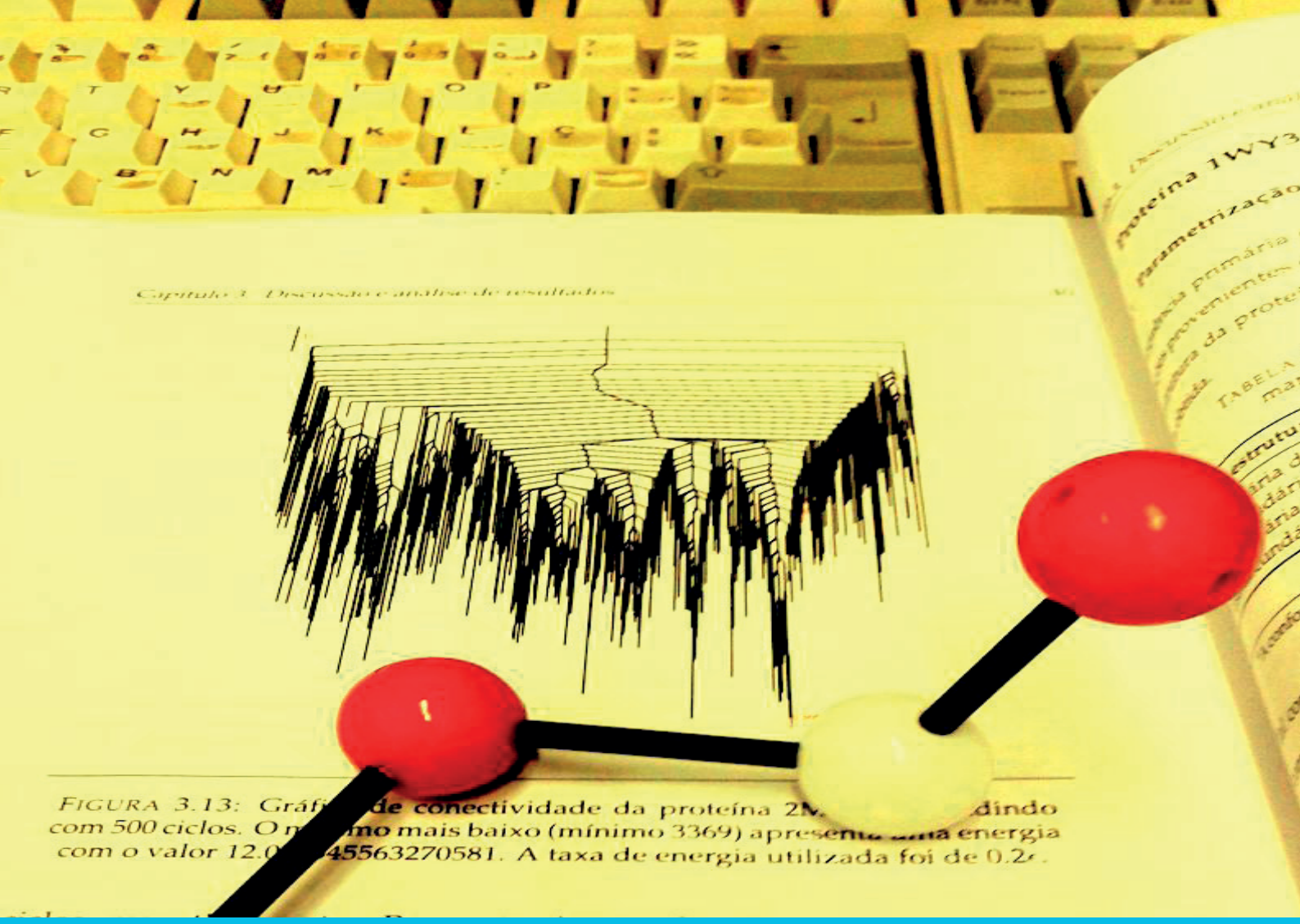


Wasina Ribeiro dos Santos Fins

ESTUDO COMPUTACIONAL DA ESTRUTURA E ENERGÉTICA DE MODELOS SIMPLES DE PROTEÍNAS

Mestrado em Química

Departamento de Química

FCTUC

Setembro de 2016

UNIVERSIDADE DE COIMBRA

Wasina Ribeiro dos Santos Fins

ESTUDO COMPUTACIONAL DA ESTRUTURA E ENERGÉTICA DE MODELOS SIMPLES DE PROTEÍNAS

Dissertação apresentada para provas de Mestrado em Química, Área de especialização em Química Avançada e Industrial

Orientador: Professor Doutor Jorge Manuel Campos Marques

Setembro, 2016

Universidade de Coimbra

Agradecimentos

Os meus agradecimentos vão ao meu orientador Professor Jorge Marques pelo seu incessante e incondicional apoio ao longo da elaboração do presente trabalho, aos meus pais, José António Fins e Maria Judite Dias dos Santos, aos meus irmãos, aos meus amigos e a todos que direta ou indiretamente permitiram que este trabalho fosse realizado.

Índice

Lista de figuras

Lista de tabelas

Abreviaturas

1L2Y	Proteína PDB com domínio estrutural "trp-cage"
1TK7	Aglomerado PDB com domínios estruturais "ww-domain"
1WY3	Proteína PDB com domínio estrutural vilina
2JOF	Proteína PDB com domínio estrutural "trp-cage"
2K6N	Proteína PDB com domínio estrutural vilina
2LL5	Proteína PDB com domínio estrutural "trp-cage"
2LSE	Proteína PDB com domínio estrutural com quatro hélices-α
2LTY	Aglomerado PDB com domínio estrutural "ww-domain"
2MWE	Proteína PDB com domínio estrutural "ww-domain"
2N8S	Proteína PDB com domínio estrutural "ww-domain"
2RJX	Proteína PDB com domínio estrutural vilina
3MYC	Proteína PDB com domínio estrutural vilina
3TOL	Aglomerado PDB com duas proteínas com domínio estrutural com quatro hélices-α
3U3B	Proteína PDB com domínio estrutural com quatro hélices-α
3UC7	Proteína globular PDB com domínios estruturais "trp-cage"
4ZKH	Proteína globular PDB com vários domínios estruturais com quatro hélices-α
4ZLW	Proteína globular PDB com vários domínios estruturais com quatro hélices-α
5AHT	Proteína PDB com domínio estrutural "ww-domain"
5FJD	Aglomerado PDB com duas proteínas com quatro hélices-α
5I1O	Proteína globular PDB com domínios estruturais vilina
5I1S	Proteína globular PDB com domínios estruturais vilina
Ala (A)	Alanine
Arg (R)	Arginine

Asn (N)	Asparagine
Asp (D)	Aspartic acid
BH	Basin-Hopping
BLN	Resíduos hidrofóbicos (B), hidrofílicos (L) e neutros (N) do modelo "coarse-grained" utilizado no presente trabalho
CFSSP	Chou & Fasman Secondary Structure Prediction Server
Cys (C)	Cysteine
DisconnectionDPS	Programa utilizado para construção de gráficos de conectividade
DSSP	Dictionary of Secondary Structure of Proteins
FASTA	Ficheiro com uma pequena descrição e a sequência de aminoácidos de uma determinada proteína
FORTRAN	Formula Translation
G77	Tipo de compilador FORTRAN
Gfortran	Tipo de compilador FORTRAN
Gln (Q)	Glutamine
Glu (E)	Glutamic acid
Gly (G)	Glycine
GMIN	Programa utilizado em otimizações globais, estudos de propriedades termodinâmicas e determinação dos MFETs por utilização do método "basin-hopping"
Gminconv2	Utilitário utilizado para gerar uma distribuição contínua dos valores obtidos nos cálculos dos MFETs
His (H)	Histidine
Ile (I)	Isoleucine
Leu (L)	Leucine
Lys (K)	Lysine
Met (M)	Methionine
MFET	Mean First Encounter Time
OPTIM	Programa utilizado no estudo de otimizações geométricas e determinação de caminhos de conexão que interligam os vários mínimos gerados
PATHSAMPLE	Controlador ("driver") do OPTIM. Utiliza amostras de caminhos de

	conectividade gerados pelo OPTIM
PC1	First Principal component
PC2	Second Pincipal Component
PCA	Principal Component Analysis
PDB	Protein Data Bank
Phe (F)	Phenylalanine
Pro (P)	Proline
Rancoords	Random coordinates
Ser (S)	Serine
Thr (T)	Threonine
Trp (W)	Tryptophan
TRP-Cage	Domínimo estrutural com uma hélices-α, uma hélice-3/10 e uma parte pequena com resíduos não atribuídos, os quais se apresentam em forma de uma pequena gaiola
TS	Tansition State
TS	Tansition State
Tyr (Y)	Tyrosine
Val (V)	Valine
VMD	Visual Molecular Dynamics
WW-Domain	Domínimo estrutural com três folhas-β antiparalelas

Resumo

A estrutura das proteínas assume grande importância na sua função celular. Embora muitas das proteínas globulares “dobrem” para um estado nativo bem definido, a complexidade estrutural é muito grande, ou seja, apresentam grandes "frustrações" geométricas devido ao número de graus de liberdade envolvidos. É habitual existirem outras estruturas de baixa energia separadas por barreiras muito elevadas, mas que possuem padrões de contacto entre pares de resíduos bem diferentes dos do estado nativo. Naturalmente, a ocorrência de tais estruturas conduz a uma alteração na função da proteína. Um estudo detalhado da estrutura e energética destes sistemas constitui, assim, uma peça fundamental para compreender o processo de enrolamento e desenrolamento de proteínas.

Neste trabalho foram feitas caracterizações energéticas e estruturais de modelos simples de proteínas, construção e análise de gráficos de conectividade ("disconnectivity graphs"). Os gráficos de conectividade permitiram analisar as relações energéticas entre os mínimos e estados de transição dos diferentes isómeros gerados em cada sistema estudado. Foram ainda feitos estudos para caracterização dos MFET (tempos médios para encontrar o mínimo global) associados a estes sistemas.

Para modelar os sistemas de proteínas foi utilizado um modelo de potencial "coarse-grained" (modelo a escala maior do que a atómica) com três tipos de resíduos - modelo BLN. As caracterizações estruturais dos sistemas de proteínas estudados foram feitas tendo em conta a manipulação dos parâmetros que caracterizam os seus motivos estruturais no modelo de potencial BLN (parâmetros associados aos ângulos diedros e as interações entre pares de resíduos que não apresentam ligações peptídicas essencialmente) em função das dobragens em seus estados nativos. Compreender os motivos estruturais das proteínas nos permite compreender seus domínios, bem como as suas próprias unidades funcionais (estruturas espaciais tridimensionais - estruturas terciárias) e as estruturas globulares (conglomerados de estruturas terciárias - estruturas quaternárias).

Os mínimos globais e outros mínimos com baixas energias foram encontrados por

otimizações "Basin-Hopping" (método que utiliza simulações de Monte Carlos) com utilização do programa GMIN. Os caminhos de conexão entre os mínimos foram encontrados com a utilização do programa OPTIM e melhorados por criação de bases de dados de pontos estacionários e expansão da mesma por utilização do programa PATHSAMPLE.

De forma geral, os resultados obtidos mostraram que o tamanho e a complexidade estrutural dos diferentes sistemas estudados estão diretamente relacionados com as suas frustrações geométricas, as quais têm influência sobre os tempos médios para encontrar o mínimo global (MFET). Por outro lado, foi possível ter boas aproximações para os sistemas modelados com a utilização do modelo de potencial BLN, o que nos permite ter uma ideia dos perfis de energia apresentados pelas proteínas referentes a estes sistemas.

Abstract

The protein structure has great importance in its cellular function. Although many of the globular proteins "fold" into a well defined native state, the structural complexity is very high, i.e., exhibit large geometric "frustrations" due to the number of degrees of freedom involved. Normally there are other low energy structures separated by high barriers but which have contact patterns between pairs of residue very different from the native state. Naturally, the occurrence of these structures leads to a change in protein function. A detailed study of the structure and energy of these systems constitutes therefore a critical piece to understand the process of winding and unwinding proteins.

In this work were made energetic and structural characterizations of simple protein models, construction and analysis of disconnectivity graphs. The disconnectivity graphs were used to study the energetic relationship between minimum and transition states of the various generated isomers in each analyzed system. There were also made studies to characterize the MFET (Mean First Encounter Time) associated with these systems.

To model the protein systems was used a "coarse-grained" potential model (model on a larger scale than the atomic) with three types of residues - BLN model. The structural characterizations of the studied protein systems were made having regard to the handling of the parameters that characterize their structural motifs in the BLN potential model (parameters associated with the dihedral angles and the interactions between residue pairs that do not present peptide bonds essentially) according to the foldings in their native states. Understanding the structural motifs of proteins allows us to understand their domains and their own functional units (three-dimensional spatial structures - tertiary structures) and globular structures (clusters of tertiary structures - quaternary structures).

The global minimum and other minimum with low energies were found by "Basin-Hopping" optimization (method that uses Monte Carlos simulations) with the use of GMIN program. The connecting paths between the minimum were found with the use of OPTIM program and improved by creation and expanding of stationary points databases with PATHSAMPLE program.

Overall, the results showed that the size and structural complexity of the different systems studied are directly related to their geometrical frustrations, which have influence on the average time to find the global minimum (MFET). On the other hand, it was possible to have good approximations for the modeled systems using the BLN potential model, which allows us to have an idea of the energy profiles presented by the proteins related to these systems.

Capítulo 1

Introdução

As proteínas são responsáveis por muitas das funções biológicas essenciais para vida, as quais estão diretamente associadas à estrutura destas. Em função do tipo de representação estrutural, as proteínas são classificadas por *estrutura primária*, *estrutura secundária*, *estrutura terciária* e *estrutura quaternária*. A estrutura primária apresenta a sequência de aminoácidos característicos da proteína; a estrutura secundária apresenta-se sob a forma de folhas$-\beta$, hélices$-\alpha$, dobras, etc.; a estrutura terciária apresenta arranjos estéreo-espaciais de conformações secundárias; e a estrutura quaternária corresponde a conglomerados de conformações terciárias, ou seja, agrupamento de várias proteínas interligadas (Branden and Tooze, 1999); na figura 1.1 são apresentadas ilustrações da sequência de aminoácidos (estrutura primária), da estrutura secundária, terciária e quaternária em conformações com hélices$-\alpha$ (Branden and Tooze, 1999).

As estruturas secundárias combinam-se para formar grupos estruturais mais complexos: motivos e domínios estruturais (Branden and Tooze, 1999; Richardson, 1981). Tais estruturas são determinantes para caracterização das conformações tridimensionais[1] das proteínas (Chiang et al., 2007; Branden and Tooze, 1999; Richardson, 1981) e consequentemente as suas propriedades funcionais (Branden and Tooze, 1999).

As proteínas tendem a constituir diversos motivos estruturais, como é o caso dos motivos $\alpha-\beta-\alpha$, $\beta-\alpha-\beta$, "$\beta-$*hairpin*", "*greek-key*", etc., a partir dos quais podem-se formar diversos domínios estruturais, ou seja, unidades estruturais mais estáveis e independentes, como é o caso dos domínios *vilina*, "*ww-domain*", "*rtp-cage*", etc., por sua vez, estes podem

[1] Estruturas terciárias.

combinar-se para formar estruturas ainda mais complexas (estruturas terciárias ou estruturas quaternárias) (Richardson, 1981; Branden and Tooze, 1999; Zhou et al., 2014; Ding, Buldyrev, and Dokholyan, 2005).

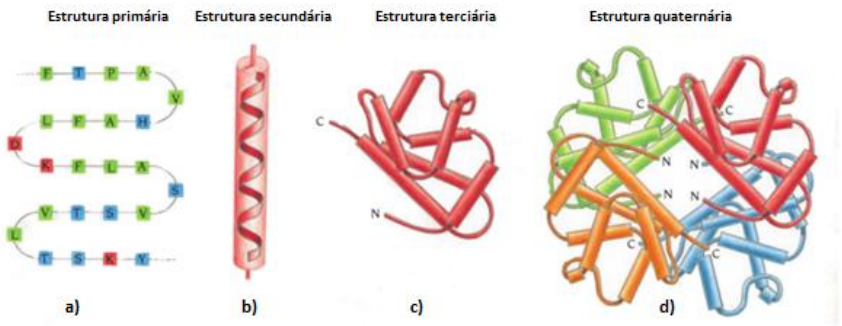


FIGURA 1.1: Sequência de aminoácidos ou estrutura primária (**a)**), conformação hélice-α de uma proteína ou estrutura secundária (**b)**), conjunto de várias estruturas hélice-α distribuídas tridimensionalmente (**c)**) e conglomerado de várias estruturas terciárias distribuídas tridimensionalmente (**d)**). Figura adaptada da referência (Branden and Tooze, 1999).

Na figura 1.2 são apresentados os diagramas topológicos dos motivos estruturais "greekey" e $\beta-\alpha-\beta$. O motivo "greek-key" apresenta quatro folhas-β orientadas antiparalelamente e o motivo $\beta-\alpha-\beta$ apresenta duas folhas paralelas interligadas por uma hélice$-\alpha$. A direção das folhas-β vai do terminal N para o terminal C da proteína (Branden and Tooze, 1999) (mais detalhes quanto aos terminais da proteína são apresentados no capítulo 2).

Apesar da existência de diversas proteínas com diferentes sequências de aminoácidos (estruturas primárias), muitas delas apresentam macro estruturas semelhantes (estruturas secundárias, motivos ou domínios estruturais semelhantes) (Anderson and Rost, 2009; Guo and Thirumalai, 1996; Serpell, 2000), o que torna interessante estudar tais estruturas de modo a melhorar a compreensão sobre as suas unidades funcionais.

Por exemplo, análises de caracterização estrutural de domínios "ww-domain" podem ser feitas para estudar a formação de fibras amilóides a nível dos órgãos ou tecidos (Serpell, 2000; Anderson and Rost, 2009). As fibras amiloides são sistemas insolúveis formados normalmente a partir de proteínas solúveis, ou seja, por alteração da sua conformação

(por exemplo, com conformações em hélice) predominantemente por folhas$-\beta$ (Serpell, 2000). É de salientar que as fibras amilóides são responsáveis por doenças graves como Alzheimer, encefalopatias espongiformes e diabetes do tipo II (Serpell, 2000).

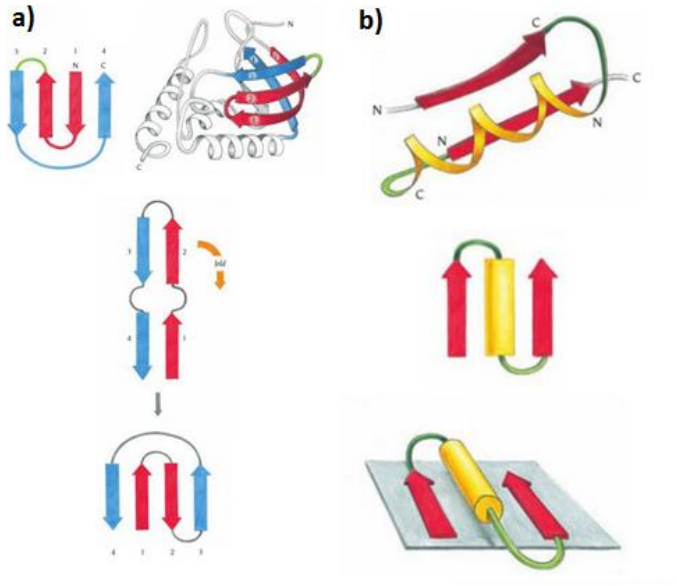


FIGURA 1.2: Diagramas topológicos para o motivo estrutural "greek-key"(**a)**) e para o motivo estrutural $\beta-\alpha-\beta$ (**b)**). Figura adaptada da referência (Branden and Tooze, 1999).

Por outro lado, o domínio estrutural vilina pode ser observado em proteínas com funções de mobilidade e contração, ou seja, proteínas com características funcionais de agregação, separação, nucleação ou de cobertura que se acumulam predominantemente nas microvilosidades das células epiteliais absorventes dos intestinos dos mamíferos (Klahre et al., 2000).

No presente trabalho foram estudados diversos sistemas com diferentes domínios estruturais, o caso do domínio vilina correspondente a estrutura PDB 1WY3, o "ww-domain" referente a conformação PDB 2MWE, o "trp-cage" apresentado no sistema PDB 1L2Y e o um domínio constituído por um agregado com quatro hélices-α, domínio estrutural referido por Guo e Thirumalai (Guo and Thirumalai, 1996) e observado na estrutura PDB 2LSE. Apesar da reduzida dimensão, estes sistemas constituem importantes elementos estruturais, que fazem parte de proteínas mais complexas. O foco do presente trabalho está voltado para os estudos de otimização estrutural tendo em conta a sequência de

aminoácidos e as estruturas secundárias destes sistemas.

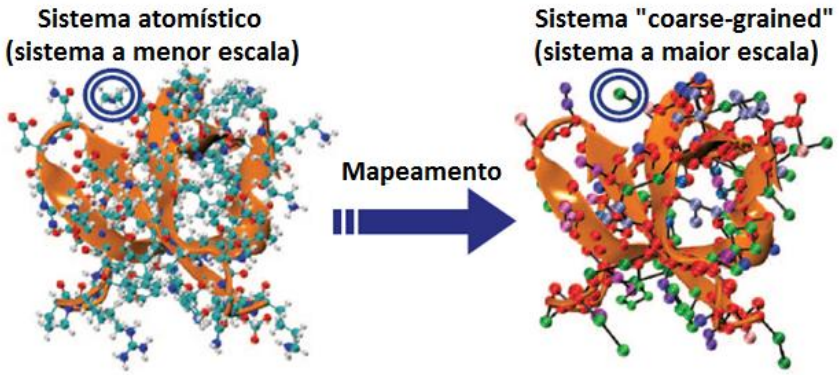


FIGURA 1.3: Mapeamento de um sistema atomístico a um sistema "coarse-grained": sistema atomístico, a esquerda; sistema "coarse-grained", a direita. Pode-se verificar que um resíduo no modelo "coarse-grained" corresponde a um conjunto de átomos do modelo atomístico depois de feito o mapeamento. Figura adaptada da referência (Noid, 2013).

Duas das abordagens computacionais utilizadas nos estudos de otimização global de sistemas de proteínas são a abordagem *atomística* e a abordagem *"coarse-grained"*.

A primeira abordagem refere-se a modelos computacionais em que são considerados todos os átomos correspondentes ao sistema molecular em estudo (Noid, 2013) – abordagem computacionalmente mais dispendiosa. Já na segunda abordagem, a estrutura primária da molécula corresponde a uma sequência ou agrupamento de monómeros, em que cada um deles é representado por uma "esfera" e resulta do mapeamento dos átomos em um aminoácido (Noid, 2013; Tozzini, 2005; Sharma, Ding, and Dokholyan, 2008); a figura 1.3 ilustra o mapeamento de átomos à resíduos "coarse-grained". Esta última abordagem tem vindo a ganhar bastante aceitação na comunidade científica pelo facto de reduzir o número de graus de liberdade relacionados com as interações entre partículas, reduzindo o tempo de cálculo durante os processos de otimização (Brasiello et al., 2010; Noid, 2013), bem como por representar de forma satisfatória os pontos estacionários (mínimos energéticos e estados de transição) comparativamente aos sistemas atomísticos (Tozzini, 2005; Sharma, Ding, and Dokholyan, 2008).

Os modelos "coarse-grained" podem ser construídos de acordo com diferentes tipos de

abordagens (Tozzini, 2005; Rudzinski and Noid, 2015; Noid, 2013): abordagem do *modelo de rede elástica*, *modelo de Go*[2], *"top-down"*, *"bottom-up"* e *"knowledge-based"*. O modelo "coarse-grained" de rede elástica é um modelo que apresenta ligações elásticas entre os aminoácidos presentes na cadeia peptídica, enquanto no modelo de Go são removidas todas interações não nativas, sendo portanto um modelo com menos frustrações geométricas, ou seja, todas interações atrativas entre pares de resíduos não ligados na estrutura nativa são anuladas, o que leva a um colapso mais rápido para o estado nativo; assim, no modelo de rede elástica e no modelo de Go um aminoácido corresponde normalmente a um resíduo contendo apenas um centro de interação (Noid, 2013; Tozzini, 2005). As ligações peptídicas entre cada par de resíduos no modelo de rede elástica podem ser mantidas rígidas (Tozzini, 2005; Brown, Fawzi, and Head-Gordon, 2003).

De modo a tornarem-se transferíveis para diferentes sistemas de proteínas, os modelos "coarse-grained" têm sido classificados em três classes principais de acordo com os centros de interação de cada aminoácido (Tozzini, 2005; Sharma, Ding, and Dokholyan, 2008): sistemas com um centro de interação (*"One bead models"*) - em que um aminoácido corresponde a um resíduo com um centro de interação; sistemas com dois centros de interação (*"two bead models"*) - em que um aminoácido corresponde a dois resíduos ou dois centros interação; e sistemas com quatro ou seis centros de interação (*Four-six bead models*) - em que um aminoácido representa quatro ou seis resíduos com quatro ou seis pontos de interação.

É óbvio que à medida que saímos de uma classe para outra (por exemplo de uma classe de modelo com vários centros de interação para uma classe de modelo com apenas um centro de interação) perde-se alguma coerência, por esta razão é utilizada a abordagem "knowledge-based" para as classes de modelo "coarse-grained" em que um aminoácido representa apenas um centro de interação (Tozzini, 2005). A abordagem "knowlege-based" é uma abordagem que se baseia em dados experimentais para caracterizar determinado sistema de proteína (Noid, 2013).

Os modelos "coarse-grained" com um centro de interação têm se revelado extremamente úteis para a caracterização de dobragens das proteínas; por outro lado, os modelos em

[2]Modelo desenvolvido por Nobuhiro Go e colaboradores (Taketomi, Ueda, and Go, 1975).

que cada aminoácido é representado por mais de um centro de interação têm sido úteis para modelar as interações entre proteínas ou ácidos nucleicos (Tozzini, 2005). Segundo Sharma, Ding e Dokholyan (Sharma, Ding, and Dokholyan, 2008), os modelos com um centro de interação apresentam bons resultados termodinâmicos e os que têm mais centros de interação apresentam maiores resoluções (mais detalhes estruturais).

A abordagem "top-down" é uma estratégia empírica que se baseia em dados experimentais observáveis a uma escala macroscópica, proporcionando baixas resoluções (menos detalhes estruturais) no processo de caracterização dos sistemas de proteínas, mas porém rápidas descrições termodinâmicas e estruturais (Noid, 2013).

No modelo "top-down" não há propriamente uma distinção entre cada tipo de aminoácido. Os aminoácidos podem ser classificados como "zonas de interação" em função das suas características, por exemplo, um conjunto de aminoácidos hidrofóbicos constituiria um local ou ponto de interação hidrofóbico. Por outro lado, as abordagens "bottom-up", proporcionam melhores descrições fundamentais dos materiais reais, pois estas apresentam modelos de potenciais mais detalhados, ou seja, com maiores aproximações aos sistemas atomísticos. (Rudzinski and Noid, 2015; Noid, 2013).

