

***Extracção de Conhecimento de Dados obtidos por Simulação da
Desnaturação Proteica***

por
Elisabeth Silva Fernandes

Dissertação de Mestrado em Análise de Dados e Sistemas de Apoio à Decisão

Orientador: Professor Alípio Jorge
Co-Orientador: Professor Rui Brito, Universidade de Coimbra

Faculdade de Economia
Universidade de Porto

2007

“If there is a hurricane, you always see the signs of it in the sky for days ahead, if you are at sea. They do not see it ashore because they do not know what to look for, he thought.”

Ernest Hemingway, In *The Old Man and the Sea* (1952)

Nota Biográfica

Elisabeth Silva Fernandes é natural de Caracas, Venezuela, onde residiu até aos 6 anos. Em 1988 iniciou o seu percurso escolar em Amares. Em 2005 concluiu a licenciatura em Matemática no Ramo Científico da Matemática Aplicada pela Faculdade de Ciências da Universidade do Porto. No mesmo ano inscreve-se no Mestrado em Análise de Dados e Sistemas de Apoio à Decisão.

Agradecimentos

No final desta dissertação gostaria de agradecer a todos os que me ajudaram a cumprir este objectivo pessoal, sem eles não teria sido possível realizar este trabalho.

Ao Professor Rui Brito agradeço a pronta disponibilidade, para esclarecimentos de dúvidas, e os conselhos científicos muito importantes para a compreensão do problema em estudo.

À Professora Paula Brito pela atenção e disponibilidade com que me recebeu para esclarecer dúvidas importantes para a resolução do problema.

Ao Professor Alípio o meu sincero muito obrigada pela dedicação com que orientou todo o trabalho, a pronta disponibilidade, os conselhos, a paciência e todo o apoio que foi fundamental para a realização deste trabalho.

Aos meus pais e às minhas irmãs pela compreensão, quando não estive presente em momentos importantes da família, e pelo apoio nas horas em que desanimei. Ao meu namorado pela paciência com que me ouviu falar horas a fio da dissertação e pelas palavras de incentivo.

E finalmente, gostaria de agradecer o facto deste trabalho fazer parte do projecto, 2007/2010 - PTDC/BIA-PRO/72838/2006, “*Procura de regras de alto nível em dobragem e desdobragem de proteínas: das doenças amiloidogénicas à previsão da estrutura proteica*”.

Resumo

Nesta dissertação é apresentado um processo de Extração de Conhecimento de dados obtidos a partir de simulações da desnaturação de uma proteína. A proteína em estudo, designada por Transtirretina (TTR), é causadora da “Doença dos Pezinhos” (cujo nome científico é polineuropatia amiloidótica familiar (PAF)).

A propriedade molecular em estudo é a variação da distância do aminoácido ao centro de massa. Esta medida foi registada ao longo de 10 desnaturações de TTR, 5 simulações de uma variante normal WT-TTR e 5 simulações de uma variante mutada L55P-TTR, e obtiveram-se 10 quadros de dados com dimensões elevadas. Com a aplicação de técnicas de Data Mining procurou-se extrair conhecimento útil destes dados.

A análise multidimensional consistiu, inicialmente, em obter três partições dos 127 aminoácidos da TTR, uma primeira partição que só considera as 5 corridas da TTR no estado normal, uma segunda partição que só considera a informação das 5 corridas da TTR no estado mutado, e uma terceira partição que considera toda a informação das 10 simulações (excepto do aminoácido nº55). A 1ª e 2ª partições permitiram comparar os *clusters* obtidos para cada variante de TTR e foram detectadas diferenças significativas nos *clusters*. Com a terceira partição obtiveram-se os aminoácidos representativos de cada *cluster* (14 aminoácidos) reduzindo-se assim o número de elementos em análise.

Pretendia-se ainda encontrar eventos, isto é, mudanças de comportamento significativas dos aminoácidos, e para tal estudou-se isoladamente os aminoácidos representativos. Foram procurados dois eventos que consistem na comparação dos aminoácidos dois a dois quanto às suas posições relativamente ao centro de massa. Com os eventos obtidos foram procuradas regras de associação, que relacionam os eventos e a variante de TTR com mais propensão para ocorrência de determinado evento. Para complementar o estudo foi feita uma procura de um terceiro evento por aminoácido (com os 127 aminoácidos) e foram obtidas regras de associação que indicam qual a variante de TTR com mais probabilidade de ocorrência desse evento.

Através do estudo da variação desta propriedade molecular da proteína ao longo do processo de desnaturação, obtiveram-se resultados que podem contribuir para uma melhor compreensão dos mecanismos moleculares da PAF.

Índice

Resumo.....	iv
Índice.....	v
Índice de figuras	vii
Índice de tabelas	ix
1 Introdução.....	1
1.1 Objectivos	1
1.2 Estrutura do trabalho	2
2 As Proteínas	3
2.1 Aminoácidos	4
2.2 Estruturas das Proteínas.....	5
2.3 Patologias humanas e Enrolamento proteico.....	8
2.4 O Processo de enrolamento das Proteínas	10
2.4.1 A Hipótese Termodinâmica.....	10
2.4.2 Paradoxo de Levinthal.....	11
2.4.3 Funil de <i>Folding</i>	11
2.5 Dinâmica Molecular - Simulação da desnaturação proteica	13
2.6 Propriedades Moleculares.....	15
2.7 Novas medidas em estudo	16
3 Extracção de Conhecimento de Dados (KDD)	18
3.1 Processo de Extracção de Conhecimento	19
3.1.1 Data Mining – Tarefas.....	20
3.1.2 Metodologia.....	21
3.2 Análise de clusters.....	23
3.2.1 Clustering hierárquico	26
3.3 Eventos	34
3.4 Regras.....	36
3.4.1 Regras de Associação.....	36
4 Processo de Extracção de Conhecimento de dados de simulação de Desnaturação Proteica	39
4.1 Descrição dos dados	40
4.1.1 Estudo Preliminar.....	41
4.2 Clustering hierárquico	44
4.2.1 Clustering de cada variante da proteína TTR	45
4.2.2 Partição consenso.....	50
4.2.3 Escolha dos Aminoácidos que representam cada uma das classes.....	56
4.3 Procura de Eventos.....	58
4.3.1 Comparação de aminoácidos dois a dois.....	58
4.3.2 Procura de alterações significativas por aminoácido.	65

4.4	Procura de Regras de Associação	68
4.5	Valor das regras obtidas	70
4.5.1	Como controlar o risco de falsas descobertas?	70
4.5.2	Como medir a aleatoriedade das regras obtidas?	72
5	Conclusão	74
	Bibliografia	76
	Anexos	76
	Anexo I – Análise Univariada.....	84
	A. Valores máximos e mínimos atingidos em cada corrida.....	84
	B. Gráficos do aminoácido da posição 55 ao longo das 10 simulações	86
	Anexo II – Classificação hierárquica - Resultados	88
	A. Clustering da WT-TTR	88
	A1- Optimização dos Índices de Milligan e Cooper	88
	A2- Clustering de cada corrida da WT-TTR.....	90
	A3- Partição consenso da WT-TTR	101
	B. Clustering da L55P-TTR	102
	B1 – Optimização dos Índices de Milligan e Cooper	102
	B2- Clustering de cada corrida da L55P-TTR	103
	C. Partição consenso - sem a posição 55	113
	D. Escolha dos aminoácidos Representativos (sem pos. 55)	113
	Anexo III – Eventos.....	114
	A. Comparação de aminoácidos dois a dois.....	114
	Evento Nº 1 - “troca” de posição entre dois aminoácidos	114
	Evento Nº 2 - “ Existem duas trocas de comportamento de dois aminoácidos em relação ao centro de massa da proteína”	115
	B. Procura de um evento por aminoácido.....	117
	Anexo IV- Regras de Associação.....	120
	A. Análise Gráfica da regra nº1.....	120
	Anexo V- Código em R	122

Índice de figuras

Figura 1 - Estrutura Primária	6
Figura 2 - Estrutura Secundária	6
Figura 3 - Estrutura Terciária.....	7
Figura 4 - Estrutura Quaternária	7
Figura 5 - Estrutura das proteínas	7
Figura 6 – Modelo de formação de placas de amilóide.....	8
Figura 7 – Representação Esquemática da estrutura tridimensional do tetramero WT-TTR (A) e de uma subunidade da WT-TTR (B).....	9
Figura 8 - (a) Funil de <i>Folding</i> (conceito de paisagem de energia): O caminho A indica um processo rápido de <i>folding</i> , o caminho B indica um processo de <i>folding</i> mais lento; (b) Superfície de Energia do processo de <i>folding</i>	12
Figura 9 - Ilustração das várias superfícies.....	15
Figura 10 - Variação da distância dos aminoácidos ao centro de massa.....	16
Figura 11 - Distâncias aos 3 vizinhos mais próximos do átomo X.....	17
Figura 12 – Dados 3-D	17
Figura 13 – Evolução das técnicas de Análise de Dados	18
Figura 14 – “O caminho para o conhecimento”.....	19
Figura 15 - Modelo CRISP-DM.....	22
Figura 16 - Fases de uma Análise de Clusters.	24
Figura 17 - Critério de Validação Externo	29
Figura 18 - Critério de Validação Interno.....	30
Figura 19 - Critério de Validação Relativo.....	30
Figura 20 - Tipos de Episódios	35
Figura 21-Processo de Extração de Conhecimento	40
Figura 22 - Representação do sistema usado nas simulações de Dinâmica Molecular..	41
Figura 23 – Variação da distância do carbono alpha dos aminoácidos GLY_1 (a) e VAL_71 (b) ao centro de massa da TTR ao longo de uma simulação de desnaturação da variante WT-TTR (WTRUN2).	43
Figura 24 - Matriz das classificações para cada tipo de proteína TTR	46
Figura 25 - Partição da WT-TTR e da L55P-TTR.....	49
Figura 26 - Índices de Milligan e Cooper – G2 e G3 - para a simulação WTRUN2....	50
Figura 27 - Índices de Milligan e Cooper – G2 e G3 - para a simulação WTRUN2....	51
Figura 28 - Índices de Milligan e Cooper – G2 e G3 - para a simulação WTRUN2....	51
Figura 29 - Índices de Milligan e Cooper – G2 e G3 - para a simulação WTRUN2....	51
Figura 30 - Índices de Milligan e Cooper – G2 e G3 - para a simulação WTRUN2....	52
Figura 31 - Índices de Milligan e Cooper – G2 e G3 - para a simulação WTRUN2....	52
Figura 32 - Índices de Milligan e Cooper – G2 e G3 - para a simulação WTRUN2....	53
Figura 33 - Índices de Milligan e Cooper – G2 e G3 - para a simulação WTRUN2....	53
Figura 34 - Índices de Milligan e Cooper – G2 e G3 - para a simulação WTRUN2....	53
Figura 35 - Índices de Milligan e Cooper – G2 e G3 - para a simulação WTRUN2....	54
Figura 36 – Partição consenso na sequência de aminoácidos inicial	55
Figura 37- Aminoácidos Representativos.....	57
Figura 38 – Zona de Viragem	58
Figura 39 – LEU_17 aproxima-se do centro de massa ao mesmo tempo que GLU_42 se afasta do centro de massa.....	59

Figura 40 – (a) Os 10 eventos mais frequente nas 10 simulações; (b) Os 10 eventos mais frequentes nas simulações da WT-TTR; (c) Os 10 eventos mais frequentes nas simulações da variante L55P-TTR.	60
Figura 41 – Procura do evento 1- ALA_91 troca com GLU_42 ao longo das 5 corridas da WT-TTR. Na corrida WT2 e na corrida WT6 existe troca, com estabilidade de 300 frames.	61
Figura 42 – Procura do evento 1- ALA_91 troca com GLU_42 ao longo das 5 corridas da L55P-TTR. Não se verifica a ocorrência do evento.	61
Figura 43 – Teste de Kruskal- Wallis.....	62
Figura 44 – O tempo foi dividido em três secções, procura-se a situação em que a relação de dois aminoácidos com o centro de massa é trocada duas vezes e mantêm-se ao longo de n frames de cada secção.	62
Figura 45 – Procura do evento 17 - GLU_42 troca 2x_A_ALA_91 ao longo das 5 corridas da L55P-TTR. Na corrida L55_3 não existem duas trocas de comportamento.	63
Figura 46 – Teste de Kruskal-Wallis.....	64
Figura 47 – Análise gráfica do evento “ <i>existência de uma alteração significativa no comporta-mento do aminoácido PRO_86</i> ” que se verifica nas corridas WT2, WT3, WT4, WT5, WT6, L55_1.....	66
Figura 48 – Análise gráfica do evento “ <i>existência de uma alteração significativa no comporta-mento do aminoácido PRO_86</i> ” nas corrdas onde não se verifica o evento.	66
Figura 49 – Teste Kruskal-Wallis	67
Figura 50 – Evolução da distância ao centro de massa do aminoácido LEU_55 da variante WT-TTR.	86
Figura 51 – Evolução da distância ao centro de massa do aminoácido LEU_55 da variante WT-TTR.	87
Figura 52 - Índices de Milligan e Cooper – G2 e G3 para a simulação WTRUN2	88
Figura 53 - Índices de Milligan e Cooper – G2 e G3 para a simulação WTRUN3	88
Figura 54 - Índices de Milligan e Cooper – G2 e G3 para a simulação WTRUN4	89
Figura 55 - Índices de Milligan e Cooper – G2 e G3 para a simulação WTRUN5	89
Figura 56 - Índices de Milligan e Cooper – G2 e G3 para a simulação WTRUN6	89
Figura 57 - Clustering de cada corrida da WT-TTR	90
Figura 58 - Dendrograma WT-TTR	101
Figura 59 - Índices de Milligan e Cooper – G2 e G3 para a simulação L55PRUN1 ..	102
Figura 60 - Índices de Milligan e Cooper – G2 e G3 para a simulação L55PRUN2 ..	102
Figura 61 - Índices de Milligan e Cooper – G2 e G3 para a simulação L55PRUN3 ..	102
Figura 62 - Índices de Milligan e Cooper – G2 e G3 para a simulação L55PRUN4 ..	103
Figura 63 - Índices de Milligan e Cooper – G2 e G3 para a simulação L55PRUN5 ..	103
Figura 64 - Clustering de cada corrida da L55P-TTR.....	103
Figura 65 - Dendrograma L55P-TTR.....	112
Figura 66 - Dendrograma TTR	113

Índice de tabelas

Tabela 1 - Principais funções das proteínas.....	3
Tabela 2 – Aminoácidos	5
Tabela 3 – Tarefas de Data Mining.....	21
Tabela 4 – Algumas Medidas de Interesse	38
Tabela 5 - Máximos e mínimos da distância do carbono alpha ao centro de massa.....	42
Tabela 6 – Número de <i>clusters</i> para cada simulação da WT-TTR	46
Tabela 7 - Número de <i>clusters</i> para cada simulação da L55P-TTR	46
Tabela 8 - Número de <i>clusters</i> ótimas para cada simulação.....	54
Tabela 9 - Partição Consenso.....	55
Tabela 10- Frequência do evento nº1 para cada variante de TTR	63
Tabela 11 – Eventos mais frequentes na proteína L55P-TTR e que não se verificam na WT-TTR	64
Tabela 12- Eventos mais frequentes na WT-TTR.....	64
Tabela 13 - Comportamento do aminoácido PRO_86	65
Tabela 14 – Aminoácidos com comportamentos mais variáveis.....	67
Tabela 15 - WTRUN2- Classificação hierárquica: elementos das 14 classes	90
Tabela 16 - WTRUN3- Classificação hierárquica: elementos das 13 classes	93
Tabela 17 - WTRUN4- Classificação hierárquica: elementos das 14 classes	95
Tabela 18 - WTRUN5- Classificação hierárquica: elementos das 14 classes	97
Tabela 19 - WTRUN6- Classificação hierárquica: elementos das 13 classes	99
Tabela 20 – Clustering da WT (Partição I).....	101
Tabela 21 – Clustering da L55PRUN1: 15 classes	104
Tabela 22 - Clustering da L55PRUN2: 17 classes.....	105
Tabela 23 - Clustering da L55PRUN3: 14 classes.....	107
Tabela 24 - Clustering da L55PRUN4: 13 classes.....	108
Tabela 25 - Clustering da L55PRUN5: 11 classes.....	110
Tabela 26 – Classificação L55 (PartiçãoII)	112
Tabela 27 – Aminoácidos Representativos.....	113
Tabela 28- Eventos “troca” de posição entre dois aminoácidos	114
Tabela 29- Eventos “troca” de posição entre dois aminoácidos	115
Tabela 30- Aminoácidos onde se verificou o evento “ <i>alteração significativa no comportamento do aminoácido</i> ”	117
Tabela 31 - Frequência dos eventos e janela onde ocorre a alteração mais significativa (com maior declive).....	118

1 Introdução

Actualmente são realizados inúmeros esforços no estudo e investigação das causas e curas para variadíssimas doenças. Nas últimas décadas novas doenças têm sido alvo de importantes estudos, entre as quais as doenças amilóides. Deste grupo de maleitas fazem parte doenças bem conhecidas como o Alzheimer e a Paramiloidose (mais conhecida por Doença dos Pezinhos). (Brito et. al., 2004)

Várias são as tentativas para perceber as origens das doenças amilóides. Sabe-se por exemplo que estas estão associadas a alterações conformacionais e à agregação das proteínas associadas a estas doenças. A procura da origem destas doenças tem despertado o interesse entre os bioquímicos e não só. Actualmente, este é um problema multidisciplinar no qual são utilizadas ferramentas de diferentes áreas.

Os quadros de dados em estudo nesta área têm dimensões elevadas o que exige aos estudiosos a procura de novos métodos de análise, métodos estes que permitam obter o maior conhecimento possível sem perda de informação. O que se verifica é que as técnicas de Extração de Conhecimento de Dados permitem atingir este objectivo. Com o aumento do volume de dados nas diversas áreas de estudo, os métodos de *Data Mining* têm sido refinados e adaptados às novas situações.

1.1 Objectivos

Nesta dissertação pretende-se, com a ajuda de técnicas de Extração de Conhecimento de Dados, analisar a desnaturação da proteína causadora da Paramiloidose, transtirretina (TTR). O objectivo é estudar a distância dos aminoácidos ao centro de massa da proteína ao longo da desnaturação da proteína TTR e encontrar características desta que possam contribuir, positivamente, para a futura descoberta das causas e soluções desta doença.

1.2 Estrutura do trabalho

Esta dissertação está dividida em duas partes: uma primeira onde são explicados os fundamentos teóricos necessários para a compreensão do problema em estudo; e uma segunda parte onde é apresentada toda a análise efectuada aos dados.

Do desenvolvimento teórico fazem parte os capítulos 2 e 3. No capítulo 2 são introduzidas algumas definições da biologia molecular e são apresentadas as razões de tão grande interesse nesta área da investigação. No capítulo 3 são abordadas as Técnicas de extracção de conhecimento de dados utilizadas neste estudo.

