericamente igual ao valor absoluto de ${}_0\omega_0^0$, que está associado com o número de elétrons e densidades esfericamente simétricas e isotrópicas. Enquanto o *fingerprint* $F_{0,1}$ é obtida de uma combinação de harmônicos esféricos com $l = 1$ e pode ser interpretada como a característica de uma densidade mais alongada ou dipolar. Deste modo o agrupamento divide as densidades entre aquelas que são fortemente dipolares e aquelas que possuem um caráter mais isotrópico e esférico. O grupo 0 será doravante referido como "grupo dipolar" e o grupo 1 como "grupo isotrópico".

Após o agrupamento, observou-se como as proteínas se dividem entre tais grupos segundo sua classe evolutiva, estado de dimerização e modernidade. Houve grande polarização de proteínas modernas no grupo isotrópico, com quase duas vezes mais proteínas

modernas no grupo isotrópico que no grupo dipolar. Quando leva-se em conta a classificação de classe I ou II, obtém-se exatamente os mesmos números. A maioria das proteínas do grupo isotrópico é de classe I, enquanto as proteínas de classe II estão mais uniformemente distribuídas entre o grupo dipolar e isotrópico.

Quanto ao estado de dimerização, as proteínas encontram-se ainda mais polarizadas, sugerindo fortemente que a maior parte das proteínas diméricas são dipolares e a maior parte das proteínas singulares são isotrópicas. Isto sugere que algum fator na forma da densidade de cada subunidade contém informações sobre o estado de dimerização de uma proteína. Esta é uma hipótese razoável, considerando que cada subunidade é sintetizada de forma independente das outras e a dimerização é um processo posterior, em grande parte guiado pelas interações intermoleculares entre as subunidades componentes. Todas estas informações estão sumarizado na Tabela 8.

Tabela 8 – Distribuição das densidades nos grupos em cada tipo de classificação considerado.

Grupo	Dimerização		Filogenia		Antiguidade	
	Singular	Dímero	Classe I	Classe II	Modernas	Ancestrais
Dipolar	4	14	8	10	8	10
Isotrópica	14	9	15	8	15	8

Para saber se os ancestrais e suas contrapartes modernas estavam mais próximas umas das outras foram consideradas as enzimas de classe I, divididas em 10 grupos, porque, sem contar as Lisil-tRNA aminoacil sintetases, tem-se 10 proteínas modernas e 10 ancestrais de classe 1. Depois, verificou-se em quantos grupos as duas enzimas (moderna e ancestral) foram agrupadas juntas. Os resultados estão mostrados na Tabela 9.

Tabela 9 – Quantidade de tRNA-aminoacil Sintetases de Classe 1 agrupadas pelo aminoácido substrato, indicados pelos seus códigos de uma letra.

Grupo	C	K	R	Q	E	I	L	M	W	Y	V
0	0	0	0	1	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	1	0	0
2	0	0	0	1	0	0	1	0	0	0	1
3	0	0	0	0	0	0	1	1	0	0	0
4	1	1	0	0	0	0	0	0	0	0	0
5	0	0	1	0	0	0	0	0	0	0	1
6	0	1	0	0	1	2	0	0	0	0	0
7	0	0	1	0	0	0	0	0	0	1	0
8	0	2	0	0	1	0	0	0	0	0	0
9	1	0	0	0	0	0	0	1	1	1	0

Ignorando as tRNA-lisil sintetases (código K), é evidente que apenas as Isoleucil-

tRNA sintetases (código I) encontram-se, moderna e ancestral, no mesmo grupo. Tampouco estão, em geral, proteínas modernas e ancestrais, em grupos que fazem parte diretamente do mesmo super-grupo. Pode-se seguramente concluir que as enzimas ancestrais não agrupam preferencialmente com suas equivalentes modernas.

Da mesma forma, dividiu-se o conjunto de todas as estruturas de classe II em 8 grupos. E mais uma vez é possível contar quantos grupos contêm ambas as enzimas (a moderna e a ancestral) para cada tipo de aminoácido substrato, como mostram os dados da Tabela 10.

Tabela 10 – Quantidade de tRNA-aminoacil Sintetases de Classe 2 agrupadas pelo aminoácido substrato, indicados pelos seus códigos de uma letra.

Grupo	A	N	D	H	F	S	T	K	P
0	0	0	0	0	1	0	0	0	0
1	0	0	1	0	0	0	0	1	0
2	0	0	0	1	0	0	1	0	0
3	0	0	0	1	0	0	0	0	1
4	1	0	0	0	1	2	0	0	0
5	0	1	0	0	0	0	0	0	1
6	0	0	1	0	0	0	1	0	0
7	1	1	0	0	0	0	0	1	0

Exceto que neste caso se observa que apenas para a tRNA-Seril sintetase (código S) obtém-se o agrupamento das estruturas moderna e ancestral. Todas as outras enzimas de classe II seguem o mesmo padrão daquelas de classe I, com nenhum agrupamento entre ancestrais e suas equivalentes modernas sendo favorecido.

Com isso conclui-se que existem, na verdade, dois tipos de grupos no presente conjunto de dados: (i) o grupo da densidade que é mais dipolar que isotrópica e, (ii) o grupo das densidades que são dipolares mas não de forma tão pronunciada. E que existem correlações entre a forma da densidade e, principalmente, seu estado de dimerização, mas que esta classificação quanto a forma também é capaz de mostrar preferências evolutivas, mas de forma menos pronunciada.

13 Conclusão

No presente trabalho criou-se uma ferramenta adequada para o uso cotidiano na expansão de densidades eletrônicas de formatos conhecidos e amplamente utilizados na base de Canterakis-Zernike. O algoritmo de integração apresentado foi considerado satisfatório.

Adicionalmente, desenvolveu-se um procedimento de integração e uma metodologia que permite que o parâmetro de integração mais difícil de se estabelecer, o r^* , fosse atribuído um valor a cada molécula de forma automática e razoável. Tais propriedades também podem ser úteis na obtenção de descrições matemáticas de características da densidade eletrônicas tais como número de elétrons, forma, complexidade, função de distribuição radial molecular de elétrons e campo elétrico de longo alcance.

Estruturas equilibradas de dinâmica molecular de ancestrais de tRNA-Aminoacil Transferases foram obtidas para análises posteriores e foram as primeiras estruturas de proteínas ancestrais a serem estudadas para estes fins específicos.

O agrupamento das densidades segundo suas invariantes rotacionais levou a conclusão que é possível utilizar estes descritores para comparar moléculas pequenas também, inclusive chegando a agrupamentos de moléculas que frequentemente contém moléculas quimicamente relacionadas.

Conclui-se ainda, em resultado publicado, que, embora os hamiltonianos semiempíricos PM7, PM6 e PM6-DH+ sejam todos capazes de descrever de forma semelhante a geometria de polipeptídeos, o PM7 se sai marginalmente melhor, especialmente no tocante a entalpia de formação. Devido a essa modesta vantagem, PM7 foi escolhido como hamiltoniano semiempírico para os cálculos de estrutura eletrônica de proteínas.

Por fim, conclui-se que existem indícios de que a forma da densidade, analisada segundo as invariantes rotacionais de Canterakis-Zernike, pode revelar propriedades de dimerização das proteínas, para o caso particular destas enzimas, está relacionado com o passado evolutivo. Embora estas propriedades estejam relacionadas com a origem evolutiva destas moléculas, os dados produzidos foram considerados insuficientes para confirmar ou refutar a hipótese inicial.

Além da associação com a quantidade de subunidades, o agrupamento também mostrou certo grau de discriminação com as enzimas de classe I, colocando uma grande fração delas no grupo isotrópico, enquanto as enzimas de classe II ficaram mais homogeneamente distribuídas entre os dois grupos. Isto foi tomado como indício de que as enzimas de classe II possuem uma variedade maior de formas de sua densidade, o que sugere que seja um grupo mais heterogêneo de enzimas. Isto pode ou não estar associado com o passado

evolutivo da linhagem.

Adicionalmente, existem evidências de transferência horizontal de genes ao longo da evolução da tRNA-Aminoacil Sintetases (WOESE *et al.*, 2000; CHALIOTIS *et al.*, 2017), no entanto como já mencionado a transferência horizontal de genes não é contemplada na ASR diretamente, mas espera-se que este efeito não tenha um impacto tão profundo na ancestralidade recente destas enzimas. Para estudar especificamente a transferência horizontal, seria necessário abandonar e/ou modificar seriamente a técnica de ASR, algo a que este trabalho não se propõe. Apesar destas limitações, no entanto, como já discutido na fundamentação teórica, trata-se de uma técnica amplamente utilizada na biologia evolutiva.

14 Perspectivas Futuras

Neste trabalho nos buscamos abrir as portas para uma nova forma de analisar estruturas eletrônicas de proteínas através do emprego de métodos de química quântica na investigação de um fenômeno que é bastante particular destas moléculas, que é a evolução pela seleção natural.

No decorrer do trabalho, ficou claro que as estratégias e métodos desenvolvidos para lidar com os problemas que apareceram poderiam ser, no futuro, aprimoradas e que isto seria um trabalho em si mesmo.