Outros paradigmas de modelos "coarse-grained" são as abordagens "physical-based" e "knolege-based" (as quais pouco se têm distinguido uma da outra) (Noid, 2013). A primeira permite fazer estudos com utilização de dados teóricos, por exemplo provenientes de sistemas "bottom-up". E a segunda, como já referido anteriormente, se baseia em dados de estruturas terciárias obtidas experimentalmente, por exemplo, pode-se utilizar esta abordagem com base em estruturas provenientes do PDB ("protein data bank": `http://www.rcsb.org/pdb/home/home.do`).

Na figura 1.4 podem ser visualizadas as abordagens "top-down", "bottom-up" e "knowlege-based" utilizadas na construção de modelos "coarse-grainded" tendo em conta as suas escalas e resoluções.

Nos estudos de caracterização de sistemas de proteínas têm sido utilizados diferentes tipos de modelos "coarse-grained". Alguns destes são designados por modelos *modelos AB*, i.e., integram apenas dois tipos de resíduos; um hidrofílico e outro hidrofóbico (Lee

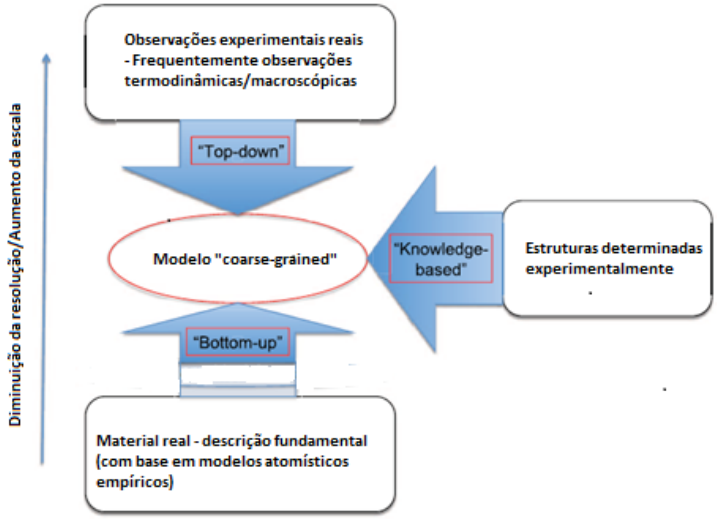


FIGURA 1.4: Abordagens "top-down", "bottom-up" e "knowlege-based" em sistemas "coarse-grained". Figura adaptada da referência (Noid, 2013).

et al., 2008; Jr., 2014); por outro lado, um dos mais estudados é o *modelo BLN*, o qual possui três tipos de resíduos (hidrofóbicos, hidrofílicos e neutros) (Honeycutt and Thirumalai, 1992; Guo and Thirumalai, 1995; Guo and Thirumalai, 1996; Sorenson and Head-Gordon, 2002b; Sorenson and Head-Gordon, 2002a; Brown, Fawzi, and Head-Gordon, 2003; Wales, 2014; Miller and Wales, 1999); o *modelo BVLN* é uma generalização do BLN e possui quatro tipos de resíduos: L e N representam resíduos repulsivos fortes e fracos, respetivamente, e B e V designam resíduos atrativos fortes e fracos, respetivamente (Yap, Fawzi, and Head-Gordon, 2008).

No presente trabalho foram feitas otimizações globais usando o modelo BLN para estabelecer a interação entre os vários aminoácidos.

O modelo BLN apresenta um potencial normalmente constituído por quatro termos, a destacar, o potencial adotado para as restrições das ligações peptídicas entre cada par de resíduos, o potencial para caracterização das restrições dos ângulos de ligação entre cada grupo de três aminoácidos, o potencial utilizado para as restrições de ângulos diedros formados para cada grupo de quatro aminoácidos sucessivamente interligados e o potencial

para parametrizar os tipos de interações que ocorrem entre os resíduos não interligados por ligações peptídicas (Honeycutt and Thirumalai, 1992; Guo and Thirumalai, 1995; Guo and Thirumalai, 1996; Sorenson and Head-Gordon, 2002b; Sorenson and Head-Gordon, 2002a; Brown, Fawzi, and Head-Gordon, 2003; Wales, 2014; Miller and Wales, 1999).

O modelo BLN tem vindo a ser bastante estudado por diversos autores por ser um modelo bastante interessante para caracterização de diversos tipos de sistemas de proteínas no ponto de vista estrutural e energético, por manipulação de parâmetros associados essencialmente as contribuições dos potenciais de torção dos ângulos diedros e do potencial das interações entre pares de resíduos não ligados por ligações peptídicas. Uma das características bastante interessantes do modelo BLN é o facto de poder integrar diferentes abordagens, como é o caso da abordagem baseada nas estruturas nativas (modelo de Go), a de rede elástica (Miller and Wales, 1999; Oakley, Wales, and Johnston, 2011) e a "knowlege-based".

Como referido, a abordagem do modelo de Go tem sido útil para uma rápida caracterização estrutural e energética de um determinado sistema de proteína por fazer-se remover todas suas interações não nativas (o que resulta em dobragens para estruturas nativas com menos frustrações). Os modelos de Go para os sistemas de resíduos BLN são construídos com base no mapa de contactos (Jr., 2014) dos pares de resíduos que interagem na estrutura nativa.

Trabalhos prévios (Miller and Wales, 1999; Oakley, Wales, and Johnston, 2011; Oakley and Roy L. Johnston, 2012) mostram a aplicação do potencial BLN para estudos de otimização global dos sistemas BLN46 e BLN69, nos quais foram utilizadas abordagens "coarse-grained" do modelo de Go e a abordagem "coarse-grained" de rede elástica[3], mostrando a primeira ser bastante útil para dobragens rápidas ao mínimo correspondente a estrutura nativa, e por outro lado, os estudos de otimização feitos com a abordagem de rede elástica apresentaram bastantes frustrações geométricas para estes sistemas, o que pode ser interessante analisar, uma vez que naturalmente muitos sistemas podem apresentar bastantes frustrações e dificuldades de evolução para a estrutura de menor energia. As superfícies

[3] Abordagem utilizada no corrente trabalho.

de energia potencial destes sistemas (sistemas modelados por modelos de Go ou modelos de rede elástica) podem ser representadas e analisadas por utilização de gráficos de conectividade ("disconectivity graphs"). Os gráficos de conectividade são gráficos que apresentam as relações energéticas entre os pontos estacionários (mínimos energéticos e estados de transição).

Para estimar o tempo de cálculo para obtenção do mínimo global de um determinado sistema têm sido calculados os tempos médios (MFET, do inglês - "Mean first encounter time") para encontrar este mínimo em várias otimizações feitas com o programa GMIN (Oakley, Wales, and Johnston, 2011). Ter noção dos MFETs apresentados por estes sistemas ao formarem a estrutura nativa ajuda-nos a ter uma melhor compreensão do paradoxo de Levinthal (Wales, 2004) associado aos mesmos, ou seja ajuda-nos a compreender melhor a relação entre o tempo de cálculo e o conjunto de n mínimos possíveis de serem obtidos em casos de sistemas com grandes frustrações geométricas.

Vários outros autores têm feito vários estudos que envolvem a utilização do potencial BLN no que diz respeito às técnicas de otimização global para explorar as paisagens energéticas de vários sistemas biológicos (Honeycutt and Thirumalai, 1992; Guo and Thirumalai, 1995; Guo and Thirumalai, 1996; Sorenson and Head-Gordon, 2002b; Sorenson and Head-Gordon, 2002a; Brown, Fawzi, and Head-Gordon, 2003).

Este trabalho tem como objetivo caracterizar energética e estruturalmente os modelos simples de proteínas anteriormente referidos (o sistema "trp-cage" 1L2Y, o sistema vilina 1WY3, o sistema "ww-domain" 2MWE e a proteína com quatro hélices-α) e por construção de gráficos de conectividade ("disconnectivity graphs") analisar a organização dos caminhos de conectividade das suas paisagens energéticas. Constitui ainda objetivo do corrente trabalho determinar os MFET destes sistemas de proteínas. A escolha dos sistemas a prior referidos foi feita de modo a ter-se uma rápida compreensão do enrolamento à estruturas nativas partindo de estruturas desnaturadas, da caracterização energética e dos tempos médios associados a obtenção dos mínimos globais em estruturas com diferentes motivos estruturais, no caso, o vilina, "ww-domain", "trp-cage", e quatro hélices-α empacotadas.

Para serem encontrados os mínimos globais e outros mínimos de baixa energia foram

feitas otimizações "Basin-Hopping" (descritas mais adiante no capítulo 2) com utilização do programa GMIN (Wales, 2014; Wales, 2016a; Wales, 2004; Oakley and Roy L. Johnston, 2012). Para se encontrar o caminho de conectividade entre dois mínimos foram feitas otimizações "double ended search" (otimizações feitas para encontrar o caminho de conectividade entre dois mínimos passando por um ou vários pontos sela) utilizando o programa OPTIM (Wales, 2016b). Os pontos estacionários (mínimos energéticos e as energias correspondentes aos pontos de transição) são os pontos mas importantes da superfície de energia potencial (Miller and Wales, 1999; Wales, 2004), os quais podem ser determinados com a utilização do OPTIM (Oakley, Wales, and Johnston, 2011) contudo, é necessário salientar que não há garantias de encontrar todos os estados de transição ligados a um mínimo global (Miller and Wales, 1999). Os resultados obtidos foram melhorados por criação de bases de dados de pontos estacionários e expansão da mesma por utilização do programa PATHSAMPLE (Wales, 2016c). Os gráficos de conectividade foram gerados com a utilização do programa disconnectionDPS disponível em `http://www-wales.ch.cam.ac.uk/`.

Capítulo 2

Técnicas Computacionais

2.1 Modelo BLN

2.1.1 Sequência de aminoácidos na cadeia peptídica

Para compreendermos a estrutura das proteínas é necessário compreendermos primeiro quais sequências de aminoáciodos ou estruturas primárias as constituem.

As proteínas apresentam cadeias normalmente formadas por, pelo menos, um dos 20 tipos de aminoácidos conhecidos, os quais apresentam a fórmula geral $NH_2C_\alpha RHCO_2H$ (Branden and Tooze, 1999; Wales, 2004). O carbono alfa (C_α) liga-se a quatro grupos de átomos diferentes: ligação simples a um grupo amino (NH_2), a um carbono carboxílico (COOH), a um átomo de hidrogénio (H) e a um radical (R) que representa uma cadeia lateral; na Figura 2.1 são apresentadas as estruturas dos 20 tipos de aminoácidos acompanhadas dos seus respetivos nomes e abreviaturas utilizadas.

Os aminoácidos distinguem-se uns dos outros por variações observadas nos padrões da cadeia lateral (Weiner et al., 1986; Branden and Tooze, 1999; Wales, 2004). Em uma proteína, diferentes tipos de aminoácidos ligam-se entre si por meio de ligações peptídicas para formar a estrutura primária; na figura 2.2 visualiza-se a ilustração de uma sequência de aminoácidos interligados por meio de ligações peptídicas, sendo R_1, R_2 e R_3 cadeias laterais correspondentes a três aminoádos diferentes caso estas sejam diferentes. As ligações peptídicas formam-se entre o "terminal C" (terminal carbonado) de um aminoácido e o "terminal N" (terminal azotado) de outro. Como referido na Introdução, os terminais

"C" e "N" podem ser utilizados para identificar a direção das folhas-β, sendo estas representadas em diagramas topológicos como setas que vão do "terminal N" ao "terminal C" (Branden and Tooze, 1999). Os terminais "C" e "N" permitem ainda identificar motivos estruturais em sistemas de proteínas, por exemplo, no caso do motivo β-α-β, motivo com duas folhas-β dispostas paralelamente (motivo apresentado na figura 1.2 no capítulo 1, página 3), é típico a primeira folha-β ligar-se a hélice-α pelo seu "terminal C" e esta ligar-se ao "terminal N" da outra folha-β (Branden and Tooze, 1999).

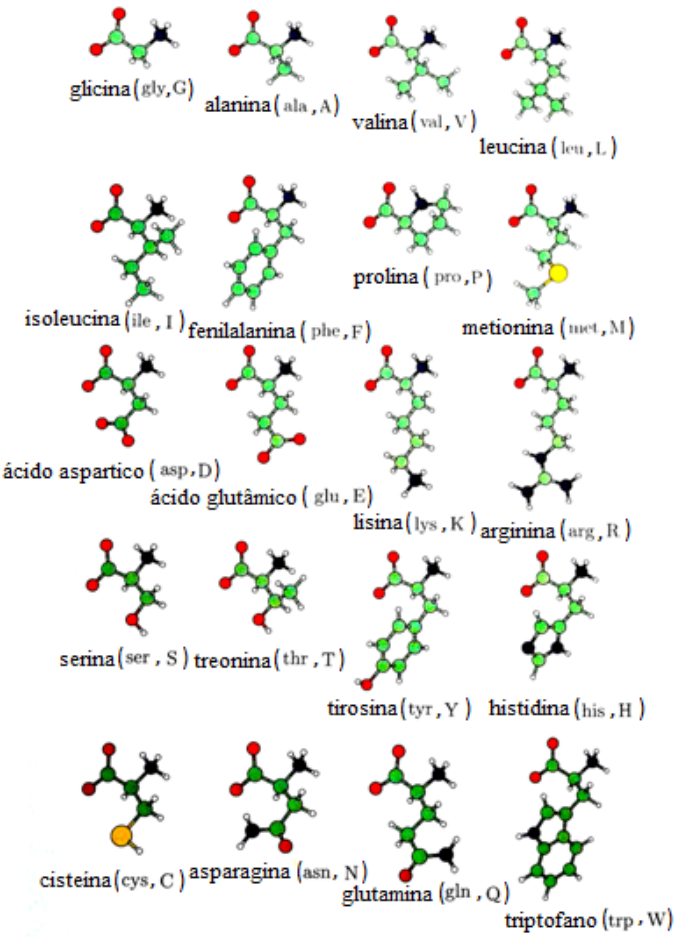


FIGURA 2.1: Representação dos 20 aminoácidos. Os acrónimos apresentados a seguir ao nome de cada aminoácido foram utilizados para fazer o mapeamento das "estruturas primárias" dos sistemas estudados experimentalmente provenientes do PDB para "estruturas primárias" de sistemas de resíduos BLN. Figura adaptada da referência (Wales, 2004).

Restrições na cadeia polipeptídica da proteína (causadas pelo carácter parcial de dupla ligação C-N[1]) resultam nos três tipos principais de estruturas secundárias referidas anteriormente no capítulo 1, folhas$-\beta$, hélices$-\alpha$ e dobras (Branden and Tooze, 1999; Anderson and Rost, 2009). Mais pormenores sobre as estruturas secundárias das proteínas são apresentados mais adiante na subsecção 2.1.3.

Para obter as estruturas primárias dos sistemas estudados no presente trabalho, foram feitos mapeamentos das suas sequências de aminoácidos para o sistema de resíduos BLN.

O mapeamento de aminoácidos a resíduos BLN foi feito com base nos dados apresentados na tabela 2.1 (Brown, Fawzi, and Head-Gordon, 2003).

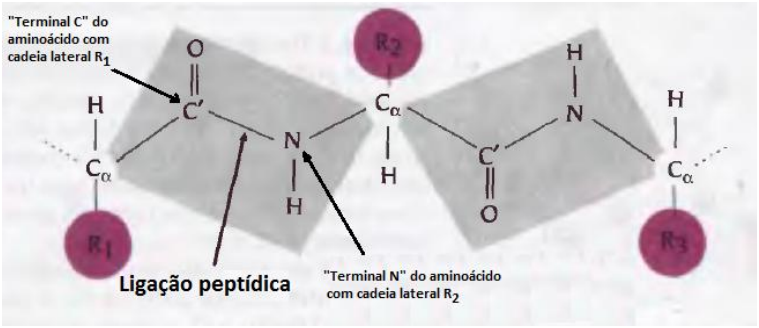


FIGURA 2.2: Esquema de ligações peptídicas entre os aminoácidos na cadeia de uma proteína. Figura adaptada da referência (Branden and Tooze, 1999).

Segundo Head-Gordon e colaboradores (Brown, Fawzi, and Head-Gordon, 2003) existem algumas incertezas associadas ao mapeamento dos aminoácidos à sequência de resíduos BLN, as quais podem levar a mutações relativamente às estruturas esperadas. Por exemplo a lisina ("Lys (K)") em função da conformação estrutural que adota a sua cadeia lateral (R) pode ser vista como um resíduo hidrofílico (quando o grupo amino da cadeia lateral está superficialmente exposto) ou como um resíduo hidrofóbico (quando a cadeia lateral se enrola num centro hidrofóbico).

A tabela 2.2 mostra a sequência de aminoácidos da proteína L, a qual foi estudada em trabalhos de Sorenson e Head-gordon (Sorenson and Head-Gordon, 2002a; Sorenson and

[1]Restrições causadas pelo efeito de ressonância da dupla ligação entre o grupo carbonilo e o par C-N ao longo da cadeia peptídica.

TABELA 2.1: Conversão dos 20 tipos de aminoácidos para resíduos BLN. Tabela adaptada da referência (Brown, Fawzi, and Head-Gordon, 2003).

Aminoádo	**Resíduo BLN**	**Aminoádo**	**Resíduo BLN**	**Aminoádo**	**Resíduo BLN**
Ala (A)	B	Trp (W)	B	Asp (D)	L
Cys (C)	B	Tyr (Y)	B	Asn (N)	L
Leu (L)	B	Pro (P)	N	His (H)	L
Ile (I)	B	Gly (G)	N	Gln (Q)	L
Phe (F)	B	Ser (S)	N	Lys (K)	L
Met (M)	B	Thr (T)	L	Arg (R)	L
Val (V)	B	Glu (E)	L		

TABELA 2.2: Mapeamento da sequência de aminoácidos da proteina L para a estrutura primária de resíduos BLN. A sequência de de aminoácidos foi retirada do PDB e mapeada para o sistema de resíduos BLN, com base nos dados apresentados na tabela 2.1.

Sequência PDB	ENKEETPETP ETDSEEEVTI KANLIFANGS TQTAEFKGTF
Est. primária BLN	LLLLLLNLLN LLLNLLLBLB LBLBBBBLNN LLLBLBLNLB
Sequência PDB	EKATSEAYAY ADTLKKDNGE YTVDVADKGY TLNIKFAG
Est. primária BLN	LLBLNLBBBB BLLBLLLLNL BLBLBBLLNB LBLBLBBN

Head-Gordon, 2000) e trabalhos de Brown, Fawzi e Head-Gordon (Brown, Fawzi, and Head-Gordon, 2003)). Na tabela 2.2 é também apresentado o mapeamento da estrutura primária da proteína L em resíduos BLN. A sequência de aminoácidos da proteína L (proteína com 78 aminoácidos) foi retirada do PDB (referência PDB 2PTL). Em trabalhos de Sorenson e Head-Gordon (Sorenson and Head-Gordon, 2002a; Sorenson and Head-Gordon, 2000) e Head-Gordon e colaboradores (Brown, Fawzi, and Head-Gordon, 2003) o número de resíduos para proteína L foi diminuído, sendo utilizado nos estudos deste sistema apenas 56 aminoácidos correspondestes as zonas com conformações em hélice-α e folhas-β.

2.1.2 Equações do potencial BLN

Como referido na introdução, o modelo BLN é descrito por quatro expressões e apresenta resíduos de três tipos, os quais são representados pelas letras B, L e N. Alguns autores (Miller and Wales, 1999; Oakley and Roy L. Johnston, 2012; Oakley, Wales, and Johnston,

2011) apresentaram o potencial BLN de acordo com a equação seguinte:

$$V_{BLN} = 1/2K_r \sum_{i=1}^{N-1}(R_{(i,i+1)} - R_e)^2 + 1/2K_\theta \sum_{i}^{N-2}(\theta_i - \theta_e)^2 + \epsilon \sum_{i}^{N-3}[A_i(1 + cos\phi_i) + B_i(1 - cos\phi_i) + C_i(1 + cos3\phi_i) + D_i(1 + cos\phi_i + \pi/4)] + \epsilon \sum_{i=1}^{N-2}\sum_{j=i+2}^{N}[S_{12}(\sigma/R_ij)^{12} + S_6(\sigma/R_ij)^6] \quad (2.1)$$

Onde R_{ij} é a separação entre os resíduos i e j e as unidades de distância e energia são dadas por σ e ϵ, respetivamente. O primeiro termo representa a energia das ligações entre os resíduos BLN na cadeia da proteína, sendo o valor de K_r=231.2 $\epsilon\sigma^{-2}$ e R_e=1 σ. O segundo termo refere-se a soma das energias de dobragem segundo os ângulos de ligação θ_i definidos por grupos de átomos em posições R_i à R_{i+2}; e valores de K_θ=20 ϵrad^{-2} e θ_e= 1.8326 rad. O terceiro termo é a soma das energias de torção segundo os ângulos diedros, ϕ_i, definidos por quartetos de resíduos em posições R_i à R_{i+3}. O último termo representa as interações que não são resultantes de ligações, ou seja, representa as interações que resultam da proximidade entre os diversos pares de resíduos não ligados entre si por ligações peptídicas.

Os valores dos parâmetros apresentados para a expressão do potencial dos ângulos diedros (valores de A, B, C e D) e os valores dos parâmetros apresentados para a expressão do potencial que representa os pares de resíduos não ligados por ligações peptídicas (S_{12} e S_6) podem ser manipulados e apresentam efeito directo sobre as conformações adotadas pela estrutura secundária da proteína. Algumas comparações entre esses parâmetros relativamente a estudos feitos por vários autores são apresentadas na subsecção 2.1.3.

2.1.3 Estrutura secundária: Implicações dos vários parâmetros na estrutura da proteína. Diferenças entre E, H e T. Modificações nos parâmetros do potencial BLN

A estrutura secundária apresenta os motivos estruturais mais básicos de uma proteína, ou seja, representa a dobragem da sequência de aminoácidos (estrutura primária) para formar conformações em hélices-α, folhas-β, dobras, etc; na figura 2.3 é apresentada uma estrutura secundária hélices-α, na qual se verifica a sequencia de aminoácidos interligados na cadeia peptídica e a formação das hélices-α pelas ligações de hidrogénio entre os átomos de hidrogénio dos grupos amino (representados pelas pequenas esferas brancas) e os átomos de oxigénio ligados aos carbonos carbonilos (C') (representados pelas esferas vermelhas).

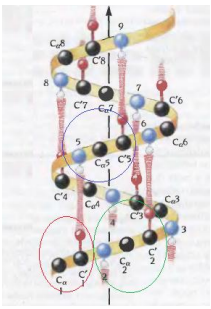


FIGURA 2.3: Estrutura secundária hélice-α. O aminoádo do $C_\alpha 1$ (marcado a vermelho) liga-se por ligação peptídica ao aminoácido do $C_\alpha 2$ (marcado a verde), por sua vez o aminoácido do $C_\alpha 5$ (marcado a azul) é atraído por ligação de hidrogénio pelo aminoácido do $C_\alpha 1$ para formar parte da hélice-α (a esfera azul corresponde ao grupo amino e C' ao carbono cabonilo. a Figura adaptada da referência (Branden and Tooze, 1999).

De acordo com Anderson e Rost (Anderson and Rost, 2009) as estruturas secundárias são basicamente classificadas em três classes: classe de estruturas que envolvem conformações em hélices (hélices-α, helices-3/10, etc.), classe de estruturas que envolvem conformações em "folhas" (folhas-β e pontes-β) e classe das estruturas que envolvem conformações em dobras ou estruturas não atribuídas. No potencial BLN, estas classes de estruturas secundárias são descritas com a utilização a parâmetros dos ângulos diedros, descritos pelas letras H, E e T, em que H descreve as conformações em hélice, E descreve as conformações estendidas e T descreve as conformações em dobra (Brown, Fawzi, and

Head-Gordon, 2003) (estas conformações são referenciadas nas tabelas 2.3, 2.5 e 2.7).

Outros parâmetros diretamente relacionados com o colapso das estruturas primárias para formar estruturas secundárias são os parâmetros associados as interações entre aos pares de resíduos não interligados por ligações peptídicas, descritos pela quarta expressão da eq.(2.1). As interações entre os resíduos são atrativas entre resíduos BB e repulsivas no caso das interações entre LL, LB, NN, NB ou NL. No potencial BLN estas interações atrativas e repulsivas são classificadas em três tipos (ver as tabelas 2.4, 2.6 e 2.8), sendo a intensidade das interações repulsivas dependente dos tipos de resíduos envolvidos (Brown, Fawzi, and Head-Gordon, 2003).

As tabelas 2.3-2.8 apresentam alguns exemplos de parâmetros para potenciais de ângulos diedros e potenciais de interações entre pares de resíduos não interligados por ligações peptídicas. Segundo vários autores, manipulações nesses parâmetros do modelo BLN permitem caracterizar energética e estruturalmente diversos sistemas de proteínas, podendo deste modo, serem identificados motivos estruturais compostos por hélices-α, folhas-β, dobras, etc.

TABELA 2.3: Parâmetros para o potencial de torção envolvendo ângulos diedros utilizados em dobragens de folhas-β.

Tipo de conformação	**A**	**B**	**C**	**D**
H	1.2	0.0	1.2	0.0
E	1.2	0.0	1.2	0.0
T	0.0	0.0	0.2	0.0

TABELA 2.4: Parâmetros para o potencial de interações entre pares de resíduos sem ligações peptídicas utilizados em dobragens de folhas-β.

Tipo de interação	**Parâmetros folhas-β**	
BB	S_{12}=1	S_6=-1
LL ou LB	S_{12}=$\frac{2}{3}$	S_6=$\frac{2}{3}$
NN, NB ou NL	S_{12}=1	S_6=0

TABELA 2.5: Parâmetros para o potencial de torção envolvendo ângulos diedros utilizados em dobragens de hélices-α.

Tipo de conformação	**A**	**B**	**C**	**D**
H	0.0	1.6	1.6	1.6
E	0.0	1.6	1.6	1.6
T	0.0	0.0	0.2	0.0

TABELA 2.6: Parâmetros para o potencial das interações entre pares de resíduos não ligados por ligações peptídicas utilizados em dobragens de hélices-α.

Tipo de interação	**Parâmetros hélices-α**	
BB	S_{12}=1.7	S_6=$-$1.7
LL ou LB	S_{12}=1.7	S_6=0
NN, NB ou NL	S_{12}=1.7	S_6=0

TABELA 2.7: Parâmetros para o potencial de torção dos ângulos diedros utilizados em dobragens mistas.

Tipo de conformação	**A**	**B**	**C**	**D**
H	0.0	1.2	1.2	1.2
E	0.9	0.0	1.2	0.0
T	0.0	0.0	0.2	0.0

TABELA 2.8: Parâmetros para o potencial das interações entre pares de resíduos não ligados por ligações peptídicas utilizados em dobragens mistas.