No capítulo 4 é apresentado todo o processo de extracção de conhecimento dos dados resultantes das simulações de desnaturação proteica. Neste capítulo são apresentadas as várias análises efectuadas aos dados e os resultados obtidos com a aplicação das técnicas de extracção de conhecimento referidas no capítulo 3.

No capítulo 5 consta a conclusão do trabalho desenvolvido. Em anexo está toda a informação necessária para complementar o estudo realizado ao longo desta dissertação.

2 As Proteínas

A palavra proteína deriva da palavra grega “proteios”, que significa “ primeiro”. Esta designação tem origem na elevada importância das proteínas na vida, uma vez que as proteínas têm um elevado peso na constituição dos organismos vivos (Morrison et. al., 1992).

As proteínas são macromoléculas que desempenham funções biológicas fundamentais para o equilíbrio dos organismos vivos. Estas funções são múltiplas e as principais encontram-se resumidas na Tabela 1.

Tabela 1 – Principais funções das proteínas

Função		Exemplo
Estrutural	Conferem rigidez e consistência aos tecidos	Colagénio - proteína de alta resistência da pele
Enzimática	Catalisam e regulam as reacções bioquímicas	Lipases - transformam os lípidos nas suas unidades
Hormonal	Exercem uma função específica sobre algum órgão ou estrutura	Insulina
Defesa	Defesa contra elementos estranhos ao organismo	Anticorpos
Transporte	Entre as várias actividades de transporte podem, por exemplo, transportar gases para as células	Hemoglobina

Segundo Morrison et. al., (1992) as proteínas dividem-se em duas grandes classes: **proteínas fibrosas**, insolúveis em água, e **proteínas globulares**, solúveis em água ou em soluções aquosas. Esta classificação está relacionada com a forma das moléculas. As proteínas fibrosas são estruturas alongadas e filamentosas. As proteínas globulares apresentam-se dobradas sobre si próprias e formam unidades compactas.

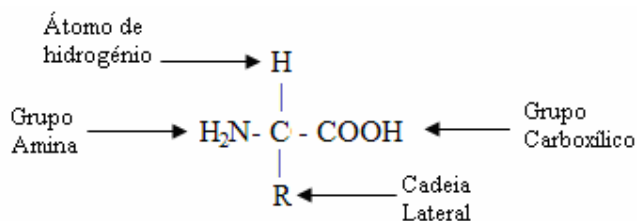
Diversos factores do meio podem levar à destruição/desnaturação da molécula da proteína, tais como a variação da temperatura, do pH, ou a presença de agentes químicos. Desta forma a proteína perde a estrutura tridimensional que lhe permitia desempenhar as suas funções no organismo. Um exemplo simples de desnaturação

ocorre com o ovo estrelado, isto é, as proteínas da clara do ovo quando sujeitas a temperaturas elevadas desnaturam e precipitam. Este é o motivo pelo qual a clara do ovo adquire a cor branca (Morrison et. al., 1992).

A perda da estrutura tridimensional de uma proteína leva à perda das suas funções, em alguns casos origina a formação de agregados fibrilares os quais se depositam nos tecidos originando diversas doenças. Para uma melhor compreensão do problema em estudo, do ponto de vista bioquímico, nas próximas secções são abordados os seguintes temas: como é constituída uma proteína, quais as suas estruturas possíveis, quais as consequências de proteínas mal formadas e quais as características moleculares de interesse para o seu estudo.

2.1 Aminoácidos

As proteínas são constituídas por **aminoácidos**. Cada uma destas unidades é um ácido orgânico $-COOH$ no qual o carbono próximo do grupo (chamado *carbono alfa*) está também ligado a um grupo $-NH_2$. O carbono alfa também se liga a uma cadeia lateral (R), que é diferente para cada aminoácido, isto é, as características de cada aminoácido variam consoante a composição química da cadeia lateral. Os vinte aminoácidos existentes na natureza foram classificados, quanto à sua cadeia lateral, em quatro grupos: **ácidos, básicos, polares e não polares**. (Robertis et. al., 1987)



Os aminoácidos diferem no tamanho e em várias propriedades físicas e químicas nomeadamente a **hidrofobicidade**, isto é, o quanto os aminoácidos evitam a água (Cejtin et. al., 2002). O *efeito hidrofóbico* tem como resultado a associação das partes hidrofóbicas (apolares) das moléculas quando em meio aquoso sendo esta interacção, por vezes, conhecida como “ligação hidrofóbica”. Os aminoácidos que tendem a interagir favoravelmente com a água são classificados como **hidrofílicos**.

Na primeira coluna da Tabela 2 é possível observar as famílias de aminoácidos, sendo esta classificação importante para na análise dos dados comparar alguns dos resultados obtidos.

Tabela 2 – Aminoácidos

Família do Aminoácido	Designação	Símbolo com 3 letras	Símbolo com 1 letra
Básico	Lisina	LYS	L
	Arginina	ARG	R
	Histidina	HIS	H
Ácido	Ácido Aspártico	ASP	D
	Ácido Glutâmico	GLU	E
Polar (hidrofílicos)	Asparagina	ASN	N
	Glutamina	GLN	Q
	Serina	SER	S
	Treonina	THR	T
	Tirosina	TYR	Y
Não Polar (hidrofóbicos)	Triptofano	TRP	W
	Glicina	GLY	G
	Valina	VAL	V
	Leucina	LEU	L
	Isoleucina	ILE	I
	Prolina	PRO	P
	Fenilalanina	PHE	F
	Metionina	MET	M
	Alanina	ALA	A
	Cisteína	CYS	C

Fonte: Alberts et. al., 1994, pp.56.

2.2 Estruturas das Proteínas

As proteínas são sintetizadas como uma sequência linear de aminoácidos. Na cadeia de aminoácidos estabelecem-se interações entre eles o que origina o enrolamento da proteína. No final, deste processo de enrolamento, a proteína adquire uma estrutura tridimensional única.

A estrutura tridimensional de uma proteína pode ser descrita em quatro níveis.

- **Estrutura primária:** Consiste na sequência linear aleatória de aminoácidos que formam uma cadeia polipeptídica. A sequência de aminoácidos tem elevada importância, pois basta a troca de dois aminoácidos, ou a substituição de um qualquer para produzir importantes alterações biológicas. Por exemplo, a doença hereditária denominada por *anemia falsiforme* tem origem na substituição de um aminoácido na molécula de hemoglobina.

–Ala–Glu–Val–Thr–Asp–Pro–Gly–

Figura 1 - Estrutura Primária

- **Estrutura secundária:** é o resultado do arranjo espacial entre aminoácidos próximos na cadeia peptídica. Dois dos principais elementos de estrutura secundária são:

- α -hélice: tipo mais comum da estrutura secundária que se caracteriza por uma estrutura helicoidal estabilizada por ligações de hidrogénio estabelecidas entre os grupos $-CO$ e $-NH$ de aminoácidos que distam entre si 4 resíduos. (Morrison et. al., 1996).

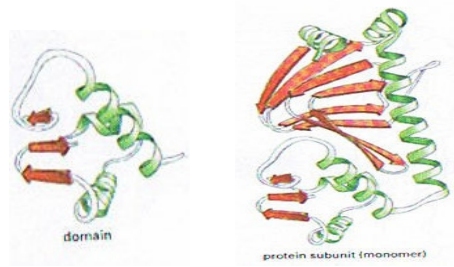
- folha β : estrutura estendida, constituída por associação lateral de segmentos peptídicos numa orientação paralela ou anti-paralela, mantida por ligações de hidrogénio.



Fonte: Alberts B. et. al., 1994.

Figura 2 - Estrutura Secundária

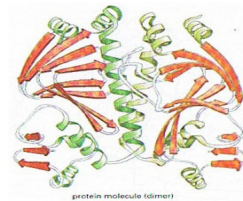
- **Estrutura terciária:** A estrutura terciária está relacionada com as dobras da cadeia proteica sobre ela mesma. Esta estrutura descreve um arranjo espacial global dos aminoácidos numa proteína, podendo estes estar bastante separados entre si na sequência linear.



Fonte: Alberts B. et. al.,1994.

Figura 3 - Estrutura Terciária

- **Estrutura quaternária:** Certas proteínas são constituídas por mais de uma cadeia proteica, a estrutura quaternária é o resultado da disposição espacial relativa de subunidades proteicas em proteínas complexas formadas por duas ou mais cadeias polipeptídicas.



Fonte: Alberts B. et. al.,1994.

Figura 4 - Estrutura Quaternária

Na Figura 5 encontram-se ilustradas diferentes estruturas que uma proteína pode tomar proteínas.

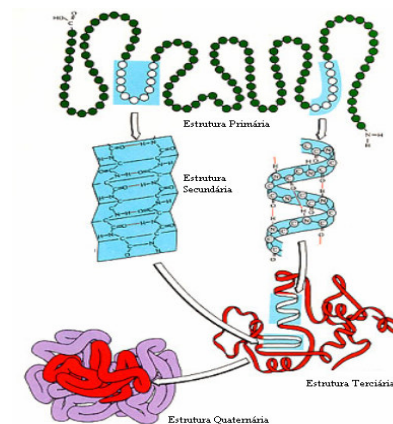


Figura 5 - Estrutura das proteínas

Cada proteína tem uma estrutura primária e tridimensional única, e como tal, cada proteína desempenha uma função biológica específica.

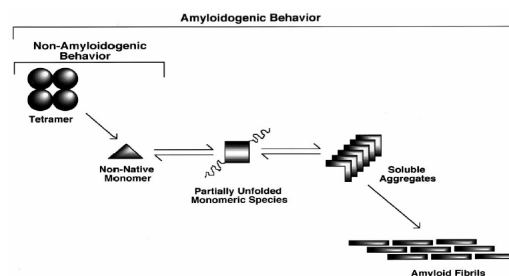
A hidrofobicidade, mencionada anteriormente, é referida por vários autores como um factor importante na organização da estrutura tridimensional de uma proteína. (Kauzmann W. 1964), (Cejtin, H. 2002), (Pace, C. N. 1996). Segundo Kauzmann W. (1964), 25 a 40% dos aminoácidos de uma cadeia paptídica são hidrofóbicos. Este tipo de aminoácidos tem pouca afinidade com a água, logo espera-se que as proteínas tendam a enrolar-se de tal forma que estes aminoácidos fiquem no interior da molécula.

Nos trabalhos de Shirley et. al., (1992) e Pace et. al., (1996) é referido que o efeito hidrofóbico tem sido considerado como uma força dominante no enrolamento das proteínas, no entanto as ligações de hidrogénio são também um factor relevante. As pontes de hidrogénio desempenham um importante papel no enrolamento das cadeias polipeptídicas e na formação da estrutura secundária.

2.3 Patologias humanas e Enrolamento proteico

As proteínas são macromoléculas com funções importantíssimas nas células, logo é fundamental que o enrolamento das proteínas seja um processo bem sucedido, pois caso contrário, pode originar alterações graves no processo biológico e por vezes irreversíveis.

Das várias patologias originadas pelo errado enrolamento proteico, neste trabalho irá ser dada especial atenção às Amiloidoses. As amiloidoses são doenças que têm origem na acumulação, sob a forma de placas, de quantidades elevadas de proteína incorrectamente enrolada (proteína insolúvel) nos tecidos. (Quintas et. al., 2001)



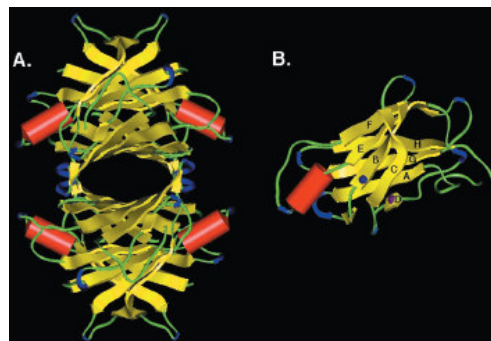
Fonte: Quintas et. al., (2001)

Figura 6 – Modelo de formação de placas de amilóide

As doenças amilóides são classificadas consoante os sinais clínicos e o tipo de proteína amilóide envolvida. Neste trabalho a doença amiloidótica em estudo é a Paramiloidose também designada por Polineuropatia Amiloidótica Familiar tipo I, de Andrade ou tipo Português, ou mais vulgarmente “Doença dos Pezinhos”. Esta doença foi descrita pela primeira vez pelo neurologista português Mário Corino de Costa Andrade. (Andrade, 1952)

A Paramiloidose é uma doença neurológica, crónica, progressiva e hereditária de transmissão *autossómica dominante*¹ que afecta o sistema nervoso. O gene responsável pela produção da proteína mutada característica desta doença localiza-se no braço longo do cromossoma 18. A proteína expressa por este gene é a Transtirretina (TTR) que é uma proteína homotetramérica, isto é, a TTR é constituída por 4 monómeros cada um dos quais constituídos por 127 aminoácidos. (Luís, 2006)

A estrutura tridimensional da TTR encontra-se representada simplificada na Figura 7.



Fonte: Quintas et. al., 1999.

Figura 7 – Representação Esquemática da estrutura tridimensional do tetramero WT-TTR (A) e de uma subunidade da WT-TTR (B)

Quando a TTR sofre uma mutação pode obter-se uma variante patogénica da proteína. A proteína mutada mais comum no tipo português é a TTR Val30Met que resulta da substituição de uma Valina na posição 30 por uma Metionina. Actualmente são conhecidas mais de 80 variantes patogénicas da TTR (Luis, 2006).

¹ Um indivíduo portador do gene desta doença (seja homozigótico ou heterozigótico) vai manifestá-la.

Uma das variantes mais amiloidogénicas é a Leu55Pro (L55P-TTR), esta variante é resultado da substituição da Leucina na posição 55 por uma Prolina. Neste trabalho, irão ser estudados dados relacionados com esta variante patogénica e com a proteína não mutada designada por *wild-type* (WT-TTR).

2.4 O Processo de enrolamento das Proteínas

Nas últimas 5 a 6 décadas foram apresentadas várias hipóteses para procurar compreender como é que uma proteína atinge a sua estrutura tridimensional nativa². Neste capítulo é apresentada uma síntese dos trabalhos desenvolvidos.

2.4.1 A Hipótese Termodinâmica

Durante a década de 50, do século XX, Christian Anfinsen procurou dar resposta às seguintes questões:

- Porque é que a proteína enrola para a estrutura nativa?
- Será ajudada por outras moléculas?
- Porque é que a estrutura nativa é única?

Para responder a estas questões Anfinsen realizou várias experiências com a proteína RNaseA que deram origem à hipótese termodinâmica (HT). Segundo a HT o enrolamento das proteínas é um processo espontâneo, acompanhado pela libertação de energia livre no qual o sistema passa a ocupar um estado de menor energia, portanto mais estável. Logo o estado mais estável no enrolamento da proteína é o estado nativo porque corresponde ao estado de energia mais baixo. (Honig, 1999)

Segundo esta hipótese, o estado nativo é único porque é determinado por uma sequência de aminoácidos única e a procura do estado nativo é aleatória.

Em 1968, Cyrus Levinthal mostrou que a hipótese termodinâmica não resolvia todas as questões no enrolamento proteico. (Levinthal, 1968)

² Conformação obtida no final do processo de enrolamento *folding* da sequência de aminoácidos.

2.4.2 Paradoxo de Levinthal

A hipótese de que o processo de enrolamento/*folding* ocorre de forma aleatória através de todas as conformações possíveis até encontrar a conformação correcta foi refutada em 1968 por Cyrus Levinthal através de um argumento simples. Considere-se uma proteína com 100 aminoácidos e cada aminoácido com 2 conformações possíveis, porque dois é o número mínimo de conformações por aminoácido. O que se obteria seriam 2^{100} conformações possíveis, entre as quais está a que corresponde ao estado nativo. Considere-se ainda que a proteína levaria 1 picosegundo a tentar cada combinação, que é o tempo de vibração térmica, então o tempo necessário para explorar todo o espaço conformacional seria $2^{100} ps$ que é aproximadamente 3.9×10^{10} anos. Mas o tempo de vida estimado do Universo é de 1.4×10^{10} anos! Levinthal concluiu que a HT não explica a escala de tempo do processo de folding das proteínas. (Levinthal, 1968) (Honig, 1999)

Para resolver este problema Levinthal, C. (1968) propôs a Hipótese Cinética, na qual defende que o estado nativo não corresponde ao estado termodinâmico mais estável, mas sim, ao estado de energia mínima que é o mais acessível do ponto de vista cinético.

Segundo Honig (1999) na década de 90, verificou-se que esta hipótese por si só também não resolvia a questão e surge uma teoria unificada, que se baseia na natureza estatística do processo (Wolynes et. al., 1995) (Dill e Chan, 1997).

2.4.3 Funil de *Folding*

Segundo a actual perspectiva de enrolamento das proteínas, este ocorre numa hiper-superfície de energia num espaço com algumas centenas ou milhares de graus de liberdade, em que as zonas de baixa energia ('vales') se encontram separadas umas das outras por barreiras de energia ('cumes'). Esta superfície é uma função de energia potencial complexa e multidimensional (Honig, 1999). Se esta superfície for muito acidentada, a proteína vai encontrar muitas barreiras de energia. Se por outro lado for muito plana, a proteína vai andar eternamente à procura do seu estado nativo. Logo a proteína precisa de uma orientação para explorar a paisagem de energia até atingir o seu

estado nativo. A topografia da paisagem de energia é em forma de funil, para que a proteína deslize de uma forma mais fácil até ao estado nativo (Figura 8).

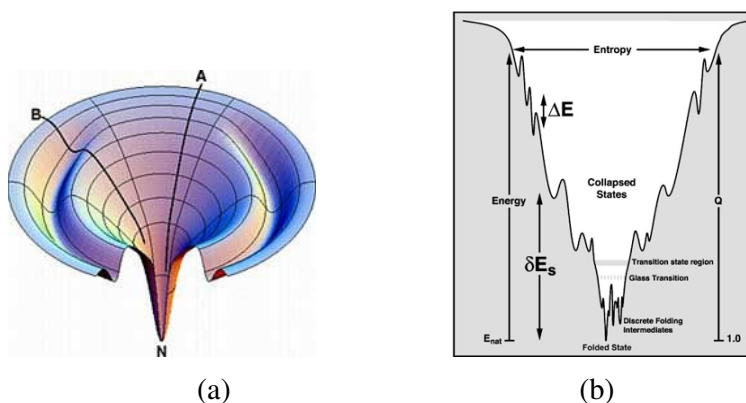


Figura 8 - (a) Funil de *Folding* (conceito de paisagem de energia): O caminho A indica um processo rápido de *folding*, o caminho B indica um processo de *folding* mais lento; (b) Superfície de Energia do processo de *folding*.

O topo do funil corresponde a todas as conformações desnaturadas possíveis, que são as conformações menos estáveis, ou seja, de maior energia. À medida que a proteína se enrola a energia vai diminuindo até atingir o mínimo global na conformação nativa. (Honig, 1999)

Esta nova perspectiva reconcilia as hipóteses termodinâmica e cinética, porque o estado nativo pode ser o estado termodinâmico mais estável, e a procura desse estado não é aleatória uma vez que existem vários caminhos de *folding* possíveis (Pande et. al., 1998). Os caminhos de *folding* implicam “uma boa definição da sequência de eventos” (Levinthal, 1968) onde “evento” é a transição de uma fase para a outra.