Além da melhoria das técnicas já empregadas aqui, abriram-se também portas para uma análise mais detalhada do formalismo e possivelmente até mesmo o desenvolvimento de novos descritores ou de metodologias de estudo de densidades eletrônicas.

Entre as perspectivas para o futuro desta linha de pesquisa temos:

- Obter bancos de dados de sequência maiores para tentar conseguir linhagens mais antigas;
- Analisar a estrutura matemática das densidades de Canterakis-Zernike para propor uma representação reduzida absoluta;
- Descrever o conceito de “ancestral” utilizando um formalismo de espaço de densidades;
- Elaborar um arcabouço teórico para discutir como caminhar no espaço de Canterakis-Zernike;
- Estudar ancestrais e proteínas modernas com os novos métodos de simetria e campo eletrostático de longo alcance;
- Estudar novas formas de se obter sequências e estruturas para mitigar os erros acumulados ao longo do protocolo estabelecido aqui;
- Continuar a colaboração com o Prof. Dr. Rafael López na criação e avaliação de descritores para densidades LCAO.

Alguns destes itens já encontram-se em andamento. Em particular, o primeiro e o último item já possuem resultados. O Prof. Dr. Sávio Torres de Farias, co-orientador do presente trabalho, já possui uma linhagem mais completa de tRNA-glicil sintetases, voltando potencialmente mais de 4 bilhões de anos no passado evolutivo da molécula. Uma parceria para o estudo da reatividade destas estruturas já está sendo discutido. Quanto

ao último item, um trabalho destinado a publicação já está sendo escrito com o Prof. López que irá discorrer sobre todo o método descrito aqui e uma metodologia semelhante utilizando polinômios diferentes dos de Canterakis-Zernike.

Referências

- ADCOCK, S. A.; MCCAMMON, J. A. Molecular dynamics: Survey of methods for simulating the activity of proteins. *Chemical Reviews*, v. 106, n. 5, p. 1589–1615, 2006. Citado na página 1.
- ADHIKARI, A. N.; FREED, K. F.; SOSNICK, T. R. Simplified protein models: Predicting folding pathways and structure using amino acid sequences. *Phys. Rev. Lett.*, v. 111, p. 28103, 2011. Citado na página 38.
- AKANUMA, S. *et al.* Experimental evidence for the thermophilicity of ancestral life. v. 110, n. 27, p. 11067–11072, 2013. Citado na página 12.
- ALBERTS, B. *et al.* *Molecular Biology of the Cell*. Sexta. [S.l.]: Garland Science, 2014. Citado 2 vezes nas páginas 5 e 10.
- ALCALDE, R. *et al.* Insights into glycol ethers - alkanol mixtures from a combined experimental and theoretical approach. *The Journal of Physical Chemistry B*, 2017. Citado na página 1.
- ANDERSON, J.; LORENZ, C.; TRAVESSET, A. General purpose molecular dynamics simulations fully implemented on graphics processing units. *J. Comput. Phys.*, v. 227, p. 5342, 2008. Citado na página 2.
- ANFENSEN, C. B. Principles that govern the folding of protein chain. *Science*, v. 181, p. 223–230, 1973. Citado na página 38.
- ANFENSEN, C. B. Principles that govern the folding of protein chains. *Science*, v. 181, n. 4096, p. 223–230, 1973. Citado na página 4.
- ANIKIN, N. A. *et al.* LocalSCF method for semiempirical quantum-chemical calculation of ultralarge biomolecules. *J. Chem. Phys.*, v. 121, p. 1266–1270, 2004. Citado na página 1.
- ANIKIN, N. A. *et al.* Method for the improved semiempirical description of intermolecular interactions of biomolecules and their fragments. *Russian Chemical Bulletin International Edition*, v. 61, p. 12–16, 2012. Citado na página 39.
- ANTAL, Z.; WARBURTON, P. L.; MEZEY, P. G. Electron density shape analysis of a family of through-space and through-bond interactions. *Phys. Chem. Chem. Phys.*, v. 16, p. 918–932, 2014. Citado na página 22.
- ASHKENAZY, H. *et al.* Fastml: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Research*, v. 40, n. W1, p. W580, 2012. Citado na página 12.
- BAR-ROGOVSKY, H. *et al.* Assessing the prediction fidelity of ancestral reconstruction by a library approach. *Protein Engineering, Design and Selection*, v. 28, n. 11, p. 507, 2015. Citado na página 12.

- BARDUCCI, A.; BUSSI, G.; PARRINELLO, M. Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Phys. Rev. Lett.*, v. 100, p. 020603, 2008. Citado na página 2.
- BEST, R. B.; HUMMER, G.; EATON, W. A. Native contacts determine protein folding mechanisms in atomistic simulations. *PNAS*, v. 110, n. 44, p. 17874–17879, 2013. Citado na página 68.
- BIASINI, M. *et al.* Swiss-model: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*, v. 42, n. W1, p. W252, 2014. Citado na página 32.
- BIKADI, Z.; HAZAI, E. Application of the pm6 semi-empirical method to modeling proteins enhances docking accuracy of autoDock. *J. Cheminf.*, v. 1, n. 15, 2009. Citado na página 39.
- BRADLEY, P.; MISURA, K. M.; BAKER, D. Toward high-resolution de novo structure prediction for small proteins. *Science*, v. 309, p. 1868–1871, 2005. Citado na página 38.
- BROOKS, C. L.; KARPLUS, M.; PETTITT, B. M. Proteins: A theoretical perspective of dynamics, structure, and thermodynamics. *Adv. Chem. Phys.*, v. 71, p. 1–259, 1988. Citado na página 41.
- BROWN, P. *et al.* Massively multicore parallelization of Kohn-Sham theory. *J. Comput. Chem.*, v. 31, p. 2008, 2010. Citado na página 2.
- BRUNK, E.; ROTHLIBERGER, U. Mixed quantum mechanical/molecular mechanical molecular dynamics simulations of biological systems in ground and electronically excited states. *Chemical Reviews*, v. 115, n. 12, p. 6217–6263, 2015. Citado na página 1.
- CANTERAKIS, N. 3d zernike moments and zernike affine invariants for 3d image analysis and recognition. In: *In 11th Scandinavian Conf. on Image Analysis*. [S.l.: s.n.], 1999. p. 85–93. Citado 2 vezes nas páginas 20 e 52.
- CASE, D. A. *et al.* *Amber 14*. [S.l.], 2014. Citado na página 35.
- CHALIOTIS, A. *et al.* The complex evolutionary history of aminoacyl-trna synthetases. *Nucleic Acids Research*, v. 45, n. 3, p. 1059, 2017. Citado na página 74.
- CHANG, T.; GAO, H. Size-dependent elastic properties of a single-walled carbon nanotube via a molecular mechanics model. *Journal of the Mechanics and Physics of Solids*, v. 51, n. 6, p. 1059–1074, 2003. Citado na página 1.
- CHATTERJEE, S.; BHATTACHARYYA, M.; VISHVESHWARA, S. Network properties of protein-decoy structures. *J. Biomol. Struct. Dyn.*, v. 29, n. 6, p. 606–622, 2012. Citado na página 38.
- COMEAU, S. R. *et al.* Cluspro: a fully automated algorithm for protein-protein docking. *Nucleic Acids Research*, v. 32, p. W96–9, 2004. Citado 3 vezes nas páginas 32, 33 e 34.
- CORMEN, T. H. *et al.* *Introduction to Algorithms*. Terceira. [S.l.]: MIT Press, 2009. Citado na página 22.