Tipo de interação	**Parâmetros dobragens mistas**	
BB	S_{12}=1	S_6=$-$1
LL ou LB	S_{12}=$\frac{1}{3}$	S_6=$\frac{1}{3}$
NN, NB ou NL	S_{12}=1	S_6=0

TABELA 2.9: Resumo dos parâmetros do potencial de torção de ângulos diedros e de interações entres resíduos não interligados por ligações peptídicas utilizados por vários autores em estudos de dobragens de vários sistemas de proteínas com diferentes conformações secundárias.

Dobragem	**Parâmetros**[a]	**Autores**	**Sistemas estudados**
hélices-α	Tabela 2.5 e 2.6	(Guo and Thirumalai, 1996)	Proteína com 4 hélices
Folhas-β	Tabela 2.3 e 2.4	(Oakley, Wales, and Johnston, 2011) (Oakley and Roy L. Johnston, 2012) (Oakley et al., 2013) (Miller and Wales, 1999) (Guo and Thirumalai, 1995)	Modelo BLN46 Modelo BLN69
Mista	Tabela 2.7 e 2.8	(Guo and Thirumalai, 1995) (Wales, 2016a) (Brown, Fawzi, and Head-Gordon, 2003)	Proteína L Proteína G
		(Sorenson and Head-Gordon, 2002b) (Sorenson and Head-Gordon, 2002a)	Ubitiquina Proteína L Proteína G

[a]Parâmetros do potencial de torção de ângulos diedros e de interações entres resíduos não interligados por ligações peptídicas apresentados nas tabelas 2.3-2.8.

Os dados apresentados nas tabelas anteriormente referidas, representam parâmetros do potencial de torção de ângulos diedros e de interações entres resíduos não interligados por ligações peptídicas referenciados por diferentes autores em estudos de caracterização de diversos sistemas de proteínas. A tabela 2.9 apresenta um resumo de parâmetros utilizados nos estudos de caraterização de alguns sistemas de proteínas em função dos seus tipos de dobragens[2] estudados por alguns autores.

Para caracterização de dobragens à estruturas nativas nos modelos de proteina BLN46 e BLN69 (modelos de proteínas caracterizados por folhas-β), têm sido utilizados tipicamente os parâmetros presentes nas tabelas 2.3 e 2.4. Observa-se que foram utilizados os potenciais de torção de ângulos diedros e de interações entre partículas não ligadas por ligações peptídicas das tabelas 2.5 e 2.6 para caracterização de dobragens à estruturas nativas em conformações com hélices-α. No caso de sistemas de proteínas com dobragens mistas (dobragens que envolvem diferentes tipos de conformações, por exemplo, hélices-α e folhas-β), pode-se verificar a utilização dos parâmetros presentes nas tabelas

[2]Dobragens à conformações hélices-α, folhas-β ou mistas.

TABELA 2.10: Sequência de resíduos BLN e estruturas secundárias dos modelos de proteína BLN46, BLN69, proteína G e L (Wales, 2016a; Oakley, Wales, and Johnston, 2011; Sorenson and Head-Gordon, 2002a; Sorenson and Head-Gordon, 2000; Brown, Fawzi, and Head-Gordon, 2003).

Nome	**Estrutura primária**	**Estrutura secundária**
BLN46	BBBBBBBBBNNNLBLBLBLBNNNBB BBBBBBBNNNLBLBLBLBLBL	EEEEEEETTTTEEEEEEETTTTEEE EEEEETTTTEEEEEEEEE
BLN69	BBBBBBBBBNNNLBLBLBLBNNNBB BBBBBBBNNNLBLBLBLBNNNBBBB BBBBBNNNLBLBLBLBLBL	EEEEEEETTTTEEEEEEETTTTEEE EEEEETTTTEEEEEEETTTTEEEEE EEETTTTEEEEEEEEE
Prot. G	LBLBLBLBBNNNLBBLBLBBBNNNL LBLLLBLLBNBBBLBBBBNNNLBBL BLBLBL	EEEEEETEHTHEEEEEEEEHHEHHH HHHHHHHHEHTEEEEEEETTTEEEEE EEE
Prot. L	LBLBLBLBBNNNNBBBLBLBBBNNNL LBLLBBLLBNBLBLBLBLNNNLBBL BLBBBL	EEEEEETEHTHEEEEEEEEHHEHHH HHHHHHHHEHTEEEEEEETTTEEEEE EEE

2.7 e 2.8.

Nos estudos da proteína Ubitiquina, L e G apresentados na tabela 2.9, foram utilizados C=1.2 como parâmetro do potencial de torção de ângulos diedros que envolvem conformações em dobras (T) (Sorenson and Head-Gordon, 2002a; Sorenson and Head-Gordon, 2002b).

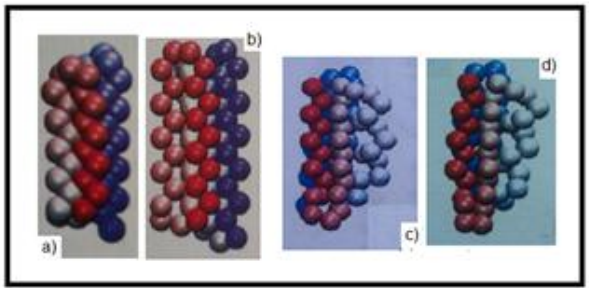


FIGURA 2.4: Modelos de estruturas terciárias dos sistemas BLN46, BLN69, Proteína G e L: a) modelo de proteína BLN46 (Oakley and Roy L. Johnston, 2012); b) modelo de proteína BLN69 (Oakley, Wales, and Johnston, 2011); c) modelo de proteína G (Wales and Head-Gordon, 2012); d) modelo de proteína L (Wales and Head-Gordon, 2012). A proteína BLN46 dobra-se num motivo estrutural com 4 folhas-β e a BLN69 em um com 6 folhas-β. Por outro lado a proteína G e a L apresentam dobragens mistas compostas por duas estruturas secundárias "β-hairpin" interligadas por uma hélice-α.

A tabela 2.10 mostra a sequência de resíduos BLN e a estrutura secundária dos modelos das proteínas BLN46, BLN69, L e G (Oakley, Wales, and Johnston, 2011; Brown, Fawzi,

and Head-Gordon, 2003). As estruturas terciárias destes sistemas de proteínas podem ser observadas na figura 2.4.

2.2 Estratégias adotadas para parametrização das estruturas secundárias

Para caracterização das estruturas secundárias foram adotados como padrão os parâmetros dos potenciais de torção de ângulos diedros e de resíduos não ligados por ligações peptídicas apresentados em estudos anteriores (parâmetros referidos na tabela 2.9, na secção 2.1). No processo de caracterização das estruturas secundárias dos sistemas estudados, foram tidas em conta duas fases: i) construção ou mapeamento das estruturas secundárias partindo da sequência de aminoácidos ou estruturas primárias; iia) melhoria das conformações adotadas pelas estruturas secundárias após a sua construção, por manipulação dos parâmetros A, B, C e D associados a cada uma das estruturas secundárias do modelo BLN (hélices (H), estruturas estendidas (E) e dobras (T)); e/ou iib) melhoramento das conformações adotadas pelas estruturas secundárias após a sua construção, por manipulações dos parâmetros que controlam o colapso hidrofóbico (BB) ou repulsões (BL, LL, NB, NL ou NN) entre os pares de resíduos que não apresentam ligações peptídicas – parâmetros S_{12} e S_6. Deste modo, para parametrização das estruturas secundárias dos vários sistemas estudados no presente trabalho foram adotadas diversas estratégias que descrevemos a seguir.

2.2.1 Utilização de ferramentas de mapeamento de estruturas primárias à estruturas secundárias

Existem diversas ferramentas auxiliares disponíveis online que nos permitem fazer o mapeamento de estruturas secundárias adotadas por uma proteína partindo da sua sequência de aminoácidos. No presente trabalho foram utilizadas a ferramenta de predição de estruturas secundárias CFSSP (servidor de predição de estruturas secundárias de Chou e Fasman)(Chou and Fasman, 1974a; Chou and Fasman, 1974b) disponível em `http://cho-fas.sourceforge.net/` e as ferramentas de predição de estruturas

secundárias de Porter referida por Hoffmann e colaboradores (Hoffmann et al., 2014) disponível em http://distillf.ucd.ie/porterpaleale/. Com estas ferramentas foi possível prever de forma rápida as estruturas secundárias adotadas pelos sistemas estudados, porém, foi necessário logo depois mapear estas para o modelo de resíduos BLN (para tal foram adotados alguns critérios apresentados na subsecção 2.2.2). Para a caracterização da estrutura secundárias com a utilização destas ferramentas auxiliares normalmente é necessário um ficheiro do tipo FASTA[3] (este ficheiro pode ser baixado diretamente do PDB), ou simplesmente a sequência de aminoácidos em formato FASTA.

2.2.2 Utilização da estrutura secundária proveniente do PDB

O modelo de potencial BLN é um modelo bastante simples, contemplando, como já referido, apenas três tipos de conformações secundárias. Do PDB é possível obter a sequência de aminoácidos e estruturas secundárias resultantes de estudos experimentais de um determinado sistema, podendo estes dados serem utilizados para a construção da estrutura primária e estrutura secundária do modelo de proteína BLN. Portanto, a estrutura secundária proveniente do PDB apresenta a atribuição DSSP (atribuição do Dicionário de Estruturas Secundárias das Proteínas), a qual comporta vários tipos de conformações secundárias, por exemplo as hélices podem ser classificadas em três "estados" diferentes, correspondentes a três tipos de hélices diferentes, as quais são representadas pelas letras G (hélice-3/10), H (hélices-α) e I (hélices-π), as estruturas estendidas são classificadas em dois "estados" diferentes representadas pelas letras E (no caso de folhas-β) e B (no caso de pontes-β) e as demais conformações são representadas pelas letras T ou S (no caso de dobras) ou “espaço em branco” no caso de ser uma estrutura secundária sem atribuição (Anderson and Rost, 2009). Estes aspetos podem criar algumas complicações ao fazer-se o mapeamento da estrutura secundária do PDB para a estrutura secundária do modelo BLN. Nesse sentido para mapear as estruturas secundárias provenientes do PDB para o modelo BLN foram adotados alguns critérios como[4]: i) correspondência entre

[3] Ficheiro com uma pequena descrição e a sequência de aminoácidos de um determinado sistema de proteína, onde cada aminoácido é representado pela sua letra correspondente como apresentado na figura 2.1, por exemplo para o caso do aminoácido alanina é representado pela letra A.

[4] Estes critérios podem ser utilizados em qualquer uma das estratégias de mapeamento da estrutura secundária adotadas.

as estruturas secundárias representadas pelas letras H, E e T provenientes do PDB com as estruturas secundárias representadas pelas letras H, E e T do modelo BLN, respetivamente; ii) atribuição de qualquer uma das conformações do modelo BLN representadas pelas H, E e T às conformações do PDB representadas pelas letras S ou por "espaços brancos" em função das geometrias tridimensionais observadas; iii) caso necessário, alteração de qualquer estrutura secundária proveniente do PDB (H, "espaço em branco", S, etc.) para qualquer uma das estruturas secundárias representadas pelas letras H, E e T do modelo BLN, por exemplo para alguns casos a conformação hélice-3/10 (G) é melhor representada pelas conformações T do que H no modelo BLN. Estes critérios não são absolutos, foram apenas adotados para servirem de diretrizes para caracterizar as estruturas secundárias dos sistemas estudados.

2.2.3 Método EHT[X]

Nos casos em que a estrutura secundária de uma proteína modelada com o modelo BLN seja caracterizada apenas por dois tipos estruturas secundárias, ou seja, caso a proteína seja caracterizada apenas por H e T, H e E ou T e E, respetivamente, pode ser utilizado o método HET[X]. O método HET[X] é um método que é proposto neste trabalho para melhorar as conformações das estruturas secundárias, onde X é tido como o terceiro tipo de estrutura secundária e corresponde a uma conformação em hélice (H), estendida (E) ou dobra (T). Por alteração dos parâmetros de torção dos ângulos diedros para o nosso X - parâmetros A, B, C e D da eq. 2.1 – pode-se manipular isoladamente um ou mais conjuntos de quartetos de resíduos que apresentem a estrutura secundária X.

2.2.4 Mapeamento manual das estruturas secundárias

Por análise da geometria das conformações dada no formato PDB de um determinado sistema é possível fazer um mapeamento manual da respetiva estrutura secundária partindo da sua sequência de aminoácidos. Embora pareça ser uma estratégia exaustiva, é bastante prática no caso de proteínas pequenas (Hoffmann et al., 2014) e apresenta

bons resultados. O mapeamento manual ou seja, caracterização feita por análise da estrutura terciária disponibilizada pelo PDB pode ser facilmente combinado com as outras estratégias adotadas.

As estratégias utilizadas para o mapeamento das estruturas secundárias dos sistemas estudados foram melhoradas por manipulação dos valores dos parâmetros do potencial de torção de ângulos diedros e dos pares de resíduos que não apresentam ligações peptídicas. As manipulações dos valores do primeiro conjunto de parâmetros (parâmetros do potencial de torção dos ângulos diedros) foram feitas com base na construção gráficos que relacionam a energia e os ângulos diedros formados, em função dos parâmetros A, B, C e D utilizados nas conformações H, E, T ou X, os quais foram bastante úteis para modelar as estruturas secundárias dos sistemas estudados (alguns gráficos "Energia/Ângulos diedros formados" são apresentados mais adiante no capítulo 3).

2.3 Optimização global e gráficos de conectividade

Para caracterização dos modelos de proteínas, foram utilizados os seguintes programas computacionais disponíveis na página do Grupo de David Wales ("Wales group home page": http://www-wales.ch.cam.ac.uk/; http://www-wales.ch.cam.ac.uk/software.html): a) GMIN, Programa utilizado em otimizações globais e em estudos de propriedades termodinâmicas por utilização do método "Basin-Hopping" (BH) (Wales, 2016a). O método "Basin-Hopping" é um método que utiliza minimizações de Monte Carlo para fazer otimizações globais, deste modo são consideradas variações na superfície de energia potencial em que são geradas várias bacias de atração com todos mínimos locais (Wales and Doye, 1997). Foram ainda utilizados os programas utilitários rancoords e gminconv2 para auxiliar o GMIN no estudo dos MFETs. O MFET é o tempo necessário para localizar o mínimo global partindo de configurações aleatórias iniciais; b) OPTIM, programa utilizado para otimizações geométricas dos sistemas de proteínas estudados e cálculos dos caminhos que interligam os vários isómeros gerados (Wales, 2016b); c) PATHSAMPLE, utilizado como um controlador do programa OPTIM, o qual

permite criar e expandir bases de dados de mínimos de energia e seus respetivos estados de transição. O PATHSAMPLE utiliza amostras de caminhos que ligam os diferentes mínimos geradas pelo OPTIM (Wales, 2016c); d)DisconnectionDPS: Programa utilizado para construção e análise de gráficos de conectividade, para tal, lê ficheiros de "output" gerados pelo PATHSAMPLE.

Na figura 2.5 é ilustrado o diagrama de fluxo utilizado na obtenção dos gráficos de conectividade: 1) otimização global; 2) conexão entre dois mínimos de energia; 3) criação da base de dados de pontos estacionários do PATHSAMPLE; 4) construção dos gráficos de conectividade para os sistemas estudados.

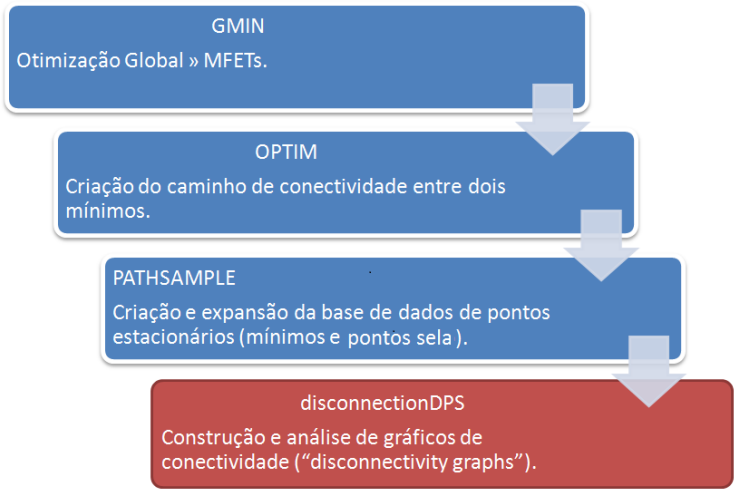


FIGURA 2.5: Processos de otimização global, conexão entre dois mínimos de energia, criação da base de dados de pontos estacionários do PATHSAMPLE e construção dos gráficos de conectividade utilizados nos estudos de caracterização dos sistemas estudados.

Em estudos de Wales e colaboradores (Oakley, Wales, and Johnston, 2011; Miller and Wales, 1999; Wales and Dewsbury, 2004) sobre os muito estudados sistemas de proteínas BLN46 e BLN69, foram feitas otimizações globais com o programa GMIN utilizando o método "basin-hopping" (Wales, 2014). Através da construção de gráficos de conectividade foi observada uma grande frustração geométrica para estes sistemas.

A figura 2.6 mostra os gráficos de conectividade apresentados para o modelo de proteína BLN69 (proteína com bastantes frustrações geométricas) ((Oakley, Wales, and Johnston,

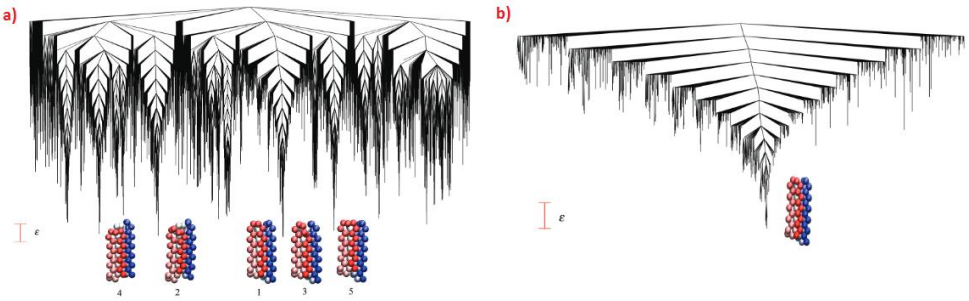


FIGURA 2.6: Gráficos de conectividade para a proteína BLN69: a) modelo de rede elástica; b) modelo de Go - figura adaptada (Oakley, Wales, and Johnston, 2011).

2011). À esquerda (fig. 2.6a)) é apresentado o gráfico de conectividade para o modelo de rede elástica. Neste modelo mantêm-se todas interações nativas e, portanto apresenta mais frustrações geométricas, as quais podem ser claramente observadas pela disposição dos vários mínimos com energias muito próximas; e a direita (fig. 2.6b)) é apresentado o gráfico de conectividade para o modelo de Go da proteína BLN69. A remoção de todas interações entre todos pares de resíduos hidrofóbicos que não apresentam interação no estado nativo (resíduos separados por distâncias superiores a 1.167 σ) (Miller and Wales, 1999; Jr., 2014) leva a um colápso mais rápido à estrutura nativa (Wales and Dewsbury, 2004) ou seja, a um gráfico de conectividade com muito menos frustrações geométricas (Jr., 2014), em que o mínimo global em estudo se evidencia claramente relativamente aos demais mínimos.

Informações relativas a instalação e utilização dos programas OPTIM, PATHSAMPLE, disconnetionDPS e GMIN estão disponibilizadas no apêndice A.

Capítulo 3

Discussão e análise de resultados

Como referido nos capítulos anteriores, no presente trabalho foram estudados diversos sistemas de proteínas, com diferentes domínios estruturais, a destacar: a) proteína com o código PDB 1L2Y, com domínio estrutural "trp-cage"; b) proteína com o código PDB 1WY3, com domínio estrutural vilina; c) proteína com o código PDB 2MWE, com domínio estrutural "ww-domain"; d) modelo de proteína constituído por quatro hélices-α, estudado por Guo e Thrirumalai (Guo and Thirumalai, 1996). As caracterizações destes sistemas foram feitas com base no diagrama de fluxo apresentado na figura 2.5 no capítulo 2, página 25.

No processo de construção das estruturas secundárias dos sistemas de proteínas estudados foram combinadas as várias estratégias apresentadas na secção 2.2 do capítulo 2, com ênfase ao mapeamento manual, de Porter e a utilização do método HET[X].

Os processos de otimização das proteínas 1L2Y, 2MWE e 1WY3 (sistemas modelados com base em informações experimentais provenientes do PDB) foram significativamente melhorados por utilização das coordenadas disponibilizadas pelo PDB como ponto de partida. No caso da proteína com quatro hélices-α, estudada por Guo e Thrirumalai, não houve qualquer dificuldade em determinar o mínimo global por ser disponibilizada a sua sequência secundária, responsável pelo seu colapso a estrutura nativa (Guo and Thirumalai, 1996).

Para as proteínas estudadas com base nas coordenadas cartesianas provenientes do PDB (proteínas 1L2Y, 2MWE e 1WY3), foi ainda necessário fazer um *escalamento*, uma vez que estas coordenadas se apresentam numa escala três vezes superior que a padrão

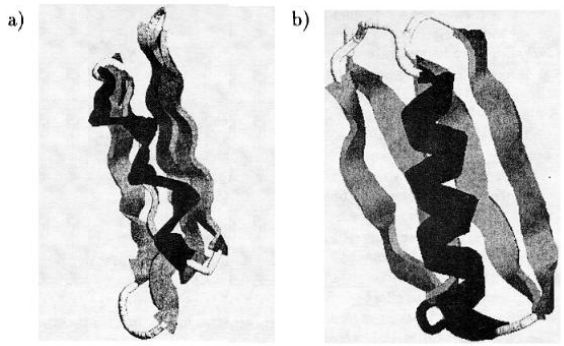


FIGURA 3.1: Proteina L: modelo BLN (**a**)) e estrutura obtida experimentalmente por ressonância magnética nuclear (**b**)). Figura adaptada da referência (Sorenson and Head-Gordon, 2000).

utilizada no modelo BLN[1] (Wales, 2016a; Miller and Wales, 1999; Brown, Fawzi, and Head-Gordon, 2003). Após o escalamento das coordenadas destes sistemas, foi feita uma *otimização local* com o GMIN de modo a convergir para a estrutura correspondente ao mínimo local mais próximo. Manipulações de parâmetros associados aos ângulos diedros por exemplo, como referido anteriormente no capítulo 2, ajudaram a diminuir a diferença entre a estrutura do mínimo local e a estrutura original do PDB. Em estudos semelhantes, Sorenson e Head-gordon (Sorenson and Head-Gordon, 2000) apresentaram semelhanças entre a estrutura da proteína L caracterizada com o modelo BLN e a determinada experimentalmente por ressonância magnética nuclear (RMN); a figura 3.1 mostra o modelo BLN (fig. 3.1a)) e a estrutura da proteína L obtida em estudos experimentais com RMN (fig. 3.1b)) apresentados por estes autores.

Feita a modelação dos sistemas baseados nas estruturas escaladas procurou-se interligá-los com estas (com a utilização do OPTIM), para tal foram utilizadas as coordenadas escaladas no ficheiro "finish" e as dos mínimos modelados no "odata" (ficheiros de entrada do programa OPTIM), o que permitiu obter nos dados de “output” as energias correspondentes a cada um deles. Estas foram utilizadas como referência (energias utilizadas como parâmetro na chaves "TARGET" no ficheiro "data", ficheiro de entrada do GMIN) para fazer *otimizações globais* com o GMIM e permitir que essas parassem uma vez

[1] A diferença de escalas foi observada por comparação das distâncias entre os resíduos interligados nas geometrias provenientes do PDB e dos resíduos interligados no modelo BLN. Os resíduos interligados no modelo BLN se separam por uma distância de 1 σ.

encontrados os mínimos pretendidos.

3.1 Proteína 1L2Y

A proteína 1L2Y foi o menor sistema estudado, contendo uma sequência de 20 resíduos que se enrolam para formar um dos domínios estruturais mais simples - domínio "trp-cage" (Neidigh, Fesinmeyer, and Andersen, 2002). O domínio "trp-cage" pode ser observado em proteínas simples como a 2JOF (fig. 3.2a) e ciclos "trp-cage" (por exemplo o ciclo "trp-cage" 2LL5 - fig. 3.2b)). Este domínio pode ser também observado em sistemas globulares. Por exemplo, os ciclos "trp-cage" podem ser observados em cristais monoclínicos formados na proteína globular 3UC7 (fig. 3.2c)). Nesta proteína 6 ciclos "trp-cage" agregam-se para forma o cristal.

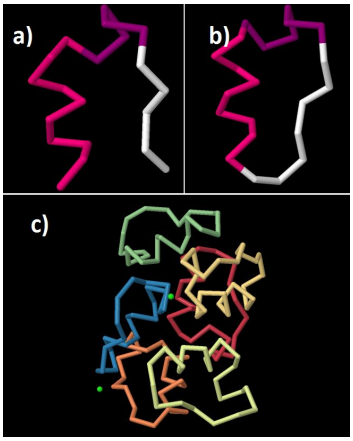


FIGURA 3.2: Estruturas das proteinas 2JOF (a), 2LL5 (b) e 3UC7 (c).

3.1.1 Parametrização do modelo BLN para a proteína 1L2Y

O sistema com o domínio estrutural "trp-cage" foi modelado com base nos dados das estruturas primárias e secundárias apresentados na tabela 3.1. Nesta tabela são apresentadas as estruturas primárias e secundárias provenientes do PDB e as resultantes do nosso mapeamento ao modelo de resíduos BLN. O mapeamento dos aminoácidos para o modelo de resíduos BLN do sistema 1L2Y e para os outros sistemas estudados foi feito com base nos dados apresentados na tabela 2.1 no capítulo 2, página 14. A sequência

TABELA 3.1: Estrutura primária e secundária do PDB e resultantes do mapeamento para o modelo de resíduos BLN da proteína 1L2Y.

Tipo de estrutura	**Sequência**
Estrutura primária do PDB	NLYIQWLKDG GPSSGRPPPS
Estrutura secundária do PDB	HHHHHHHTT GGGGT
Estrutura primária no modelo BLN	LBBBLBBLLNNNNNNLNNNN
Estrutura secundária no modelo BLN	THHHHHHHHX[a]TTTTTTTT

[a] A conformação X corresponde ao E do nosso método HET[X] (para o qual os valores dos parâmetros de torção dos ângulos diedros foram alterados.

da estrutura secundária (sequência composta pelas conformações em hélices (H), em dobras (T) e estendidas (E)) do sistema 1L2Y foi montada com base na estrutura secundária fornecida pelo PDB e melhorada com a utilização do mapeamento manual e por utilização do método HET[X].

Foram tidos como padrão os parâmetros do potencial de torção de ângulos diedros e de pares de resíduos não ligados por ligações peptídicas utilizados para dobragens mistas tratados no capítulo 2 nas tabelas 2.7 e 2.8.