Pande et. al., (1998) introduz o termo “neo-clássico” para definir esta nova perspectiva porque o processo de *folding* da proteína pode ser entendido como uma extensão da teoria clássica que considera o mecanismo estatístico do processo. No entanto, ainda não está tudo resolvido! Para proteínas pequenas pode garantir-se que se sabe quase tudo mas para proteínas grandes os mecanismos de *folding* ainda são desconhecidos.

Desde a segunda metade do século XX vários investigadores se empenham em analisar o *folding* das proteínas. Em 1999, Honig, apresenta uma análise da evolução dos estudos. Segundo este autor verificaram-se importantes avanços na previsão da estrutura das proteínas, no entanto, para Honig, este progresso depende em parte da combinação

entre a compreensão dos princípios físico-químicos e a análise sofisticada de informação válida.

Segundo Pande (1998) a natureza dos caminhos de *folding* pode, em princípio, ser completamente compreendida através de simulação directa em computador onde todos os detalhes são acessíveis. Desta forma, faz todo o sentido, obter simulações da desnaturação proteica da TTR, para encontrar características importantes desta proteína através de técnicas de extracção de conhecimento de dados.

2.5 Dinâmica Molecular - Simulação da desnaturação proteica

Analisar e compreender as relações entre a sequência e a estrutura das proteínas tornou-se uma tarefa central na “Era pós-genoma”, e terá grande impacto na genética, bioquímica e na química farmacêutica. A descrição detalhada do processo de *folding* ajuda a compreender melhor a origem de diversas doenças (Pande et. al., 2003).

Trabalhos recentes mostram a importância da simulação da desnaturação proteica para os estudos da Biologia Molecular. No entanto, as simulações de Dinâmica Molecular apresentam o problema de elevado custo de recursos computacionais.

A Dinâmica Molecular pode ser usada para simular a dinâmica de estruturas conhecidas, ou para descrever uma estrutura de uma sequência a partir da minimização da energia potencial.

A Dinâmica Molecular clássica está baseada na segunda lei de Newton ou nas *equações de movimento de Newton*. Nestas equações cada átomo é considerado um ponto com massa m_i :

$$m_i \ddot{r}_i = F_i, \quad i = 1, 2, \dots, N \quad \ddot{r}_i = \frac{d^2 r_i}{dt^2} \quad (2.5.1)$$

onde $r_i = (x_i, y_i, z_i)$ é o vector das coordenadas cartesianas do i-ésimo átomo, $\ddot{r}_i$ é a respectiva aceleração, F_i é o vector de forças que actuam no i-ésimo átomo, e N é o número de átomos. (Scheraga et. al., 2007)

Para aproximar os mecanismos naturais entre os átomos, tais como colisões e fricções, é utilizada a dinâmica de Langevin. Isto é, à equação (2.5.1) são adicionados dois novos termos: a força potencial que actua sob o átomo i , $-\nabla_{r_i} U$, e o vector de forças aleatórias, $R_i(t)$, resultantes da colisão entre o átomo i e as molecular do solvente. Desta forma obtém-se a **Equação de Langevin** (2.5.2).

$$F_i = -\nabla_{r_i} U(r_1, r_2, \dots, r_N) - m_i \gamma_i \dot{r}_i + R_i(t), \quad i = 1, 2, \dots, N \quad (2.5.2)$$

Em que, $\dot{r}_i$ é a velocidade do átomo i e γ_i é o coeficiente de fricção.

Nas simulações é assumido que o vector $R_i(t)$ tem média zero e as forças aleatórias em tempos diferentes são não correlacionadas. Considere-se ω a frequência característica do sistema, quando $\gamma \gg 2\omega$ a equação (2.5.2) reduz-se a uma equação diferencial de primeira ordem:

$$0 = -\nabla_{r_i} U(r_1, r_2, \dots, r_N) - m_i \gamma_i \dot{r}_i + R_i(t), \quad i = 1, 2, \dots, N \quad (2.5.3)$$

Integrando as equações de movimento obtém-se uma trajectória que descreve as posições, velocidades e acelerações dos átomos, e como eles variam ao longo do tempo. Esta trajectória é uma série de sub estados entre o estado inicial e o final, separados por um pequeno intervalo de tempo (por exemplo, picosegundo)

Os campos de forças mais utilizados, para calcular as forças potencias da equação (2.5.2), na descrição de sistemas proteicos, são CHARMM (Brooks et. al., 1983), AMBER (Pearlman et al., 1995), GROMOS (Berendsen et. al., 1995) e CVFF (Ewing et. al., 1999). A escolha de um destes campos é feita consoante os custos computacionais implicados na descrição do sistema e a precisão na representação das forças que actuam no sistema (Scheraga et. al., 2007).

As Simulações de Dinâmica Molecular para obtenção dos dados em estudo neste trabalho foram realizadas através do programa NAMD (Kalé et. al., 1999), usando a versão 27 do campo de forças CHARMM.

Para analisar a evolução dos resíduos de aminoácidos ao longo da simulação serão apresentadas algumas propriedades úteis na análise de trajectórias de Dinâmica Molecular.

2.6 Propriedades Moleculares

As propriedades moleculares definidas nesta secção permitem analisar as trajectórias obtidas por simulações de Dinâmica Molecular, neste caso, simulações de desnaturação da TTR. Através dos valores obtidos por algumas destas medidas foram encontrados resultados que se podem revelar de extrema importância para a compreensão da Biologia Molecular do Sistema.

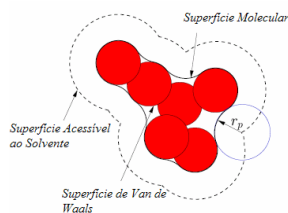
Várias são as propriedades moleculares que podem ser calculadas ao longo de uma simulação, nomeadamente, a Raiz do Desvio Quadrático Médio, o Raio de Giração, a estrutura secundária, as ligações de hidrogénio, e os contactos nativos. (Brito et. al., 2004)

Uma das mais conhecidas medidas variação da conformação molecular é a Raiz do Desvio Quadrático Médio, cuja expressão matemática é a seguinte:

$$RMSD = \sqrt{\sum_{i=1}^n \frac{(x_i(t) - x_i(t_0))^2}{n}} \quad (2.4.1)$$

onde $x_i(t)$ é a posição do i -ésimo átomo no instante t e n é o número de átomos.

Uma outra medida usada é a área da superfície molecular acessível ao solvente, designada por SASA. A superfície de uma molécula determina várias das suas propriedades físicas e químicas, o que torna a superfície molecular uma característica interessante e importante.



Adaptado de Wei et. al., 2004

Figura 9 - Ilustração das várias superfícies

A superfície acessível ao solvente (SASA) é determinada usando uma “esfera de prova” (molécula de água com 1.4 Å de raio) que “rola” à volta da molécula. A superfície

imaginária descrita pelo centro da esfera de prova fornece uma estimativa da superfície molecular acessível ao solvente (SASA).

A SASA é **Global** quando é calculada a superfície global da proteína que está exposta ao solvente. Quando se calcula esta superfície para cada aminoácido temos a SASA **por resíduo**.

Noutros trabalhos, no âmbito deste projecto, foram analisadas séries temporais que descrevem os valores de SASA ao longo da desnaturação da proteína TTR (Azevedo et. al, 2005) (Ferreira, et. al., 2007). Neste trabalho decidiu-se estudar uma nova medida para que, sendo o problema abordado de outro ponto de vista, se possam tirar novas conclusões e complementar todo o trabalho desenvolvido anteriormente.

2.7 Novas medidas em estudo

Neste trabalho a medida em estudo é a **Variação da Distância do Carbono Alfa de cada resíduo ao Centro de Massa da Proteína** ao longo da desnaturação.

O que se pretende com esta medida é obter um quadro de dados com 127 séries temporais, em que cada série descreve a variação da distância do carbono α de um aminoácido ao Centro de Massa da Proteína, ao longo da desnaturação da proteína TTR.

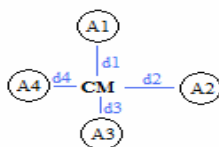


Figura 10 - Variação da distância dos aminoácidos ao centro de massa

Para trabalho futuro, foi ainda proposta uma outra medida que consiste em calcular, para cada aminoácido, os 3 aminoácidos mais próximos da seguinte forma:

- Para todos os átomos, a_i , dos aminoácidos, A_i , são calculadas as distâncias entre átomos.
- Para cada aminoácido A_i são registados os 3 valores mais baixos, que correspondem aos três aminoácidos com átomos mais próximos de A_i .

Por exemplo, como vemos na Figura 11, se para o aminoácido A, que possui o átomo X as distâncias mais curtas forem d_1 , d_2 e d_3 , então o aminoácido A está mais próximo dos aminoácidos A1, A2 e A3.

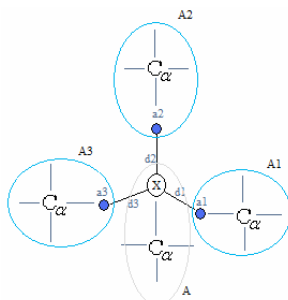


Figura 11 - Distâncias aos 3 vizinhos mais próximos do átomo X.

As distâncias serão calculadas para cada instante t de uma simulação, logo para cada aminoácido no instante t_i vamos obter um vector $(A_1, A_2, A_3, d_1, d_2, d_3)$. Este conjunto de dados constitui um “cubo” de dados, para cada simulação, da forma:

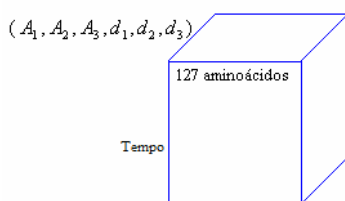


Figura 12 – Dados 3-D

Devido à elevada dimensão dos dados obtidos por esta medida e à dificuldade em obtê-los, partiu-se para o estudo da Variação da distância dos aminoácidos ao centro de massa.

Inicialmente pensou-se que a análise da *Variação da Distância dos aminoácidos ao Centro de Massa* poderiam não complementar a informação obtida, em trabalhos anteriores, com a análise dos valores da SASA. Mas após alguma análise exploratória concluiu-se que seria interessante analisar a evolução das Distâncias dos Aminoácidos ao Centro de Massa da Proteína. Para esta análise foi necessário aplicar técnicas de Data Mining que permitissem extrair conhecimento interessante.

3 Extracção de Conhecimento de Dados (KDD)

Segundo Fayyad et. al., (1996) a Extracção de Conhecimento em Bases de Dados (KDD) é o processo não trivial de identificar padrões válidos, novos, potencialmente úteis e compreensíveis nos dados. O *Data Mining* é um passo da KDD que consiste num tipo particular de algoritmos que, sob aceitáveis limitações de eficiência computacional, produzem uma enumeração particular de padrões dos dados.

Com o desenvolvimento e a disseminação dos computadores, a informação compilada cresceu exponencialmente. Os analistas tiveram a oportunidade de organizar e analisar toda a informação que apareceu rapidamente. Os métodos tradicionais de análise dos dados, nos vários sectores (industrial, empresarial, financeiro, farmacêutico, etc), tornaram-se lentos e dispendiosos.

Em 1996, Fayyad et. al., afirmam que “é necessária uma nova geração de técnicas e ferramentas computacionais para extrair conhecimento útil de grandes bases de dados.” No trabalho de Liu et. al., (2002) é analisada a utilidade das aplicações de Data Mining. Segundo este autor as aplicações de KDD trouxeram benefícios significativos aos negócios, reduzindo custos, aumentando os lucros e elevando a qualidade dos serviços.

No início do século apresenta-se um cenário em que os dados que surgem apresentam elevadas dimensões, existe uma rápida construção de ferramentas de análise e os resultados são apresentados de forma perceptível ao utilizador de tal forma que este possa tirar o melhor proveito do conhecimento obtido. Na Figura 13 é possível observar a evolução das técnicas de Extracção de Conhecimento para acompanhar as necessidades crescentes de tratamento de informação.

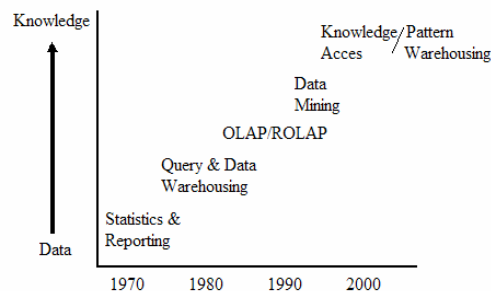
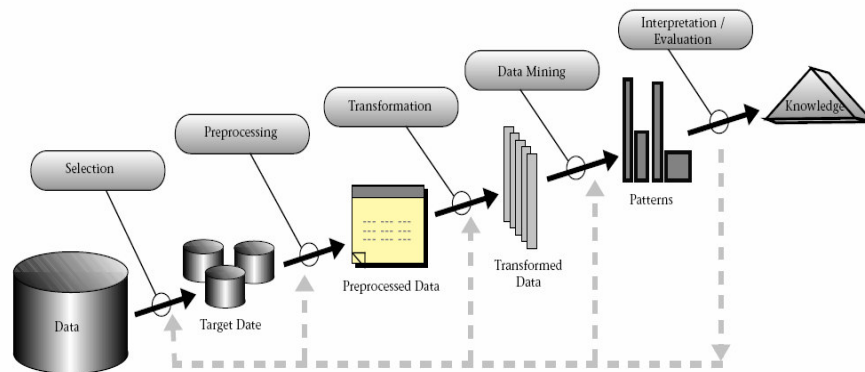


Figura 13 – Evolução das técnicas de Análise de Dados

Para analisar os dados em estudo nesta dissertação foram aplicadas técnicas de *Data Mining* que permitam extrair conhecimento útil, isto é, conhecimento que contribua para o avanço em direcção ao diagnóstico ou tratamento das doenças relacionadas com o deficiente enrolamento da TTR.

3.1 Processo de Extração de Conhecimento

O processo da extração de conhecimento é um processo interactivo e iterativo que envolve vários passos. (Fayyad et. al., 1996) Na Figura 14 podem observar-se as etapas seguidas para a obtenção de conhecimento.



Fonte: Adaptado de Fayyad et. al., 1996(2).

Figura 14 – “O caminho para o conhecimento”.

Segundo Fayyad et. al., (1996) o processo de obtenção de conhecimento começa com a compreensão do problema em estudo, dos dados. Segue-se uma fase de preparação dos dados, designada por “*Preprocessing*”, onde os dados em estudo são tratados, isto é, são aplicadas operações básicas, por exemplo, extração de ruído ou *outliers*. Os dados em estudo podem ser transformados ou reduzidos para, por exemplo, diminuir o número de variáveis em consideração (“*Transformation*”).

Na etapa de “*Data Mining*” decide-se qual a função de *data mining* a aplicar aos dados. Define-se o objectivo do modelo obtido pelo algoritmo de *data mining* e selecciona-se o método e os parâmetros apropriados para o problema em estudo. Os resultados obtidos são apresentados com formas de representação próprias para cada tipo de método.

A etapa “*Interpretation/Evaluation*” inclui interpretar os resultados obtidos, eliminar informação redundante, traduzir os resultados em linguagem acessível aos leitores e possivelmente voltar a etapas anteriores (pode, por exemplo, chegar-se à fase de avaliação e decidir-se voltar a transformar os dados para aplicar um outro método).

3.1.1 *Data Mining* – Tarefas

Ao longo do Processo de Extracção de Conhecimento de dados uma das etapas que importa destacar é a etapa de “*Data Mining*”. Consoante o tipo de dados em estudo e o objectivo da sua análise, podem definir-se diferentes tarefas de *Data Mining*, as quais podem ser classificadas em duas categorias:

- *descriptive data mining*: consiste em descrever e apresentar características gerais dos dados;
- *predictive data mining*: consiste na construção de modelos, realização de inferências e/ou predição do comportamento de novos conjuntos de dados.

As principais tarefas de Data Mining são *Classificação*, *Clustering*, *Regressão*, *Previsão*, *Associação* (Berry, 2000). Na Tabela 3 são apresentadas estas tarefas, e algumas das técnicas mais populares em Data Mining, que foram definidas em função do objectivo do processo de extracção de conhecimento.

Tarefa	Técnicas	Definição/Objectivo
Classificação	Árvores de Decisão; Regras de Classificação	Classificar novos indivíduos/variáveis a partir da classificação de indivíduos/variáveis dos dados de treino.
Clustering	Classificação hierárquica; K-means	Definir grupos nos indivíduos/variáveis dos dados baseados numa métrica de semelhança.
Regressão	Regressão Linear; Regressão Local; Árvores de Regressão; Redes Neurais; Algoritmos Genéticos	Desenvolver um modelo que descreva o comportamento dos indivíduos/variáveis em estudo.
Previsão	Regressão Linear; Regressão Local; Árvores de Regressão; Redes Neurais; Algoritmos Genéticos; Árvores de Decisão	Prever o valor de um novo item a partir do modelo obtido.
Associação	Regras de Associação	Descobrir relações de associação entre um conjunto de itens.

Tabela 3 – Tarefas de Data Mining

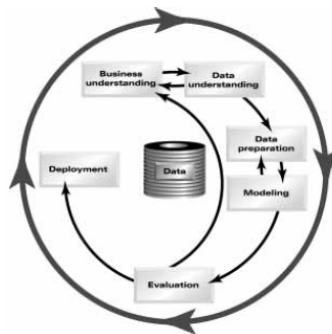
3.1.2 Metodologia

A falta de metodologia na obtenção do conhecimento pode ser um entrave à obtenção do mesmo, e para resolver este problema, no ano 2000, foi apresentado o modelo CRISP-DM (Cross-Industry Standard Process for Data Mining), (Shearer, 2000) que pretende ser um modelo standard, isto é, pretende-se que a partir deste modelo o utilizador apresente o processo de KDD para o seu problema.

Segundo a metodologia CRISP-DM a implementação de um processo KDD pode ser desenvolvido ao longo de seis etapas.

Como mostra a Figura 15 as fases do modelo CRISP-DM são as seguintes:

- **Compreensão da actividade (*Business Understanding*)**: esclarecer os objectivos e os requisitos sob o ponto de vista do negócio; é definido um plano para atingir os objectivos.
- **Compreensão dos dados (*Data Understanding*)**: como o nome indica, procura-se entender os dados, isto é, pretende-se que nesta fase se encontrem problemas nos dados, subconjuntos interessantes e problemas de qualidade nos dados.
- **Preparação dos Dados (*Data Preparation*)**: das tarefas de preparação de dados fazem parte selecção de tabelas, registos e atributos, transformação e limpeza de dados.
- **Modelação (*Modeling*)**: aplicação de modelos aos dados e refinação de parametros. Por vezes é necessário adaptar os dados aos modelos que se pretendem aplicar, logo é frequentemente necessário voltar à fase de preparação dos dados.
- **Avaliação (*Evaluation*)**: avaliação do modelo aplicado para verificar se este atinge os objectivos do negócio.
- **Acção/ Produção (*Deployment*)**: utilização e aplicação do modelo, aqui o analista pode, por exemplo, recomendar acções mediante os resultados obtidos.