- DAS, R. *et al.* Structure prediction for CASP7 targets using extensive all-atom refinement with rosetta@home. *Prot. Struc. Func. Bioinfo.*, v. 69, p. 118–128, 2007. Citado na página 38.
- DENG, J. *et al.* Highly active and durable non-precious-metal catalysts encapsulated in carbon nanotubes for hydrogen evolution reaction. *Energy Environ. Sci.*, The Royal Society of Chemistry, v. 7, p. 1919–1923, 2014. Citado na página 1.
- DILL, K. A.; MCCALLUM, J. L. Protein-folding problem, 50 years on. *Science*, v. 338, p. 1042–1046, 2012. Citado 2 vezes nas páginas 4 e 38.
- DIXON, S. L.; MERZ, K. M. J. Semiempirical molecular orbital calculations with linear system size scaling. *The Journal of Chemical Physics*, v. 104, n. 17, p. 6643–6649, 1996. Citado na página 1.
- DROZDETSKIY, A. *et al.* Jpred4: a protein secondary structure prediction server. *Nucleic Acids Research*, v. 43, n. W1, p. W389, 2015. Citado na página 32.
- DUTRA, J. D. L. *et al.* Sparkle/PM7 lanthanide parameters for the modeling of complexes and materials. *J. Chem. Theory Comput.*, v. 9, p. 3333–3341, 2013. Citado na página 39.
- EMAMIAN, S. Toward understanding regioselectivity and molecular mechanism in the synthesis of cf2h-containing 2-pyrazolines: A molecular electron density theory study. *Journal of Fluorine Chemistry*, v. 199, p. 77 – 91, 2017. Citado na página 1.
- F, G. A. J.; MILLER, J. H.; SUZUKI, D. T. *An Introduction to Genetic Analysis*. Setima. [S.l.]: W. H. Freeman, 2000. Citado 2 vezes nas páginas 9 e 10.
- FANFRLIK, P. *et al.* A reliable docking/scoring scheme based on the semiempirical quantum mechanical PM6-DH2 method accurately covering dispersion and H-bonding: HIV-1 protease with 22 ligands. *J. Phys. Chem. B*, v. 39, p. 12666–12678, 2010. Citado na página 39.
- FARIAS, S. T. Suggested phylogeny of trnas based on the construction of ancestral sequences. *J. of Theo. Bio.*, v. 335, p. 245–248, 2013. Citado 2 vezes nas páginas 31 e 32.
- FARIAS, S. T.; GUIMARAES, R. C. Aminoacyl-trna synthetase classes and groups in prokaryotes. *J. of Theo. Bio.*, v. 250, p. 221–229, 2008. Citado 2 vezes nas páginas 31 e 32.
- FERRARIO, M. *et al.* Molecular dynamics simulation of aqueous mixtures: Methanol, acetone, and ammonia. *The Journal of Chemical Physics*, v. 93, n. 7, p. 5156–5166, 1990. Citado na página 1.
- FISHER, D. 3d-shotgun: a novel, cooperative, fold-recognition meta-predictor. *Proteins*, v. 51, p. 434–444, 2003. Citado na página 38.
- FORTERRE, P. The universal tree of life: an update. *Front Microbiol*, Frontiers Media S.A., v. 6, p. 717, jul 2015. Citado na página 10.
- FRIEDRICHS, M. S. *et al.* Accelerating molecular dynamic simulation on graphics processing units. *J. Comput. Chem.*, v. 30, p. 864, 2009. Citado na página 2.

FRISCH, M. J. *et al.* *Gaussian 09, Revision A.02*. 2009. Citado 3 vezes nas páginas 22, 25 e 53.

FUKUSHIMA, K.; WADA, M.; M., S. An insight into the general relationship between the three dimensional structures of enzymes and their electronic wavefunctions: Implication for the prediction of functional sites of enzymes. *Proteins*, v. 71, p. 1940–1954, 2008. Citado na página 2.

FURNESS, J. W. *et al.* Electron localisation function in current density functional theory. *Molecular Physics*, v. 114, n. 7-8, p. 1415–1422, 2016. Citado na página 1.

GEERLINGS, P. *et al.* The woodward-hoffmann rules reinterpreted by conceptual density functional theory. *Accounts of Chemical Research*, v. 45, n. 5, p. 683–695, 2012. Citado na página 1.

GEERLINGS, P.; PROFT, F. D. Conceptual dft: the chemical relevance of higher response functions. *Phys. Chem. Chem. Phys.*, v. 10, p. 3028–3042, 2008. Citado na página 1.

GLANSDORFF, N.; YING, X.; LABEDAN, B. The last universal common ancestor: emergence, constitution and genetic legacy of an elusive forerunner. *Biology Direct*, v. 3, p. 29, 2008. Citado na página 10.

GOTZ, K. *et al.* Comparative analysis of electron density and electron localization function for dinuclear manganese complexes with bridging boron and carbon-centered ligands. *Chemistry A European Journal*, WILEY-VCH Verlag, v. 15, n. 3, p. 623–632, 2009. Citado na página 1.

GRABOWSKI, S. J. QTAIM characteristics of halogen bond and related interactions. *The Journal of Physical Chemistry A*, v. 116, n. 7, p. 1838–1845, 2012. Citado na página 1.

GRABOWSKI, S. J.; UGALDE, J. M. Bond paths show preferable interactions: Ab initio and qtaim studies on the x-h...pi hydrogen bond. *The J. Phys. Chem. A*, v. 114, n. 26, p. 7223–7229, 2010. Citado na página 1.

GRIMME, S. Accurate description of van der waals complexes by density functional theory including empirical corrections. *J. Comput. Chem.*, v. 25, p. 1463–1473, 2004. Citado na página 39.

GRIMME, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.*, v. 27, p. 1787–1799, 2006. Citado na página 39.

GRIMME, S. *et al.* A consistent and accurate ab initio parametrization of density functional dispersion correction (dft-d) for the 94 elements h-pu. *J. Chem. Phys.*, v. 132, n. 15, p. 154104–1, 2010. Citado na página 39.

GROMIHA, M. M.; YUGANDHAR, K.; JEMIMAH, S. Protein–protein interactions: scoring schemes and binding affinity. *Current Opinion in Structural Biology*, v. 44, p. 31 – 38, 2017. Citado na página 33.

GUO, Z. *et al.* Intrinsic reaction coordinate analysis of the activation of ch4 by molybdenum atoms: A density functional theory study of the crossing seams of the potential energy surfaces. *Organometallics*, v. 27, n. 2, p. 181–188, 2008. Citado na página 1.

- HANSON-SMITH, V.; KOLACZKOWSKI, B.; T JOSEPH, W. Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Molecular Biology and Evolution*, v. 27, n. 9, p. 1988, 2010. Citado na página 12.
- HARTL, F. U.; HAYER-HARTL, M. Converging concepts of protein folding in vitro and in vivo. *Nat. Struct. Mol. Biol.*, v. 16, n. 6, p. 574–581, 2009. Citado na página 4.
- HENRY, E. R.; BEST, R. B.; EATON, W. A. Comparing a simple theoretical model for protein folding with all-atom molecular dynamics simulations. *PNAS*, v. 110, n. 44, p. 17880–17885, 2013. Citado na página 68.
- HERNANDEZ-ESPARZA, R. *et al.* Grid-based algorithm to search critical points, in the electron density, accelerated by graphics processing units. *Journal of Computational Chemistry*, v. 35, n. 31, p. 2272–2278, 2014. Citado na página 2.
- JABLONSKI, M.; PALUSIAK, M. Nature of a hydride–halogen bond. a sapt, qtaim and nbo-based study. *J. of Phys. Chem. A*, v. 116, n. 9, p. 2322–2332, 2012. Citado na página 1.
- JAKO, E. *et al.* In silico detection of trna sequence features characteristic to aminoacyl-trna synthetase class membership. *Nucleic Acids Research*, v. 35, n. 16, p. 5593, 2007. Citado 2 vezes nas páginas 31 e 32.
- JANATI-FARD, F.; REZA, M. H. H.; MONHEMI, H. Investigation of structural stability and enzymatic activity of glucose oxidase and its subunits. *Journal of Molecular Catalysis B: Enzymatic*, v. 134, p. 16 – 24, 2016. Citado na página 37.
- K, T. Electron-pair shell density approximation applied to inner and outer electron radii of atoms. *Journal of Physics B: Atomic, Molecular and Optical Physics*, v. 47, n. 13, p. 135001, 2014. Citado na página 22.
- KAGAMI, L. P. *et al.* Identification of a novel putative inhibitor of the plasmodium falciparum purine nucleoside phosphorylase: exploring the purine salvage pathway to design new antimalarial drugs. *Molecular Diversity*, p. 1–19, 2017. Citado na página 1.
- KARPLUS, K. Sam-t08, hmm-based protein structure prediction. *Nucleic Acids Research*, v. 37, n. suppl2, p. W492, 2009. Citado na página 32.
- KARPLUS, M.; ICHIYE, T.; PETTITT, B. M. Configurational entropy of native proteins. *Biophysics Journal*, v. 52, n. 6, p. 1083–1085, 1987. Citado na página 41.
- KARPLUS, M.; MCCAMMON, J. A. Molecular dynamics simulations of biomolecules. *Nat Struct Mol Biol*, v. 9, n. 9, p. 646–652, Sep 2002. Citado na página 1.
- KINDRATENKO, V. *et al.* High-performance computing with accelerators. *Comput. Sci. Eng.*, v. 12, p. 12, 2010. Citado na página 2.
- KITAURA, K. *et al.* Fragment molecular orbital method: an approximate computational method for large molecules. *Chemical Physics Letters*, v. 313, p. 701–706, 1999. Citado na página 1.
- KLAMT, A.; SCHUURMANN, A. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc., Perkin Trans. 2*, p. 799–805, 1993. Citado 2 vezes nas páginas 40 e 43.