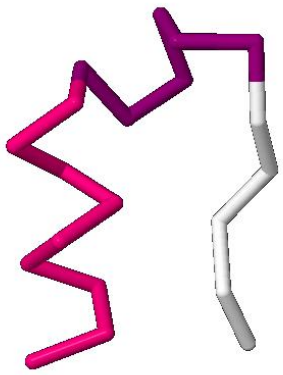


FIGURA 3.3: Estrutura da proteina 1L2Y visualizada no PDB. Distinguem-se da esquerda a direita uma estrutura secundária hélice-α (em cor rosa), uma hélice-3/10 (em cor roxa) e conformações não atribuídas (em cor branca) (Anderson and Rost, 2009).

Como referido, a caracterização do sistema 1L2Y foi feita com base nas coordenadas escaladas da estrutura obtida experimentalmente e apresentada no PDB. Na figura 3.3 é apresentada a estrutura da proteína 1L2Y visualizada a partir do PDB.

TABELA 3.2: Parâmetros para o potencial de torção dos ângulos diedros utilizados para modelação da proteína 1L2Y. Os valores dos parâmetros A, B, C e D utilizados para a conformação E (X) têm um efeito sobre o quarteto de resíduos 9-12 (no modelo da proteína 1L2Y).

Tipo de conformação	**A**	**B**	**C**	**D**
H	0.0	1.2	1.2	1.2
X	0.0	1.2	0.0	0.0
T	0.0	0.0	0.2	0.0

Na tabela 3.2 são apresentados os parâmetros de torção de ângulos dietros utilizados. Alterações dos parâmetros das conformações estendidas, aqui referidas como X, foram efetuadas recorrendo à análise dos perfis de energia potencial apresentados na figura 3.4.

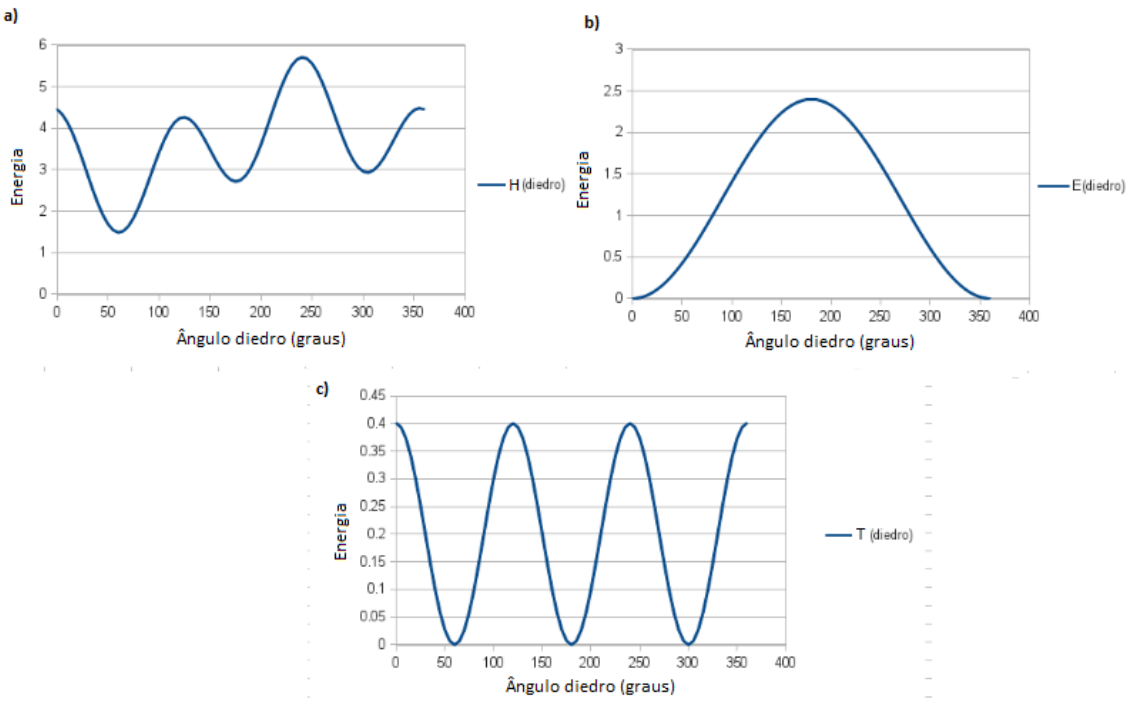


FIGURA 3.4: Ângulos diedros formados para cada conjunto de quatro resíduos na proteína 1L2Y: manipulações dos valores dos parâmetros A, B, C e D das conformações H (a)), E (b)) e T (c)) (gráficos criados com base na expressão do potencial BLN para a parte que controla os ângulos diedros).

Os gráficos apresentados na figura 3.4 mostram o efeito dos valores dos parâmetros A, B, C e D, da eq. (2.1) sobre as conformações H, E (X) e T adotadas para cada quarteto de resíduos no sistema 1L2Y. Para a conformação em hélice há uma tendência de formação

de ângulos diedros de 60º, 180º ou 300º (fig. 3.4a)). Os ângulos diedros que podem ser adotados pela conformação em hélice apresentam vários mínimos, sendo o de 60º o mais estável. A figura 3.4b) mostra uma tendência de formação de ângulos diedros de 0º ou de 360º para a conformação estendida (E (X)). Nota-se que, neste caso, X corresponde a uma conformação, a qual tem um efeito direto sobre o quarteto de resíduos 9-12. Por outro lado as conformações em dobra apresentam as mesmas tendências de formação de ângulos diedros apresentadas para as hélices, porém, com mínimos com energias iguais e mais baixas (fig. 3.4c)).

O mínimo modelado obtido para a proteína 1L2Y é apresentado na figura 3.5a) (mínimo com a energia 12.912648). Utilizando do método HET, original, obteve-se a estrutura apresentada na figura 3.5b) com energia igual a 18.424453[2]. Nesta última pode-se visualizar um claro afastamento da parte neutra correspondente aos resíduos 17-20 por efeitos de repulsão com a parte hidrofóbica da hélice-α. O método HET[X] permitiu assim encontrar uma melhor aproximação à estrutura escalada proveniente do PDB. Porém, com a utilização deste método houve alguma distorção na hélice-α.

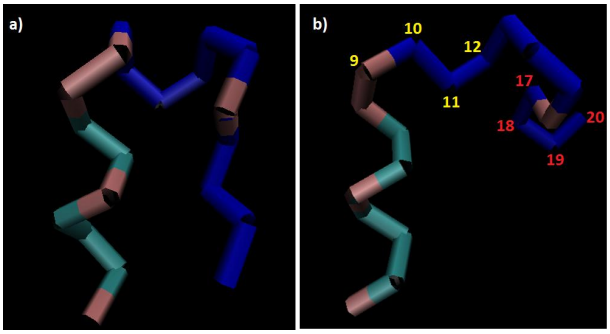


FIGURA 3.5: Modelos BLN obtidos para a proteína 1L2Y: a) usando o método HET[X]; b) com o método HET. Os resíduos (9-12) do ângulo diedro afetado pelo X do método HET[X] estão numerados em ordem crescente em amarelo na fig. (b). Os resíduos neutros 17-20 (numerado em vermelho) são repelidos pelos resíduos hidrofóbicos presentes na hélice-α.

[2]As estruturas apresentados na figura 3.5 foram obtidas por otimização local.

3.1.2 Resultados referentes à paisagem energética da proteína 1L2Y

Uma vez estabelecido o modelo para a proteína 1L2Y, gerou-se o gráfico de conectividade apresentado na figura 3.6, onde o mínimo modelado (mínimo 2) foi conectado ao mínimo correspondente a energia 12.484355 (mínimo 1) na otimização com o PATHSAMPLE. Por sua vez estes mínimos ligam-se ao 3 com valor de energia 12.674485. Comparando os valores das energias destes três mínimos locais gerados, se verificam poucas diferenças entre eles relativamente aos demais sistemas estudados, o que é refletido no desvio padrão 0.214595 (desvio padrão obtido para os três mínimos de energia referidos). Este gráfico de conectividade mostra alguma facilidade de transição do mínimo modelado para o mínimo 1, uma vez que há uma barreira energética baixa (TS, com valor de energia 12.928326) que os separa. Esta tendência pode ser também observada caso o sentido da isomerização seja para o mínimo 3, pelo facto da barreira que o separa dos demais também não ser muito alta.

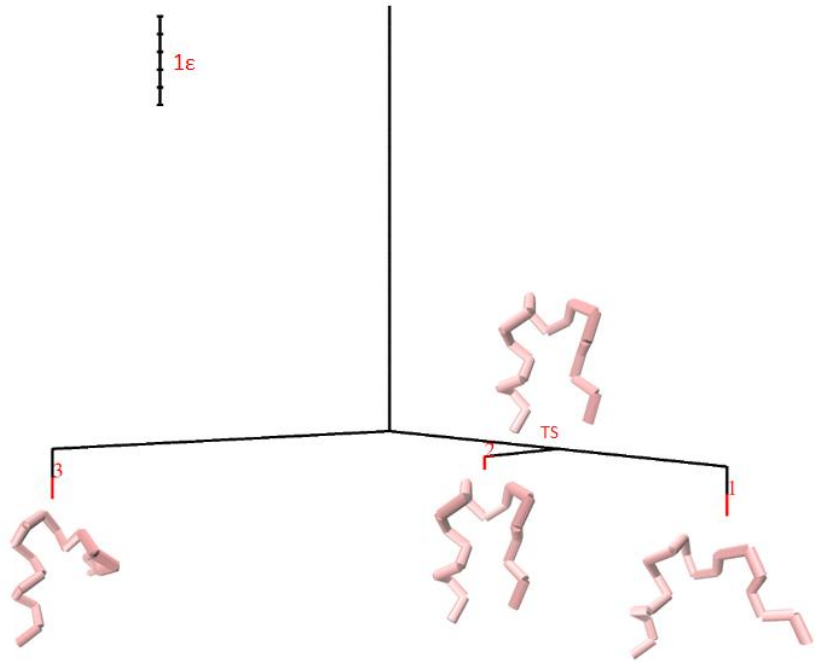


FIGURA 3.6: Gráfico de conectividade obtido para o mínimo modelado da proteína 1L2Y: mínimo modelado (**2**); mínimo a ele com o PATHSAMPLE (**1**); mínimo resultante da conectividade entre o **2** e o **1** (**3**). As estruturas apresentadas para cada um dos mínimos foram reproduzidas com o VMD. O espaçamento de energia utilizado foi de 0.2ϵ.

Poucas diferenças de energia foram também observadas para o gráfico de conectividade expandido com 50 ciclos (gráfico apresentado na figura 3.7), porém, neste se verifica maiores frustrações estruturais (mínimos de pouca energia separados por barreiras energéticas altas). Neste gráfico de conectividade podem ser observadas essencialmente duas bacias de atração, numa das quais se encontram os mínimos observados no gráfico de conectividade anterior mais o mínimo com energia 12.677977 (mínimo 236) e, outra onde observam-se os mínimos de valores de energias 12.434109, 12.240553, 12.079013, 11.892286, 11.957609 e 11.730815 representados pelos números 279, 346, 294, 295, 237 e 278, respetivamente. O desvio padrão 0.390088 obtido para os mínimos apresentados no gráfico 3.7 comprova que estes apresentam poucas diferenças de energias. Neste gráfico notam-se duas barreiras energéticas altas que passam pelos estados de transição TS1 e TS2, os quais pressupõem algumas dificuldades de relaxação nas transições em que é necessário passar por estes pontos.

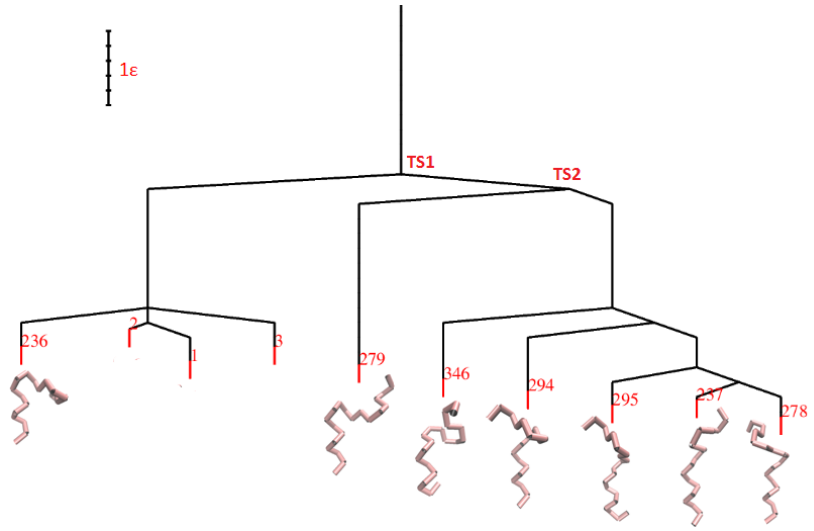


FIGURA 3.7: Gráfico de conectividade expandindo com 50 ciclos. O espaçamento de energia utilizado foi de 0.2ϵ.

As tendências acima referidas (mínimos com poucas diferenças de energias e surgimento de barreiras energéticas altas) são observadas na maioria dos mínimos no gráfico de conectividade expandido e melhorado com 500 ciclos (figura 3.9), no qual visualizam-se alguns mínimos com energias bem mais profundas, no caso, o mínimo global com valor de energia 9.694151 (mínimo 4810), o de energia 10.015942 (mínimo 1281) e poucos

outros.

Segundo Wales e colaboradores (Miller and Wales, 1999; Wales, 2004) o padrão exibido pela bacia de atração em que se encontra o mínimo modelado apresenta relaxações não muito eficientes para o mínimo global, a qual denominou de *"banyan tree"* pelo facto da diferença de energia entre os mínimos ser inferior à barreira energética que os separa. No gráfico apresentado na figura 3.9 observa-se claramente o padrão de *"banyan tree"* associado a zona em que se encontra o mínimo modelado (zona marcada com um contorno esverdeado) o que nos dá uma ideia clara da frustração estrutural que pode ser apresentada pelo sistema 1L2Y. A zona em que se encontra o mínimo modelado é enfatizada na figura 3.9.

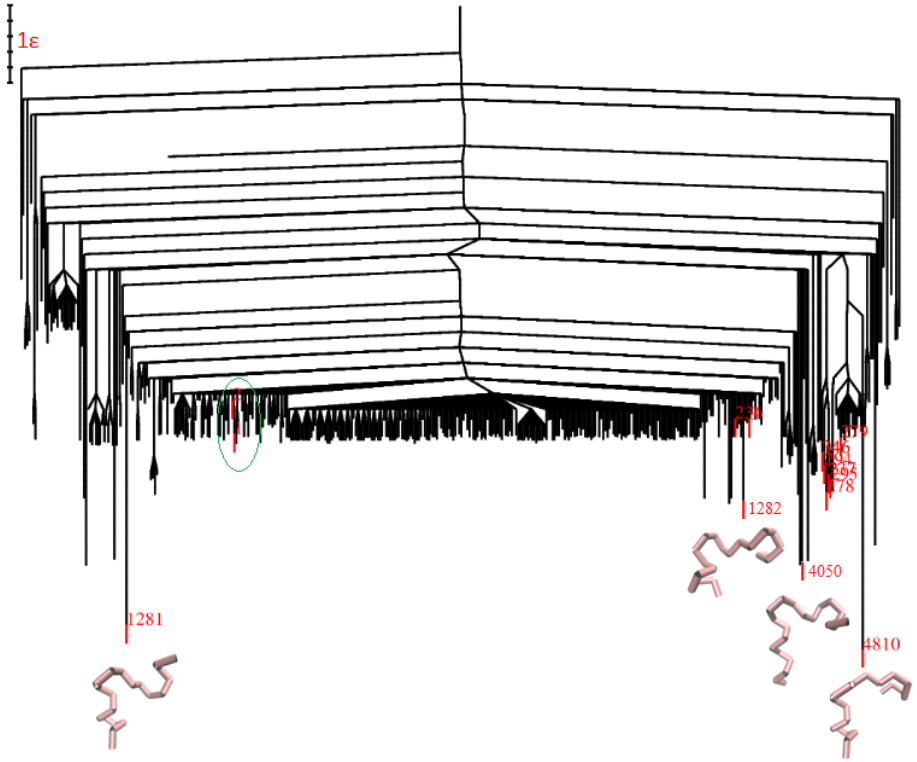


FIGURA 3.8: Gráfico de conectividade expandindo com 500 ciclos: mínimo 1 e 2 estão marcados por um contorno esverdeado; O mínimo modelado (mínimo 2) se encontra num conjunto de mínimos com energias muito próximas, os quais dão uma ideia da frustração estrutural que pode ser apresentada pelo sistema 1L2Y; para este sistema não se verificou muitas semelhanças estruturais entre o mínimo global encontrado (mínimo 4810) e o modelado. Neste gráfico foram conectados um total de 434 mínimos e 3609 estados de transição. Foi utilizado um espaçamento de 0.2ϵ.

De forma geral o gráfico expandido para 500 ciclos mostra que apesar das barreiras baixas observadas nos gráficos anteriores, existem grandes frustrações geométricas associadas a transição dos vários isómeros ao mínimo global. Estes aspetos nos dão uma ideia do paradoxo de Levinthal associado a este sistema, uma vez que podem haver caminhos de conectividade diferentes e tempos distintos para encontrar o mínimo global, tendo em conta os vários perfis energéticos dos mínimos conectados (mais adiante no fim deste capítulo são apresentados os tempos médios obtidos (MFETs) para encontrar o mínimo global deste e dos demais sistemas estudados para várias otimizações com o GMIN).

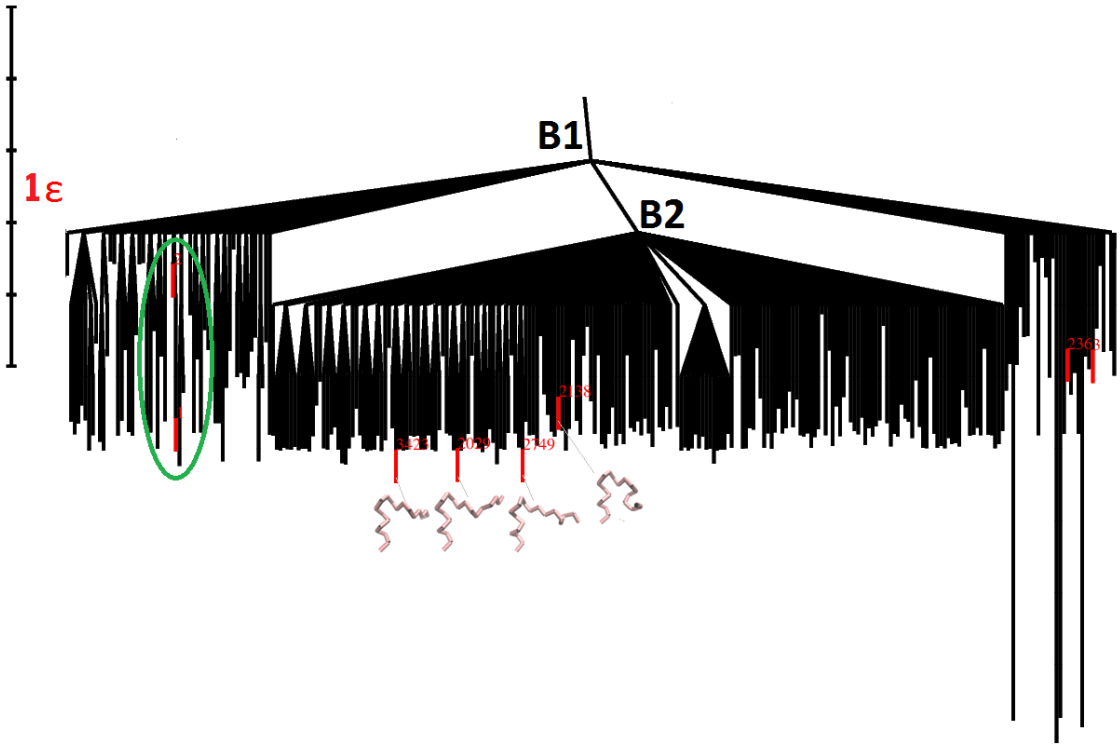


FIGURA 3.9: Gráfico de conectividade expandindo com 500 ciclos com ênfase à zona em que se encontra o mínimo modelado e o mínimo 1 (zona marcada com contorno verde na bacia de atração **B1**). Para além do mínimo modelado notam-se vários mínimos com o padrão "banyan tree", a maior parte deles presentes na bacia de atração **B2**.

3.2 Proteína 2MWE

A proteína 2MWE apresenta um domínio estrutural do tipo "ww-domain", o qual contém basicamente três folhas-β dispostas antiparalelamente (Serpell, 2000; Brown, Fawzi, and Head-Gordon, 2003). O motivo "ww-domain" é observado em várias proteínas simples, tais como a 5AHT (fig. 3.10a) e a 2N8S(fig. 3.10b), e em proteínas mais complexas como a 2LTY (fig. 3.10c) e a 1TK7 (fig. 3.10d).

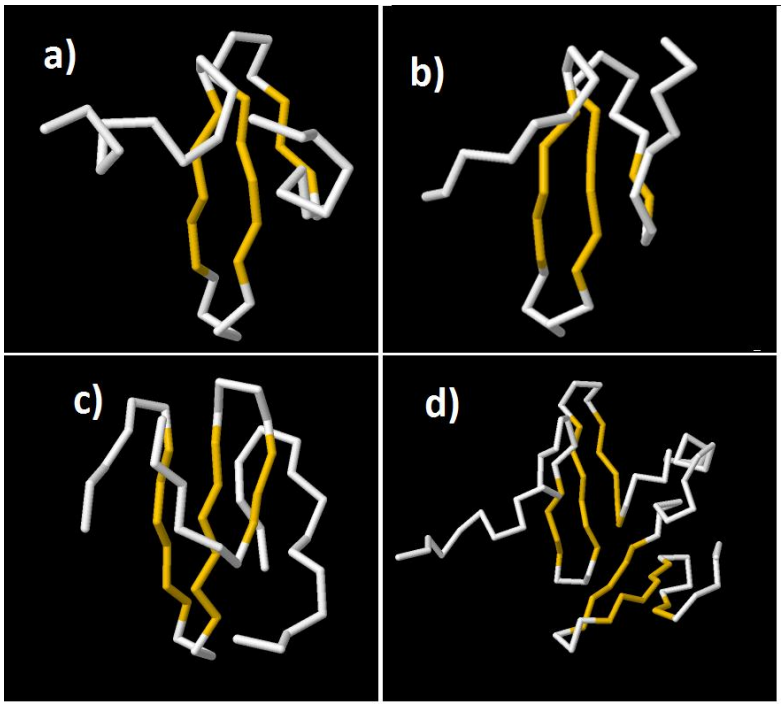


FIGURA 3.10: Estruturas das proteinas 5AHT (a), 2N8S (b), 2LTY (c) e 1TK7 (d).

3.2.1 Parametrização do modelo BLN para a proteína 2MWE

Como referido anteriormente, no processo de caracterização do sistema 2MWE partiu-se de igual modo de informações provenientes do PDB (sequência de aminoácidos, estrutura secundária e coordenadas escaladas).

TABELA 3.3: Estrutura primária e secundária do PDB e resultantes do mapeamento para o modelo de resíduos BLN da a proteína 2MWE.

Tipo de estrutura	**Sequência**
Estrutura primária do PDB	SEWTERKTAD GKTYYYNNRT AESTWEKP
Estrutura secundária do PDB	SEEEEETT TEEEEEETTT TEEE S
Estrutura primária no modelo BLN	NLBLLLLLBLNLLBBBLLLLBLNLBLLN
Estrutura secundária no modelo BLN[a]	TTTEEEEHHHTTEEEEEETTHTEEE
Estrutura secundária no modelo BLN[b]	TTTEEEETTTTTEEEEETTTTTEEE

[a]Mapeamento manual.
[b]Mapeamento utilizando o método de Porter.

Na tabela 3.3 são apresentadas as estruturas primária e secundária provenientes do PDB e as resultantes do processo de mapeamento para o sistema de resíduos BLN. Para este sistema obteve-se melhores resultados (maiores semelhanças com o sistema PDB) para o mapeamento da estrutura secundária utilizando o método de Porter e o mapeamento manual. O modelo final utilizado para este sistema foi o modelado com o método de mapeamento manual; na figura 3.11 é apresentada estrutura da proteína 2MWE visualizada no PDB.

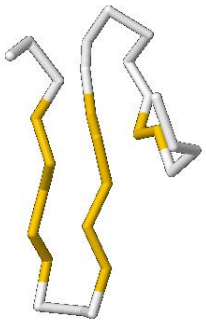


FIGURA 3.11: Proteina 2MWE: Estrutura visualizada no PDB (as três folhas-β são apresentadas em amarelo).

Na figura 3.12 são apresentadas as conformações adotadas pelos ângulos diedros, em função de dos valores dos parâmetros A, B, C e D apresentados na tabela 3.4 para as estruturas H, E e T na proteína 2MWE. Para este sistema foram necessárias alterações dos

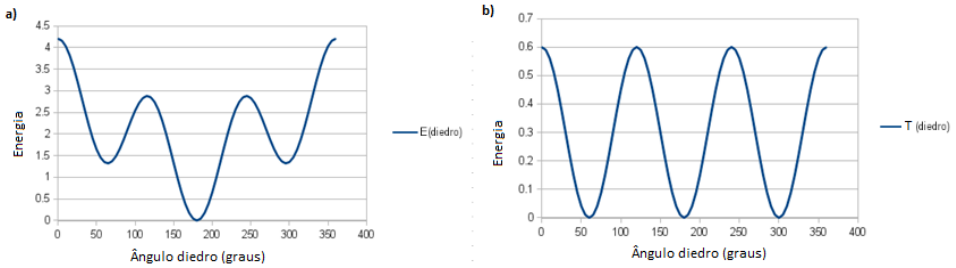


FIGURA 3.12: Ângulos diedros formados para cada conjunto de quatro resíduos na proteína 2MWE por manipulações dos valores dos parâmetros A, B, C e D das conformações E (a)) e T (B)).

TABELA 3.4: Parâmetros para o potencial de torção dos ângulos diedros utilizados para modelação da proteína 2MWE.

Tipo de conformação	**A**	**B**	**C**	**D**
H	0.0	1.2	1.2	1.2
E	0.9	0.0	1.2	0.0
T	0.0	0.0	0.3	0.0

parâmetros A (das conformações estendidas) e C (das conformações em dobra). Para as estruturas secundárias em hélices, foram utilizados os valores dos parâmetros utilizados no exemplo anterior, para os quais há maior tendência de formação de ângulos diedros de 60º. No caso das conformações estendidas e em dobras, predominaram os ângulos diedros de 180º (mínimo de menor energia na fig. 3.12a)) e, 60º, 180º e 300º (correspondentes aos três mínimos na fig. 3.12b)), respetivamente. As tendências conformacionais em dobra (com facilidade de dobrar para os ângulos diedros de 60º, 180º e 300º) do sistema 2MWE, em função do valor do parâmetro C=0.3 adotado, tendem a apresentar maior rigidez em relação às dos demais sistemas, o que é patente nos mínimos de energia mais profundos observados na figura 3.12b).