Fonte: (Shearer, 2000)

Figura 15 - Modelo CRISP-DM

Para desenvolver o processo de Extração de Conhecimento desta dissertação, procurou-se seguir uma linha de raciocínio tendo em conta o processo descrito por Fayyad et. al.,(1996) e as fases da metodologia CRISP-DM.

3.2 *Análise de clusters*

Na etapa de *Data Mining*, do processo KDD desta dissertação, a primeira técnica utilizada para compreender e encontrar semelhanças entre os aminoácidos da TTR foi a análise de clusters. A análise de *clusters* é uma técnica multivariada muito usada para analisar sequências genéticas e pode ser usada em vários tipos de dados: dados genéticos, dados temporais, dados geográficos, etc.

Saporta (1990) sustenta que o objectivo dos métodos de classificação é construir uma partição de um conjunto de objectos dos quais são conhecidas as distâncias dois a dois. Os clusters obtidos devem ser o mais homogéneos possível e distintos entre si.

Nesta tese pretende-se aplicar a análise de *clusters* para explorar os dados da proteína TTR, para reduzir a dimensão dos dados e para gerar hipóteses. Estes objectivos vão de encontro com os objectivos gerais da análise de *clusters* definidos no trabalho de Branco (2004).

Objectivos da análise de clusters:

- a. **Exploração dos dados:** este é o objectivo mais imediato da análise de *clusters*, isto é, encontrar uma estrutura/partição nos dados.
- b. **Redução de dados:** Se os grupos forem homogéneos significa que cada grupo pode ser substituído pelo seu representante.
- c. **Geração de hipóteses:** Podem ser encontrados grupos não esperados, o que pode eventualmente levar a conjecturas e hipóteses para explicar a classe encontrada.
- d. **Predição:** Os elementos do mesmo grupo tendem a ter características semelhantes, este facto torna possível prever propriedades de um outro elemento quando comparado com os elementos de um grupo.

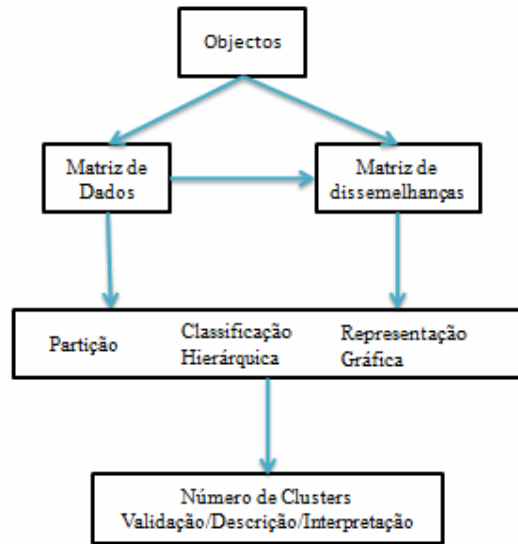
Para obter as classes finais são seguidos alguns passos importantes da análise de clusters, na Figura 16 são apresentadas as fases de uma análise de clusters adaptadas do trabalho de Gordon (1999)

1. Selecção de Objectos
2. Selecção de Variáveis
3. Transformação de Variáveis

4. Construção da medida de dissimilaridade/semelhança

5. Escolha do método a aplicar aos dados

6. Discussão e apresentação dos resultados



Fonte: Adaptado de Gordon (1999)

Figura 16 - Fases de uma Análise de Clusters.

A análise de clusters tem como ponto de partida um quadro de dados inicial ou um quadro de proximidades.

- (i) O quadro de dados, geralmente, é da forma “indivíduos vs variáveis”.

$$X = \begin{matrix} & \begin{matrix} 1 & \dots & j & \dots & p \end{matrix} \\ \begin{matrix} 1 \\ \vdots \\ i \\ \vdots \\ n \end{matrix} & \begin{bmatrix} & & & & \\ & & & & \\ & & & & \\ & & x_i^j & & \\ & & & & \\ & & & & \end{bmatrix} \end{matrix}$$

Dado o conjunto E de elementos a classificar, o elemento x_i^j é o valor que o indivíduo $j \in E$ assume na variável i .

As variáveis podem ser classificadas consoante as suas características. Para as variáveis podem ser definidos os seguintes tipos:

- *Variáveis numéricas ou quantitativas* – o conjunto de chegada da aplicação associada a cada variável aleatória é $\mathfrak{R}$, as variáveis aleatórias induzem preordens.
- *Variáveis nominais ou ordinais* – a cada variável aleatória está associado um conjunto de estados, ou modalidades, e a um objecto corresponde um e um só estado de cada variável.
- *Variáveis binárias* – é um caso particular das variáveis nominais ou ordinais, que ocorre quando o número de estados possíveis de uma variável é igual a dois.

- (ii) O quadro de proximidades é também designado por função de comparação. Como refere Sousa, (2000) define-se função de comparação entre elementos de E como uma aplicação $\gamma: E \times E \rightarrow \mathfrak{R}_0^+$. Esta função pode ser do tipo semelhança, ou de dissemelhança, entre pares de objectos, ou variáveis, permitindo quantificar o seu estado de semelhança ou de dissemelhança.

Uma dissemelhança satisfaz as seguintes propriedades: (Gordon, 1999:15)

1. $d(x, y) = d(y, x), \forall x, y \in E$
2. $d(x, x) = 0, \forall x \in E$
3. $d(x, y) = 0 \Rightarrow x = y$

Uma semelhança satisfaz as seguintes propriedades: (Saporta, 1990)

1. $s_{ij} = s_{ji}, \forall i, j$
2. $s_{ij} \geq 0, \forall i, j$
3. s_{ij} é tanto maior quanto maior for a semelhança entre objectos.

Em muitos estudos a medida de proximidade mais fácil de calcular é a semelhança entre objectos. Em geral, pode obter-se uma dissemelhança d_{ij} a partir da semelhança s_{ij} , estabelecendo uma relação entre a semelhança e a dissemelhança dos mesmos objectos, por exemplo, $d_{ij} = k - s_{ij}$ (k constante adequada). (Gordon, 1999:15)

Para cada tipo de dados existem várias medidas de proximidade definidas e propostas por diversos autores nomeadamente Gordon (1999), Hartigan (1975), Saporta (1990).

3.2.1 Clustering hierárquico

O clustering hierárquico consiste em obter um conjunto de partições de E em classes cada vez menos finas obtidas a partir de reagrupamentos sucessivos de partes. Um clustering hierárquico representa-se por um *dendrograma* ou *árvore de classificação*.

Existem dois tipos de métodos de clustering hierárquico: os métodos ascendentes ou aglomerativos e os descendentes ou divisíveis.

O **método ascendente, ou aglomerativo**, considera, inicialmente, cada indivíduo como sendo uma classe e vai agrupando-os de acordo com as suas proximidades. O algoritmo pára quando se obtiver uma única classe.

O **método descendente, ou divisível**, consiste em aplicar dicotomias sucessivas ao conjunto E até se obter os elementos todos isolados.

A. Método Aglomerativo

O método aglomerativo é um método muito usado, sendo este o método utilizado nesta dissertação. Os passos seguidos num algoritmo de Classificação Hierárquica Ascendente (CHA) como refere Sousa (2000) são:

Passo 0: Considerar os objectos iniciais (*conjunto E*), isto é, definir a partição mais fina. Definir a função de comparação entre os objectos e entre classes.

Passo i : ($i = 1, \dots, k$)

Determinar o par ou pares de classes P_{i-1} que verificam o critério de agregação, isto é, obter os pares que verificam o mínimo de $\{d(A, B) : A, B \in P_{i-1}\}$.

Actualizar a partição, $P_i = \{P_{i-1} \setminus (\{A\} \cup \{B\})\} \cup \{A \cup B\}$.

Actualizar os valores da matriz de proximidades entre partes de E .

Uma abordagem um pouco diferente do algoritmo é apresentada em Hartigan (1975).

No algoritmo de CHA podem ser introduzidos critérios de paragem, por exemplo, impor o número de classes da última partição, ou limitar o valor da função de comparação entre partes.

Ao longo do algoritmo é necessário calcular as distâncias entre as classes que se vão formando e é neste ponto que os algoritmos aglomerativos diferem. Os métodos mais usados são: Ligação Simples (*Single linkage*), Ligação Completa (*Complete linkage*), Ligação Média (*Average linkage*), Método centróide e finalmente o Método de Ward.

B. Critérios de Agregação

Neste trabalho são usados o método da Ligação Média e o método de Ward.

Ligação Média

No método da Ligação Média a distância entre classes é distância média dos pares de elementos.

$$d_{AB} = \frac{\sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d_{ij}}{n_A n_B}$$

Esta ideia de distância é simples, este método foi um dos primeiros a ser construído, é um método robusto adequado para isolar classes de forma esférica ou elipsoidal. Este

método actua sobre a matriz de proximidades, enquanto que o método seguinte é mais complexo e actua sobre a matriz de dados.

Método de Ward

No método de Ward (Ward, 1963) as classes são formadas de modo a minimizar a soma dos quadrados dos erros. Segundo Saporta (1990) este é o critério de agregação mais indicado para distâncias euclidianas. Os elementos de E são considerados como uma nuvem de pontos num espaço $\Re^p$ e o critério de agregação consiste em minimizar a variação de inércia interclasse. Como os dados em estudo, neste trabalho, são as distâncias dos aminoácidos ao centro de massa da TTR, faz todo o sentido utilizar este critério de agregação.

Se se considerarem dois clusters A e B, este método minimiza o incremento da soma dos quadrados, ou seja, minimiza a perda de informação resultante da aglutinação.

Seja $SW_C - (SW_A + SW_B)$ o incremento da soma dos quadrados quando A e B são

aglutinados, onde $SW_A = \sum_{i \in A} \sum_{j=1}^p (x_{ijA} - \bar{x}_{jA})^2$ é a soma dos quadrados dos erros dentro do

cluster A, $SW_B = \sum_{i \in B} \sum_{j=1}^p (x_{ijB} - \bar{x}_{jB})^2$ é a soma dos quadrados dos erros dentro do cluster

B e $SW_C = \sum_{i \in C} \sum_{j=1}^p (x_{ijC} - \bar{x}_{jC})^2$ é a soma dos quadrados do cluster $C = A \cup B$.

Notação: x_{ijA} (x_{ijB}) é a observação do objecto i do cluster A (B) na variável j , $\bar{x}_{jA}$ e $\bar{x}_{jB}$ são as médias da variável j nos clusters A e B . (Branco, 2004)

Em cada iteração o algoritmo calcula todos os pares possíveis de clusters e calcula o valor do incremento para cada união. O algoritmo selecciona os pares que minimizam o valor do incremento, isto é, escolhe os pares que minimizam a perda de informação.

Após obtenção das classes a questão que se coloca é se os *clusters* obtidos nos fornecem a melhor partição do espaço.

Para algoritmos hierárquicos alguns autores recorrem a métodos gráficos. Segundo Maroco (2003) através do gráfico do índice de difusão versus número de *clusters* é possível escolher o número de *clusters*. Grandes alterações no nível de fusão correspondem à aglutinação de grupos muito diferentes, logo estas alterações podem indicar o número ideal de *clusters*.

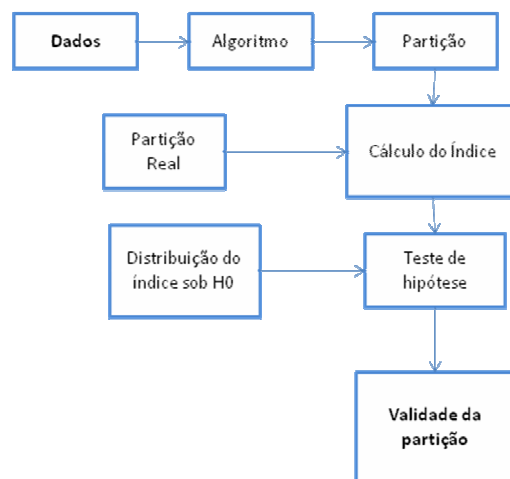
De acordo com Pal et. al., (1995), citado por Silva (2005), se o algoritmo de classificação tiver como único parâmetro de entrada o número de classes, faz-se variar de 2 até ao valor máximo $k_{\max}$, podendo considerar-se $k_{\max} \leq \sqrt{n}$.

No ponto seguinte são abordadas algumas medidas que permitem a escolha do número de *clusters*.

C. Escolha do número de clusters

Segundo Gordon (1999), Halkidi (2001), Jakel (2004) e Kovacs (2003) as medidas de validação de *clusters* podem ser agrupadas segundo 3 tipos de critérios. Mediante a estratégia utilizada para validar a partição obtida temos:

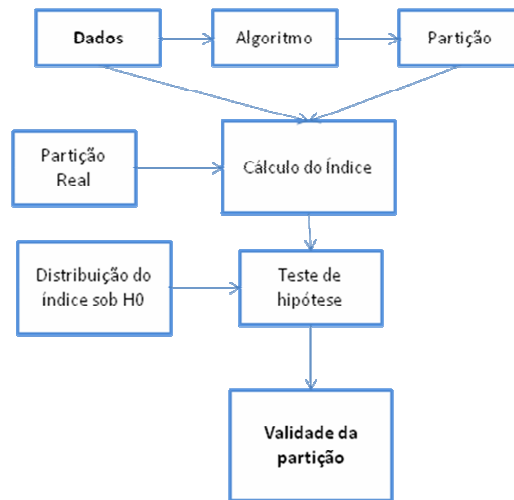
- **CrITÉrios Externos:** avaliam-se os resultados mediante informação externa fornecida a priori e não utilizada na análise. Esta análise é baseada na hipótese nula de aleatoriedade dos dados, para testar esta hipótese são utilizados testes estatísticos. Pode usar-se, por exemplo, técnicas de Monte Carlo para validar a análise de *clusters*.



Adaptado de Faceli et. al., (2005)

Figura 17 - Critério de Validação Externo

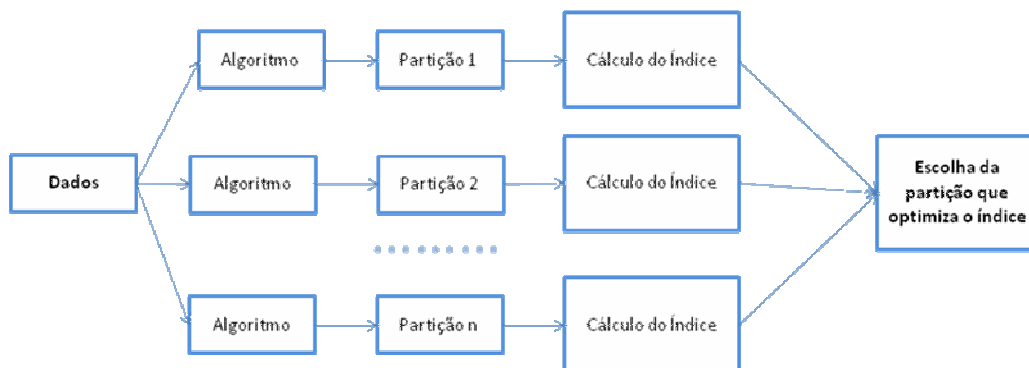
- **Critérios Internos:** Medem a qualidade do clustering com base nos dados originais (matriz de proximidades). Neste critério são usadas medidas que se baseiam em quantidades e características inerentes aos dados e ao esquema de classificação. (Halkidi, 2001)



Adaptado de Faceli et. al., (2005)

Figura 18 - Critério de Validação Interno

- **Critérios Relativos:** comparam diferentes estruturas, construídas a partir dos mesmos objectos, para encontrar o melhor clustering a partir de diferentes resultados. Pode ainda ser usado para comparar diversos algoritmos de agrupamento.



Adaptado de Faceli et. al., (2005)

Figura 19 - Critério de Validação Relativo

A escolha da melhor partição dos objectos em estudo é uma questão analisada por diversos autores e têm sido apresentados muitos métodos de escolha do número de classes. Halkidi et. al., (2000) sustentam que os critérios externos e internos são baseados em métodos estatísticos mas têm como desvantagem a elevada complexidade computacional.

A forma como o índice é aplicado é que define se o critério de utilização do índice é interno ou relativo. A aplicação de um índice com um critério relativo consiste no cálculo do seu valor para vários clusterings, o “melhor” clustering vai ser o que optimiza o índice. Para efectuar o clustering dos aminoácidos da TTR é necessário definir como escolher o número de classes.

Alguns dos índices mais usados com o critério relativo são os seguintes:

- Estatística de Hubert (Jain e Dubes, 1988)
- Família da Índices de Dunn (Halkidi et. al., 2001)
- Índice de Davies-Bouldin (Jain e Dubes, 1988)
- Estatística de Silhouettes (Rousseeuw, 1987)
- Índice de Calinski-Harabasz (Calinski e Harabasz, 1974)
- Índice de Krzanowski e Lai (Krzanowski e Lai, 1985)

“Índices de Milligan e Cooper”

Devido ao elevado número de índices e regras de paragem disponíveis, em 1985, Milligan e Cooper apresentaram um estudo comparativo detalhado de 30 índices. Neste estudo, através de uma avaliação de Monte Carlo são analisados os índices em *data sets* artificiais, foram aplicados quatro algoritmos de agrupamento hierárquico e através do critério externo encontraram-se os índices que recuperavam a verdadeira estrutura de grupo. Neste estudo foram escolhidos 5 índices, e desde então, estes índices passaram a ser designados na literatura por “Índices de Milligan e Cooper”.

Segundo Milligan & Cooper (1985) os “melhores” índices são os seguintes:

– **Goodman & Kruskal (G2)**

Segundo Gordon (1999) este índice é muito utilizado em estudos de classificação. Após obtenção dos *clusters* são feitas comparações entre as dissemelhanças *intraclusters* e *interclusters*. A comparação diz-se concordante (respectivamente discordante) se a dissemelhança dentro dos *clusters* é mais baixa (resp. elevada) que a dissemelhança *interclusters*. O índice é então dado por:

$$G_2 = \frac{s(+)-s(-)}{s(+)+s(-)}$$

onde $s(+)$ e $s(-)$ representam o número de pares concordantes e discordantes respectivamente, envolvendo os valores das matrizes de dissemelhança e ultramétricas. O valor máximo desta medida indica o número de *clusters* a reter.

Em Gordon (1999) este índice tem a designação de “G2” sendo esta a mesma designação utilizada na biblioteca “fpc” do R (Hennig, 2004), que é utilizado no capítulo 4.

– **Índice-C (G3)**

Este índice é dado por:

$$\frac{D(c)-D_{\min}}{D_{\max}-D_{\min}}$$

Onde D é a soma das dissemelhanças intra-classes. O valor mínimo desta medida indica o número de classes óptimo. Este índice é definido em Gordon (1999) como sendo o índice $G3$ e esta mesma nomenclatura é usada na biblioteca “fpc” do R (Hennig, 2004), que é utilizado no capítulo 4.