- KLEPEIS, J. L. *et al.* Long timescale molecular dynamics simulations of protein structure and function. *Current Opinion in Structural Biology*, v. 19, n. 2, p. 120–127, 2009. Citado na página 1.
- KOONIN, E. V. Carl Woese's vision of cellular evolution and the domains of life. *RNA Biol*, Landes Bioscience, v. 11, n. 3, p. 197–204, mar 2014. Citado na página 10.
- KORTH, M. Third-generation hydrogen-bonding corrections for semiempirical QM methods and force fields. *J. Chem. Theo. Chem.*, v. 6, p. 3808–3816, 2010. Citado na página 39.
- KORTH, M. *et al.* A transferable h-bonding correction for semiempirical quantum-chemical methods. *J. Chem. Theo. Chem.*, v. 6, p. 344–352, 2010. Citado na página 39.
- KUMAR, A. *et al.* Damqt 2.1.0: A new version of the damqt package enabled with the topographical analysis of electron density and electrostatic potential in molecules. *Journal of Computational Chemistry*, v. 36, n. 31, p. 2350–2359, 2015. Citado na página 24.
- LAIO, A.; GERVASIO, F. L. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics*, v. 71, n. 12, p. 126601, 2008. Citado na página 2.
- LANIER, K. A.; WILLIAMS, L. D. The Origin of Life: Models and Data. *J Mol Evol*, Springer US, New York, v. 84, n. 2, p. 85–92, 2017. Citado na página 10.
- LEVITT, M.; WARSCHEL, A.; KARPLUS, M. *The Nobel Prize in Chemistry 2013*. 2013. Disponível em: <http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2013/>. Citado na página 3.
- LILIENFELD, O. A. von *et al.* Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties. *International Journal of Quantum Chemistry*, v. 115, n. 16, p. 1084–1093, 2015. Citado na página 22.
- LOBANOV, M. Y.; BOGATYREVA, N. S.; GALZITSKAYA, O. V. Radius of gyration as an indicator of protein structure compactness. *Molecular Biology*, v. 42, n. 4, p. 623–628, 2008. Citado na página 37.
- LUAN, B.; ZHOU, R. A novel self-activation mechanism of candida antarctica lipase b. *Phys. Chem. Chem. Phys.*, 2017. Citado na página 1.
- LUEHR, N.; JIN, A. G. B.; MARTINEZ, T. J. Ab initio interactive molecular dynamics on graphical processing units (GPUs). *Journal of Chemical Theory and Computation*, v. 11, n. 10, p. 4536–4544, 2015. Citado na página 2.
- LUEHR, N.; UFIMTSEV, I. S.; MARTÍNEZ, T. J. Dynamic precision for electron repulsion integral evaluation on graphical processing units (GPUs). *J. Chem. Theory Comput.*, v. 7, p. 949, 2011. Citado na página 2.
- MAEDA SAND HARABUCHI, Y. *et al.* Intrinsic reaction coordinate: Calculation, bifurcation, and automated search. *International Journal of Quantum Chemistry*, v. 115, n. 5, p. 258–269, 2014. Citado na página 2.

- MAIA, J. D. C. *et al.* GPU linear algebra libraries and GPGPU programming for accelerating MOPAC semiempirical quantum chemistry calculations. *J. Chem. Theory Comput.*, v. 8, p. 3072–3081, 2012. Citado 3 vezes nas páginas 2, 39 e 40.
- MCINERNEY, J.; PISANI, D.; OCONNELL, M. J. The ring of life hypothesis for eukaryote origins is supported by multiple kinds of data. *Philos Trans R Soc Lond B Biol Sci*, The Royal Society, v. 370, n. 1678, p. 20140323, 2015. Citado na página 10.
- MCNAMARA, J. P.; HILLIER, I. H. Semi-empirical molecular orbital methods including dispersion corrections for the accurate prediction of the full range of intermolecular interactions in biomolecules. *Phys Chem Chem Phys.*, v. 9, p. 2362–2370, 2007. Citado na página 39.
- MIKULSKIS, P. *et al.* A semiempirical approach to ligand-binding affinities: Dependence on the hamiltonian and corrections. *J. Comp. Chem.*, v. 33, n. 12, p. 1179–1189, 2012. Citado na página 39.
- MORALES-BAYUELO, A.; MATUTE, R. A.; CABALLERO, J. Understanding the comparative molecular field analysis (comfa) in terms of molecular quantum similarity and dft-based reactivity descriptors. *Journal of Molecular Modeling*, v. 21, n. 6, p. 156, 2015. Citado na página 22.
- NIKOLAIENKO, T. Y.; BULAVIN, L. A.; HOVORUN, D. M. Bridging qtaim with vibrational spectroscopy: the energy of intramolecular hydrogen bonds in dna-related biomolecules. *Phys. Chem. Chem. Phys.*, v. 14, p. 7441–7447, 2012. Citado na página 1.
- NOVOTNI, M.; KLEIN, R. 3d zernike descriptors for content based shape retrieval. In: *The 8th ACM Symposium on Solid Modeling and Applications*. [S.l.: s.n.], 2003. Citado 3 vezes nas páginas 21, 22 e 57.
- PALSHIN, V. *et al.* X-ray absorption spectroscopy, simulation and modeling of si-dlc films. *Journal of Materials Science*, v. 37, n. 8, p. 1535–1539, 2002. Citado na página 1.
- PANDEY, R. B.; JACOBS, D. J.; FARMER, B. L. Preferential binding effects on protein structure and dynamics revealed by coarse-grained monte carlo simulation. *The Journal of Chemical Physics*, v. 146, n. 19, p. 195101, 2017. Citado na página 2.
- PAULING, L.; ZUCKERKANDL, E. Chemical paleogenetics. molecular restoration studies of extinct forms of life. *Acta Chemica Scandinavica*, v. 17, p. 9–16, 1963. Citado na página 11.
- PETTERSEN, E. F. *et al.* UCSF chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.*, v. 25, p. 1605–1612, 2004. Citado na página 40.
- PINTER, B. *et al.* Halogen bonding from a hard and soft acids and bases perspective: Investigation by using density functional theory reactivity indices. *Chemistry A European Journal*, v. 19, n. 2, p. 519–530, 2013. Citado na página 1.
- POLYAK, I. B. E.; SEN, K.; THIEL, W. A microiterative intrinsic reaction coordinate method for large QM/MM systems. *Phys. Chem. Chem. Phys.*, The Royal Society of Chemistry, v. 15, p. 14188–14195, 2013. Citado na página 1.

- PUZZARINI, C.; STANTON, J. F.; GAUSS, J. Quantum-chemical calculation of spectroscopic parameters for rotational spectroscopy. *International Reviews in Physical Chemistry*, v. 29, n. 2, p. 273–367, 2010. Citado na página 1.
- REZAC, J. *et al.* Semiempirical quantum chemical pm6 method augmented by dispersion and H-bonding correction terms reliably describes various types of noncovalent complexes. *J. Chem. Theory Comput.*, v. 5, p. 1749–1760, 2009. Citado na página 39.
- REZAC, J.; HOBZA, P. Advanced corrections of hydrogen bonding and dispersion for semiempirical quantum mechanical methods. *J. Chem. Theo. Chem.*, v. 8, p. 141–151, 2012. Citado na página 39.
- RICO, J. F. *et al.* Translation of sto charge distributions. *J. of Comp. Chem.*, v. 26, n. 8, p. 846–855, 2005. Citado na página 17.
- RIVAS, O. M. M.; REY, A. D. Molecular dynamics of dilute binary chromonic liquid crystal mixtures. *Mol. Syst. Des. Eng.*, 2017. Citado na página 1.
- ROCHA, G. B. *et al.* Rm1: A reparameterization of am1 for h, c, n, o, p, s, f, cl, br, and i. *Journal of Computational Chemistry*, v. 27, n. 10, p. 1101–1111, 2006. Citado 2 vezes nas páginas 24 e 56.
- ROSE, G. D. *et al.* A backbone-based theory of protein folding. *Proc. Natl. Acad. Sci.*, v. 103, p. 16623–16633, 2006. Citado 2 vezes nas páginas 37 e 38.
- ROSE, P. W. *et al.* The rcsb protein data bank: integrative view of protein, gene and 3d structural information. *Nucleic Acids Research*, v. 45, n. D1, p. D271, 2017. Citado na página 34.
- RYCKAERT, J.; CICCOTTI, G.; C, B. H. J. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comp. Physics*, v. 23, p. 327–341, 1977. Citado na página 36.
- SAEL, L. *et al.* Fast protein tertiary structure retrieval based on global surface shape similarity. *Prot. Struct. Func. and Bioinf.*, v. 72, n. 4, p. 1259–1273, 2008. Citado 2 vezes nas páginas 32 e 57.
- SAKATA, K. Electron reorganization along the intrinsic reaction coordinate in 1,3-dipolar cycloaddition. *The Journal of Physical Chemistry A*, v. 104, n. 44, p. 10001–10008, 2000. Citado na página 1.
- SAMUDRALA, R.; LEVITT, M. Decoys r us - a database of incorrect protein conformations to improve protein structure prediction. *Proteins Science*, v. 9, p. 1399–1401, 2000. Citado na página 38.
- SEARS, A.; BATRA, R. C. Macroscopic properties of carbon nanotubes from molecular mechanics simulations. *Phys. Rev. B*, v. 69, p. 235406, Jun 2004. Citado na página 1.
- SELLERS, B. D. *et al.* Towards better refinement of comparative models: predicting loops in inexact environments. *Proteins*, v. 72, p. 959–971, 2008. Citado na página 38.
- SENN, H. M.; THIEL, W. Qm/mm methods for biomolecular systems. *Angewandte Chemie International*, WILEY-VCH Verlag, v. 48, n. 7, p. 1198–1229, 2009. Citado na página 1.