Para os pares de resíduos não ligados por ligações peptídicas foram usados os valores dos parâmetros utilizados em dobragens mistas apresentados no capítulo anterior.

Na figura 3.13 são apresentados os modelos finais obtidos para a proteína 2MWE resultantes do mapeamento manual (fig. 3.13a)) e do mapeamento com método de Porter (fig.

3.13b)), para os quais é possível notar algumas semelhanças relativamente a estrutura escalada proveniente do PDB (ver a figura 3.11).

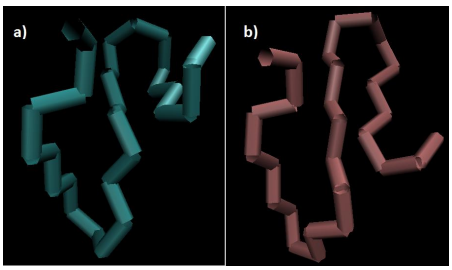


FIGURA 3.13: Estruturas modeladas para a proteína 2MWE: mapeamento manual (**a)**) e método de Porter (**b)**).

3.2.2 Resultados referentes à paisagem energética da proteína 2MWE

As diferenças de energia apresentadas pelos mínimos da proteína 2MWE foram maiores, relativamente ao sistema 1L2Y, como pode ser observado no gráfico de conectividade obtido para este sistema (ver a figura 3.14).

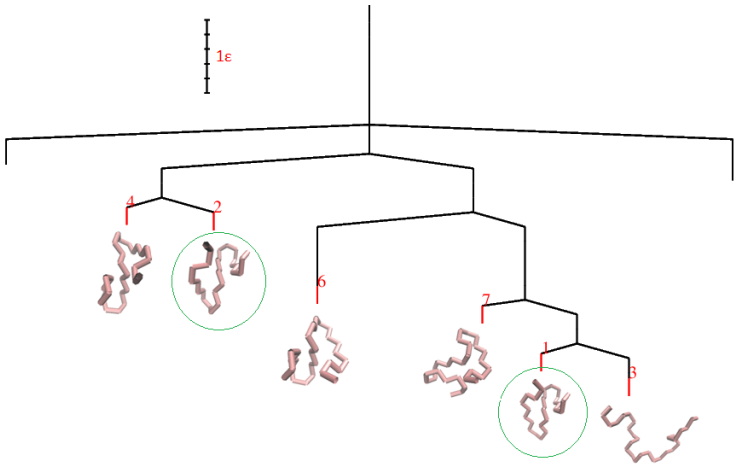


FIGURA 3.14: Gráfico de conectividade obtido para o mínimo modelado da proteina 2MWE: a estrutura do mínimo modelado (**2**) e estrutura do mínimo a ele conectado com o PATHSAMPLE (**1**) estão marcadas por um contorno verde. As demais estruturas geradas então representadas pelos números 3, 4, 6 e 7, das quais a 6 apresenta maior barreira energética ao ligar-se ao mínimo mais baixo. Foi utilizado o espaçamento de energia de 0.2ϵ.

Neste gráfico o mínimo modelado, com o valor de energia 20.194729 (mínimo 2, marcado com contorno esverdeado) e, o a ele conectado com o PATHSAMPLE, com valor de energia 18.262078 (mínimo 1, também marcado com contorno esverdeado), apresentam uma diferença de energia claramente superior em relação a apresentada pelo sistema 1L2Y visto no ponto anterior (secção 3.1, fig. 3.6) e barreiras energéticas relativamente baixas. Apesar da "armadilha" energética (barreira energética alta) apresentada pelo mínimo de energia 19.183436 (mínimo 6), nota-se uma tendência de relaxação mais ou menos eficiente para o mínimo mais baixo, no caso o mínimo com energia 17.920067 (mínimo 3).

Calculando o desvio padrão de todas energias salientadas a vermelho presentes na bacia de atração dos mínimos conectados com o PATHSAMPLE (mínimo 2 e 1), obteve-se o valor 0.96817 que, nos dá uma ideia da separação energética entre elas e enfatiza relaxações eficientes para o mínimo mais baixo (os mínimos representados pelos números 4 e 7, correspondem as energias 20.264931 e 18.917402, respetivamente).

No gráfico de conectividade expandido para 500 ciclos, ilustrado na figura 3.15, observa-se melhor o padrão dos mínimos gerados relativamente a um novo mínimo mais baixo encontrado, sendo este ainda mais profundo que o obtido no gráfico de conectividade anterior. Deste modo há uma maior diferença de energia deste mínimo em relação ao mínimo modelado e barreiras não muito altas que pode implicar maior facilidade de relaxação. Porém, apesar do mínimo modelado não apresentar uma barreira energética muito alta ao relaxar para o mais baixo (mínimo global, com o valor de energia 12.073546 - mínimo correspondente ao número 3369 no gráfico de conetividade expandido) (padrão que Wales denominou por *"folha de palmeira"* (Wales, 2004), no qual há uma relaxação mais eficiente para o mínimo global), existem muitos mínimos com energias não muito distantes deste e, com barreiras altas, o que Wales chamou de *"folhas de salgueiros"* (Wales, 2004), os quais podem constituir alguma "frustração" estrutural.

Apesar das "frustrações" estruturais observadas, o gráfico de conectividade expandido com 500 ciclos obtido para o referente sistema, apresenta padrões que sugerem maior eficiência de relaxação do mínimo modelado para o mínimo global do que no sistema 1L2Y.

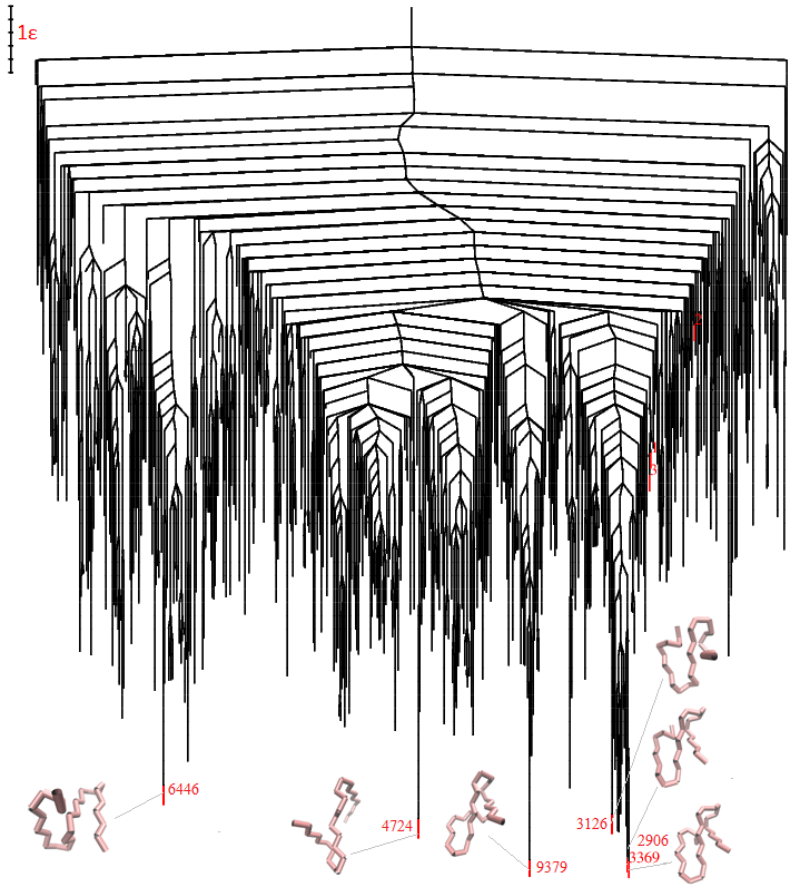


FIGURA 3.15: Gráfico de conectividade da proteína 2MWE expandindo com 500 ciclos. Apesar da diferença energética entre o mínimo modelado (**2**) e o global (**3369**) as suas estruturas são muito parecidas contrariamente ao verificado no sistema 1L2Y. Para este sistema foram conectados um total de 551 mínimos e 6016 estados de transição. Foi utilizado o espaçamento de energia de 0.2ϵ.

3.3 Proteína 1WY3

O motivo vilina faz parte das microvilosidades do citoesqueleto em eucariotas superiores e inferiores, o qual está envolvido em várias funções associadas a mobilidade e contração (Friederich et al., 1999). Este motivo pode ser encontrado nas proteínas simples como a 3MYC (fig. 3.16a)), a 2RJX (fig. 3.16b)) e a 2K6N (fig. 3.16c)) ou nos sistemas globulares 5I1O (fig. 3.16d)) e 5I1S (fig. 3.16e)).

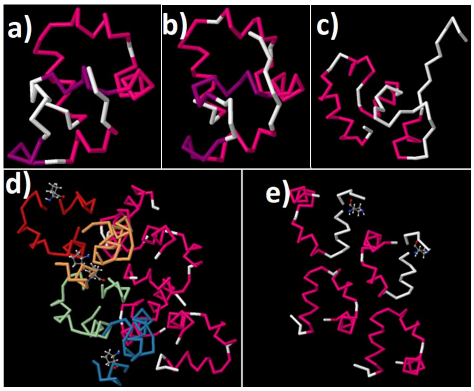


FIGURA 3.16: Estruturas das proteinas 3MYC (a), 2RJX (b), 2K6N (c), 5I1O (d) e 5I1S (e).

3.3.1 Parametrização do modelo BLN para a proteína 1WY3

TABELA 3.5: Estrutura primária e secundária do PDB e resultantes do mapeamento para o modelo de resíduos BLN da proteína 1WY3

Tipo de estrutura	**Sequência**
Est. primária do PDB	LSDEDFKAVF GMTRSAFANL PLWLQQHLKK EKGLF
Est. secundária do PDB	HHHHHHHH SSHHHHHHS HHHHHHHHH HHT
Est. primária (modelo BLN)	BNLLLBLBBBNBLLNBBBLBNBBBLLLBLLLLNBB
Est. sec. (mod. BLN/mét. HET[X])	TTHHHHHHHHHHX[a]HHHHHHHHHHTTHHHHHHH

[a] A conformação X corresponde ao E do método HET[X], o qual torce o ângulo diedro dos resíduos 13-16.

A sequência primária e a secundária do PDB que serviram de base para modelação da proteína 1WY3 e as mapeadas para o sistemas de resíduos BLN são apresentadas na tabela 3.5. Para a construção da estrutura secundária foram utilizadas, de forma combinada, as diferentes estratégias anteriormente referidas, de forma análoga ao estudo feito para o sistema 1L2Y, com ênfase ao mapeamento manual. Para este sistema também foi utilizado o método HET[X]. Uma vez que não há estruturas estendidas neste sistema, a conformação E correspondeu ao nosso X. Alterações dos valores dos parâmetros do nosso X foram feitas de modo a torcer o nosso sistema, procurando proporcionar o afastamento duma das hélices-α para a frente do plano, procedimento semelhante ao feito para o sistema 1L2Y.

Tal como nos dois sistemas referidos nos pontos anteriores (secções 3.1 e 3.2), para este

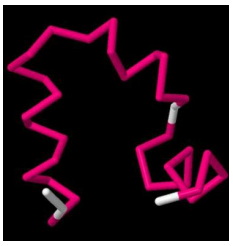


FIGURA 3.17: Proteina 1WY3: Estrutura visualizada no PDB. As 3 hélices-α do motivo vilina estão marcadas em tom rosa e em branco as partes não atribuídas.

sistema partiu-se das coordenadas escaladas de modo a minimizar os cálculos para obtenção do mínimo modelado; na figura 3.17 é apresentada a estrutura da proteína 1WY3 visualizada pelo PDB.

TABELA 3.6: Parâmetros para o potencial de torção dos ângulos diedros utilizados para modelação da proteína 1WY3.

Tipo de conformação	**A**	**B**	**C**	**D**
H	0.0	1.2	1.2	1.2
E	0.8	0.0	0.1	0.0
T	0.0	0.0	0.2	0.0

TABELA 3.7: Parâmetros para o potencial das interações entre pares de resíduos não ligados por ligações peptídicas utilizados na modelação da proteína 1WY3.

Tipo de interação	**Parâmetros utilizados**
BB	S_{12}=0.3 S_6=$-$0.3
LL ou LB	S_{12}=0.93333333333333 S_6=$\frac{1}{3}$
NN, NB ou NL	S_{12}=1 S_6=0

O sistema 1WY3 foi difícil de modelar devido (em grande parte) ao colapso hidrofóbico apresentado pelos resíduos 12, 16, 17 e 18. Para suavizar este efeito foram utilizados valores mais pequenos para o parâmetro de contacto hidrofóbico (BB) e aumentou-se a repulsão entre os resíduos hidrofílicos (repulsão apresentada pelas interações LL e LB) de modo a ter-se maior afastamento entre os resíduos 13 e 19 presentes na zona de colapso referida. Não foram possíveis maiores afastamentos pelo facto destes parâmetros não estarem associados apenas a alguns resíduos específicos, mas a todos de um determinado

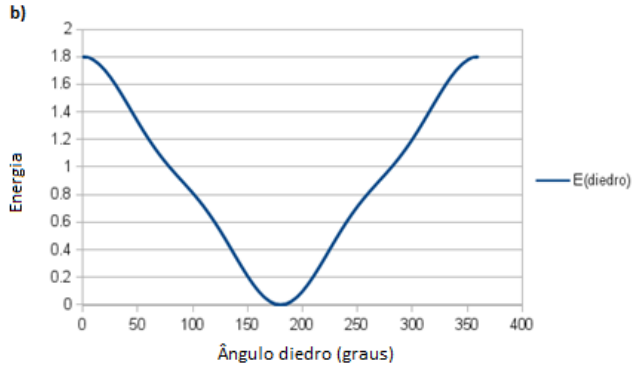


FIGURA 3.18: Ângulo diedro formado para o conjunto de quatro resíduos afetados pelo X do método HET[X] na proteína 1WY3.

tipo, o que levaria à destruição das outras estruturas secundárias obtidas; nas tabelas 3.6 e 3.7 são apresentados os valores dos parâmetros do potencial de torção dos ângulos diedros e pares de resíduos não ligados por ligações peptídicas (para os quais aumentou-se o valor de S_{12} para os resíduos LL ou LB e diminuiu-se os valores de S_{12} e S_6 para os resíduos BB) utilizados na modelação da proteína 1WY3.

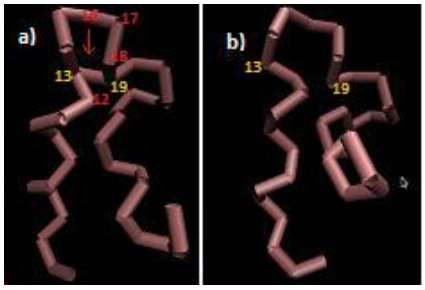


FIGURA 3.19: Estruturas modeladas para a proteína 1WY3: estrutura obtida tendo como parâmetros padrão os utilizados em dobragens mistas (**a)**) (apresentados no capítulo 2) e a gerada por alteração destes, de modo a aumentar a repulsão do resíduos 13 e 19 e, diminuir a atração do 12, 16, 17 e 18, (**b)**) (a zona apontada pela seta vermelha na estrutura (**a)**) não corresponde a uma ligação e sim a aproximação no espaço entre 13 e 19, causada pelo colapso hidrofóbico).

Os valores dos parâmetros do X (E) do método HET[X] representam uma tendência de torção de 180º para o ângulo diedro formado pelos resíduos 13-16 (resíduos do ângulo diedro afetados pelo X do método HET[X]) de acordo com o gráfico apresentado na figura

3.18. As tendências de formação de ângulos diedros das conformações H e T para este sistema foram as mesmas apresentadas pelo 1L2Y.

Na figura 3.19 é apresentada a estrutura obtida com a utilização dos valores dos parâmetros S_{12} e S_6 padrão (fig. 3.19a), na qual observa-se maior atração entre os resíduos hidrofóbicos 12, 16, 17 e 18 (estrutura correspondente ao valor de energia 40.573150) e, a melhorada por alteração destes (fig. 3.19b) (estrutura correspondente ao valor de energia 53.703798)[3]. A estrutura apresentada na figura 3.19b assemelha-se mais ao modelo experimental descrito no PDB para a proteína 1WY3, por haver espaçamento ligeiramente maior entre os resíduos 12, 16, 17 e 18 (pela diminuição da atração dos resíduos hidrofóbicos) e entre os resíduos 13 e 19 (por aumento da repulsão entre os resíduos hidrofílicos).

3.3.2 Resultados referentes à paisagem energética para proteína 1WY3

O gráfico de conectividade obtido para a conexão do mínimo obtido por otimização local a partir da estrutura PDB escalada reflete uma relaxação eficiente deste aos mínimos mais baixos, pela clara diferença energética e baixa barreira que passa por TS observada (padrão "folha de palmeira").

Na figura 3.20 é apresentado o gráfico de conectividade obtido para o sistema 1WY3, onde o sistema modelado corresponde ao mínimo 2, o qual foi conectado com o PATHSAMPLE ao mínimo com energia 51.465087 (mínimo 1). Ambos os mínimos (mínimo 2 e 1) estão marcados por um contorno verde.

Analisando os caminhos de conectividade apresentados na figura 3.20, é um pouco difícil saber qual o padrão predominante, se "folha de palmeira", "folha de salgueiro" ou "banyan tree". Por exemplo, apesar do mínimo modelado apresentar uma baixa barreira e grande diferença energética face à estrutura de menor energia (padrão "folha de palmeira"), este e o mínimo de energia 53.953518 (mínimo 4) apresentam pouca diferença de energia e a barreira que os separa reflete o padrão "banyan tree"; por outro lado, o mínimo 1 e o de energia 51.148624190256157 (mínimo 3) apresentam uma ligeira diferença de energia e uma barreira energética que pressupõe o padrão "folha de salgueiro".

[3] Estrutura final utilizada.

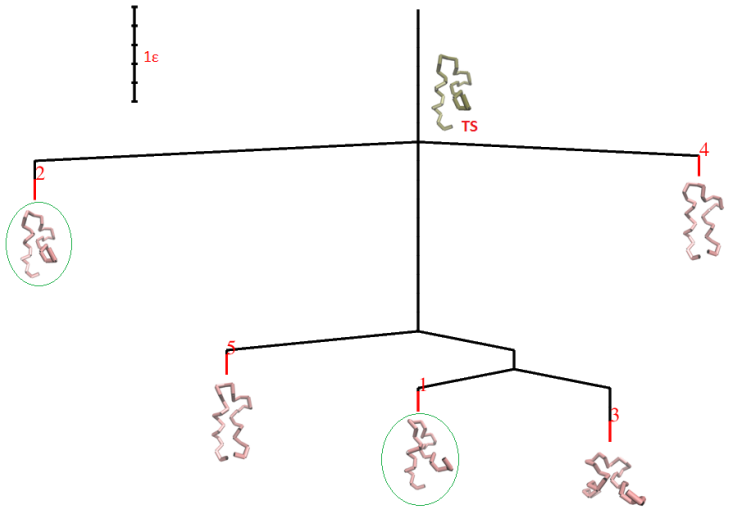


FIGURA 3.20: Gráfico de conectividade obtido para o mínimo modelado para a proteína 1WY3: As estruturas correspondentes ao mínimo modelado (**2**) e o mínimo a ele conectado (**1**) com o PATHSAMPLE estão marcados por um contorno esverdeado. O mínimo modelado liga-se aos demais passando por um TS não muito alto. O espaçamento de energia utilizado foi de 0.2ϵ.

Pelo desvio padrão 1.308379 associado as energias apresentadas, dá para ter uma ideia das diferenças entre elas em relação aos sistemas apresentados nas secções anteriores (sistema 1L2Y e 2MWE), as quais são relativamente altas.

Porém a eficiência de relaxação do sistema modelado para os mínimos mais baixos é mais facilmente visualizada por análise dos gráficos de conectividade expandidos com o PATHSAMPLE.

O mínimo modelado e os demais observados no gráfico de conectividade anterior estão conectados a mínimos ainda mais profundos como se pode observar no gráfico de conectividade expandido para 500 ciclos na figura 3.21, dos quais estão separados por barreiras energéticas bem mais altas. Os padrões observados nestas novas conexões assemelham-se de forma geral a "salgueiros", ou seja, mínimos próximos ao mais baixo separados por barreiras relativamente altas. Estes aspetos enfatizam uma relaxação menos eficaz do sistema modelado ao mínimo mais baixo (mínimo de energia 46.673145, correspondente ao número 7108 no gráfico de conectividade expandido), uma vez que as barreiras de energia altas podem constituir "armadilhas" energéticas que podem dificultar a isomerização

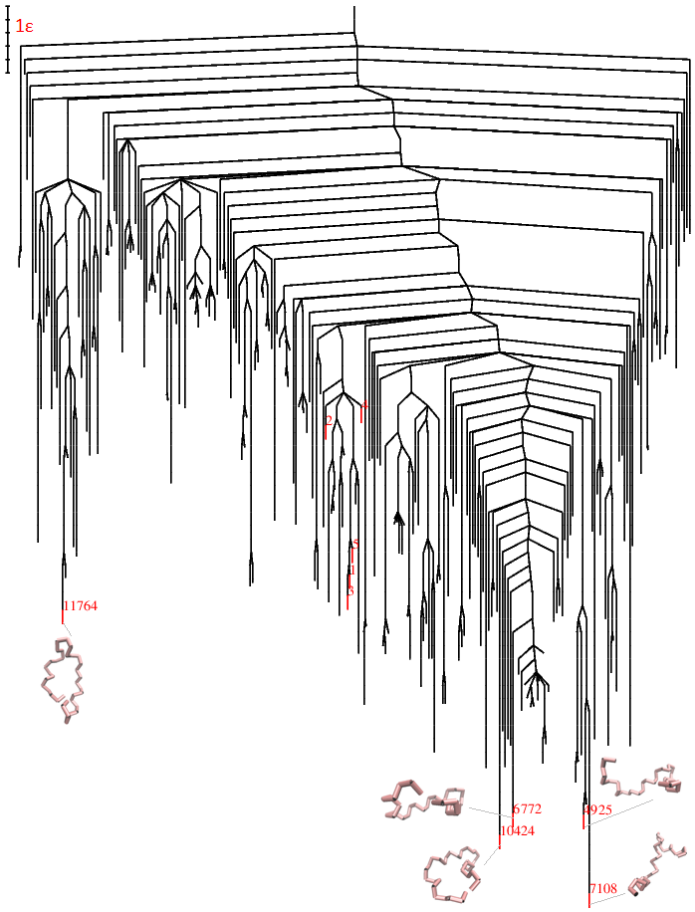


FIGURA 3.21: Gráfico de conectividade da proteína 1WY3 expandindo com 500 ciclos. Para este sistema não se verificou muitas semelhanças entre a estrutura e energética do mínimo modelado (**2**) em relação a dos mínimos de energia mais baixa (mínimo 7108, 10424, etc.). Neste gráfico foram conectados um total de 248 mínimos e 6777 estados de transição. O espaçamento de energia utilizado foi de 0.2ϵ.

(por exemplo as barreiras associadas aos mínimos 4925, 10424 e 6772). Porém nota-se a existência de alguns mínimos com energias bem mais altas que a do mínimo global (padrão "folha de palmeira"), para os quais esperam-se relaxações mais eficientes.

Relacionando os dois gráficos de conectividade obtidos para a proteína 1WY3 pode-se verificar que existe alguma facilidade de relaxação do sistema modelado para o mínimo a ele conectado com o PATHSAMPLE, porém a complexidade do mecanismo de isomerização aumenta à medida que nos deslocamos para estruturas com energias mais baixas.

3.4 Proteína com quatro hélices-α

Segundo Guo e Thrirumalai (Guo and Thirumalai, 1996) o domínio estrutural apresentado pela proteína com quatro hélices-α é bastante abundante na dobragem de várias proteínas[4]. Este domínio estrutural é observado nas proteínas simples 3U3B (fig, 3.22a)), 2LSE (fig, 3.22b)) (compostas por uma unidade funcional), 3TOL (fig, 3.22c)) e 5FJD (fig, 3.22d)) (compostas por duas unidades funcionais), como também em proteínas globulares bem mais complexas (compostas por agregados de várias proteínas com quatro hélices-α), por exemplo, a 4ZLW (fig, 3.22e)) e a 4ZKH (fig, 3.22f)).

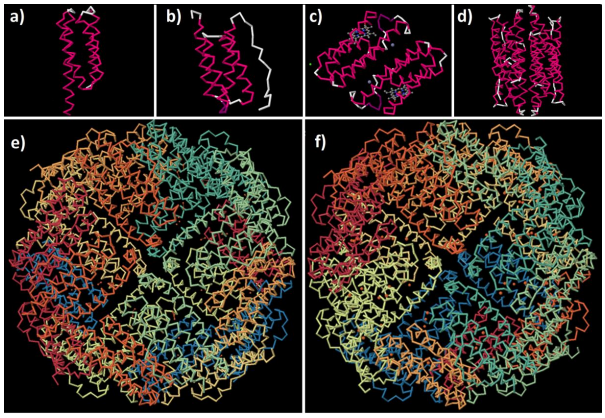


FIGURA 3.22: Estruturas das proteínas 3U3B (a), 2LSE (b), 3TOL (c), 5FJD (d), 4ZLW (e) e 4ZKH (f).

3.4.1 Parametrização do modelo BLN para a Proteína com quatro hélices-α

O sistema da proteína com quatro hélices-α foi relativamente mais simples de modelar em relação aos anteriores, sendo o mapeamento da sequência de aminoácidos e a construção da estrutura secundária feitos com base em dados prévios apresentados por Guo e Thrirumalai (Guo and Thirumalai, 1996); na tabela 3.8 é apresentada a estrutura primária e secundária utilizadas na modelação da proteína com quatro hélices-α.

[4] As quatro hélices-α desta proteína se apresentam empacotadas na sua estrutura nativa.

TABELA 3.8: Estrutura primária e secundária do PDB e resultantes do mapeamento para o modelo de resíduos BLN da proteína de quatro hélices-α.

Tipo de estrutura	**Sequência**
Est. primária (modelo BLN)	LLBLLBBLLBLLBBLLNNNLLBLLBBLLBLLBBLLNN
	NLLBLLBBLLBLLBBLLNNNLLBLLBBLLBLLBBLL
Est. secundária (modelo BLN)	HHHHHHHHHHHHHHHTTTTTTTHHHHHHHHHHHHT
	TTTTTTHHHHHHHHHHHHTTTTTTTHHHHHHHHHH

Para modelação do sistema da proteína com quatro hélices-α foram utilizados os parâmetros das tabelas 2.7 e 2.8 propostos por Guo e Thrirumalai, apresentados no capítulo 2.