- **Índice de Calinski & Harabasz** definido por:

$$CH(k) = \frac{\frac{B(k)}{(k-1)}}{\frac{W(k)}{(n-k)}}$$

Onde $W(k) = \sum_{l=1}^k \sum_{x \in C_r} d^2(x, g_r)$ e $B(k) = \sum_{i,j=1, i \neq j}^k dist^2(C_i, C_j)$ sendo g_r o centro de gravidade da classe r e $dist(C_i, C_j)$ é a dissimilaridade entre classes. Escolhe-se o número de classes que maximiza este índice.

- **Índice de Duda & Hart:**

Segundo Silva (2005) este índice é baseado na razão entre a soma dos quadrados das dissimilaridades intraclasse das duas classes candidatas a serem reunidas (W_2), e a soma dos quadrados das dissimilaridades da classe resultante da reunião (W_1). A hipótese de uma classe é rejeitada, ou seja, as duas classes candidatas não se reúnem, se a razão for menor que um dado parâmetro.

$$W_2 / W_1 < 1 - 2/(\pi p) - z[2(1 - 8/(\pi^2 p))/(mp)]^{1/2}$$

- **Índice de Beale** definido por:

$$F \equiv \left(\frac{W_1 - W_2}{W_2} \right) / \left(\left(\frac{m-1}{m-2} \right) 2^{2/p} - 1 \right)$$

Se o valor de F for significativamente elevado rejeita-se a hipótese nula (isto é, não se divide o cluster). F segue uma distribuição $F_{p, (m-2)p}$. (Gordon, 1999)

Este índice aplica-se a métodos hierárquicos descendentes (Silva, 2005).

Uma descrição detalhada destes Índices pode ser encontrada em Gordon (1999) e Silva (2005).

Segundo Sarle (1993) os critérios escolhidos por Milligan e Cooper são apropriados apenas para *clusters* compactos ou levemente alongados, de preferência aproximadamente multivariados normais.

Em 2005, Bertrand e Bell Mufti definiram um índice de estabilidade estatístico da partição (designado por BB) e compraram-no com dois índices de Milligan e Cooper, chegam à conclusão que o índice BB, por eles proposto, para os dois quadros de dados utilizados no estudo, fornece melhores resultados.

Nesta dissertação foram utilizadas medidas de validação segundo o critério relativo. Devido à sua elevada credibilidade, para a escolha da “melhor” partição dos aminoácidos, foram calculados e otimizados os índices G2 e G3 de Milligan e Cooper, otiveram-se grupos de aminoácidos e foi possível comparar e analisar os clusters obtidos, como será mostrado no capítulo 4.2 *Clustering hierárquico*.

3.3 *Eventos*

Um dos pontos que se pretende desenvolver, neste trabalho, é encontrar características e acontecimentos interessantes nas séries temporais dos aminoácidos em estudo. Desta forma, estando em estudo um conjunto de dados volumoso, ao nível de uma propriedade dos aminoácidos, pretende-se obter características das simulações para obter conhecimento a outro nível.

No estudo de séries temporais são, por vezes, procuradas alterações do comportamento das séries e acontecimentos raros dentro do contexto em análise. Nesta dissertação procura-se, a partir das séries que descrevem o comportamento dos aminoácidos ao longo da desnaturação, encontrar eventos, isto é, encontrar uma ocorrência ou uma eventualidade no comportamento do aminoácido que se destaque relativamente ao seu comportamento geral ao longo da simulação.

Mannila et. al.,(1995,1997) definem uma **sequência de eventos** como uma sequência de acções de utilizadores num sistema e definem um **episódio** como sendo uma colecção de eventos que ocorrem relativamente próximos uns dos outros numa dada ordem.

Estes autores consideraram três tipos de episódios como mostra a figura seguinte:

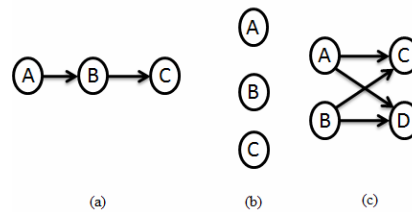


Figura 20 - Tipos de Episódios

- (a) **Episódio sequencial:** sequência de eventos, em que entre os eventos (A, B, C) podem ocorrer outros eventos.
- (b) **Episódio em paralelo:** neste caso não existem exigências quanto à ordem dos três eventos.
- (c) **Episódio não-sequencial e não paralelo:** Neste episódio ocorrem os eventos A e B que precedem os eventos C e D, não existe ordem entre A e B (ou C e D), as quatro ocorrências devem ser próximas.

Um problema básico em analisar sequências de eventos está em encontrar episódios frequentes, isto é, dada uma classe de episódios e uma sequência de eventos, o problema consiste em encontrar todos os episódios que acontecem frequentemente numa dada sequência.

Nos trabalhos de Mannila e Toivonen (1995,1997) é considerado que um evento está sempre associado ao tempo de ocorrer, logo uma sequência de eventos é um conjunto do tipo $S = \langle (A_1, t_1), \dots, (A_n, t_n) \rangle$ em que A_i pertence ao conjunto de eventos E_0 (conjunto de eventos tipo pré-definido) e t_i é o instante em que A_i ocorre.

Em 1999, Guralnik e Srivastava, apresentam uma definição mais abrangente de evento. Este autores definem um evento como sendo uma mudança significativamente qualitativa de um fenómeno dinâmico que mantém o comportamento alterado tempo suficiente para ser considerado um acontecimento raro.

No seu trabalho, Guralnik e Srivastava (1999) procuram identificar os pontos da escala temporal em que a mudança de comportamento ocorre. O chamado *problema da procura de “pontos de mudança” em séries temporais* consiste em determinar o número

de pontos de troca de comportamento e escolher uma função que encaixe entre os diferentes pontos de troca. Guralnik e Srivastava (1999) propõem um algoritmo para encontrar os “pontos de mudança”, isto é, pretendem encontrar eventos interessantes em séries temporais.

Nesta dissertação pretende-se procurar, nas séries temporais dos aminoácidos, mudanças de comportamento significativas ao longo de determinado tempo, isto é, procuram-se eventos frequentes e relevantes no comportamento dos aminoácidos da proteína TTR.

Depois de encontrar os eventos frequentes podem ser usadas regras para descrever relações entre os eventos e o tipo de proteína com maior propensão para ocorrência de um evento. Pretende-se ainda utilizar as relações obtidas para prever o tipo de proteína onde ocorre determinado evento.

3.4 Regras

Em 1998, Das et. al., apresentam um trabalho onde abordam o problema da procura de regras em séries temporais. Neste trabalho as séries temporais são discretizadas e o algoritmo calcula regras do tipo: se ocorre A então B ocorre no tempo T , $\left(A \xrightarrow{T} B \right)$.

Onde A e B são letras do alfabeto obtidas pela discretização das séries. Um trabalho análogo foi desenvolvido por Azevedo et. al., (2005) com valores de SASA discretizados e obtidos por simulação da desnaturação da TTR.

3.4.1 Regras de Associação

As regras de associação foram definidas por Agrawal, R., et al. em 1993, da seguinte forma:

Seja $I = \{i_1, \dots, i_m\}$ um conjunto de itens distintos, e D um conjunto de transacções, onde cada transacção T é um conjunto de itens tal que $T \subseteq I$. Associado a cada transacção está um identificador único, designado por TID. Uma Regra de Associação é uma implicação do tipo $X \rightarrow Y$, onde $X \subset I$, $Y \subset I$, e $X \cap Y = \emptyset$. (X é o

antecedente e Y é o consequente). Uma regra $X \Rightarrow Y$ tem confiança c se $c\%$ das transacções em D que contêm X também contêm Y , e suporte s se $s\%$ das transacções em D contêm $X \cup Y$.

O antecedente e o consequente de uma regra podem conter mais do que um item. Para medir o interesse de uma regra podem ser usadas medidas objectivas e medidas subjectivas. Segundo Silberschatz et.al., (1996) das medidas objectivas fazem parte, por exemplo, o Suporte e a Confiança.

O suporte mede a associação entre o antecedente X e o consequente Y , não relaciona possíveis dependências entre X e Y e é dado por:

$$\text{suporte} = P(X \cap Y)$$

A confiança é a frequência relativa da regra, isto é, é a frequência de ocorrência do consequente quando se verifica o antecedente.

$$\text{confiança} = P(Y | X) = \frac{P(X \cap Y)}{P(X)}$$

No entanto, o suporte pode não ser interessante porque suportes elevados podem resultar de padrões triviais. No caso de X e Y serem independentes a confiança da regra é $P(Y)$ e se $P(Y)$ for superior à confiança mínima obtem-se uma relação entre acontecimentos independentes! (Velo, 2003)

Devido a estes problemas com o suporte e com a confiança foram estudadas e definidas outras medidas. Na Tabela 4 são apresentadas três medidas de interesse de uma regra, por exemplo, o lift que permite avaliar o grau de dependência do consequente em relação ao antecedente.

Diversas medidas de interesse foram definidas e comparadas nos trabalhos de: Azevedo (2007), Bayardo (1999), Liu (2000), Lenca (2006), Hahsler (2007) e Omiecinski (2003).

Designação	Fórmula	Referência
Lift	$\frac{P(X \cap Y)}{P(X)P(Y)}$	Brin, 1997 (a)
Conviction	$\frac{P(X)P(\bar{Y})}{P(X \cap \bar{Y})}$	Brin, 1997 (b)
Ganascia	$2 * P(X Y) - 1$	Ganascia, 1991

Tabela 4 – Algumas Medidas de Interesse

Quanto à escolha da medida correcta para cada tipo de dados em Tan (2002) é apresentado um estudo comparativo de 20 medidas onde são descritas as propriedades das medidas e é apresentado um algoritmo de escolha de medidas. Neste estudo é ainda desmonstrado que não existem medidas consistentemente melhores que outras para todos os casos.

Em 1993 é apresentado o primeiro algoritmo para a procura de Regras de Associação, designado por AIS (Agrawal et. al., 1993). Desde o aparecimento deste algoritmo vários foram os métodos desenvolvidos para melhorar a procura de regras, como por exemplo:

- Apriori (AprioriTid, AprioriHybrid) (Agrawal, 1994);
- Partition Technique (Savasere, 1995);
- Sampling Technique (Toivinen, 1996 (a));
- Método DIC (Brin et. al., 1997);

Nesta dissertação é utilizado o programa CAREN (Azevedo, 2003) no qual as regras são geradas tendo em conta o Algoritmo Apriori (Agrawal et. al., 1994). Este algoritmo está dividido em duas partes:

- na primeira parte são extraídos os itens frequentes que têm suporte superior ao suporte mínimo fixado.

- na segunda são procuradas as regras que tenham um grau de confiança mínimo e tais que todas as combinações de itens tenham um suporte mínimo fixado à partida.

Uma descrição mais pormenorizada deste algoritmos pode ser vista em Agrawal e Srikant (1994) e Azevedo (2003). Trabalhos mais recentes tentam desenvolver e aplicar algoritmos de regras de associação específicos para dados temporais (Mennis and Liu 2003; Sarker et al. 2003; Dafas e Garcez 2005).

4 Processo de Extração de Conhecimento de dados de simulação de Desnaturação Proteica

Com o objectivo de encontrar características relevantes do processo de desenrolamento da proteína TTR foi desenvolvido um processo de extração de conhecimento, dos dados obtidos por simulação do desenrolamento da TTR, que é apresentado neste capítulo.

Numa primeira abordagem, efectuou-se uma análise exploratória dos dados de 10 simulações para visualizar o comportamento geral das séries. Depois partiu-se para uma análise multivariada. Como em estudo estão duas variantes de TTR (a *wild-type* e a variante patogénica L55P) obtiveram-se duas partições dos aminoácidos, uma partição que tem em conta a informação contida nos dados das simulações da variante WT-TTR (partição I) e uma segunda partição que tem em conta a informação obtida por simulação da desnaturação da variante L55P-TTR (partição II).

De seguida, desenvolveu-se um processo mais elaborado para, a partir de todos os dados obtidos por simulação (da variante WT-TTR e da variante L55P-TTR), escolher aminoácidos que representem o comportamento da proteína TTR nas 10 simulações. Isto é, a partir dos dados obtidos por simulação dos 127 aminoácidos obteve-se uma partição destes que considera toda a informação das 10 simulações (partição consenso – partição que se obtém a partir de várias partições do mesmo conjunto de objectos). Para esta partição foi calculado o aminoácido representativo de cada cluster e reduziu-se a dimensão dos dados em estudo de 127 aminoácidos para 15 aminoácidos representativos.

Na fase de procura de eventos foram comparados os aminoácidos representativos dois a dois, procuraram-se dois tipos eventos diferentes ao longo das 10 simulações e obtiveram-se frequências de ocorrência dos dois eventos. Esta fase foi ainda complementada com a procura de um terceiro evento nas 10 simulações, a procura deste evento foi efectuada ao longo das séries dos 127 aminoácidos.

Após obter as ocorrências de cada evento foram procuradas regras de associação que associam a ocorrência do evento ao tipo de proteína.

Na Figura 21 é apresentado esquematicamente todo o processo de extração de conhecimento desenvolvido.

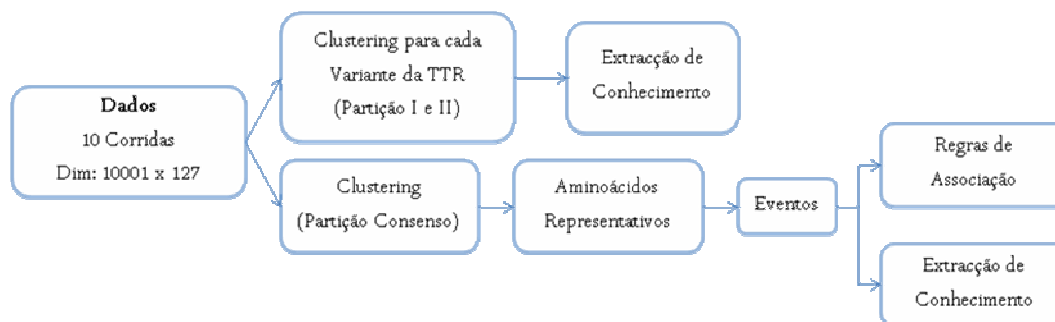


Figura 21-Processo de Extracção de Conhecimento

4.1 Descrição dos dados

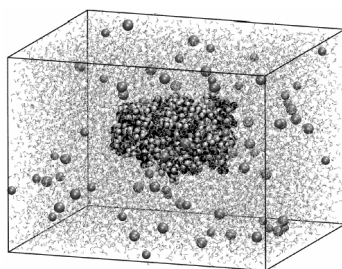
Os dados em estudo foram obtidos a partir de simulações da desnaturação da proteína TTR e foram fornecidos por Rui Brito cujo trabalho “*Protein Folding and Unfolding Simulations: A New Challenge for Data Mining*” foi o ponto de partida desta tese.

Os valores da Variação da Distância do Carbono Alpha de cada resíduo ao Centro de Massa da proteína foram obtidos ao longo de simulações de Dinâmica Molecular do Desdobramento Proteico (MDPUS) da proteína TTR (Brito et. al., 2004).

Os passos seguidos foram os seguintes:

- Obtiveram-se as coordenadas da proteína TTR a partir da sua estrutura cristalográfica e adicionam-se átomos de hidrogénio. Representaram-se todos os átomos explicitamente (1917 átomos da proteína).
- As moléculas de água foram controladas e introduzidas através dos programas Dowser (Zhang, 1996) e Solvate (Grubmiller, 1996). Com o programa Solvate foram também introduzidos 145 mM de iões de NaCl à volta de proteína.

O sistema final, ilustrado na Figura 22, é constituído por 1917 átomos da proteína, 14418 moléculas de água, 45 iões de sódio e 40 iões de cloro, no total o sistema possui 45256 átomos.



Fonte: Brito (2004)

Figura 22 - Representação do sistema usado nas simulações de Dinâmica Molecular

O sistema foi elevado até à temperatura desejada sob a dinâmica de Langevin (dinâmica que tenta aproximar os mecanismos naturais através de modelos simples).

Os parâmetros usados ao longo das simulações foram de 1 atm de pressão constante (com a aplicação do método de Langevin (Feller et al., 1995)) e volume constante. Na simulação de controlo o sistema foi sujeito a uma temperatura de 310 K e nas restantes simulações a temperatura foi de 500 K (Brito, 2004).

Neste estudo foram simuladas várias desnaturações proteicas da proteína TTR no seu estado natural WT-TTR e da proteína mutada L55P-TTR. Os valores da Distância do Carbono Alpha da cada resíduo ao Centro de Massa foram registados e em estudo temos 5 quadros de dados com valores da WT-TTR (WTRUN2, WTRUN3, WTRUN4, WTRUN5, WTRUN6) e 5 quadros de dados com valores da L55P-TTR. (L55PRUN1, L55PRUN2, L55PRUN3, L55PRUN4, L55PRUN5).

4.1.1 Estudo Preliminar

Seja n é o número de frames e p o número de aminoácidos em estudo, o quadro de dados que se obtêm contém $n \times p$ observações e tem n linhas e p colunas. Cada coluna é considerada como um ponto no espaço de dimensão n munido de uma estrutura euclideana. As coordenadas de cada ponto são obtidas a partir dos valores dos p aminoácidos da observação correspondente.

$$X = \begin{matrix} & \begin{matrix} 1 & \dots & j & \dots & p \end{matrix} \\ \begin{matrix} 1 \\ \vdots \\ i \\ \vdots \\ n \end{matrix} & \begin{bmatrix} & & & & \\ & & & & \\ & & & & \\ & & x_i^j & & \\ & & & & \\ & & & & \end{bmatrix} \end{matrix}$$

Neste trabalho a matriz de dados tem 127 aminoácidos (43% hidrofóbicos) e 10001 registos, onde x_i^j é o valor que o aminoácido j assume na frame i . As estruturas são guardadas a cada picosegundo, isto é, 10000 em 10 nanosegundos de simulação mais a estrutura inicial (tempo 0).

Numa primeira análise efectuou-se um estudo unidimensional dos aminoácidos com o objectivo de analisar o seu comportamento ao longo da simulação da desnaturação da TTR.

Na Tabela 5 são apresentados os valores máximos e mínimos atingidos nas dez simulações.