- SHIROTA, M.; ISHIDA, T.; KINOSHITA, K. Absolute quality evaluation of protein model structures using statistical potentials with respect to the native and reference states. *Proteins*, v. 79, n. 5, p. 1550–1563, 2011. Citado na página 38.
- SILVI, B. Pressure effect on electron localization in solid lithium. *Structural Chemistry*, p. 1–9, 2017. Citado na página 1.
- SMITH, T. F.; HARTMAN, H. The evolution of class II aminoacyl-trna synthetases and the first code. *FEBS Letters*, v. 589, n. 23, p. 3499–3507, 2015. Citado 2 vezes nas páginas 31 e 32.
- STEWART, J. P. P. Application of localized molecular orbitals to the solution of semiempirical self-consistent field equations. *Int. J. Quant. Chem.*, v. 58, p. 133–146, 1996. Citado 3 vezes nas páginas 1, 40 e 43.
- STEWART, J. P. P. Application of the PM6 method to modeling proteins. *J. Mol. Model.*, v. 15, p. 765–805, 2009. Citado 3 vezes nas páginas 22, 39 e 40.
- STEWART, J. P. P. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *J. Mol. Model.*, v. 19, p. 1–32, 2013. Citado na página 39.
- TEE, W.; RIPEN, A. M.; MOHAMAD, S. B. The conformational dynamics of h2-h3n and s2-h6 in gating ligand entry into the buried binding cavity of vitamin d receptor. *Scientific Reports*, v. 6, p. 35937, 2016. Citado na página 37.
- TEMELSO, B. *et al.* Structural analysis of alpha-fetoprotein (AFP)-like peptides with anti-breast-cancer properties. *J. Phys. Chem. B*, v. 118, p. 4514–4526, 2014. Citado na página 39.
- THEOBALD, D. L. A formal test of the theory of universal common ancestry. *Nature*, v. 465, p. 219–222, 2010. Citado na página 10.
- UFIMTSEV, I. S.; MARTINEZ, T. J. Quantum chemistry on graphical processing units. 2. direct Self-Consistent-Fsield (SCF) implementation. *J. Chem. Theory Comput.*, v. 5, p. 3138, 2009. Citado na página 2.
- URQUIZA-CARVALHO, G. A.; FRAGOSO, W. D.; ROCHA, G. B. Assessment of semiempirical enthalpy of formation in solution as an effective energy function to discriminate native-like structures in protein decoy sets. *Journal of Computational Chemistry*, v. 37, n. 21, p. 1962–1972, 2016. Citado 2 vezes nas páginas 42 e 63.
- VALAS, R. E.; BOURNE, P. E. The origin of a derived superkingdom: how a gram-positive bacterium crossed the desert to become an archaeon. *Biol Direct*, v. 6, p. 16, 2011. Citado na página 10.
- VALLEAU, S. *et al.* Absence of selection for quantum coherence in the fenna-matthews-olson complex: a combined evolutionary and excitonic study. 2017. Citado 2 vezes nas páginas 6 e 12.
- VANGA, S. K.; SINGH, A.; RAGHAVAN, V. Effect of thermal and electric field treatment on the conformation of ara h 6 peanut protein allergen. *Innovative Food Science and Emerging Technologies*, v. 30, p. 79–88, Aug 2015. Citado na página 37.

- WARSHHEL, A.; LEVITT, M. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of Molecular Biology*, v. 103, n. 2, p. 227 – 249, 1976. Citado na página 1.
- WELACK, S. *et al.* Single-electron counting spectroscopy: Simulation study of porphyrin in a molecular junction. *Nano Letters*, v. 8, n. 4, p. 1137–1141, 2008. Citado na página 1.
- WILSON, C. *et al.* Using ancient protein kinases to unravel a modern cancer drug's mechanism. *Science*, v. 347, n. 6224, p. 882–886, 2015. Citado na página 12.
- WOESE, C. The universal ancestor. *PNAS*, v. 95, p. 6854–6859, 1998. Citado na página 10.
- WOESE, C. R.; FOX, G. E. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, v. 74, n. 11, p. 5088–5090, 1977. Citado 2 vezes nas páginas xi e 11.
- WOESE, C. R. *et al.* Aminoacyl-trna synthetases, the genetic code, and the evolutionary process. *Microbiology and Molecular Biology Reviews*, v. 64, n. 1, p. 202–236, 2000. Citado na página 74.
- WU, S.; SKOLNIK, J.; ZHANG, Y. Ab initio modeling of small proteins by iterative tasser simulations. *BMC Biology*, v. 5, p. 17, 2007. Citado 3 vezes nas páginas 32, 38 e 40.
- WU, X. *et al.* Molecular dynamics simulation of room-temperature ionic liquid mixture of [bmim][bf4] and acetonitrile by a refined force field. *Phys. Chem. Chem. Phys.*, v. 7, p. 2771–2779, 2005. Citado na página 1.
- XU, D.; WILLIAMSON, M. J.; WALKER, R. C. Advancements in molecular dynamics simulations of biomolecules on graphical processing units. *Annu. Rep. Comput. Chem.*, v. 6, p. 1, 2010. Citado na página 2.
- XU, D.; ZHANG, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins*, v. 80, p. 1715, 2012. Citado na página 38.
- YAN, Y. *et al.* Addressing recent docking challenges: A hybrid strategy to integrate template-based and free protein-protein docking. *Proteins: Structure, Function, and Bioinformatics*, 2017. Citado na página 33.
- YAO, N. Y.; O'DONNELL, M. E. Evolution of Replication Machines. *Crit Rev Biochem Mol Biol*, v. 51, n. 3, p. 135–149, 2016. Citado na página 10.
- YASUDA, S. *et al.* Free-energy function for discriminating the native fold of a protein from misfolded decoys. *Proteins*, v. 79, n. 7, p. 2161–2171, 2011. Citado na página 38.
- YI, X. *et al.* Ab initio study of the electronic structure, magnetic and dynamics properties in limo2 and namo2 (m=v and cr) as electrode cathode material for batteries. *Solid State Ionics*, v. 307, p. 26–34, 2017. Citado na página 1.
- YILMAZER, N. D. *et al.* Benchmark of electronic structure methods for protein–ligand interactions based on high-level reference data. *J. Theo. Comp. Chem.*, v. 14, n. 01, p. 169–175, 2015. Citado na página 39.

- YILMAZER, N. D. *et al.* Benchmark of electronic structure methods for protein-ligand interactions based on high-level reference data. *J. Theo. Comp. Chem.*, v. 14, n. 1, p. 1540001–1540013, 2015. Citado na página 39.
- YILMAZER, N. D.; KORTH, M. Comparison of molecular mechanics, semi-empirical quantum mechanical, and density functional theory methods for scoring protein–ligand interactions. *J. Phys. Chem. B*, v. 117, n. 27, p. 8075–8084, 2013. Citado na página 39.
- YILMAZER, N. D.; M, K. Enhanced semiempirical qm methods for biomolecular interactions. *Comp. Struct. Biotechn. J.*, v. 13, p. 169–175, 2015. Citado na página 39.
- YIN, S.; DOKHOLYAN, N. V. Fingerprint based structure retrieval using electron density. *Proteins*, v. 79, n. 1002-1009, 2011. Citado 4 vezes nas páginas 5, 21, 22 e 26.
- ZERNIKE, F. Diffraction theory of the knife-edge test and its improved form, the phase-contrast method. *Monthly Notices of the Royal Astronomical Society*, v. 94, p. 377–384, 1934. Citado na página 20.
- ZHANG, J.; ZHANG, Y. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PLoS One*, v. 5, p. e15386, 2010. Citado 2 vezes nas páginas 38 e 39.
- ZHANG, Y. Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, v. 18, n. 3, p. 342–348, 2008. ISSN 0959-440X. Citado na página 32.
- ZHANG, Y.; KOLINSKI, A.; SKOLNICK, J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophys. J.*, v. 85, p. 1145–1164, 2003. Citado 2 vezes nas páginas 32 e 38.
- ZHANG, Y.; SKOLNICK, J. Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc. Natl. Acad. Sci. USA*, v. 101, p. 7594–7599, 2004. Citado 2 vezes nas páginas 32 e 38.
- ZHAO, Y. *Criterion Functions for Document Clustering: Experiments and Analysis*. Tese (Doutorado em Ciência da Computação) — University of Minnesota, 2003. Citado na página 26.
- ZHAO, Y.; KARYPIS, G.; FAYYAD, U. Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, v. 10, n. 2, p. 141–168, 2005. Citado na página 26.
- ZHOU, J. *et al.* Svrcaf: an integrated score function for detecting native protein structures among decoys. *Proteins*, v. 82, n. 4, p. 556–564, 2014. Citado na página 38.
- ZINN, E. *et al.* In silico reconstruction of the viral evolutionary lineage yields a potent gene therapy vector. *Cell Reports*, v. 12, n. 6, p. 1056 – 1068, 2015. Citado na página 12.
- ZUCKERKANDL, E.; PAULING, L. Molecules as documents of evolutionary history. *Journal of Theoretical Biology*, v. 8, n. 2, p. 357 – 366, 1965. ISSN 0022-5193. Citado na página 11.