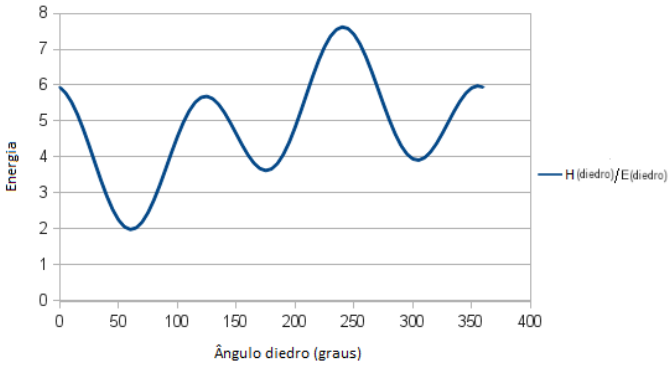


FIGURA 3.23: Mínimos de energia dos ângulos diedros das conformações H e E na proteína com quatro hélices-α.

As influências do potencial de torção dos ângulos diedros em função dos valores dos parâmetros A, B, C e D (A = 0; B=C=D=1.6) sobre as conformações H e E são apresentas na figura 3.23, para as quais o mínimo mais estável corresponde a um ângulo diedro de 60º. Para as conformações em dobra utilizou-se os parâmetros usados para os sistemas 1L2Y e 1WY3. A tendência de formação de ângulos diedros com 60º para as conformações H e E da proteína com quatro hélices-α é mais estável que a correspondente nos sistemas anteriormente analisados, pelo valor de energia mais profundo apresentado pelo mínimo mais baixo (ver a figura 3.23).

O processo de otimização global deste sistema foi feito partindo de coordenadas aleatórias[5], para o qual se obteve um mínimo com o valor de energia 31.602812; na figura 3.24 são apresentadas a estrutura do mínimo obtido por otimização global feita com o GMIN (3.24a)) e a estrutura correspondente ao estado nativo do sistema da proteína de quatro hélices-α proposta por Guo e Thrirumalai (3.24b)), para as quais se observam bastantes semelhanças.

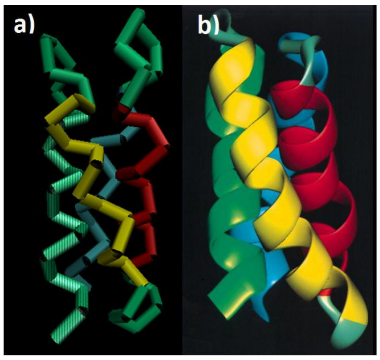


FIGURA 3.24: Estrutura modelada para a proteína com quatro hélices-α **(a)**) e sistema proposto por Guo e Thrirumalai (**b)**).

3.4.2 Resultados referentes à paisagem energética para Proteína com quatro hélices-α

Para o mínimo modelado da proteína com quatro hélices-α verificaram-se diversos padrões (energias próximas separadas por barreiras pequenas, diferenças de energias grandes separadas por barreiras pequenas, diferenças de energias relativamente pequenas e grandes barreiras, etc.). Porém por análise do gráfico de conectividade obtido para este sistema (gráfico apresentado na figura 3.25) observa-se que o mínimo modelado (mínimo com o valor de energia 31.602812, correspondente ao número 21 no gráfico) pode apresentar uma relaxação não muito eficiente para o mínimo a ele conectado com o PATHSAMPLE (mínimo com o valor de energia 31.654112, correspondente ao número 17 no gráfico), devido a barreira energética alta que os separa, apesar da pequena diferença de energia entre eles.

[5]No processo de otimização global deste sistema não partiu-se de coordenadas escaladas do PDB.

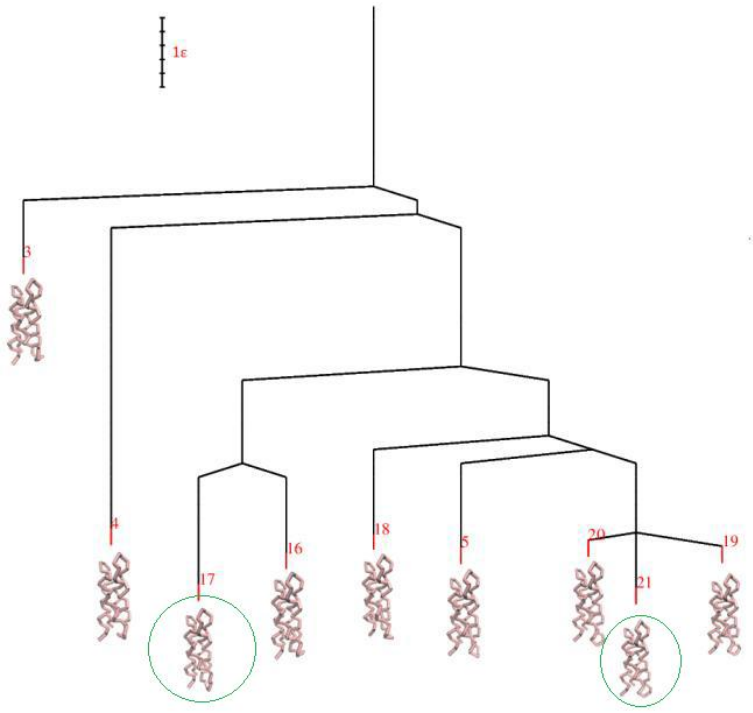


FIGURA 3.25: Gráfico de conectividade obtido para o mínimo modelado para a proteína com quatro hélices-α: o mínimo modelado (**21**) e o mínimo a ele conectado com o PATHSAMPLE (**17**) estão marcados por um contorno em tom verde. O espaçamento de energia utilizado foi de 0.2ϵ.

O valor 0.306726 para o desvio padrão das energias observadas no gráfico de conectividade apresentado na figura 3.25 (excluindo os mínimos de energias 36.376142 e 32.459288 representados pelos números 3 e 4), comprova que pode haver alguma dificuldade de transição do mínimo modelado para o isómero pretendido (por exemplo para o mínimo 17) face às barreiras energéticas altas. Os mínimos com os números 5, 16, 18, 19 e 20 (apresentados na figura 3.25), correspondem as energias 32.184564, 32.099925, 32.376229, 32.203473 e 32.290294, respetivamente.

Um dos aspetos interessantes a referir é o facto de para o sistema da proteína de quatro hélices-α não haver grandes variações estruturais entre os mínimos interligados quando comparado com os sistemas analisados nas secções anteriores.

O padrão predominante adotado pelos mínimos da proteína com quatro hélices-α relativamente ao mínimo global (padrão que combina "folhas de palmeira" e "folhas de

salgueiro) pode ser logo observado no gráfico de conectividade expandido para 50 ciclos apresentado na figura 3.26. Embora se possam ver um número relativamente grande de bacias do tipo "palmeira", o mínimo modelado, o mínimo a ele conectado na otimização com o PATHSAMPLE, e alguns outros, as suas bacias de atração assemelham-se a "folhas de salgueiro".

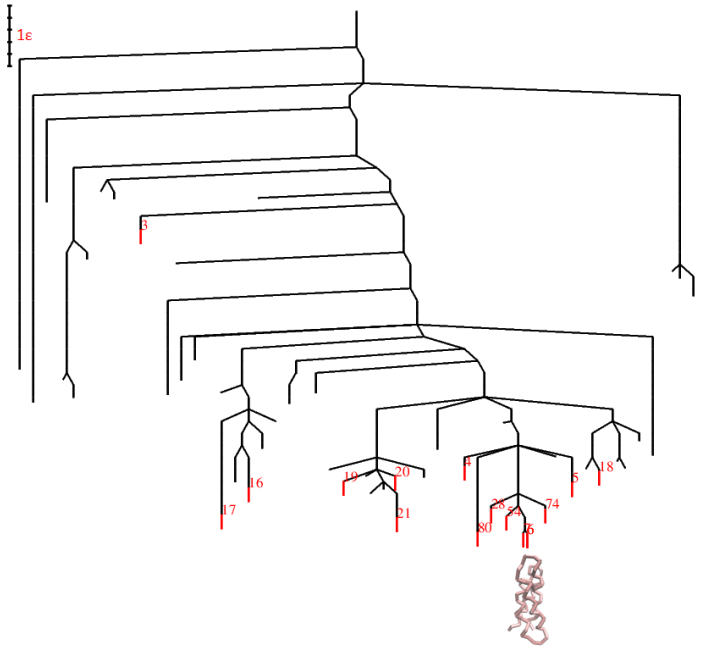


FIGURA 3.26: Gráfico de conectividade da proteína com quatro hélices-α expandindo com 50 ciclos. O espaçamento de energia utilizado foi de 0.2ϵ.

Outro aspeto que pode ser observado é o facto de para este sistema o mínimo modelado não se distanciar muito em termos energéticos e estruturais do mínimo global encontrado (mínimo com valor de energia 31.336770, representado pelo número 6), o que, em função da barreira energética que os separa, enfatiza a sua estabilidade quando comparado com os demais sistemas estudados.

O padrão "folhas de palmeira" é enfatizado no gráfico de conetividade expandido para 500 ciclos, apresentado na figura 3.27, onde se pode visualizar o mínimo global evidenciado em relação aos outros mínimos, o que pressupõe relaxamentos mais eficazes para a estrutura nativa. Porém, de forma análoga ao sistema visto no ponto anterior (sistema 1WY3, cujo padrão do gráfico de conectividade expandido apresenta "folhas de palmeira" bem como "folhas de salgueiro"), é necessário não esquecer que existem também vários

mínimos com barreiras altas que podem constituir "armadilhas" energéticas e levar a tempos de relaxação mais longos. Para além destes aspetos, para este sistema espera-se haver grande frustração estrutural pelo número muito elevado de mínimos conectados (710 mínimos) relativamente aos restantes analisados.

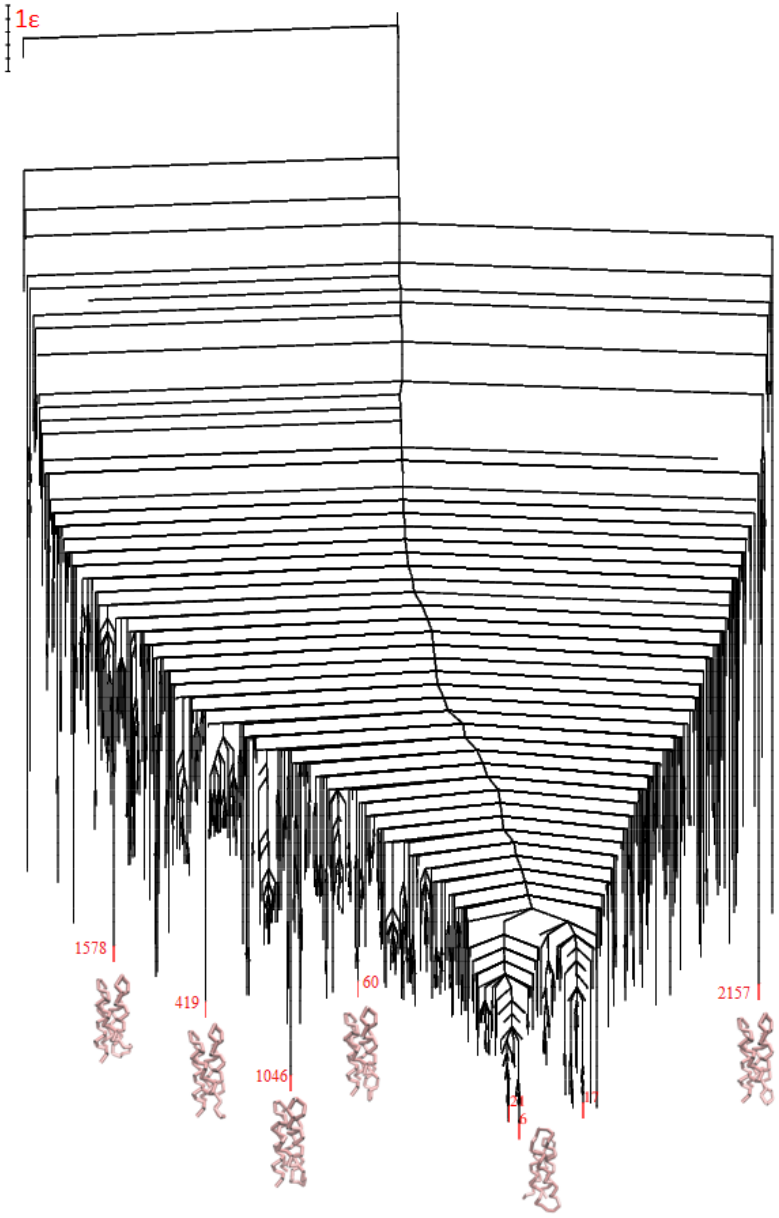


FIGURA 3.27: Gráfico de conectividade da proteína com quatro hélices-α expandindo com 500 ciclos. Foram conectados no total 710 mínimos e 1791 estados de transição. O espaçamento de energia utilizado foi de 0.2ϵ.

3.5 Tempos médios para encontrar os mínimos globais (MFET)

Na figura 3.28 é apresentado o gráfico com as distribuições de todos os mínimos gerados na base de dados do PATHSAMPLE[6] para cada um dos sistemas estudados, no qual, podemos analisar as zonas em que se encontram os mínimos conectados nos gráficos de conectividade apresentados nas secções anteriores.

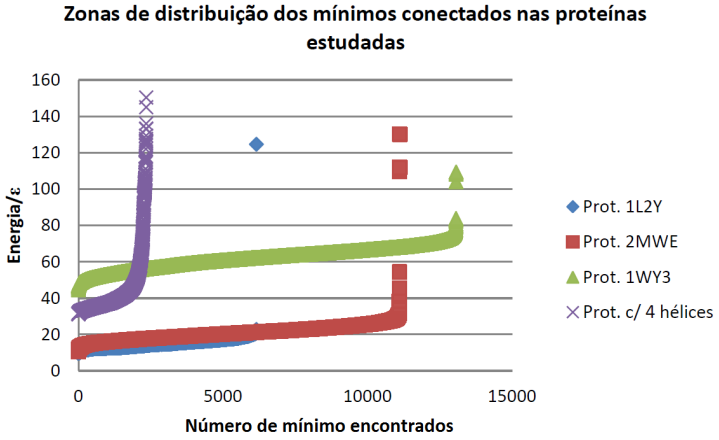


FIGURA 3.28: Gráfico com a distribuição de todos os mínimos gerados na base de dados de pontos estacionários de PATHSAMPLE para cada um dos sistemas estudados.

Por exemplo, podemos verificar nesse gráfico que para a proteína com quatro hélices-α há uma zona relativamente pequena (caso dividirmos imaginariamente a área de distribuição em duas, uma paralela ao eixo x e outra ao y) onde se encontram o mínimo modelado, o global e outros com baixas energias (mínimos com valor de energia por volta de 30.0 ϵ - nesta zona se encontram os vários mínimos com energia próxima do global e que apresentam tendências a formar o padrão "folhas de salgueiro) e, outra relativamente maior, que reflete a existência de mínimos com diferenças de energia bem mais altas, o que enfatiza o padrão “folhas de palmeira” anteriormente referido para este sistema (na secção anterior). Este último sugere um relaxamento mais eficaz ao mínimo global. Porém para este sistema obteve-se um MFET muito grande para um total de 43 otimizações com o GMIN se compararmos com os demais sistemas estudados, como pode se

[6]No total foram gerados 6160 mínimos para a proteína 1L2Y, 11119 para a 2MWE, 2348 para a proteína com quatro hélices-α e 13074 para a 1WY3.

observar na tabela 3.9, o que nos leva a concluir estar associado ao seu maior tamanho e complexidade estrutural, os quais levam a maiores frustrações estruturais (vários mínimos de baixa energia separados por barreiras altas). A proteína com quatro hélices-α apresenta o domínio estrutural mais complexo (quatro hélices-α empacotadas), constituído por um total de 73 resíduos.

TABELA 3.9: Tempos médios para encontrar o mínimo global dos sistemas estudados para várias otimizações com o GMIN.

Sistema	**MFET (s)**	**Desv. padrão (s)**	**Otim. c/ o GMIN**	**# de resíduos**
Prot. 1L2Y	0.33	0.23	100	20
Prot. 1WY3	2.90	1.16	100	35
Prot. 2MWE	0.19	0.13	100	28
Prot. c/ 4 hélices-α	31874.33	21802.17	43	73

Os restantes sistemas (proteína 1L2Y, 1WY3 e 2MWE) apresentam padrões relativos a distribuição energética dos seus mínimos semelhantes, como pode ser observado na figura 3.28.

Para a proteína 1L2Y, no gráfico apresentado na figura 3.28, com exceção da energia correspondente ao valor 124.58158 ϵ (mínimo não conectado nos gráficos deste sistema apresentados na secção 3.1), nota-se uma distribuição energética com uma área praticamente paralela ao eixo x, o que pressupõe diferenças de energias não muito grandes e enfatiza o padrão "banyan tree" como referido na secção 3.1, para o qual se espera uma relaxação menos eficiente. O valor maior do MFET apresentado para o sistema 1L2Y (0.33 s) comparando com o do sistema 2MWE[7] (0.19 s), enfatiza essa relaxação menos eficiente.

As zonas de distribuição energética dos sistemas 2MWE e do 1WY3 apresentadas na figura 3.28 são muito parecidas, apesar de se encontram em áreas de energia diferentes. Por análise dos gráficos de conectividade expandidos para 500 ciclos (apresentados nas secções anteriores) destes dois sistemas, apesar dos padrões ligeiramente semelhantes (várias barreiras energéticas altas), se verifica uma barreira energética para a transição

[7]Sistema com zona de distribuição dos mínimos mais semelhante a do sistema 1L2Y (ver a fig. 3.28; por outro lado o sistema 2MWE apresenta número de resíduos mais próximo ao da proteína 1L2Y.

"mínimo modelado"-"mínimo global" maior para a proteína 1WY3. Sugere-se que estes padrões estão diretamente relacionados com as diferentes complexidades estruturais e números de resíduos destes sistemas e, têm influência sobre os MFETs gerados. O sistema 1WY3 apresenta maior número de resíduos (ver a tabela 3.9) e um domínio estrutural mais complexo (domínio vilina, com três hélices-α) em relação ao 2MWE (sistema com domínio estrutural "ww-domain" composto praticamente por "folhas" estendidas). Para a proteína 1WY3 obteve-se um MFET superior ao da proteína 2MWE, como se pode visualizar na tabela 3.9.

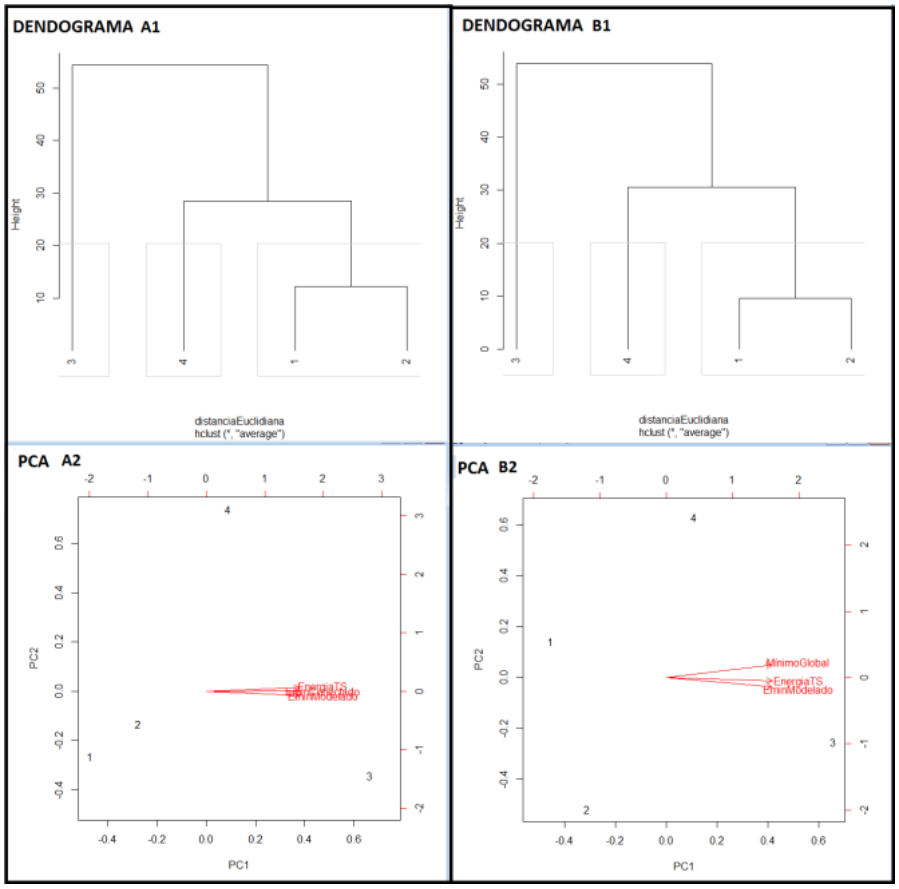


FIGURA 3.29: Dendogramas e gráficos PCA feitos para as conexões "mínimo-TS-mínimo": 1)"mínimo modelado - estado de transição - mínimo conectado ao mínimo modelado com o PATHSAMPLE" (dendograma A1 e gráfico PCA A2); 2) "mínimo modelado - estado de transição - mínimo global"(dendograma B1 e gráfico PCA B2). Os números 1, 2, 3 e 4, correspondem aos perfis energéticos (em função dos critérios apresentados nas tabelas 3.10 e 3.11) dos sistemas 1L2Y, 2MWE, 1WY3 e proteína com quatro hélices-α, respetivamente.

Análises estatísticas das componentes principais (PCA) e utilização do método hierárquico de agrupamento de dados de ligação média permitiram comparar cada um destes sistemas em função dos critérios apresentados nas tabelas 3.10 e 3.11, de modo a se perceber melhor a relação entre seus MFETs e frustrações geométricas. Foram analisadas as conexões "mínimo-TS-mínimo" seguintes: 1) "mínimo modelado - estado de transição - mínimo conectado ao mínimo modelado com o PATHSAMPLE" e; 2) "mínimo modelado - estado de transição - mínimo global".

TABELA 3.10: Valores dos mínimos e estados de transição utilizados para construção do dendograma e gráfico PCA da conexão "mínimo modelado - estado de transição - mínimo conectado ao mínimo modelado com o PATHSAMPLE".

Sistema	**Mínimo modelado**	**TS**	**Mín. conect. PATH.**
Prot. 1L2Y	12.91264763	12.92832566	12.48435444
Prot. 2MWE	20.1947293	20.8127517	18.26207781
Prot. 1WY3	53.70379826	54.09610048	51.46508738
Prot. c/ 4 hélices-α	31.60281206	34.75003792	31.65411212

TABELA 3.11: Valores dos mínimos e estados de transição utilizados para construção do dendograma e gráfico PCA da conexão "mínimo modelado - estado de transição - mínimo global".

Sistema	**Mínimo modelado**	**TS**	**Mínimo global**
Prot. 1L2Y	12.91264763	14.99791665	9.694150715
Prot. 2MWE	20.1947293	20.8127517	12.07354556
Prot. 1WY3	53.70379826	55.4015866	46.67314528
Prot. c/ 4 hélices-α	31.60281206	34.75003792	31.33676974

Para as conexões "mínimo modelado - estado de transição - mínimo conectado ao mínimo modelado com o PATHSAMPLE" foram gerados o dendograma A1 e o gráfico PCA A2 (ver a figura 3.29). Analisando o dendograma A1 e o gráfico PCA A2, em função dos critérios apresentados na tabela 3.10, se observam os seguintes aspetos:

- A proteína com quatro hélices (4)[8] apresenta a barreira energética e MFET mais altos. A barreira energética nesta conexão "mínimo-TS-mínimo" da proteína com quatro hélices-α tem o valor de energia 3.147226. O MFET obtido para 43 otimizações com o GMIN é de 31874.33 s para um desvio padrão alto (21802.17 s) se compararmos com os demais obtidos para os restantes sistemas (ver a tabela 3.9), o qual sugere tempos distintos para se encontrar o mínimo global e, portanto maior frustração estrutural. Analisando o dendograma A1, se verifica que, em função dos critérios estudados, esta proteína corresponde a um grupo isolado, mais próximo dos sistemas 1L2Y e 2MWE. O isolamento do perfil de energia da proteína com quatro hélices-α é também verificado no gráfico PCA A2 para as componentes principais PC1 e PC2 (primeira e segunda componentes principais), no qual é possível visualizar também uma maior proximidade deste aos sistemas 2MWE e 1L2y (correspondentes aos números 2 e 1, respetivamente).

- A proteína 1WY3 (3) apresenta a segunda barreira energética mais alta (com o valor 0.392302) e o segundo MFET mais alto (2.90 s). O dendograma A1 mostra também que este sistema se encontra isolado dos restantes (em função da análise multicriterial feita com método hierárquico). A análise das componentes principais PC1 e PC2 (visualizadas no gráfico PCA A2) enfatizam essa separação do sistema 1WY3 (tendo em conta os critérios de energia analisados) em relação aos restantes, porém, mostram uma maior proximidade deste ao sistema 2MWE.

- A proteína 1L2Y apresenta a menor barreira energética (com o valor 0.015678) e MFET baixo (como pode ser verificado na tabela 3.9). O dendograma A1, bem como o gráfico da análise das componentes principais A2 mostram maior proximidade do perfil de energia desta proteína ao do sistema 2MWE como verificado anteriormente na análise da figura 3.28. No gráfico PCA A2, verifica-se ainda, que o sistema 1L2Y apresenta um perfil de energia que se distancia muito da proteína com quatro hélices-α devido ao baixo valor da variável TS nesta conexão "mínimo-TS-mínimo" (para a conexão "mínimo modelado - estado de transição - mínimo global" deste

[8] Os números 1, 2, 3 e 4 apresentados na figura 3.29 representam os perfis energéticos das proteínas 1L2Y, 2MWE, 1WY2 e proteína com quatro hélices-α, respetivamente, em função dos critérios de energia "mínimo-TS-mínimo" analisados.

sistema TS é bem maior e, se verifica uma aproximação maior - ver o gráfico PCA B2 na figura 3.29).

- A proteína 2MWE também apresenta uma barreira energética baixa (0.618022) e o MFET mais baixo (ver a tabela 3.9).

Para a conexão "mínimo modelado - estado de transição - mínimo global", feita com base nos critérios apresentados na tabela 3.11, obteve-se padrões semelhantes para o dendograma (B1) e para o gráfico PCA (B2) (ver a figura 3.29), para os quais enfatizam-se os seguintes aspetos:

- Na conexão "mínimo modelado - mínimo global" da a proteína com quatro hélices-α TS mantém-se constante e apesar do mínimo global ser ligeiramente mais baixo não há muita alteração na posição do perfil de energia desta apresentado no gráfico PCA B2 comparando com o do PCA A2.