Tabela 5 - Máximos e mínimos da distância do carbono alpha ao centro de massa

Corrida	Mínimo($\text{\AA}$)	Aminoácido	Tempo(ps)	Máximo($\text{\AA}$)	Aminoácido	Tempo(ps)
WTRUN 2	0,183	VAL_71	7023	43,284	GLY_1	6370
WTRUN 3	1,487	VAL_71	9891	40,44	GLY_1	1592
WTRUN 4	0,953	LEU_55	8678	40,053	GLY_1	1996
WTRUN 5	0,54	CYS_10 (*)	8468	50,332	GLU_127	6516
WTRUN 6	1,723	VAL_32	8096	40,139	GLY_1	7077
L55PRUN 1	0,139	ALA_37	9559	45,725	GLU_127	7686
L55PRUN 2	0,14	CYS_10	9049	38,376	GLU_127	9544
L55PRUN 3	1,02	VAL_71	9851	40,393	GLU_127	4085
L55PRUN 4	1,256	VAL_32	3687	38,678	GLY_1	6794
L55PRUN 5	1,14	VAL_14	4015	37,794	GLY_1	9824

(*) nesta corrida o aa. VAL_71 atinge o valor 0,547 $\text{\AA}$

Quando se observa um filme da simulação da desnaturação proteica verifica-se, visualmente, que as pontas da cadeia polipeptídica parecem perdidas no espaço, isto é, entram e saem do “novelo” proteico como se fossem elementos esquecidos pela proteína.

Quando se analisam os dados da simulação verifica-se, como era de esperar, que os valores máximos são atingidos pelos aminoácidos das regiões terminais da cadeia polipeptídica. Na WT-TTR é a GLY_1 (hidrofóbico) que atinge mais vezes o valor máximo enquanto que na L55P-TTR é a GLU_127. Os valores máximos na WT-TTR são, nestas simulações, sempre superiores a 40.

Quanto aos aminoácidos que atingem o valor mínimo, na WT-TTR temos a VAL_71 que se destaca como sendo o aminoácido que atinge mais vezes o valor mínimo de distância ao centro de massa. Na Figura 23 é possível observar o comportamento dos aminoácidos GLY_1 e VAL_71 ao longo da desnaturação da TTR

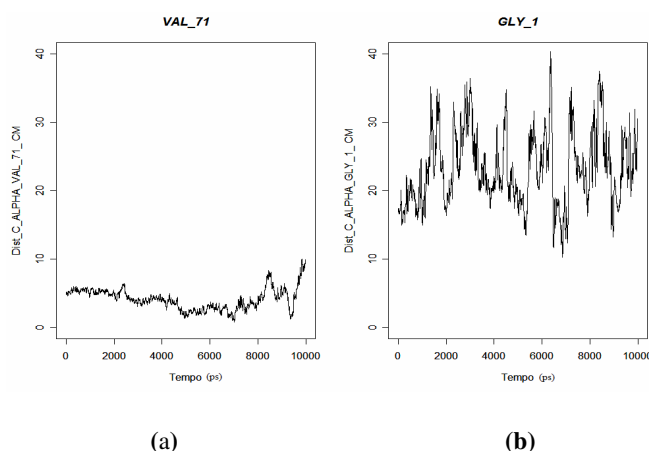


Figura 23 – Variação da distância do carbono alpha dos aminoácidos VAL_71 (a) e GLY_1 (b) ao centro de massa da TTR ao longo de uma simulação de desnaturação da variante WT-TTR (WTRUN2).

Ao longo da simulação WTRUN2, a GLY_1 mantém valores sempre superiores a 10 e apresenta um comportamento irregular, com frequentes afastamentos e aproximações ao centro de massa. A Val_71 apresenta sempre valores inferiores a 10⁰Å, o que indica que este resíduo se mantém sempre muito próximo do centro de massa ao longo da simulação.

No *Anexo I -A* estão representados os gráficos com a localização de cada aminoácido relativamente ao centro de massa nos instantes em que são atingidos os máximos e mínimos para cada corrida.

Quanto aos instantes, em que são atingidos os valores extremos, importa referir o seguinte:

- na WT-TTR o valor mínimo é sempre atingido depois dos 7000 ps, enquanto que na L55P-TTR existem duas simulações onde o mínimo é atingido mais cedo (entre os 3500 ps e 4550 ps);

- na L55P-TTR o valor máximo é atingido depois dos 6500 ps, em 4 das 5 corridas, enquanto que na WT-TTR o máximo é sempre atingido antes dos 7100 ps, sendo que em duas corridas a GLY_1 afasta-se do centro de massa no início da corrida (1592 ps e 1996 ps).

No *Anexo I-B* pode ainda observar-se o comportamento do aminoácido da posição 55 nas duas variantes da proteína. Na L55P-TTR, em 3 das 5 corridas a PRO_55, tende a afastar-se do Centro de Massa (toma valores superiores a $20^{\circ}\text{\AA}$). Na WT-TTR, em 3 das 5 corridas a LEU_55 mantêm valores no intervalo $[0,10]^{\circ}\text{\AA}$.

4.2 *Clustering hierárquico*

Uma forma de analisar grandes volumes de dados consiste em classificá-los em grupos que revelem padrões entre si. No trabalho “*A Closer Look on Protein Unfolding Simulations through Hierarchical Clustering*” (Ferreira et al., 2007), os autores apresentam a aplicação do clustering hierárquico a um quadro de dados de variação de SASA ao longo de diversas simulações de desnaturação da WT-TTR. A medida em estudo foi a SASA e a construção do dendrograma foi enriquecida com informação dos aminoácidos (distância entre os aminoácidos na estrutura primária, a distância espacial entre resíduos ao longo da desnaturação e a propensão para a hidrofobicidade).

Nesta dissertação um dos objectivos da aplicação do clustering hierárquico é identificar e comparar os *clusters* obtidos para cada variante de TTR. Pretende-se partir das 5 corridas de cada variante da TTR e obter duas partições dos aminoácidos que tenham em conta a informação das cinco corridas da WT-TTR, obtendo-se a Partição I, e a informação das cinco corridas da L55P-TTR, obtendo-se a Partição II. Um outro objectivo é obter uma partição dos aminoácidos que considere toda a informação obtida pelas 10 simulações (Partição consenso). A partir das classes obtidas pretende-se ainda escolher o aminoácido representativo por classe, desta forma pode-se reduzir a complexidade do problema passando de 127 aminoácidos em estudo para 15 que representam o comportamento geral da TTR.

Cada corrida em estudo é um conjunto de 127 séries temporais e o que se pretende é agrupar séries temporais, como referem Caiado et al.,(2006) o clustering de séries temporais tornou-se uma área importante de pesquisa em várias áreas. Nos seus trabalhos, Caiado et al. estudam, por exemplo, como agrupar séries temporais com e sem o mesmo comprimento. Em Caiado et al.,(2006) é apresentada uma métrica para o clustering de séries temporais, que permite distinguir séries estacionárias de séries não-estacionárias. Esta métrica é baseada no periodograma normalizado e é utilizado o algoritmo de clustering hierárquico.

Nesta dissertação as séries apresentam todas o mesmo comprimento e o objectivo é agrupar os aminoácidos relativamente à propriedade em estudo, isto é, os aminoácidos que ficam no mesmo grupo, em princípio têm tendência para, ao longo da desnaturação, apresentarem valores parecidos de distância ao centro de massa.

Para calcular e obter as classes foram utilizados as seguintes bibliotecas do R *statistical package*: *class*, *MASS*, *fpc*. (Hennig, 2004)

4.2.1 Clustering de cada variante da proteína TTR

Nesta secção é apresentado como foram obtidas duas partições dos 127 aminoácidos da proteína TTR. Considerem-se as cinco simulações da variante WT-TTR, para cada corrida obteve-se uma partição dos aminoácidos. Para escolher o número de *clusters* foram calculados e optimizados os índices G2 e G3 de Milligan e Cooper. Com as 5 partições foi construída uma partição consenso.

Nestas classificações a medida de dissimilaridade usada foi a distância Euclidiana porque as variáveis são quantitativas contínuas e todas obtidas da mesma medida. Quanto ao critério de agregação utilizou-se o método de Ward, indicado por vários autores como sendo o melhor método para distâncias Euclidianas.

No *Anexo II-A1* é apresentada a escolha do número de classes para cada clustering.

Tabela 6 – Número de *clusters* para cada simulação da WT-TTR

	Nº de <i>clusters</i>
WTRUN2	14
WTRUN3	13
WTRUN4	14
WTRUN5	14
WTRUN6	13

Considerando as cinco simulações da variante L55P-TTR, para cada corrida obteve-se uma partição dos aminoácidos. Para escolher o número de *clusters* foram novamente calculados e otimizados os índices G2 e G3 de Milligan e Cooper. Com as 5 partições foi construída uma partição consenso. No *Anexo II-B1* é apresentada a escolha do número de classes para cada clustering.

O número de classes obtido para cada clustering da L55P-TTR é dado na Tabela 7:

Tabela 7 - Número de *clusters* para cada simulação da L55P-TTR

	Nº de <i>clusters</i>
L55PRUN1	15
L55PRUN2	17
L55PRUN3	14
L55PRUN4	13
L55PRUN5	11

A partir destas classificações pretende-se obter uma classificação para cada tipo de proteína, isto é, pretende-se utilizar as 5 classificações da WT-TTR e obter uma partição única dos resíduos. Para a L55P-TTR o processo é análogo. Desta forma pretende-se diminuir o erro de classificação que se obteria se fosse só considerada uma corrida de cada tipo de proteína. O que se obteve foram duas matrizes com 127 aminoácidos e 5 classificações para cada resíduo, logo estas matrizes têm de dimensões 127 x 5, como se pode observar na Figura 24.

127 resíduos
Clustering nº 1
Clustering nº 2
Clustering nº 3
Clustering nº 4
Clustering nº 5

Figura 24 - Matriz das classificações para cada tipo de proteína TTR

A partir desta matriz pretende-se obter uma partição dos aminoácidos que considera as classificações das cinco simulações. Para tal foi necessário definir uma medida de dissimilaridade e aplicar a classificação hierárquica sob a matriz de classificações.

Medida de dissimilaridade

Partindo do princípio que os resíduos que se mantêm a distâncias semelhantes do centro de massa são aqueles que ficam mais vezes na mesma classe, pretende-se uma medida de dissimilaridade na qual:

- a dissimilaridade é nula para aminoácidos que ficam sempre na mesma classe;
- a dissimilaridade é máxima para aminoácidos que nunca ficam na mesma classe.

Como os aminoácidos com comportamentos semelhantes serão aqueles que mais vezes ficaram na mesma classe, a medida de proximidade, aqui definida, compara os resíduos dois a dois e conta o número de vezes que estes ficam na mesma classe, tomando valores entre 0 e 5. A aplicação definida é $D : \mathcal{R}^5 \times \mathcal{R}^5 \rightarrow \{0,1,2,3,4,5\}$.

Considerem-se, por exemplo, dois resíduos cujos vectores na matriz das classificações são os seguintes:

$$aa_1 = (3,2,1,1,1) \quad aa_2 = (3,4,3,2,4)$$

Estes aminoácidos só ficaram no mesmo cluster no primeiro clustering (partição da primeira simulação), neste caso o valor de dissimilaridade é 4. Se no primeiro aminoácido o vector fosse $aa_1' = (3,2,3,1,1)$ a dissimilaridade seria 3.

Verificação das propriedades de uma dissimilaridade:

1. $d_{ij} \geq 0$, a dissimilaridade entre dois resíduos é sempre maior ou igual a zero.
2. $d_{ij} = d_{ji}, \forall i, j$. No exemplo anterior, a dissimilaridade entre aa_1 e aa_2 é 4, que é igual à dissimilaridade entre aa_2 e aa_1 que também é 4.

3. $d_{ij} = 0 \Rightarrow x_i = x_j$ a dissimilaridade é nula quando os dois vectores das classificações forem iguais.

Os valores da matriz de dissimilaridades são frequências absolutas no intervalo [0,5].

Partição consenso

Na introdução do Capítulo 3 já foi apresentada uma breve definição de partição consenso, nesta secção pretende-se definir com mais exactidão em que consiste uma partição consenso. Chama-se “Partição consenso” à partição obtida a partir de um conjunto de diferentes partições obtidas dos mesmos indivíduos. Dadas $t \geq 2$ partições $T_r (r = 1, \dots, t)$ dos mesmos n objectos, a partição consenso sintetiza a informação contida no conjunto de dados original e fornece um bom sumário das relações entre os objectos. Os clusterings originais devem ser obtidos por análise dos objectos utilizando t diferentes clusterings. (Gordon, 1999)

Neste ponto pretende-se obter uma partição dos aminoácidos para cada tipo de proteína considerando as 5 partições de cada uma, logo o que se pretende é obter uma partição consenso para WT-TTR e outra para a L55P-TTR. Para obter estas partições foi aplicado o clustering hierárquico à matriz das classificações e o critério utilizado foi o da Ligação Média.

O resultado do clustering sobre a matriz das classificações da WT-TTR é uma partição dos 127 aminoácidos (partição I), que considera a informação das cinco simulações da WT-TTR (Figura 25). A partição obtida divide os 127 aminoácidos em 11 grupos (ver Tabela 20, Anexo II-A3).

Analogamente se obtêm uma partição dos 127 aminoácidos (partição II) mas considerando a informação das cinco simulações da L55P-TTR. Isto é, aplicando a classificação hierárquica à matriz das classificações da L55P-TTR (Figura 25), com a medida de semelhança definida anteriormente e o critério de Ligação Média. A partição obtida divide os 127 aminoácidos em 15 grupos (Tabela 26, Anexo II-B3).

Algumas considerações

As partições obtidas não têm o mesmo número de *clusters*, é interessante verificar o seguinte:

- existem pequenos subgrupos de aminoácidos que se mantêm sempre juntos em todas as classificações.

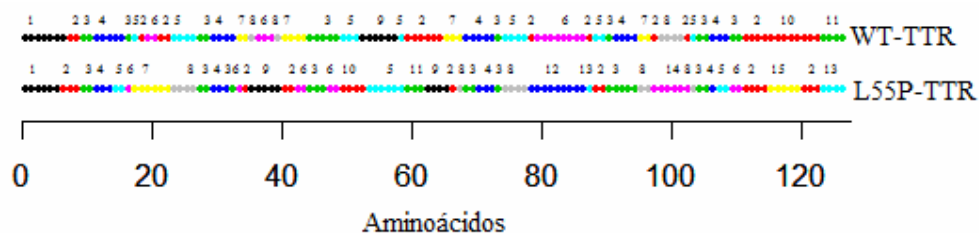


Figura 25 - Partição da WT-TTR e da L55P-TTR

- existem aminoácidos distantes na estrutura primária que ficam no mesmo cluster. Por exemplo, nas duas partições obtidas, podem encontrar-se a SER_8 e LYS_9 no mesmo cluster da SER_112, PRO_113 e TYR_114 (*Anexo II*, Tabela 20 e Tabela 26).
- na partição WT-TTR a LEU_55 ficou numa classe com elementos próximos na estrutura primária (como GLY_53, GLU_54, HIS_56, GLY_57 e LEU_58) (ver *Anexo II – Classificação hierárquica - Resultados-A3 - Partição consenso da WT-TTR*, Tabela 20, *cluster 9*) enquanto que na partição L55P-TTR a PRO_55 ficou na mesma classe que aminoácidos distantes na estrutura primária (por exemplo, LYS_15, ALA_109) (ver *Anexo II-B3*, Tabela 26, *cluster 5*)
- a sequência TYR_116, SER_117, THR_118, THR_119, ALA_120 constitui um grupo isolado na partição da L55P-TTR (*Anexo II-B3*, Tabela 26) enquanto que na partição da WT-TTR (*Anexo II-A3*, Tabela 20) estão no mesmo cluster que VAL_121, VAL_122 e THR_123.
- Os aminoácidos LEU_12, MET_13, VAL_14, VAL_30, VAL_32, VAL_71, ILE_73, ILE_107 ficaram nos mesmos *cluster* nas duas partições.
- a sequência ASN_124, PRO_125, LYS_126 e GLU_127 na partição I constitui um cluster isolado e na partição II estão associados a HIS_88.

Até este ponto o objectivo era encontrar uma partição dos aminoácidos para cada tipo de proteína e comparar resultados. No ponto seguinte pretende-se obter uma partição (partição III) dos aminoácidos que considere a informação das 10 corridas, tendo em conta que na posição 55 o aminoácido é diferente para as duas proteínas.

4.2.2 Partição consenso

Nesta secção pretende-se obter uma partição dos aminoácidos que considere a informação das 10 simulações. Inicialmente efectuou-se o clustering para cada uma das corridas sem o aminoácido da posição 55, porque na posição 55 na WT-TTR está uma leucina e na L55-TTR está uma prolina. A escolha do número de classes segue o mesmo raciocínio que na secção anterior.

Para a escolha do número de classes, em cada corrida, foram calculados e optimizados os Índice de Validação das classes, G1 e G2, de Milligan e Cooper. Os índices foram calculados num intervalo maior que o apresentado, no entanto só foram apresentados no intervalo 10 a 20 porque um número superior a 20 classes ficam muito afastado do critério $k_{\max} \leq \sqrt{n}$ (Pal et al., 1995), e menos que 10 classes não optimizou os índices em nenhum caso. O critério de agregação usado, para cada cluster hierárquico, foi o método de Ward e como medida de proximidade usou-se a distância euclideana.

Na Figura 28 o índice G3 apresenta a quebra mais acentuada de 13 para 14 classes, o que indica que 14 será um bom número de classes, como no índice G2 de 13 para 14 existe uma aumento elevado de G2, a escolha do número de classes para WTRUN2 foi de 14.

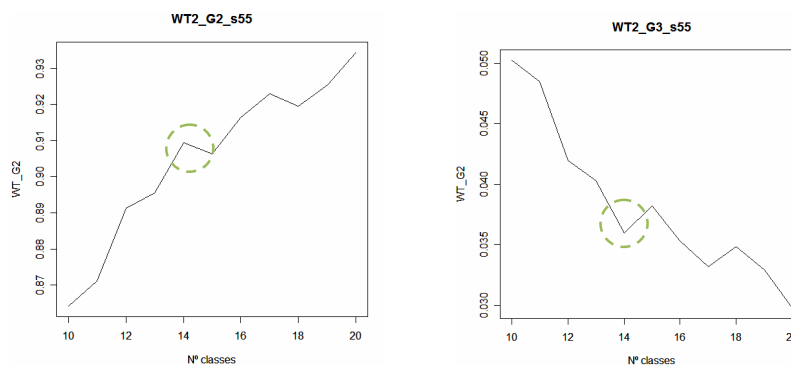


Figura 26 - Índices de Milligan e Cooper – G2 e G3 - para a simulação WTRUN2

Na corrida WTRUN3 o índice G2 atinge um valor óptimo para 13 classes e o índice G3 atinge o valor mínimo para 13 classes também, logo o número de classes escolhido para particionar os 126 aminoácidos nesta corrida é 13.

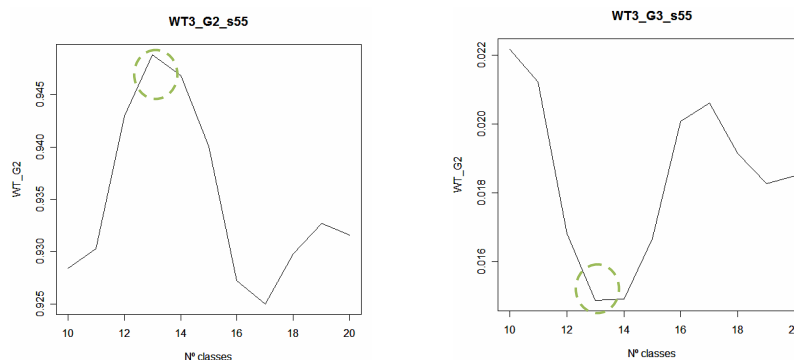


Figura 27 - Índices de Milligan e Cooper – G2 e G3 - para a simulação WTRUN2

Na corrida WTRUN4 o índice G2 atinge um máximo local expressivo para 14 classes e o índice G3 verifica um mínimo local para o mesmo número de classes.