Apêndices

Aqui nos Apêndices comentários são tecidos ligando algumas propriedades da expansão de Canterakis-Zernike com propriedades que se possam calcular de densidades expressas como expansões deste tipo. Até onde foi investigado e da conclusão deste trabalho, não foi encontrada bibliografia pertinente na qual tais propriedades foram discutidas em um contexto envolvendo densidades eletrônicas moleculares descritas na base de Canterakis-Zernike.

Para discutir os temas a frente, será introduzida uma nova notação. Seja uma densidade eletrônica escrita na base de Canterakis-Zernike dada por:

$$\rho(\mathbf{r}) = \sum_{k=0}^K \sum_{l=0}^L \sum_{m=-l}^l {}_k\omega_l^m {}_k\tilde{Z}_l^m(\mathbf{r}). \quad (1)$$

Denotam-se suas projeções através de um símbolo ρ , indexado por k , l ou m . Os índices que não aparecem no símbolo da projeção representam aqueles para os quais o somatório acontece. Dessa forma, tem-se que as projeções de um índice são:

- ${}_k\rho(\mathbf{r}) = \sum_l \sum_m {}_k\omega_l^m {}_k\tilde{Z}_l^m(\mathbf{r})$
- $\rho_l(\mathbf{r}) = \sum_k \sum_m {}_k\omega_l^m {}_k\tilde{Z}_l^m(\mathbf{r})$
- $\rho^m(\mathbf{r}) = \sum_k \sum_{l \geq |m|} {}_k\omega_l^m {}_k\tilde{Z}_l^m(\mathbf{r})$

As de dois índices são:

- ${}_k\rho_l(\mathbf{r}) = \sum_m {}_k\omega_l^m {}_k\tilde{Z}_l^m(\mathbf{r})$
- ${}_k\rho^m(\mathbf{r}) = \sum_{l \geq |m|} {}_k\omega_l^m {}_k\tilde{Z}_l^m(\mathbf{r})$
- $\rho_l^m(\mathbf{r}) = \sum_k {}_k\omega_l^m {}_k\tilde{Z}_l^m(\mathbf{r})$

E, para propósitos de completude da notação, define-se também a projeção de zero índices que é a própria densidade eletrônica ρ e as projeções de três índices ${}_k\rho_l^m(\mathbf{r}) = {}_k\omega_l^m {}_k\tilde{Z}_l^m(\mathbf{r})$, que nada mais são que versões escalonadas das funções de base ${}_k\tilde{Z}_l^m(\mathbf{r})$.

Também se definem as densidades complementares ${}_k\bar{\rho}_l^m(\mathbf{r}) = \rho(\mathbf{r}) - {}_k\rho_l^m(\mathbf{r})$ que são apenas aquilo que resta da densidade original após a remoção da projeção. Isto faz com que a densidade total possa ser escrita como a soma de uma projeção e seu complemento.

A notação $W[{}_k\rho_l^m(\mathbf{r})]$ é utilizada ainda para denotar os coeficientes de uma projeção em forma de vetor.

Função de Distribuição Radial da Densidade Eletrônica

Previamente utilizamos o fato de que integrar a expansão de Canterakis-Zernike em um esfera cancela todos os termos com exceção do primeiro, uma relação expressa na Equação 2:

$$\sqrt{V(r^*)}_0 \omega_0^0(r^*) = \int_{\Omega} \rho_0^0 d\Omega \quad (2)$$

nota-se, por definição, que a projeção $\rho_0^0(\mathbf{r})$ possui a parte angular igual a 1, o que significa que ela só é dependente da coordenada radial tal que $\rho_0^0 = \sum_k k \omega_0^0 \tilde{Z}_{k0}(r)$. Neste caso a Equação (2) reduz para $\sqrt{V(r^*)}_0 \omega_0^0(r^*) = \int_r^{r^*} \rho_0^0 r^2 dr$. Uma vez que $\omega_0^0 \in W[\rho_0^0]$ então $\int_r \rho_0^0 r^2 dr$ será diferente de 0 e exatamente igual a $\sqrt{V(r^*)}_0 \omega_0^0$.

Levando em conta o fato de $\rho_0^0(r)$ ser rotacionalmente invariante, pode-se ver que $\rho_0^0(r)$ na verdade descreve a densidade total de elétrons contida em cada uma das infinitesimalmente finas camadas concêntricas arranjadas em volta do centro da densidade. Então $\rho_0^0(r) = G(r)$, onde $G(r)$ é a função de distribuição radial de elétrons em volta da origem do sistema de coordenadas.

O mesmo pode ser dito de cada uma das projeções ${}_k \rho_0^0(r) \forall k > 0$ uma vez que elas descrevem como uma densidade completamente isotrópica é "moldada" como argila, tomando densidade de uma camada e colocando em outra, sem criar ou destruir densidade. Adicionalmente, como ρ_0^0 não muda com rotações, o vetor de seus coeficientes $W[\rho_0^0]$ também é invariante. Chamamos $W[\rho_0^0]$ o vetor $\vec{Z}_{k0}$.

Uma vez que, por construção, a densidade eletrônica pode ser escrita como a soma de uma projeção e seu complemento, temos que

$$\rho(\mathbf{r}) = G(r) + \bar{G}(\mathbf{r}) \quad (3)$$

Sabe-se que $\int_{\Omega} \rho d\Omega = \int_{\Omega} G d\Omega$, então disso conclui-se que $\int_{\Omega} \bar{G} d\Omega = 0$. Isto significa que cada aumento que $\bar{G}$ causa na densidade local $\rho(\mathbf{r})$ nas vizinhanças de algum ponto $\mathbf{r} \in \Omega$ é perfeitamente contrabalançado por um decréscimo de igual proporção a um ponto diferente $\mathbf{r}' \in \Omega$. Enquanto $G(r)$ descreve como uma densidade perfeitamente isotrópica é rearranjada em um conjunto de camadas em volta da origem, $\bar{G}(\mathbf{r})$ descreve como a isotropia de cada camada em quebrada ainda mais para formar regiões discretas de densidade eletrônica concentrada ao ausente, simplesmente mudando densidade de lugar, sem criar ou destruir qualquer coisa.

Momentos de Multipolo da Densidade de Canterakis-Zernike

O potencial eletrostático de longo alcance $U(\mathbf{r})$ pode ser escrito como uma expansão do tipo expresso na Equação (4).

$$U(\mathbf{r}) = \sum_l \sum_{m=-l}^l \frac{1}{r^{l+1}} Q_l^m Y_l^m(\theta, \phi) \quad (4)$$

Os termos $\frac{1}{r^{l+1}} Y_l^m(\theta, \phi)$ são multipolos e os coeficientes Q_l^m são os momentos de multipolos. Multipolos são estruturas úteis no estudo de potenciais elétricos complexos, como o gerado por uma densidade eletrônica. Ao invés de calcular o potencial em cada ponto do espaço, uma possível abordagem é decompor o campo em termos de multipolos e a expansão resultante deve conter características gerais do campo e possibilitar sua reconstrução.

Multipolos são estruturas que possuem um campo elétrico bem conhecido. Um monopolo ($l = 0$) se parece com o potencial elétrico gerado por uma distribuição uniforme de carga, dipolos ($l = 1$) se comportam como uma seta com um pólo norte e um pólo sul. Tem-se, pois, estruturas de complexidade crescente, porém bem descritas. O princípio por trás da expansão de multipolos é decompor um potencial em termos de estruturas conhecidas, com um número de pólos igual a 2^l . Em essência, os momentos de multipolos descrevem o quanto um sistema de descrição difícil se comporta como uma soma de outros sistemas de descrição mais exata. Assim, o primeiro momento descreve o quanto a molécula se comporta como uma esfera de carga uniforme, os momentos de dipolo descrevem o quanto a densidade eletrônica se comporta como um dipolo ao longo de cada eixo e dessa forma sucessivamente para os termos de maior ordem.

Em geral a descrição em termos de multipolos é mais exata para descrever a influência do potencial elétrico para além dos limites da molécula, onde os detalhes dos potenciais eletrostáticos de cada átomo individual se simplifica com a distância. De forma análoga, por exemplo, ao campo magnético da Terra. Muito próximo da superfície do planeta, as linhas de campo magnético possuem detalhes e nuances que podem dificultar sua descrição exata, mas de longe, no espaço, a Terra pode ser aproximada como um grande dipolo magnético, e a influência deste dipolo nas adjacências do planeta torna-se uma aproximação razoável para o campo magnético real. Por essa razão, os multipolos são particularmente importantes quando se está tentando descrever forças intermoleculares.