- No sistema 1WY3, verifica-se uma barreira energética maior na conexão "mínimo modelado - mínimo global" (com o valor de energia 1.697788) em relação a "mínimo modelado - mínimo conectado ao mínimo modelado com o PATHSAMPLE" e, portanto maior aproximação do perfil de energia deste ao da proteína com quatro hélices-α (ver o gráfico PCA B na figura 3.29), apesar do mínimo global encontrado ser mais baixo.

- No caso do sistema 1L2Y, a barreira energética que liga o mínimo modelado ao mínimo global é bem maior que a verificada na conexão anterior, a qual apresenta o valor de energia 2.085269, o que o aproximou mais o seu perfil energético ao da proteína com quatro hélices-α (proteína com MFET mais alto), como pode-se verificar no gráfico da análise das componentes principais B2. Este aspeto justifica o facto do MFET deste sistema ser maior que o apresentado pelo 2MWE.

- O perfil de energia da proteína 2MWE apresentado no gráfico PCA B2 mostra mostra um maior afastamento deste em relação ao da proteína de quatro hélices, aspeto verificado pelo facto de TS se manter constante e a diferença de energia entre os mínimos conectados ser maior que a observada na conexão "mínimo modelado - mínimo conectado na primeira otimização com o PATHSAMPLE"; nesta conexão

se verificou uma diferença de energia entre os mínimos conectados de 8.121184 e na anterior de 1.932652. Estes aspetos ajudam a compreender o MFET mais baixo obtido para este sistema.

Os aspetos acima apresentados enfatizam uma separação dos perfis de energia dos sistemas 1WY3 e proteína com quatro hélices-α (3 e 4) em relação aos demais, os quais apresentam os MFETs com maior valor. Por outro lado mostram maior proximidade entre os perfis de energia dos sistemas 1L2Y e 2MWE (1 e 2) que, apresentam valores do MFET próximos e mais baixos. Assim se verificam três grupos de perfis, como apresentado nos dendogramas A1 e B1, um do sistema da proteína com quatro hélices-α, outro do 1WY3 e o dos sistemas 1L2Y e 2MWE. Por análise dos gráficos das componentes principais (A2 e B2) se verificou aproximações dos perfis de energia dos demais sistemas ao da proteína com quatro hélices-α a medida que TS aumenta e distanciamento quando este diminui ou quanto mais baixos foram os mínimos globais. Obviamente mais análises de conexões ("mínimo-TS-mínimo") seriam necessárias para compreender melhor a relação entre os tempos de otimização obtidos para encontrar o mínimo global com a frustração estrutural apresentada por cada sistema.

Os desvios padrão e os MFETs das proteínas 2MWE, 1L2Y e 1WY3 (desvios padrão e MFETs não muito altos - apresentados na tabela 3.9), sugerem menores frustrações estruturais e pouca variação dos caminhos que ligam o mínimo modelado ao mínimo global nestes sistemas. Para a proteína com quatro hélices-α, como referido anteriormente, os valores altos do desvio padrão e do MFET sugerem maiores frustrações estruturais e consequentemente maiores variações dos caminhos que ligam o mínimo modelado ao mínimo global.

Capítulo 4

Conclusões

Apesar dos métodos de modelação dos sistemas estudados (modelação "coarse-grained" com o modelo BLN) corresponderem a uma escala de baixa resolução estrutural, foi possível modelar a estrutura secundária de algumas proteínas simples; as estratégias adotadas para modelar os sistemas analisados permitiram ter boas aproximações estruturais caso compararmos com as semelhanças verificadas na estrutura da proteína L obtida experimentalmente e a modelada com o modelo BLN apresentada por Sorenson e Head-Gordon (Sorenson and Head-Gordon, 2000). Os sistemas modelados com base nas proteínas descritas no PDB apresentaram alguma dificuldade na construção das suas estruturas secundárias, em específico o 1WY3, para o qual se esperava ter um maior afastamento entre as hélices que formam o motivo vilina. Para melhorar a estrutura destes sistemas, sugere-se ter um modelo de potencial BLN com maior número de parâmetros que descrevem os ângulos diedros, para além de H, T e E; ter por exemplo, a hélice G (hélice-3/10), a dobra S (para dobras ligeiras) e parâmetros para as conformações não atribuídas. Por outro lado, outro estudo que pode ser feito para melhorar a caracterização dos sistemas modelados é a utilização do modelo BVLN descrito por Head-Gordon e colaboradores (Yap, Fawzi, and Head-Gordon, 2008), para o qual tem-se mais um tipo de resíduo (resíduo V, utilizado para descrever as atrações hidrofóbicas fracas entre pares de resíduos que não apresentam ligações peptídicas). Contudo a parametrização utilizada na modelação dos sistemas analisados pode servir como referência em estudos futuros.

Segundo Wales (Wales, 2004) os caminhos de conectividade são importantes na determinação da dinâmica de relaxação de um determinado sistema, tal aspeto foi observado nos gráficos de conectividade encontrados para os sistemas de proteínas estudados, onde foi

possível analisar transições energéticas dos sistemas modelados relativamente a vários outros mínimos com os quais se conectam e ter informações sobre a frustração estrutural associada a dobragem destes ao estado nativo.

Em função dos padrões apresentados pelos mínimos nos gráficos de conectividade obtidos (padrões em "folha de palmeira", "folha de salgueiro", "banyan tree" e algumas combinações destes), foi possível ter uma ideia geral da eficiência de relaxação destes ao mínimo global.

Os gráficos de conectividade gerados mostraram que os sistemas estudados apresentam diferentes frustrações estruturais, por exemplo, o sistema da proteína com quatro hélices-α, para além de apresentar a barreira mais alta que conecta o mínimo modelado ao mínimo global, apresenta vários mínimos próximos a este com barreiras energéticas altas. Aspetos semelhantes foram observados para a proteína 1WY3. Por outro lado a proteína 1L2Y também apresenta barreira energética alta na conexão entre o mínimo modelado e o mínimo global. Este sistema apresenta maior predominância do padrão "banyan tree" e, portanto, maior frustração estrutural que o sistema com número de resíduos mais próximo a ele (sistema 2MWE). No caso da proteína 2MWE, constataram-se menores barreiras energéticas na conexão "mínimo modelado"-"mínimo global", apesar de haver vários outros mínimos baixos com barreiras altas, para este sistema houve menos frustração estrutural, o que é comprovado pelo valor mais baixo do MFET obtido.

Para a proteína com quatro hélices-α e a 2MWE (modelo com domínio "ww-domain") se verificaram menos variações das geometrias dos isómeros gerados em relação aos demais sistemas analisados. Este aspeto sugere haver padrões de contacto entre os pares de resíduos destes, próximos aos dos seus estados nativos e pode levar a concluir que para estes sistemas pode haver menores variações nas suas atividades funcionais. Para a proteína com quatro hélices-α se verificou que o mínimo modelado se encontra conectado a vários mínimos de baixa energia separados por barreiras energéticas altas, o que mostra também a maior estabilidade deste em relação aos dos demais sistemas. Por outro lado, para a proteína 2MWE notou-se grande diferença energética entre o mínimo modelado e os mínimos de energia mais baixa, porém, estabilidade do motivo estrutural "ww.domain", pelas estruturas destes serem semelhantes; estes aspetos podem estar relacionados com

facto das fibras amiloides (sistemas responsáveis por doenças como o Alzheimer e constituído por domínios "ww-domain") serem sistemas bastante insolúveis.

Os padrões apresentados nos gráficos de conetividade gerados, estão diretamente relacionados com os tempos médios obtidos para se encontrar o mínimo global em cada sistema estudado (sendo o MFET maior quanto maior a frustração estrutural apresentada), estes aspetos foram comprovados por análises estatísticas com o método hierárquico de análise de dados por ligação média e por análises PCA. Estes métodos mostraram (tendo como critérios a energia dos mínimos conectados e a dos TS" que os ligam) separações dos perfis de energia destes sistemas, os quais se relacionam com os diferentes MFETs obtidos. Alguns dos aspetos observados foram: i) Proximidade dos perfis de energia da proteína 1L2Y e da 2MWE (análise dos dendogramas obtidos e dos gráficos de análise das componentes principais); ii) proximidade do sistema 2MWE ao 1WY3 (análise dos gráficos PCA); iii) perfis de energia da proteína com quatro hélice-α e a 1WY3 mais isolados dos demais (análise dos gráficos PCA); iv) tendência de aproximação dos perfis dos restantes sistemas ao da proteína com quatro hélices-α a medida que a barreira de energia aumenta e, afastamento quando esta diminui.

Os aspetos acima apresentados e os MFETs obtidos mostraram que a frustração energética dos sistemas estudados foi maior para os que possuem maior desvio padrão gerado para os tempos de obtenção do mínimo global. Deste modo a proteína com quatro hélices-α apresenta maior frustração estrutural, seguida da 1WY3, da 1L2Y e da 2MWE. Por outro lado notou-se uma relação entre o tipo de domínio estrutural de cada modelo de proteína analisado com sua frustração estrutural, sendo os sistemas de maior MFET os de geometrias mais complexas.

Porém, para uma melhor caracterização da relação entre os MFETs gerados e as frustrações estruturais apresentadas pelos sistemas estudados, sugere-se que sejam feitas análises PCA para mais conexões "mínimo-TS-mínimo", em específico para o sistema com maior frustração estrutural, para o qual há maior possibilidade do mínimo modelado passar por diferentes caminhos de conectividade para chegar ao mínimo global.

Apêndice A

GMIN, OPTIM, PATHSAMPLE, DisconnectionDPS e outros utilitários: Compilação e noções básicas de utilização

A.1 compilação dos programas utilizados

Para instalação do GMIN, OPTIM, PATHSAMPLE e do disconnectionDPS utilizou-se o *pacote "source"*, versão de 2015, disponível em http://www-wales.ch.cam.ac.uk/software.html. O pacote "source" contém vários aplicativos disponibilizados pelo grupo de David Wales.

Mais informações podem ser encontradas na página de exemplos disponibilizada pelo grupo de David Wales, https://github.com/wales-group/examples.

Por questão de organização, a compilação destes aplicativos foi feita dentro das directorias "GMIN", "OPTIM", "PATHSAMPLE" e "DISCONNECT", disponibilizadas no pacote "source".

Foi utilizado como sistema operativo para a instalação dos programas acima referidos o Linux Fedora.

A.1.1 compilação do GMIN com o compilador gfortran e utilitários rancoords e gminconv2

Para compilar o GMIN com o compilador FORTRAN, seguiu-se os passos seguintes: **Primeiro passo:** Criação da directoria "build" e subdirectoria "gfortran" na directoria "GMIN". Para tal, na directoria "GMIN" já existente, utilizou-se o seguinte comando no "terminal":

mkdir –p build/gfortran

Segundo passo: utilização do comando que permitirá a compilação do "makefile" (ficheiro existente na subdirectoria "source" da directoria "GMIN") na directoria "gfortran" com utilização do *"cmake"* (pacote disponibilizado na "raiz" do pacote "source"):

FC=gfortran cmake ../../source

Nota: O pacote "source" e a directoria "source" são descrições diferentes. Um é o pacote todo do software, com os vários programas e, a directoria "source" é uma subdirectorias da directoria "GMIN". Dentro da directoria "GMIN" há a directoria "bin", a qual não foi utilizada. Estes aspetos foram considerados de forma análoga para a instalação do OPTIM e do PATHSAMPLE.

Terceiro passo: instalação do programa GMIN com utilização do comando *make -jx* no "teminal", onde "x" tem a ver com a capacidade de processamento da CPU (mais informações consultar https://github.com/wales-group/examples):

make –j4

Para a instalação dos utilitários rancoords e gminconv2 foram utilizados os ficheiros rancoords.f e gminconv2.f disponíveis na página de exemplos do grupo de David Wales (https://github.com/wales-group/examples), e compilaram-se os seguintes comandos no "terminal":

a) Compilação do aplicativo rancoords:

gfortran –o rancoords rancoords.f

b) Compilacao do aplicativo gminconv2:

gfortran –f fixed-line-length-132 –o gminconv2 gminconv2.f

Nota: Na compilação dos utilitários rancoords e do gminconv2 é necessário ter em conta o tipo de compilador FORTRAN existente. Por exemplo caso o compilador FORTRAN for o g77 os códigos ("scripts") presentes nos ficheiros rancoords.f e gminconv2.f devem rigorosamente começar na coluna 7 ou superior (como está nos ficheiros disponibilizados), caso contrário os utilitários não são compilados. Por outro lado, por exemplo, a

linha de código representada abaixo (presente no ficheiro rancoords.f) dá erro de compilação caso o asterisco estiver numa posição igual ou superior a da coluna 7, nesse caso deve-se ter o asterisco na coluna 6 para compilar o utilitário:
**NTB=32, ... (Na corrente instalação esta linha de código corresponde a 42 do ficheiro rancoords.f).*

Para o ficheiro gminconv2.f (ficheiro utilizado na instalação do utilitário gminconv2) deve-se evitar ter os números que referem-se a continuação das linhas de código na coluna 7 ou posição superior, caso contrário o compilador não compilará o utilitário.

A.1.2 Instalação do OPTIM com o compilador gfortran

Para instalação do programa OPTIM teve-se em conta os passos seguintes:

Primeiro passo: Criação da directoria "build" e subdirectoria "gfortran" na directoria "OPTIM" de forma análoga ao referido na instalação anterior. Para tal, é utilizado dentro da directoria "OPTIM" o seguinte comando:

"mkdir –p build/gfortran"

Segundo passo: compilação do ficheiro "makefile" existente na subdirectoria "source" (sudirectoria da directoria "OPTIM") na directoria "gfortran" com a utilização do comando:

FC=gfortran cmake ../../source

Terceiro passo: compilação compilação do programa OPTIM com o comando:

make –j4

A.1.3 Instalação do PATHSAMPLE com o compilador gfortran

Na instalação do PATHSAMPLE, de forma análoga aso programas anteriores, seguiu-se os passos seguintes:

Primeiro passo: Criação da directoria "build" e subdirectoria "gfortran" na directoria "PATHSAMPLE". Uma vez na directoria "PATHSAMPLE" utilizou-se no "terminal" o seguinte comando:

"mkdir –p build/gfortran"

Segundo passo: utilização do comando para a compilação do "makefile" (ficheiro presente na subdirectoria "source" da directoria "PATHSAMPLE") na directoria "gfortran" com utilização do pacote "cmake":

FC=gfortran cmake ../../source

Terceiro passo: compilação do software:

make –j4

A.1.4 Instalação do disconnectionDPS

Para a instalação do programa disconnectionDPS seguiu-se os passos seguintes:

Primeiro passo: Ir a subdirectoria "source" dentro da directoria "DISCONNECT":

cd DISCONNECT/source

Segundo passo: Instalar o programa disconnectionDPS:

gfortran –o disconnectionDPS disconnectionDPS.f90

A.2 Noções básicas de utilização dos programas GMIN, OPTIM, PATHSAMPLE e disconnectionDPS

A.2.1 Utilização dos programas GMIN, OPTIM, PATHSAMPLE e disconnectionDPS em Modelos BLN

Para correr o GMIN utilizando a abordagem "coarse-grained" (modelo BLN), foram necessários os seguintes ficheiros de "input" (ficheiros de entrada) (Wales, 2016a): a) "BLNSequence" (ficheiro que contém a sequência de resíduos BLN, a estrutura secundária

e os parâmetros para o potencial de torção dos ângulos diedros e de interações entre pares de resíduos não ligados por ligações peptídicas); b) "coords" (ficheiro com as coordenadas iniciais do sistema a estudar); c) data (ficheiro com várias chaves utilizadas pelo programa GMIN); d) "Contactmap" (Ficheiro com os pares de resíduos que entram em contacto nas ligações nativas; só é utilizado para o modelo de Go); e) o executável do GMIN (executado no terminal com o comando "./GMIN"). No caso de cálculos longos é sugerível correr o GMIN ou os outros executáveis anteriormente referidos em "background". Para correr o GMIN em "background" basta utilizar o comando "nohup ./GMIN&" no terminal. Para os outros executáveis (OPTIM, PATHSAMPLE), utiliza-se de forma análoga o comando "nohup ./EXECUTÁVEL&". Para anular um processo que esteja a correr em "background" basta aceder aos processos em "background" com o comando "top", verificar o número do processo (PID), sair dos processos em "background" com o comando "q" e executar o comando "kill -9 PID" no terminal, onde PID é o número do processo.

Para executar o OPTIM é necessário apenas o ficheiro "odata" e o executável OPTIM (Wales, 2016b), porem em certos casos pode-se utilizar outros ficheiros em função do que se pretende calcular. Por exemplo em situações em que se pretende estudar os caminhos de conectividade entre dois mínimos, é necessário definir as coordenadas destes (mínimo inicial (reagente) e mínimo final (produto)). Nesses casos, para além do ficheiro "odata" pode ser utilizado o ficheiro "finish" que terá as coordenadas do produto. As coordenadas do reagente são inseridas diretamente no ficheiro "odata" logo após a chave "POINTS" - para mais detalhes consultar a documentação do OPTIM (Wales, 2016b). As coordenadas do reagente e produtos[1] colocadas nos ficheiros "odata" e "finish" respectivamente podem ser calculadas em otimizações feitas com o GMIN, sendo geradas no ficheiro "lowest", ficheiro de "output". O ficheiro "odata" é um ficheiro que utiliza várias chaves utilizadas pelo programa OPTIM, as quais podem ser manipuladas em função dos estudos a serem realizados. Este ficheiro é semelhante ao ficheiro "data" utilizado pelo GMIN.

No caso de estudos de otimização com o programa OPTIM com a utilização da abordagem "coarse-graned", tal como no programa GMIN, é necessário o ficheiro "BLNsequence" como ficheiro auxiliar. Tanto no ficheiro "data" (ficheiro de "input" do programa

[1] Coordenadas do mínimo inicial e do final no caso do nosso estudo, por terem sido analisadas isomerizações e não propriamente reações.

GMIN) como no "odata" é inserida a chave "BLN *rkr rkt*", onde *rkr* e *rkt*, são parâmetros referentes as constantes de restrições de ligação entre dois resíduos ligados por ligação peptídica e ângulos de ligação para cada grupo de três resíduos, respetivamente (Wales, 2016b). No caso de estudos de modelos de Go, utiliza-se a chave "BLNGO k_r k_θ λ" (nos ficheiros "data" e "odata"), onde k_r refere-se a constante de restrições de ligação, k_θ a constante de restrições dos ângulos de ligação, e λ é um parâmetro opcional utilizado para manipular a força das interações não nativas (Wales, 2016b; Oakley and Roy L. Johnston, 2012). No estudo de "Go-models" é ainda necessário o ficehiro "contactmap" como ficheiro auxiliar, tal como nas otimizações feitas com o GMIN (Wales, 2016a). O "contactmap" é um ficheiro que contém pares de resíduos em contacto no estado nativo da proteína (Wales, 2016a; Wales, 2016b; Jr., 2014).

O PATHSAMPLE é nada mais que um utilitário do OPTIM, ou seja, precisa do OPTIM a correr em "background" para poder funcionar. Alguns dos ficheiros necessários para correr o PATHSAMPLE são (Wales, 2016c): a) "pathdata": ficheiro semelhante ao ficheiro "data" e "odata", com várias chaves (parâmetros) para manipular as amostras de pontos estacionários. b) Ficheiros "odata.*": o PATHSAMPLE utiliza vários ficheiros "odata" que "correm" com o OPTIM em "background". Os ficheiros "odata" podem ser utilizados com diferentes extensões em função dos cálculos que se pretendem efetuar ("odata.connect", ficheiro utilizado para estudar caminhos de conectividade entre mínimos, ou seja, pode ser utilizado para encontrar estados de transição ligados a dois mínimos; odata.path, ficheiro utilizado para estudar caminhos de conectividade entre dois mínimos partindo de um estado de transição; odata.tssearsh, realiza pesquisas partindo de um mínimo para determinar o estado de transição a ele ligado). c) "min.A", "min.B" e o "min.data": Apresentam os mínimos correspondentes aos reagentes (mínimos presentes no ficheiro "min.A"), os mínimos correspondentes aos produtos (mínimos presentes no ficheiro min.B) e o conjunto de todos os mínimos gerados no processo de conectividade entre os reagentes e o produtos (mínimos presentes no ficheiro min.data) [2]. min.A corresponde aos mínimos dos reagentes se a chave "DIRECTION" no ficheiro "pathdata" estiver definida para a direção AB, caso contrário corresponderá aos mínimos dos produtos.

[2] No ficheiro min.data são gerados os mínimos presentes nos ficheiros "min.A" e "min.B" e todos os mínimos gerados na conexão destes.

Os ficheiros "min.A" e "min.B" apresentam informações referentes ao número de mínimos existentes nos reagentes ou produtos, ou seja, o número de mínimos a serem conectados e, o número da linha ou linhas em que esses se encontram no ficheiro min.data. É necessário que os mínimos (ou mínimo) em min.A e em min.B estejam bem especificados relativamente ao ficheiro min.data. d) ts.data: Ficheiro que contém os estados de transição que conectam os mínimos gerados no ficheiro "min.data". Cada energia de estado de transição é apresentada em uma linha, o que também é verificado de forma análoga no ficheiro "min.data" (numerando as linhas de cima para baixo seguindo a ordem dos números naturais). e) "points.min": ficheiro onde são geradas as coordenadas cartesianas dos mínimos presentes no ficheiro "min.data". f) points.ts: ficheiro onde são geradas as coordenadas cartesianas dos estados de transição presentes no ficheiro "ts.data". Os ficheiros min.data e ts.data são ficheiros de "output" do PATHSAMPLE utilizados pelo programa disconnectionDPS para criar os gráficos de conectividade; g) "path.info.startup": Os ficheiros min.A, min.B e os ficheiros min.data, ts.data, points.min e points.ts (sendo estes quatro referidos como base de dados de pontos estacionários do PATHSAMPLE[3]) podem ser calculados pelo PATHSAMPLE com a utilização do ficheiro "path.info". O ficheiro "path.info" é um ficheiro de "output" gerado no processo de otimização com o OPTIM, o qual contém informação de conexão dos mínimos que se pretende conectar (conexões mínimo-TS-mínimo). No processo de otimização com o PATHSAMPLE pode-se adicionar ao ficheiro "path.info" a extensão ".startup" e referenciar-se o ficheiro "path.info" com a nova extensão (por exemplo: "path.info.startup") como parâmetro para a chave STARTFROMPATH no ficheiro "pathdata", podendo as linhas que especificam a posição dos mínimos conectados no ficheiro min.data serem referenciadas como números aleatórios (exemplo 1 e 2) logo a seguir na mesma chave. As posições das linhas que contêm os mínimos conectados nos ficheiros min.A e min.B podem depois ser editadas em função das posições dos mínimos no ficheiro min.data gerado no processo de otimização com a chaves STARTFROMPATH. Pode ser utilizado o ficheiro de "output" "optim.out" gerado pelo OPTIM para editar os números das linhas dos mínimos conectados para as posições corretas nos ficheiros "min.A" e "min.B".

[3]Convém fazer um backup da base de dados de pontos estacionários do PATHSAMPLE, uma vez que ela pode se atualizar em função dos cálculos feitos.

Os "outputs" do PATHSAMPLE podem variar em função dos cálculos que se pretende fazer, tal como é o caso dos outputs do GMIN e do OPTIM, nesse sentido são feitas manipulações dos ficheiros que funcionam com várias chaves (ficheiro "data", "odata", e ficheiro "pathdada"). É necessário ter em atenção que para os cálculos de otimização utilizando o PATHSAMPLE é necessário ter em conta dois ficheiros controlados por chaves, ficheiros pathdata e o ficheiro odata.*, os quais são importantes para definir o tipo de cálculo e output que se espera do PATHSAMPLE.

O programa disconnectionDPS foi utilizado para visualizar ficheiros de output gerados pelos PATHSAMPLE. O DisconnectionDPS precisa do ficheiro min.data, do ficheiro ts.da (ficheiros da base de dados de pontos estacionários do PATHSAMPLE) e do ficheiro dinfo para poder correr. O ficheiro dinfo tal como o data, odata e o pathdata anteriormente referidos, é um ficheiro que funciona com chaves, as quais podem ser manipuladas em função dos "outputs" que se esperam para os gráficos de conectividade. Uma vez executado o disconnectionDPS geram-se ficheiros de "outputs" como o ficheiro tree.ps, que contém o gráfico de disconectividade e o ficheiro "disconnection", ficheiro com os mínimos que não estabelecem conexão.

A.2.2 Exemplos de ficheiros de "input" utilizados e "output"" obtidos no processo de caracterização dos sistemas de proteínas estudados

Neste ponto é tomada como exemplo a proteína de quatro hélices$-\alpha$ para apresentar alguns ficheiros de "input" utilizados e "output" gerados nos processos de caracterização deste sistema, os quais podem ser observados nas figuras A.1-A.10.