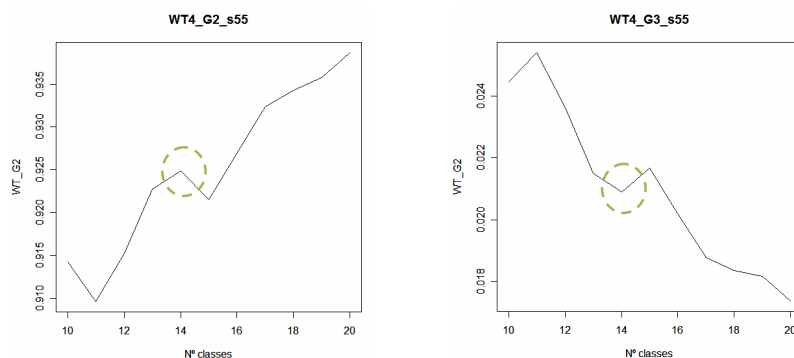


Figura 28 - Índices de Milligan e Cooper – G2 e G3 - para a simulação WTRUN2

Na corrida WTRUN5 a escolha do número de classes foi 14 por ser um máximo local do índice G2 e um mínimo local do índice G3.

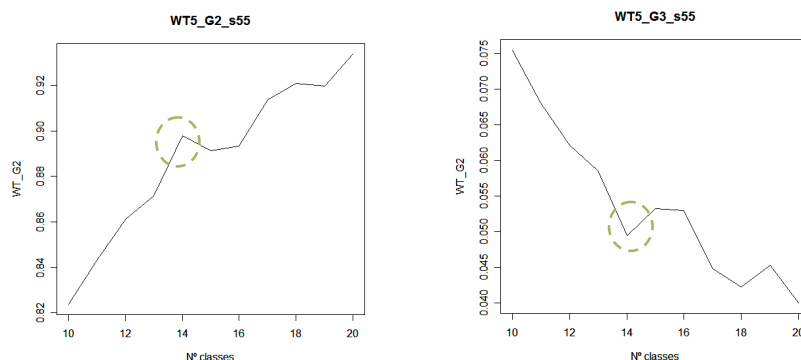


Figura 29 - Índices de Milligan e Cooper – G2 e G3 - para a simulação WTRUN2

Na corrida WTRUN6 os índices apresentam um comportamento regular. No entanto para 13 classes verifica-se uma oscilação, sendo 13 o número de classes escolhido.

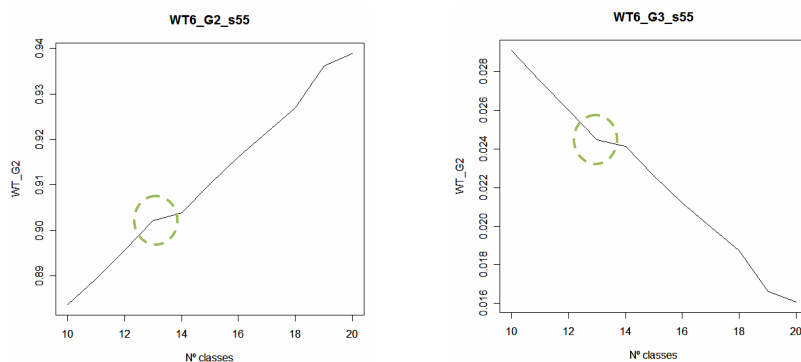


Figura 30 - Índices de Milligan e Cooper – G2 e G3 - para a simulação WTRUN2

Na variante WT-TTR verifica-se que o facto de ser retirado o aminoácido da posição 55 não afectou o número de classes escolhidas, sendo o número de classes de cada corrida igual ao número de classes escolhidas para obter a partição consenso da WT-TTR (Tabela 6).

Na corrida L55PRUN1 existe um máximo local do índice G2 para 15 classes, no índice G3 existe um mínimo para o mesmo número de classes, logo foi escolhida uma partição com 15 classes dos 126 aminoácidos.

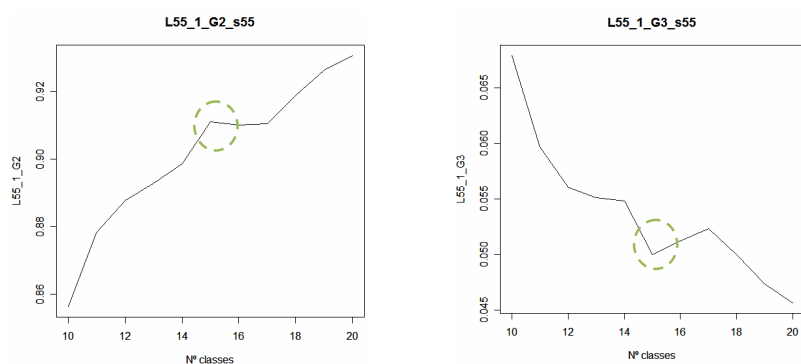


Figura 31 - Índices de Milligan e Cooper – G2 e G3 - para a simulação WTRUN2

Na corrida L55PRUN2 a escolha do número de classes foi 16, uma vez que existe uma pequena oscilação.

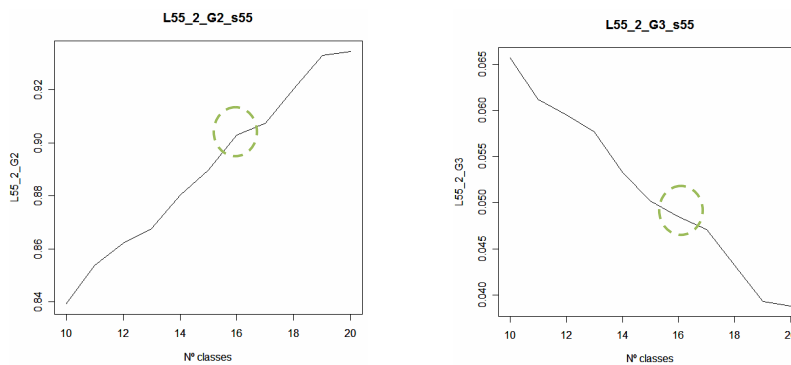


Figura 32 - Índices de Milligan e Cooper – G2 e G3 - para a simulação WTRUN2

Na corrida L55PRUN3 o número de classes escolhido foi 13.

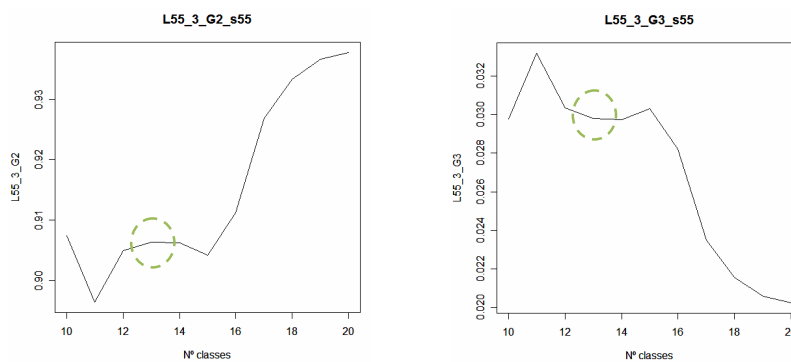


Figura 33 - Índices de Milligan e Cooper – G2 e G3 - para a simulação WTRUN2

Na corrida L55PRUN4 o número de classes escolhido foi 13.

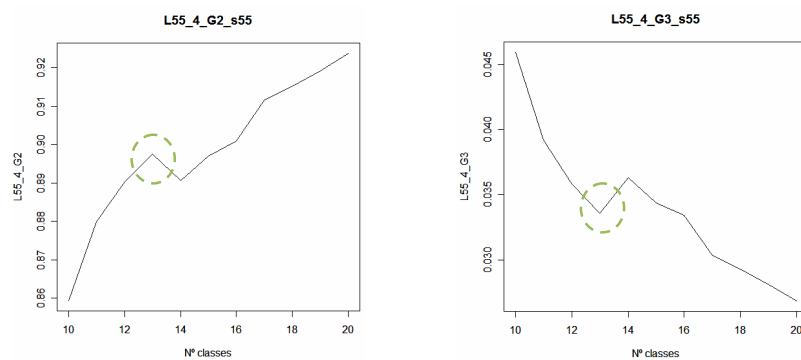


Figura 34 - Índices de Milligan e Cooper – G2 e G3 - para a simulação WTRUN2

Na corrida L55PRUN5 o número de classes escolhido foi 11.

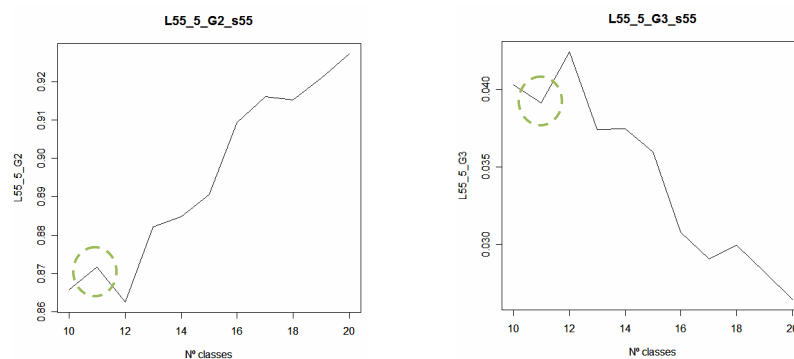


Figura 35 - Índices de Milligan e Cooper – G2 e G3 - para a simulação WTRUN2

Na Tabela 8 é apresentado o número de *clusters* que optimiza os índices nas diferentes corridas.

Tabela 8 - Número de *clusters* óptimas para cada simulação

	Nº de <i>clusters</i>		Nº de <i>clusters</i>
WTRUN2	14	L55PRUN1	15
WTRUN3	13	L55PRUN2	16
WTRUN4	14	L55PRUN3	13
WTRUN5	14	L55PRUN4	13
WTRUN6	13	L55PRUN5	11

Na segunda e na terceira corrida de L55P-TTR o número de clusters é diferente ao número de clusters obtidos anteriormente (Tabela 7). Desta forma verifica-se que ao retirar o aminoácido da posição 55 o número de classes varia na variante L55P-TTR, logo este aminoácido influencia o comportamento dos outros aminoácidos.

Para obter a partição final considerou-se uma matriz (126 por 10) com os dez clusterings obtidos e efectuou-se um novo clustering sob esta matriz. Isto é, pretende-se a partição que vai obter a informação “consensual” das 10 partições.

O critério de agregação usado foi o da Ligação Média e a medida de proximidade usada é análoga à dissimilaridade definida anteriormente, mas neste caso, toma valores entre 0 e 10 ($D: \mathcal{R}^5 \times \mathcal{R}^5 \rightarrow \{0,1,\dots,10\}$). A dissimilaridade entre dois aminoácidos toma o valor 10 quando estes nunca estiveram no mesmo cluster, e é nula quando aos dois resíduos lhes foi sempre atribuído o mesmo cluster.

A partição consenso (partição III) foi obtida por corte do dendrograma em $h=10$ (*Anexo II-C*, Figura 66) e divide os aminoácidos em 14 classes.

Tabela 9 - Partição Consenso

1	2	3	4	5	6	7	8	9	10	11	12	13	14
GLY_1	GLU_7	CYS_10	LEU_12	ASP_18	ALA_19	VAL_20	VAL_28	ARG_34	ALA_36	GLU_62	SER_112	TYR_116	ASN_124
PRO_2	SER_8	PRO_11	MET_13	PRO_24	GLY_22	ARG_21	ALA_29	PRO_43	THR_40	GLU_63	PRO_113	SER_117	PRO_125
THR_3	LYS_9	LEU_17	VAL_14	ALA_25	SER_23	ALA_37	ALA_45	PHE_44	VAL_65	PHE_64	TYR_114	THR_118	LYS_126
GLY_4	LYS_35	LYS_48	LYS_15	ILE_26	THR_60	ASP_38	SER_46	ILE_68	GLU_66	TRP_79	SER_115	THR_119	GLU_127
THR_5	TRP_41	THR_49	VAL_16	ASN_27	GLU_61	ASP_39	GLY_47	THR_96	ASN_98	HSD_88	VAL_121	ALA_120	
GLY_6	GLU_42	HSD_56	VAL_30	SER_50	SER_77	LYS_80	ASP_74	ALA_97	ASP_99		VAL_122		
	GLY_67	GLY_57	HSD_31	GLU_51	TYR_78	ALA_81	ALA_91		SER_100		THR_123		
	LYS_76	LEU_58	VAL_32	SER_52		LEU_82	GLU_92		GLY_101				
	GLU_89	TYR_105	PHE_33	THR_59		GLY_83	VAL_93		PRO_102				
	HSD_90	THR_106	GLY_53	THR_75		ILE_84	VAL_94		ARG_103				
		LEU_110	GLU_54	ARG_104		SER_85	PHE_95						
		LEU_111	TYR_69			PRO_86							
			LYS_70			PHE_87							
			VAL_71										
			GLU_72										
			ILE_73										
			ILE_107										
			ALA_108										
			ALA_109										

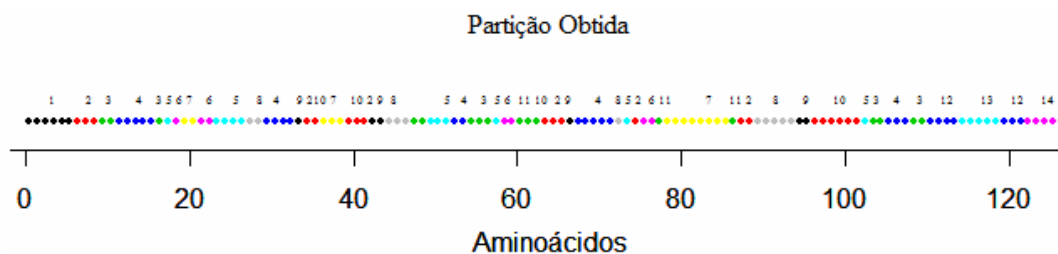


Figura 36 – Partição consenso na sequência de aminoácidos inicial

Nos *clusters* obtidos verifica-se que:

- os *clusters* apresentam dimensões variadas, sendo o cluster mais pequeno constituído pelos 4 últimos aminoácidos da cadeia.

- o cluster 4 possui 19 aminoácidos dos quais 13 são hidrofóbicos e de posições afastadas na cadeia inicial (por exemplo, LEU_12 e ALA_109).
- existem aminoácidos afastados na estrutura primária que estão no mesmo cluster, por exemplo CYS_10 e LEU_111 (*cluster* 3).

4.2.3 Escolha dos Aminoácidos que representam cada uma das classes

Quando o número de objectos em estudo é elevado, e após obter a partição dos objectos, a análise do comportamento geral de cada cluster pode ser simplificada através da escolha de elementos representativos dos *clusters*.

O elemento representativo não garante que este tem as mesmas características que os representados, no entanto, como os aminoácidos estão ligados em cadeia irá haver uma tendência para que o comportamento de um aminoácido seja acompanhado pelos aminoácidos vizinhos.

Para a escolha dos aminoácidos representativos poder-se-ia ter subdividido cada cluster mediante a organização dos aminoácidos na cadeia, e escolher um aminoácido de cada subgrupo (obter-se-iam mais de 30 representativos). Mas o que se pretende é distinguir as duas variantes da proteína. Logo o grupo de aminoácidos representativos deve ser pequeno uma vez que só estão em análise dez corridas.

Neste estudo o elemento representativo de cada cluster é o elemento mais próximo do centróide (para uma medida de distância adequada aos objectos em estudo).

O conjunto final de aminoácidos, que representa o comportamento geral dos aminoácidos ao longo das 10 simulações, foi obtido da seguinte forma:

- para cada corrida foram calculados os aminoácidos representativos de cada classe, isto é, os aminoácidos mais próximos do centróide de cada classe.
- para cada classe o aminoácido escolhido como representativo é o que mais vezes foi representativo nas 10 corridas. Por exemplo, se na classe 1, ao longo das 10 corridas, o

aminoácido A_i foi representativo 7 vezes então este é escolhido como sendo o aminoácido representativo da classe (ver *AnexoII - B*, Tabela 27).

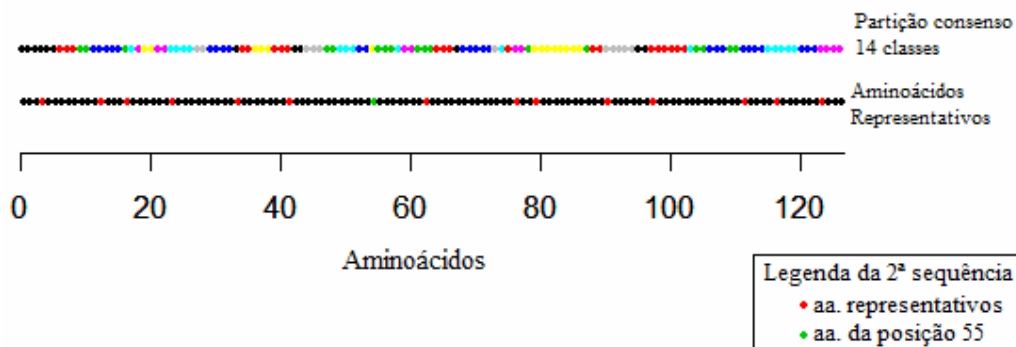


Figura 37- Aminoácidos Representativos

Os aminoácidos considerados para a procura dos eventos são os aminoácidos representativos e o resíduo da posição 55, porque como uma diferença entre os dois tipos de proteína é o aminoácido da posição 55 pode ser interessante procurar eventos com este aminoácido. Em seguida são enumerados os aminoácidos representativos da proteína TTR:

- GLY_4, MET_13, LEU_17, PRO_24, ARG_34, GLU_42, GLU_63, SER_77, LYS_80, ALA_91, ASN_98, SER_112, SER_117, ASN_124, e POS_55.

Seja qual for o número de indivíduos em estudo, a escolha dos elementos representativos dos *clusters* permite obter uma análise geral do fenómeno em estudo e simplifica um pouco o problema. Corre-se o risco de perdermos algum pormenor na análise, que pode vir a indentificar-se muito relevante para a situação em questão. No entanto, a filosofia seguida nesta tese foi de que perante um problema complexo devemos torná-lo mais simples, para que exista um ponto de partida e uma linha de raciocínio completa e bem definida.