Naturalmente, o potencial elétrico gerado pela densidade eletrônica de uma molécula

é uma quantidade importante e bastante característica da molécula. Os momentos da expansão de multipolos Q_l^m de uma distribuição de carga $\rho(\mathbf{r})$ são dados pela Equação (5).

$$Q_l^m \sim \int_{\Omega} r^l \rho(\mathbf{r}) Y_l^m(\theta, \phi) d\Omega = \langle r^l Y_l^m(\theta, \phi) | \rho(\mathbf{r}) \rangle \quad (5)$$

A densidade eletrônica é escrita em termos de bases atômicas:

$$\rho = \sum_k \sum_l \sum_m {}_k\omega_l^m \tilde{Z}_{kl}(r) N_l^m Y_l^m(\theta, \phi),$$

onde as funções $\tilde{Z}_{kl}$ são escritas como mostra a Equação (5.15). $\mathbb{P}$ é o espaço recíproco de ρ em Ω na base de Canterakis-Zernike. Da mesma forma, pode-se dizer que $\mathbb{Y}$ é o espaço recíproco do potencial U na base dos multipolos.

Relembrando a Equação (9.1) e também que ${}_k\tilde{Z}_l^m(\mathbf{r}) = \tilde{Z}_{kl}(r) N_l^m Y_l^m(\theta, \phi)$, conclui-se que:

$$\langle {}_0\tilde{Z}_l^m | \rho \rangle = \sqrt{\frac{2l+3}{r^{*2l+3}}} N_l^m \langle r^l Y_l^m | \rho \rangle = \sqrt{\frac{2l+3}{r^{*2l+3}}} N_l^m Q_l^m. \quad (6)$$

Como a densidade eletrônica é expressa em um conjunto de base ortonormal, então $\langle {}_0\tilde{Z}_l^m | \rho \rangle = {}_0\omega_l^m$ e, assim:

$$Q_l^m = {}_0\omega_l^m \sqrt{\frac{r^{*2l+3}}{2l+3}} \frac{1}{N_l^m}. \quad (7)$$

A Equação (7) pode ser usada para escrever o potencial como uma função dos momentos de Canterakis-Zernike ao substituí-la ela na Equação (4), assim conectando os espaços recíprocos $\mathbb{P}$ e $\mathbb{Y}$ e, por consequência, $\rho(\mathbf{r})$ a $U(\mathbf{r})$. Dessa forma, apenas através da expansão da densidade em termos dos momentos de Canterakis-Zernike tem-se, sem que nenhum outro procedimento numérico extensivo seja necessário, informações sobre a eletrostática de longo alcance da densidade eletrônica através da Equação (8).

$$U(\mathbf{r}) = \sum_l \sum_{m=l} {}_0\omega_l^m \sqrt{\frac{r^{*2l+3}}{2l+3}} \frac{1}{r^l} Y_l^m(\theta, \phi). \quad (8)$$

Subspaços de $\mathbb{P}$ e $\mathbb{Y}$

Cada uma das projeções de ρ gera um espaço de funções denominado $\mathbb{P}$ indexado da mesma forma que a projeção que lhe deu origem. Assim, a projeção ${}_k\rho(\mathbf{r})$, por exemplo, gera o espaço ${}_k\mathbb{P}$ e, da mesma forma, para outras projeções e seus complementos e o espaço $\mathbb{P}$ é o espaço de todas as densidades possíveis de serem escritas com a expansão de Canterakis-Zernike. É evidente que todos os espaços gerados pelas projeções e seus complementos são subspaços de $\mathbb{P}$.

Tem-se um interesse particular nos subspaços ${}_k\mathbb{P}_l$ porque os harmônicos esféricos são uma base do grupo de rotação tridimensional $SO(3)$. Isto significa que para qualquer rotação de uma função $f(\theta, \phi) = \sum_l \sum_{m=-l}^l c_l^m Y_l^m(\theta, \phi)$, a função rotacionada tem uma representação na mesma base de harmônicos esféricos. Isto significa que quando ρ é rotacionado, qualquer projeção definida nos subspaços ${}_k\mathbb{P}_l$ continuará a ser definida no mesmo subespaço, sem que seja necessário a adição de funções de base adicionais ao espaço. Isto tem a consequência de tornar a norma do vetor $W[{}_k\rho_l]$ invariante sob rotação 3D.

Cada subespaço ${}_k\mathbb{P}_l$ é um espaço $(2l + 1)$ -dimensional e existem K espaços deste tipo para cada l . Por causa disto, existem K subspaços de dimensão 1, que são na verdade escalares, subspaços do tipo ${}_k\mathbb{P}_0^0$. Uma vez que as projeções nos subspaços ${}_k\mathbb{P}_0^0$ não tem dependência com a parte angular, sendo rotacionalmente invariantes.

Uma vez que o conjunto de base no qual ρ é descrito é ortonormal, o conjunto de base de cada subespaço $(2l + 1)$ -dimensional também será. Isto faz com que a representação $W[{}_k\rho_l]$ seja topologicamente equivalente a um vetor no espaço $\mathbb{R}^{2l+1}$, efetivamente mapeando o vetor $({}_k\omega_l^{-l}, {}_k\omega_l^{-l+1}, \dots, {}_k\omega_l^0, \dots, {}_k\omega_l^l)$ em ${}_k\mathbb{P}_l$ no vetor $(x_0, x_1, \dots, x_{2l+1})$, escrito em um sistema de coordenadas cartesianas em $\mathbb{R}^{2l+1}$.

Da mesma forma que o sistema de coordenadas cartesianas utilizado para representar pontos em $\mathbb{R}^3$ pode ser convertido no sistema de coordenadas esféricas, existem sistemas de coordenadas *hiperesféricas* que generalizam as coordenadas polares e esféricas em 2D e 3D para equivalente de dimensionalidade geral. Um ponto em ${}_k\mathbb{P}_l$ expresso em coordenadas hiperesféricas terá uma coordenada $r \in [0, \infty)$, que será sua coordenada radial e generaliza a distância da origem do sistema de coordenada. O ponto também terá uma coordenada azimutal, ou simplesmente azimute $\phi \in [0, 2\pi]$, que generaliza a noção de circundar o equador da hiper-esfera e $(2l - 1)$ latitudes $\theta_i \in [0, \pi] \forall i \in [0, 2l - 1]$ que generalizam a noção de inclinação do ponto na superfície da esfera em relação ao eixo que inclui o centro e o pólo norte.

Durante a rotação de ρ em Ω , as normas dos vetores $W[{}_k\rho_l]$ devem permanecer as mesmas porque os Harmônicos Esféricos são a base do grupo $SO(3)$. Nas coordenadas

hiperesféricas isto significa que todas as coordenadas radiais devem permanecer as mesmas para cada representação em cada um dos subespaços ${}_k\mathbb{P}_l$. Manter a coordenada radial e variar apenas as coordenadas angulares é a definição de uma rotação em torno do sistema de coordenadas. As rotações dos vetores $W[{}_k\rho_l] \in {}_k\mathbb{P}_l$ ao redor de seus respectivos sistemas de coordenadas são denominadas *Rotações Recíprocas*. Obviamente elas são chamadas assim porque os vetores sendo rotacionados são definidos em subespaços do espaço recíproco de ρ .

Durante a rotação de ρ em Ω a representação $W[{}_k\rho_l] \in {}_k\mathbb{P}_l$ é permitida rotacionar em torno da origem de seu sistema de coordenadas hipersféricas, mas não é permitido que o ponto definido pelo vetor varie sua distância da origem. O que acontece é que todas as representações vetoriais em ${}_k\mathbb{P}_l$ serão rotacionadas em torno de sua origem de forma que a combinação final seja a densidade rotacionada em Ω . Existem muito mais formas de se rotacionar a densidade nos subespaços recíprocos do que no espaço real, o que significa que de todas as possíveis rotações recíprocas que podem ser aplicadas à densidade ρ , apenas algumas combinações particulares delas corresponderão a rotações reais de ρ em Ω .

Cada coordenada radial terá, agora, um k e um l específicos associados com elas que não mudará quando a densidade for rotacionada. Eles podem ser organizados em uma matriz $\mathbf{F}$ cujos vetores coluna contêm as coordenadas radiais de todos os subespaços ${}_k\mathbb{P}_l$ de mesma dimensionalidade. Assim, o vetor coluna $\vec{F}_{k0}$ é o vetor contendo todos os $|W[{}_k\rho_0]| \forall k \geq 0$. Estes vetores são rotacionalmente invariantes, assim como toda a matriz. Comparações entre tais quantidades em densidades diferentes forma uma base sólida para a comparação de densidades de forma rotacionalmente invariante.

Uma análise similar poder ser feita para Y , exceto que os multipolos não dependem do índice k , o que significa que existe um único subespaço Y_l de dimensionalidade $(2l + 1)$ para cada l , e que, portanto, existe uma única coordenada radial nos subespaços recíprocos associados a cada l . Todas essas coordenadas radiais podem ser combinadas em um único vetor $\vec{Q}$, também rotacionalmente invariante.