BLNsequence

```
comment: $S_{12}>0$ and $S_6<0$ for B-B, L-L and L-B, N-L and N-B and N-N
1.7D0 -1.7D0
1.7D0 0.0D0
1.7D0 0.0D0
comment: coefficients A, B, C, D
comment: for Helical, Extended and Turn residues in order, four per line
0.0D0 1.6D0 1.6D0 1.6D0
0.0D0 1.6D0 1.6D0 1.6D0
0.0D0 0.0D0 0.2D0 0.0D0
LLBLLBBLLBLLBBLLNNNLLBLLBBLLBLLBBLLNNNLLBLLBBLLBLLBBLLNNNLLBLLBBLLBLLBBLL
HHHHHHHHHHHHHHHTTTTTTTHHHHHHHHHHHHHTTTTTTTHHHHHHHHHHHHHTTTTTTTHHHHHHHHHH
```

FIGURA A.1: Ficheiro BLNsequence utilizando na modelação da proteína de quatro hélices$-\alpha$.

data

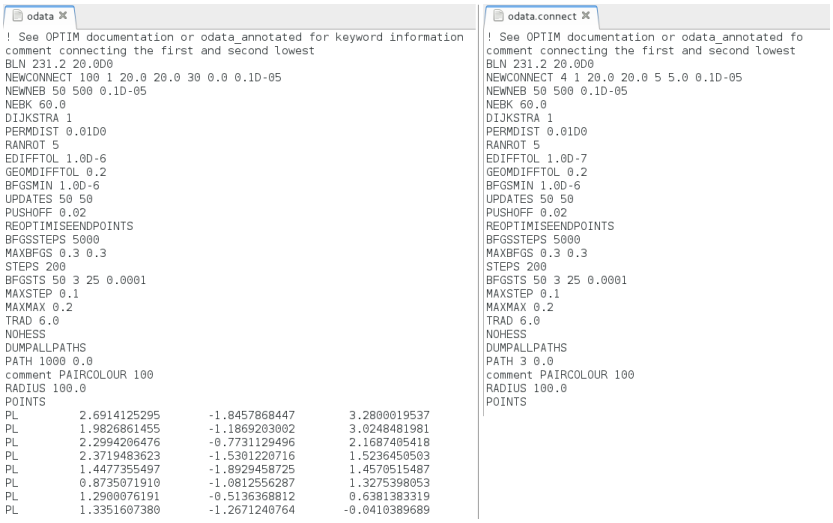
```
comment SAVE (10)
BLN 231.2 20.0D0
TIGHTCONV  0.0000001
UPDATES 50
DGUESS  10.0
RADIUS 100.0
SAVE 10
CENTRE
EDIFF 0.001
STEPS 2000000 1.0
MAXBFGS 0.3
MAXIT 10000 10000
TEMPERATURE 2.3
NEWRESTART 3500
AVOID 11.5 20 F
STEP 0.65
SLOPPYCONV 0.0075
RANSEED 525
```

FIGURA A.2: Ficheiro data utilizando na modelação da proteína de quatro hélices$-\alpha$.

lowest

```
         73
Energy of minimum      1=     31.60281206     first found at step  1506928 after           248405194 function calls
LL         2.6914125295       -1.8457868447        3.2800019537
LL         1.9826861455       -1.1869203002        3.0248481981
BL         2.2994206476       -0.7731129496        2.1687405418
LL         2.3719483623       -1.5301220716        1.5236450503
LL         1.4477355497       -1.8929458725        1.4570515487
BL         0.8735071910       -1.0812556287        1.3275398053
BL         1.2900076191       -0.5136368812        0.6381383319
LL         1.3351607380       -1.2671240764       -0.0410389689
LL         0.3722886758       -1.4934128445       -0.1341360875
BL         0.0393300498       -0.5846249584       -0.3966011895
LL         0.5403875145       -0.5784386725       -1.2769084250
LL         0.1101334153       -1.2791612693       -1.8419131984
BL        -0.8608701586       -1.0059201961       -1.8653117064
BL        -0.9478873862       -0.0855039348       -2.2278244589
LL        -0.5370025647       -0.1258233001       -3.1414828016
LL        -1.0880278157       -0.7783680211       -3.6619520950
NL        -2.0026461502       -0.3736870879       -3.7184472234
NL        -2.5841085169       -1.1130427636       -4.0571104253
NL        -2.8859131488       -1.6188432386       -3.2457925640
LL        -2.1778355818       -1.5098529591       -2.5437038602
LL        -2.2187796241       -0.5451034139       -2.2689077641
BL        -1.7349894357       -0.4780861988       -1.3828372059
LL        -2.2346541079       -1.2154762230       -0.9047081321
LL        -1.5089352018       -1.8707532211       -0.8270381132
BL        -0.8141393231       -1.1621607483       -0.7852829595
BL        -1.1024331723       -0.4692603035       -0.1579836925
LL        -1.2313157743       -0.9233578995        0.7347048680
LL        -0.2991246243       -1.2093757826        0.8711047123
BL         0.1848703257       -0.3240072259        0.7651101202
LL        -0.0594184564        0.1686008201        1.6142009913
LL         0.4929255475       -0.3101170689        2.2742479540
```

FIGURA A.3: Parte do ficheiro lowest gerado na otimização global da proteína de quatro hélices$-\alpha$.



odata

```
! See OPTIM documentation or odata_annotated for keyword information
comment connecting the first and second lowest
BLN 231.2 20.0D0
NEWCONNECT 100 1 20.0 20.0 30 0.0 0.1D-05
NEWNEB 50 500 0.1D-05
NEBK 60.0
DIJKSTRA 1
PERMDIST 0.01D0
RANROT 5
EDIFFTOL 1.0D-6
GEOMDIFFTOL 0.2
BFGSMIN 1.0D-6
UPDATES 50 50
PUSHOFF 0.02
REOPTIMISEENDPOINTS
BFGSSTEPS 5000
MAXBFGS 0.3 0.3
STEPS 200
BFGSTS 50 3 25 0.0001
MAXSTEP 0.1
MAXMAX 0.2
TRAD 6.0
NOHESS
DUMPALLPATHS
PATH 1000 0.0
comment PAIRCOLOUR 100
RADIUS 100.0
POINTS
PL        2.6914125295      -1.8457868447       3.2800019537
PL        1.9826861455      -1.1869203002       3.0248481981
PL        2.2994206476      -0.7731129496       2.1687405418
PL        2.3719483623      -1.5301220716       1.5236450503
PL        1.4477355497      -1.8929458725       1.4570515487
PL        0.8735071910      -1.0812556287       1.3275398053
PL        1.2900076191      -0.5136368812       0.6381383319
PL        1.3351607380      -1.2671240764      -0.0410389689
```

odata.connect

```
! See OPTIM documentation or odata_annotated fo
comment connecting the first and second lowest
BLN 231.2 20.0D0
NEWCONNECT 4 1 20.0 20.0 5 5.0 0.1D-05
NEWNEB 50 500 0.1D-05
NEBK 60.0
DIJKSTRA 1
PERMDIST 0.01D0
RANROT 5
EDIFFTOL 1.0D-7
GEOMDIFFTOL 0.2
BFGSMIN 1.0D-6
UPDATES 50 50
PUSHOFF 0.02
REOPTIMISEENDPOINTS
BFGSSTEPS 5000
MAXBFGS 0.3 0.3
STEPS 200
BFGSTS 50 3 25 0.0001
MAXSTEP 0.1
MAXMAX 0.2
TRAD 6.0
NOHESS
DUMPALLPATHS
PATH 3 0.0
comment PAIRCOLOUR 100
RADIUS 100.0
POINTS
```

FIGURA A.4: Ficheiro odata (a esquerda, figura com parte das coordenadas do mínimo modelado) e odata.connect (a direita) utilizados nas otimizações da proteína de quatro hélices$-\alpha$ feitas com o OPTIM e PATHSAMPLE, respetivamente. No ficheiro odata.connect as coordenadas por baixo dos vários parâmetros são retiradas.

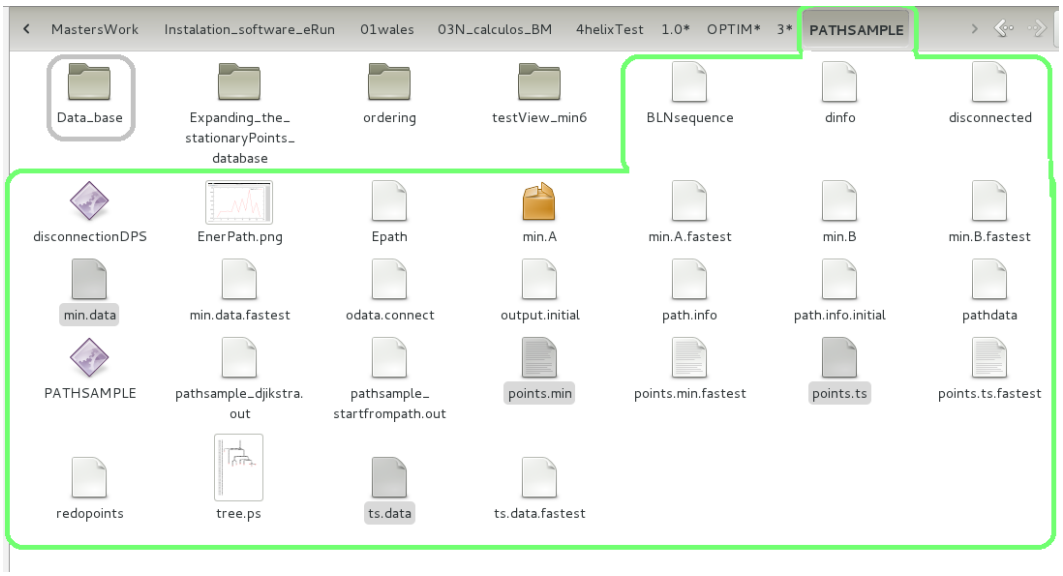
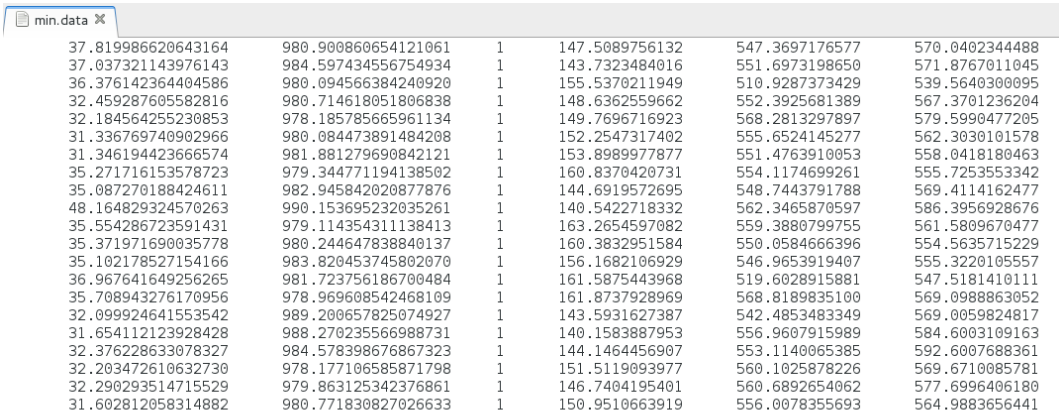


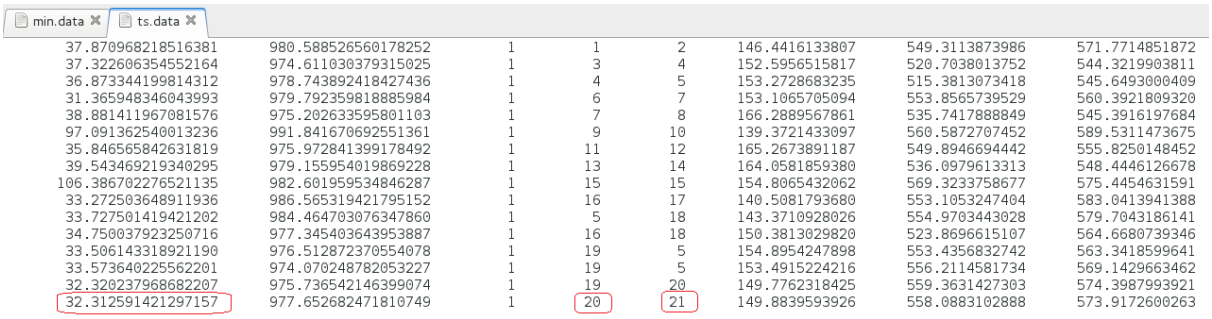
FIGURA A.5: "Output" gerado pelo PATHSAMPLE no processo de criação da base de dados de pontos estacionários. Os ficheiros guardados na base de dados (na directoria do canto superior esquerdo) estão marcados em tom cinza (ficheiro min.data, ts.data, points.min e points.ts).



min.data

37.819986620643164	980.900860654121061	1	147.5089756132	547.3697176577	570.0402344488
37.037321143976143	984.597434556754934	1	143.7323484016	551.6973198650	571.8767011045
36.376142364404586	980.094566384240920	1	155.5370211949	510.9287373429	539.5640300095
32.459287605582816	980.714618051806838	1	148.6362559662	552.3925681389	567.3701236204
32.184564255230853	978.185785665961134	1	149.7696716923	568.2813297897	579.5990477205
31.336769740902966	980.084473891484208	1	152.2547317402	555.6524145277	562.3030101578
31.346194423666574	981.881279690842121	1	153.8989977877	551.4763910053	558.0418180463
35.271716153578723	979.344771194138502	1	160.8370420731	554.1174699261	555.7253553342
35.087270188424611	982.945842020877876	1	144.6919572695	548.7443791788	569.4114162477
48.164829324570263	990.153695232035261	1	140.5422718332	562.3465870597	586.3956928676
35.554286723591431	979.114354311138413	1	163.2654597082	559.3880799755	561.5809670477
35.371971690035778	980.244647838840137	1	160.3832951584	550.0584666396	554.5635715229
35.102178527154166	983.820453745802070	1	156.1682106929	546.9653919407	555.3220105557
36.967641649256265	981.723756186700484	1	161.5875443968	519.6028915881	547.5181410111
35.708943276170956	978.969608542468109	1	161.8737928969	568.8189835100	569.0988863052
32.099924641553542	989.200657825074927	1	143.5931627387	542.4853483349	569.0059824817
31.654112123928428	988.270235566988731	1	140.1583887953	556.9607915989	584.6003109163
32.376228633078327	984.578398676867323	1	144.1464456907	553.1140065385	592.6007688361
32.203472610632730	978.177106585871798	1	151.5119093977	560.1025878226	569.6710085781
32.290293514715529	979.863125342376861	1	146.7404195401	560.6892654062	577.6996406180
31.602812058314882	980.771830827026633	1	150.9510663919	556.0078355693	564.9883656441

FIGURA A.6: Ficheiro min.data gerado na modelação da proteína de quatro hélices$-\alpha$.



min.data | ts.data

37.870968218516381	980.588526560178252	1	1	2	146.4416133807	549.3113873986	571.7714851872
37.322606354552164	974.611030379315025	1	3	4	152.5956515817	520.7038013752	544.3219903811
36.873344199814312	978.743892418427436	1	4	5	153.2728683235	515.3813073418	545.6493000409
31.365948346043993	979.792359818885984	1	6	7	153.1065705094	553.8565739529	560.3921809320
38.881411967081576	975.202633595801103	1	7	8	166.2889567861	535.7417888849	545.3916197684
97.091362540013236	991.841670692551361	1	9	10	139.3721433097	560.5872707452	589.5311473675
35.846565842631819	975.972841399178492	1	11	12	165.2673891187	549.8946694442	555.8250148452
39.543469219340295	979.155954019869228	1	13	14	164.0581859380	536.0979613313	548.4446126678
106.386702276521135	982.601959534846287	1	15	15	154.8065432062	569.3233758677	575.4454631591
33.272503648911936	986.565319421795152	1	16	17	140.5081793680	553.1053247404	583.0413941388
33.727501419421202	984.464703076347860	1	5	18	143.3710928026	554.9703443028	579.7043186141
34.750037923250716	977.345403643953887	1	16	18	150.3813029820	523.8696615107	564.6680739346
33.506143318921190	976.512872370554078	1	19	5	154.8954247898	553.4356832742	563.3418599641
33.573640225562201	974.070248782053227	1	19	5	153.4915224216	556.2114581734	569.1429663462
32.320237968682207	975.736542146399074	1	19	20	149.7762318425	559.3631427303	574.3987993921
32.312591421297157	977.652682471810749	1	20	21	149.8839593926	558.0883102888	573.9172600263

FIGURA A.7: Ficheiro ts.data gerado na modelação da proteína de quatro hélices$-\alpha$. A energia do estado de transição que conecta o mínimo 20 ao modelado (mínimo 21) está marcada a vermelho na última linha da primeira coluna.

dinfo

```
! disconnectionDPS input to generate
! For further details see dinfo_annot
! REQUIRED KEYWORDS
DELTA 0.2
FIRST 40.0
LEVELS 90
MINIMA min.data
TS ts.data
! OPTIONAL KEYWORDS
MAXTSENERGY 40.0
NCONNMIN 0
CENTREGMIN
CONNECTMIN 21
IDMIN 21
IDMIN 17
IDMIN 6
IDMIN 5
IDMIN 4
IDMIN 3
IDMIN 19
IDMIN 20
IDMIN 18
IDMIN 16
! IDENTIFY
! LABELSIZE 5
```

FIGURA A.8: Ficheiro dinfo utilizado para gerar o gráfico de conectividade da proteína de quatro hélices$-\alpha$.

pathdata

```
! PATHSAMPLE input to create an initial database from OPTIM path.info
! For further details, see the PATHSAMPLE documentation
comment using PERMDIST 0.001
EXEC /home/wasina/Documents/MastersWork/Needed/OPTIM
CPUS 1
SYSTEM PL
NATOMS 73
SEED 1
DIRECTION AB
CONNECTIONS 1
TEMPERATURE 2.3
PERMDIST 0.01D0
RANROT 5
ETOL 1.0D-6
GEOMDIFFTOL 0.2
ITOL 0.1D0

! STEP 1: creating the initial database from OPTIM path.info file
!STARTFROMPATH path.info.initial 1 2
!CYCLES 0
! STEP 2: run a Dijkstra analysis to identify the 'fastest path'
DIJKSTRA 0
CYCLES 0
```

pathdata

```
! PATHSAMPLE input to create an initial database from OPT
! For further details, see the PATHSAMPLE documentation
comment using PERMDIST 0.001
EXEC /home/wasina/Documents/MastersWork/Needed/OPTIM
CPUS 1
SYSTEM PL
NATOMS 73
SEED 1
DIRECTION AB
CONNECTIONS 1
TEMPERATURE 2.3
PERMDIST 0.01D0
RANROT 5
ETOL 1.0D-7
GEOMDIFFTOL 0.2
ITOL 0.1D0

UNTRAP 4.0 3.0
CYCLES 50
```

FIGURA A.9: Ficheiros pathdata utilizados na criação e expansão da base de dados de pontos estacionários do PATHSAMPLE. Para a expansão da base de dados foi utilizada a chaves UNTRAP.

```
#!/bin/csh
# script to collect mean first encounter time
stats for 4helixbundle73 by running 100
# GMIN runs from random initial configurations
#
# Usage: ./100_4helixbundle73_runs.csh withSYM 42
# check for correct number of arguments
set expected_args=2
if ($#argv != $expected_args) then
echo "ERROR: Missing arguements:"
echo "./100_4helixbundle73_runs.csh <working
directory name> <seed>"
exit
endif
# specify system size and container (to prevent
evaporation) radius
set radius=3.0
set natoms=73
# specify the binary to use
# if GMIN is already in your $PATH
set exec=GMIN
# to specify a specific binary
# set exec=~/workshop/binaries/GMIN
# the first argument specifies the working
directory, the second the seed to start at
set directory=$1
set seed=$2
# create the working directory and copy over the
input data file
mkdir $directory
cd $directory
cp ../data data
cp ../BLNsequence BLNsequence
# remove any old output and supress the output
rm hits >& /dev/null
# initialise the counter variable
set count=1
echo
echo "Running GMIN for 73 4helixbundle residues
from 100 random starting geometries"
echo "Random number seeds start from $2"
echo
# loop to start 100 GMIN jobs
while ($count <= 100)
# use rancoords to generate initial coordinates
echo $natoms $radius -$seed > randata
../rancoords # >& /dev/null
cp newcoords coords
# start GMIN
echo "Starting run $count"
../$exec output >& extra_out.$count
# extract the number of quenches, number of
potential calls and time taken from the output
echo `grep hit output | head -1 | sed -e 's/[a-zA-
Z]//g' -e 's/[a-zA-Z]//g' -e 's/\.//' -e 's/>//'`
\
`grep time= output | tail -1 | sed 's/.*time=//'`
>> hits
# back up initial coordinates and output
mv coords coords.$count
mv output $directory.output.$natoms.$count
# increment counter variable and seed
@ count +=1
@ seed +=1
end
# construct a smooth distribution for the time
taken (t), number of quenches (Q) and number
# of potential calls (V) and return the mean and
standard deviation for each
../gminconv2 < hits > temp ; head -1 temp > pdf
rm temp
# print the results
echo
echo "Mean and standard deviation for global
minimum first encounter time:"
echo
cat pdf
```

FIGURA A.10: Ficheiro utilitário do GMIN utilizado para calcular o MFET da proteína com quatro hélices$-\alpha$.

Bibliografia

Anderson, Claus A. and Burkhard Rost (2009). "Secondary Structure Assinment". In: *Structural Bioinformatics, Second Edition*.

Branden, C. and J. Tooze (1999). "Introduction to protein structure". In: *Garland publishing Inc. 19 Union Square West, New York, NY, 10003-3382*.

Brasiello, A. et al. (2010). "Multi-Scale Modelling And Coarse-Grained Analysis Of Triglycerides Dynamics". In: *20th European Symposium on Computer Aided Process Engineering – ESCAPE20. S. Pierucci and G. Buzzi Ferraris (Editors)*.

Brown, S., N. J. Fawzi, and T. Head-Gordon (2003). "Coarse-grained sequences for protein folding and design". In: *PNAS* 100.19, pp. 10712–10717.

Chiang, Yih-Shien et al. (2007). "New classification of supersecondary structures of sandwich-like proteins uncovers strict patterns of strand assemblage". In: *Wiley InterScience*. URL: DOI:10.1002/prot.21473.

Chou, Peter Y. and Gerald D. Fasman (1974a). "Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins". In: *Biochemistry*, 211–222.

– (1974b). "Prediction of protein conformation". In: *Biochemistry*, 222–245.

Ding, Feng, Sergey V. Buldyrev, and Nikolay V. Dokholyan (2005). "Folding Trp-Cage to NMR Resolution Native Structure Using a Coarse-Grained Protein Model". In: *Biophysical Journal* 88, 147–155.

Friederich, Evelyne et al. (1999). "Villin Function in the Organization of the Actin Cytoskeleton". In: *The Jornal of Biological Chemistry* 274.38, 26751–26760.

Guo, Z. and D. Thirumalai (1995). "Kinetics of protein folding: Nucleation mechanism, time scales, and pathways". In: *Biopolymers* 36, pp. 83–102.

– (1996). "Kinetics and thermodynamics of folding of a novo designed four-helix bundle protein". In: *J. Mol. Biol.* 263, pp. 323–343.

Hoffmann, Falk et al. (2014). "Protein Structure Prediction: Assembly of Secundary Structure Elements by Basin-Hopping". In: *ChemPhysChem* 15, pp. 3378–3390.

Honeycutt, J. D. and D. Thirumalai (1992). "The nature of folded states of globular proteins". In: *Biopolymers* 32, pp. 695–709.

Jr., Ronald D. Hills (2014). "Balancing bond, nonbond, and go-like terms in coarse-grain simulations of conformational dynamics". In: *Dennis R. Livesay (ed.), Protein Dynamics: Methods and Protocols, Methods in Molecular Biology* 1084.

Klahre, Ulrich et al. (2000). "Villin-Like Actin-Binding Proteins Are Expressed Ubiquitously in Arabidopsis". In: *Plant Physiology* 122, 35–47.

Lee, Jinwoo et al. (2008). "Re-examination of structures optimization of off-lattice protein AB models by conformational space anneling". In: *J Comput Chem*, pp. 2479–2484.

Miller, M. A. and D. J. Wales (1999). "Energy landscape of a model protein". In: *Journal of chemical physics* 111.14.

Neidigh, Jonathan W., R. Matthew Fesinmeyer, and Niels H. Andersen (2002). "Designing a 20-residue protein". In: *Nature Strutural Biology* 9.6.

Noid, W. G. (2013). "Perspective: Coarse-grained models for biomolecular systems". In: *J. Chem. Phys* 139. URL: http://dx.doi.org/10.1063/1.4818908.

Oakley, M. T., D. J. Wales, and R. L. Johnston (2011). "Energy landscape and global optimization for a frustrated model protein". In: *J. Phys. Chem. B* 115, pp. 11525–11529.

Oakley, M. T. et al. (2013). "The Effect of Nonnative Interactions on the Energy Landscapes of Frustrated Model Proteins". In: *IEEE/ACM Transations on Computational Biology and Bioinformatics* 10.6.

Oakley, Mark T. and David J. Wales a Roy L. Johnston (2012). "The effect of nonnative interations on the energy landscapes of frustrated model proteins". In: *Journal of atomic, molecular, and optical physics* 2012.

Richardson, Jane S. (1981). "The Anatomy and Taxonomy of Protein Structure". In: *Advances In Protein Chemistry* 34.

Rudzinski, J. F. and W. G. Noid (2015). "Bottom-Up Coarse-Graining of Peptide Ensembles and Helix-Coil Transitions". In: *J. Chem. Theory Comput.* 11, pp. 1278–1291.

Serpell, Louise C. (2000). "Alzheimer's amyloid fibrils: structure and assembly". In: *Biochimica et Biophysica Acta* 1502, pp. 16–30.

Sharma, Shantanu, Feng Ding, and Nikolay V. Dokholyan (2008). "Probing protein aggregation using discrete molecular dynamics". In: *Frontiers in Bioscience*.

Sorenson, J. M. and T. Head-Gordon (2002a). "Protein engineering study of protein L by simulation". In: *Jornal of Computational Biology* 9.1, pp. 35–54.

– (2002b). "Toward minimalist models of larger proteins: a ubiquitin-like protein". In: *PROTEINS: Structure, Functions and Genetics* 46, pp. 368–379.

Sorenson, Jon M. and Teresa Head-Gordon (2000). "Matching simulation and experiment: A new simplified model for simulating protein folding". In: *Journal of computational biology* 7.3/4, pp. 469–481.

Taketomi, Hiroshi, Yuzo Ueda, and Nobuhiro Go (1975). "STUDIES ON PROTEIN FOLDING, UNFOLDING AND FLUCTUATIONS BY COMPUTER SIMULATION". In: *Int. J. Peptide Protein Res.* 7, pp. 445–459.

Tozzini, Valentina (2005). "Coarse-grained models for proteins". In: *Current Opinion in Structural Biology* 15.

Wales, D. J. (2004). "Energy Landscapes Applications to Clusters, Biomolecules and Glasses". In: *Cambridge Molecular Science*.

– (2014). "GMIN User Guide. Updeted in March 14, 2014". In: URL: http://www-wales.ch.cam.ac.uk/software.html.

– (2016a). "GMIN User Guide. Updeted in February 19, 2016". In: URL: http://www-wales.ch.cam.ac.uk/software.html.

– (2016b). "OPTIM User Guide. Updeted in February 19, 2016". In: URL: http://www-wales.ch.cam.ac.uk/software.html.

– (2016c). "PATHSAMPLE User Guide. Updeted in February 19, 2016". In: URL: http://www-wales.ch.cam.ac.uk/software.html.

Wales, David J. and Peter E. J. Dewsbury (2004). "Effect of salt bridges on the energy landscape of a model protein". In: *Journal of Chemical Physics* 121.20.

Wales, David J. and Jonathan P. K. Doye (1997). "Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms". In: *J. Phys. Chem. A* 101, pp. 5111–5116.

Wales, David J. and Teresa Head-Gordon (2012). "Evolution of the potential energy landscape with static pulling force for two model proteins". In: *J. Phys. Chem. B* 116, pp. 8394–8411.

Weiner, S. J. et al. (1986). "An All Atom Force Field for Simulations of Proteins and Nucleic Acids". In: *Journal of Computational Chemistry* 7.2, pp. 230–252.

Yap, Eng-Hui, Nicolas Lux Fawzi, and Teresa Head-Gordon (2008). "A coarse-grained α-carbon protein model with anisotropic hydrogen-bonding". In: *Proteins* 70, 626–638.

Zhou, Rui et al. (2014). "Folding kinetics of WW domains with the united residue force field for bridging microscopic motions and experimental measurements". In: *PNAS* 111.51, pp. 18243–18248.