Mesmo $\mathbf{F}$ sendo rotacionalmente invariante, para uma dada densidade $\rho \in \Omega$, apenas uma pequena parcela de todas as densidades em $\mathbb{P}$ com a mesma matriz $\mathbf{F}$ representarão versões rotacionadas de ρ . É possível escrever um número infinito de densidades com mesma $\mathbf{F}$ simplesmente aplicando rotações recíprocas aos pontos definidos nos subespaços de $\mathbb{P}$. Existem simplesmente muito mais graus de liberdade no ato de rotações n -dimensionais aplicadas nos subespaços recíprocos que na rotação tridimensional da densidade. Não é sabido o que acontece à densidade quando tais graus de liberdade são explorados, mas existem algumas informações que podem ser deduzidas.

Uma vez que $\mathbf{F}$ não muda sob rotações, então a coluna 0 da matriz também não muda. Tal coluna é o vetor $\vec{Z}_{k0}$, definido previamente, relacionado à projeção $\rho_0^0(r)$ e a estrutura de camadas concêntricas em torno da origem da densidade. O que é importante

sobre isso é que, como cada componente $\vec{Z}_{k0}$ é definido em um espaço 1D, eles são suas próprias coordenadas radiais e não possuem nem latitudes nem azimutes nos quais se aplicam rotações e portanto permanecem constantes nas rotações recíprocas para explorar o espaço das densidades. Isto significa dizer que, para qualquer conjunto de rotações no espaço recíproco, o vetor $\vec{Z}_{k0}$ permanecerá o mesmo.

Uma vez que $\vec{Z}_{k0}$ permanece o mesmo durante tais transformações, e ${}_0\omega_0^0 \in \vec{Z}_{k0}$, então nem a distribuição radial de elétrons nem o número total de elétrons contido na densidade mudará. Com isso se conclui que quando se aplicam rotações recíprocas em uma densidade, a densidade transformada será diferente da primeira apenas na forma como a isotropia de cada camada infinitesimal de densidade em torno da origem é quebrada. Neste sentido, $\vec{Z}_{k0}$ é rotacionalmente invariante até mesmo em relação a rotações recíprocas. Utilizando-se rotações recíprocas é possível explorar diferentes formas de $\bar{G}(\mathbf{r})$ para cada densidade de mesmo $G(r)$, número de elétrons e matriz $\mathbf{F}$.

Forma e Simetria da Densidade de Canterakis-Zernike

Como discutido anteriormente, as funções de base ${}_k\tilde{Z}_l^m(\mathbf{r})$ são um produto de um polinômio radial e um Harmônico Esférico. As funções radiais quebram a isotropia de uma densidade eletrônica uniforme de forma radialmente simétrica e os Harmônicos Esféricos quebram a simetria de cada camada ao longo das coordenadas angulares. O modo particular em que isso acontece é mostrado na Figura 35.

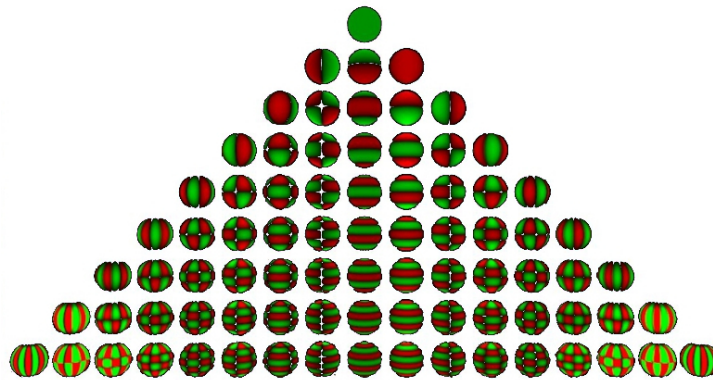


Figura 35 – Bandas de Harmônicos Esféricos até $l = 9$. As partes vermelhas indicam valores negativos e as partes verdes indicam valores positivos.

Como é possível ver na figura, a soma total de partes positivas e negativas é igual a zero para todos os Harmônicos Esféricos com exceção de $l = 0$, que é a densidade eletrônica média na esfera de integração. Esta propriedade significa que os Harmônicos Esféricos meramente modulam a densidade contida em cada camada concêntrica, sem destruir nem criar densidade no processo.

Os coeficientes ${}_k\omega_l^m$ indicam se a modulação deve ser feita da forma como se vê na Figura 35 ou na forma inversa, com as partes positivas se tornando negativas e vice-versa. Os coeficientes indicam ainda se os valores somados e subtraídos devem ser muito distantes ou muito próximos de 0, isto é, a intensidade com que a função de base está sendo aplicada na densidade total. Uma projeção ${}_k\rho_l^m(\mathbf{r})$, por exemplo, será bem sutil quando $|{}_k\omega_l^m|$ for muito pequeno, e bem intensa caso contrário. Quanto mais intensa a projeção, mais participação ela tem na forma final da densidade. Densidades com grande fração da intensidade total contidas em poucas projeções costumam ser mais simples e simétricas que aquelas em que a intensidade está mais distribuída pelas projeções.

Uma vez que as partes radiais são rotacionalmente simétricas, elas modificam seus Harmônicos Esféricos associados em cada direção radial da mesma forma, o que significa

que elas preservam alguma simetria dos Harmônicos Esféricos e a transfere para a densidade final. Os Harmônicos Esféricos, por sua vez, têm simetrias muito bem conhecidas. De fato, eles são classificados como Tesserais, Zonais ou Setorais dependendo de suas propriedades de simetria, como visto na Figura 36.

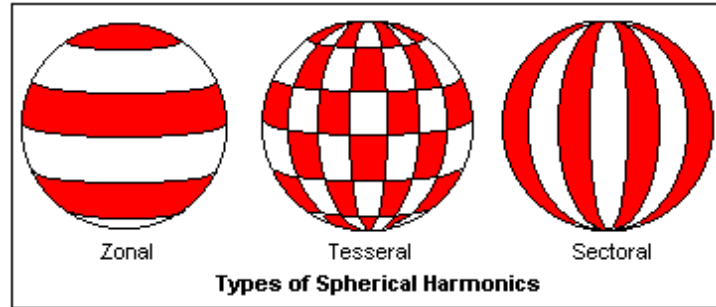


Figura 36 – Tipos de Harmônicos Esféricos segundo suas propriedades de simetria.

Durante uma rotação, $|W[k\rho_l]| = F_{kl}$ deve permanecer constante. Isto quer dizer que, se existe uma orientação para a qual $F_{kl} = W[k\rho_l^\mu]$ para algum $m = \mu$ específico, então todos os outros componentes de $W[k\rho_l]$ serão zero e $k\rho_l = k\rho_l^\mu$ nesta orientação em particular. Quando isto acontece, diz-se que ρ está *alinhada* com $k\rho_l^\mu$. Uma projeção é dita ser alinhável com a densidade se existir uma rotação que alinhe a projeção com a densidade. Se duas ou mais projeções podem ser alinhadas com a densidade simultaneamente, elas são chamadas de co-alinháveis.

Se uma densidade for alinhada com várias projeções que tem um alto F_{kl} , então é razoável assumir que tais projeções produzem um efeito decisivo na forma total da densidade. Encontrar a orientação para a qual a densidade se alinha com o maior número de projeções simultaneamente é uma forma de estudar possíveis elementos de simetria da densidade. À medida que a densidade se alinha com mais e mais projeções, é de se esperar que seu formato e sua simetria comece a espelhar elementos de simetria que todas as projeções alinhadas tem em comum.

Alinhamentos também podem ser utilizados para descobrir representações reduzidas da densidade. Um alinhamento com a projeção $k\rho_l^\mu$, por exemplo, torna $k\omega_l^m = 0 \forall m \neq \mu$. Isto significa que esses graus de liberdade da base Canterakis-Zernike podem ser ignorados para representar a molécula, criando uma representação reduzida da densidade. Quanto maior for o alinhamento com projeções de alto l , mais graus de liberdade serão desnecessários para a representação da molécula. O número de graus de liberdade utilizados pela representação reduzida é uma medida da simplicidade e simetria da densidade.

Se for definido um sistema de coordenadas tal qual visto na Figura 37 para designar uma rotação em torno de um eixo, é possível associar uma rotação a uma tripla (θ, ϕ, ψ) única.

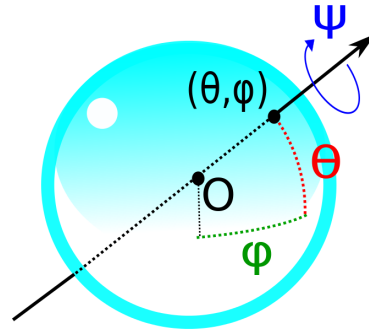


Figura 37 – Sistema de coordenadas (θ, ϕ, ψ) utilizado para especificar uma rotação em torno de um eixo.

Ao explorar o espaço das triplas possíveis, tem-se uma forma numérica de investigar as rotações e verificar quais conjuntos de alinhamentos acontecem em cada rotação em particular. Isto possibilita saber a quantidade real de graus de liberdade necessários para se especificar a forma da densidade na base de Canterakis-Zernike.