4.3 Procura de Eventos

Nesta secção são apresentados os eventos, isto é, foram analisadas as séries dos aminoácidos e procuraram-se mudanças significativas no comportamento destes ao longo das simulações. Na procura de eventos foram efectuadas duas abordagens, uma primeira onde se compara o comportamento dos aminoácidos representativos dois a dois e se procura um tipo de ocorrência; uma segunda abordagem onde é procurado um acontecimento pré-definido por aminoácido.

4.3.1 Comparação de aminoácidos dois a dois

Relacionar o comportamento dos aminoácidos dois a dois, ao longo da desnaturação, é uma tarefa que permite, por um lado, obter conhecimento do comportamento dos aminoácidos ao longo da desnaturação e permite, principalmente, verificar e comparar as ocorrências dos eventos entre as duas variantes de TTR. Para comparar os aminoácidos dois a dois foram definidos dois eventos:

- Evento Nº 1 - “troca” de posição entre dois aminoácidos.

No evento “troca” de posição entre dois aminoácidos é procurada a ocorrência da seguinte situação: um aminoácido afasta-se do centro de massa e outro aproxima-se do centro de massa, em simultâneo e durante um intervalo de tempo. Esta “troca” de comportamento entre dois aminoácidos considera-se significativa se se mantêm ao longo de, pelo menos, n frames e nesse caso temos um evento, sendo n um valor dado.

Para encontrar este tipo de eventos são calculadas novas séries temporais, a partir da diferença das séries dos aminoácidos dois a dois e procuram-se mudanças de sinal com elevada estabilidade destas novas séries. A estabilidade do sinal foi definida como sendo um número n de frames, antes e depois da zona de viragem, em que o sinal se mantém.

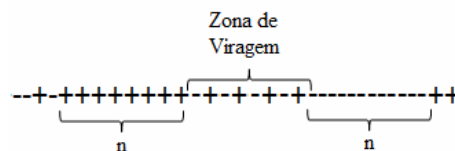


Figura 38 – Zona de Viragem

Partindo do conjunto dos aminoácidos representativos foi procurado o evento: “o aminoácido i (que está mais afastado do centro de massa que o aminoácido j ao longo de 300 frames) ‘troca’ com o aminoácido j e mantém esse comportamento ao longo de 300 frames”.

A escolha do número n de frames foi feita mediante o número de eventos encontrados, com um número superior a 300 (foram experimentados os valores 350 e 400) não se obtinham eventos em algumas corridas, com um n inferior a 300 (foram experimentados os valores 250 e 200) obtinha-se um número muito elevado de eventos e com elevada frequência. Pretende-se estudar um acontecimento interessante da proteína que ocorra mas que não seja muito frequente.

Por exemplo, na Figura 39 é possível observar que o aminoácido GLU_42 está mais afastado do centro de massa que LEU_17 (ao longo de mais de 300 frames). No decorrer do tempo da simulação LEU_17 fica mais afastada do centro de massa que GLU_42, num intervalo de 773 frames, voltado a estar a distâncias semelhantes do centro de massa.

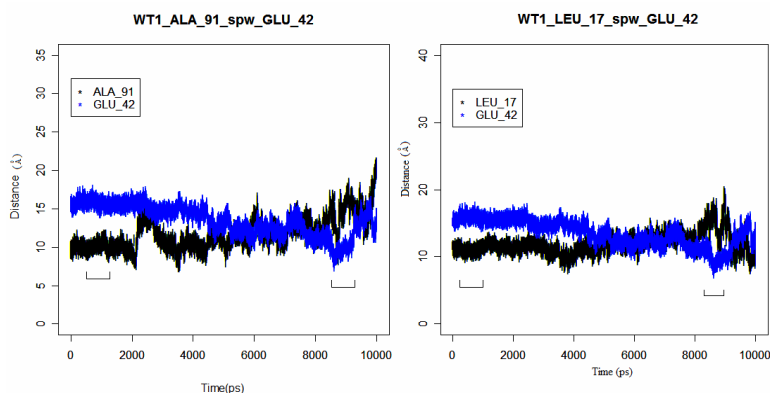


Figura 39 – LEU_17 aproxima-se do centro de massa ao mesmo tempo que GLU_42 se afasta do centro de massa, o mesmo acontece entre ALA_91 e GLU_42.

Na procura deste tipo de ocorrências foram encontrados 95 eventos. No Anexo III-A, Tabela 28, estão descritos os eventos obtidos.

Nos gráficos da Figura 40 são apresentados os 10 eventos mais frequentes (os gráficos foram obtidos a partir do Excell). Verifica-se que os eventos nº 84 - “SER_117 spw

GLU_42” e nº 90 - “SER_117 spw SER_112” acontecem nas 5 corridas da WT-TTR, enquanto que na proteína L55P-TTR só acontece três vezes.

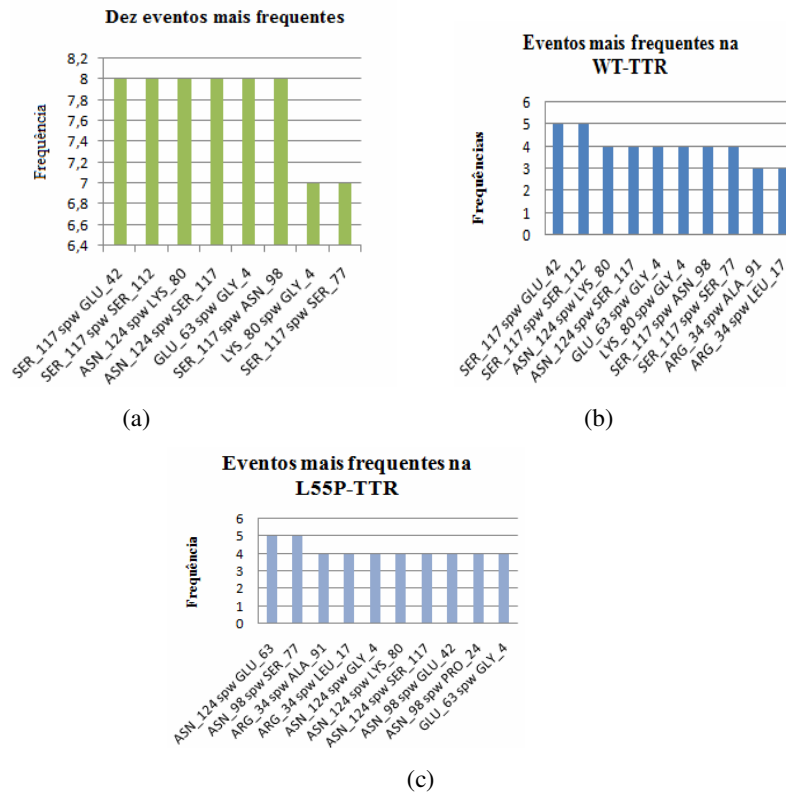


Figura 40 – (a) Os 10 eventos mais frequente nas 10 simulações; (b) Os 10 eventos mais frequentes nas simulações da WT-TTR; (c) Os 10 eventos mais frequentes nas simulações da variante L55P-TTR.

Por outro lado verifica-se que o evento 18 - “ASN_124 spw GLU_63” e o evento 33 - “ASN_98 spw SER_77” acontecem nas cinco corridas da L55P-TTR enquanto que na WT-TTR só acontece duas vezes.

Pode ainda constatar-se que o evento 41- “GLU_63 spw LYS_80” acontece em 4 corridas da L55-TTR e não acontece na WT-TTR. Os eventos 84 -”SER_117 spw GLU_42” e 90 - “SER_117 spw SER_112” ocorreram nas mesmas simulações. O evento 1 só se verifica em duas corridas da WT-TTR (em alguns gráficos da L55P-TTR existem trocas entre a ALA_91 e a GLU_42 mas não se mantêm ao longo de 300 frames (ver Figura 41 e Figura 42).

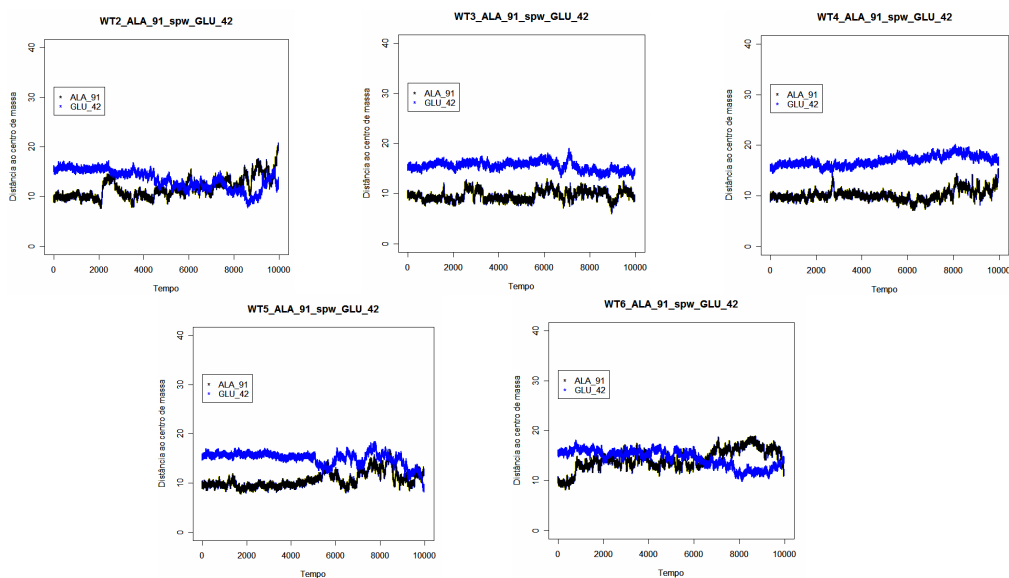


Figura 41 –Procura do evento 1- ALA_91 troca com GLU_42 ao longo das 5 corridas da WT-TTR. Na corrida WT2 e na corrida WT6 existe troca, com estabilidade de 300 frames.

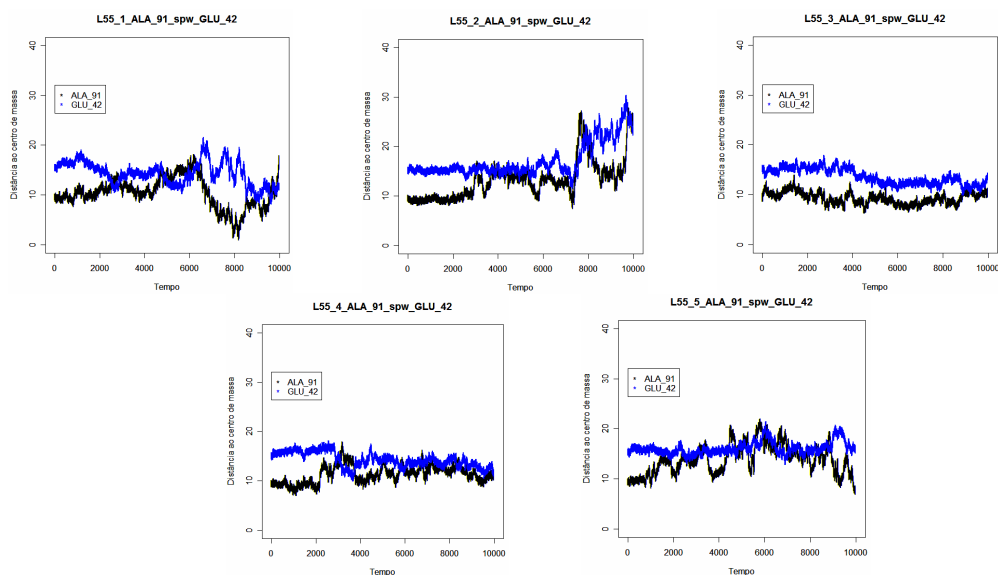


Figura 42 – Procura do evento 1- ALA_91 troca com GLU_42 ao longo das 5 corridas da L55P-TTR. Não se verifica a ocorrência do evento.

O valor médio da frequência deste evento na proteína WT-TTR é de 1,64 sendo inferior ao valor médio registado na proteína L55P-TTR de 2,48. Neste contexto, foi utilizado o teste de Kruskal-Wallis em que o objectivo era analisar se as frequências de ocorrência deste tipo de evento apresentam diferenças significativas consoante o tipo de proteína. Para a realização deste teste foi utilizada a biblioteca *stats* do programa *R- Statistical package*.

Considere-se o ensaio de:

H_0 : Os dois tipos proteínas têm a mesma distribuição

vs

H_1 : As proteínas não possuem a mesma média $E(X_1) \neq E(X_2)$

```
> kruskal.test(list(wt,155))

Kruskal-Wallis rank sum test

data: list(wt, 155)
Kruskal-Wallis chi-squared = 22.5951, df = 1, p-value = 2e-06
```

Figura 43 – Teste de Kruskal- Wallis

Como p-value obtido foi inferior a 0.01 então Rejeito H_0 com 99% de confiança, isto é, não existe evidência estatística para aceitar a hipótese de que as duas amostras independentes possuem distribuição idêntica. Logo existe uma diferença significativa na frequência dos eventos consoante o tipo de proteína.

- Evento Nº 2 - “Existem duas trocas entre os aminoácidos”

Neste evento são procurados pares de aminoácidos que “trocam” duas vezes ao longo da simulação. O tempo foi dividido em três secções e em cada secção procura-se uma janela de n frames onde um aminoácido está mais próximo do centro de massa que o outro aminoácido alternadamente.



Figura 44 – O tempo foi dividido em três secções, procura-se a situação em que a relação de dois aminoácidos com o centro de massa é trocada duas vezes e mantêm-se ao longo de n frames de cada secção.

Por exemplo, se na primeira janela o aminoácido i está mais afastado do centro de massa que o aminoácido j então temos o evento tipo **A** que se designa: $AA_i_troca_2x_A_AA_j$ (isto é, o aminoácido i troca duas vezes com o aminoácido j). Se na primeira janela o aminoácido j está mais afastado do centro de massa que o aminoácido i então temos o evento tipo **B** que se designa: $AA_i_troca_2x_B_AA_j$.

O valor de n escolhido foi 200 porque, para valores superiores a 200 não se encontravam eventos em algumas corridas (foi experimentado $n=300$) e para valores inferiores a 200 registam-se um número elevado de situações, isto é, deixamos de encontrar o evento, duas trocas de relação com o centro de massa com elevada estabilidade e obtemos ocorrências de um fenómeno com elevada frequência.

O evento **ALA_91_troca_2x_A_PRO_55** (Tabela 10) significa que o aminoácido ALA_91 ao longo de 200 frames, nas primeiras 2000 frames, está mais afastado do centro de massa que PRO_55; entre as 2000 e as 8000 frames existe um intervalo de 200 frames em que PRO_55 se mantém mais afastado do centro de massa que ALA_91; e entre as 8000 e as 10000 volta a ser ALA_91 que está mais afastado do centro de massa que PRO_55 ao longo de 200 frames. Esta situação só se verifica em duas simulações da proteína L55P-TTR.

Tabela 10- Frequência do evento nº1 para cada variante de TTR

Nº	Evento	L55	WT
1	ALA_91_troca_2x_A_PRO_55	2	0

Na Tabela 29 do *Anexo III – Eventos* é possível observar as frequências de cada evento para cada variante de TTR. Verifica-se que o evento que ocorre com mais frequência na proteína L55P-TTR é o evento 17 - GLU_42_troca_2x_A_ALA_91.

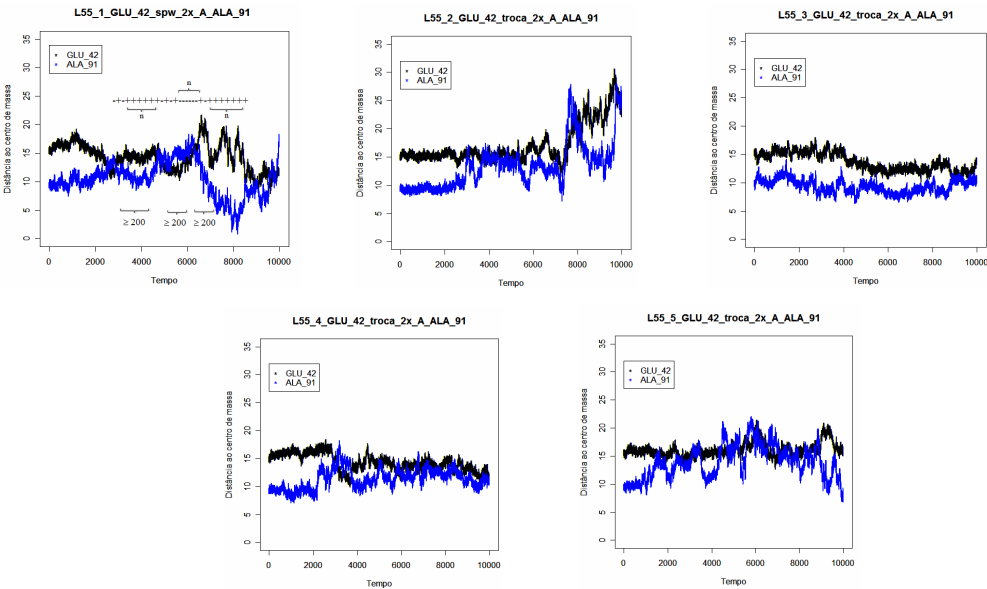


Figura 45 –Procura do evento 17 - GLU_42_troca_2x_A_ALA_91 ao longo das 5 corridas da L55P-TTR. Na corrida L55_3 não existem duas trocas de comportamento.

Os eventos que ocorrem três vezes na L55P-TTR e nenhuma vez na WT-TTR são apresentados na Tabela 11.

Tabela 11 – Eventos mais frequentes na proteína L55P-TTR e que não se verificam na WT-TTR

Nº	Evento	Freq_L55
6	ALA_91_troca_2x_B_SER_117	3
50	LEU_17_troca_2x_A_POS_55	3
55	LEU_17_troca_2x_B_PRO_24	3
58	LEU_17_troca_2x_B_SER_77	3
81	SER_77_troca_2x_A_ALA_91	3
17	GLU_42_troca_2x_A_ALA_91	4

As frequências de ocorrência destes eventos na proteína WT-TTR são baixas, sendo o seu valor máximo de três. Os eventos seguintes só ocorrem três vezes nesta proteína.

Tabela 12- Eventos mais frequentes na WT-TTR.

Nº	Evento	Freq_L55	Freq_WT
22	GLU_42_troca_2x_B_ASN_124	0	3
30	GLU_63_troca_2x_B_ASN_124	1	3
74	PRO_24_troca_2x_B_SER_112	1	3

O valor médio da frequência do evento Nº2 na proteína WT-TTR é de 0,72 sendo, mais uma vez, inferior ao valor médio registado na proteína L55P-TTR que é 1.44. Com o teste estatístico de Kruskal- Wallis pretende-se verificar se as frequências médias são significativamente diferentes entre a proteína WT-TTR e a variante L55P-TTR.

Considere-se o ensaio de:

H_0 : Os dois tipos proteínas têm a mesma distribuição

vs

H_1 : As proteínas não possuem a mesma média $E(X_1) \neq E(X_2)$

```
> kruskal.test(freq[,1],freq[,2])

Kruskal-Wallis rank sum test

data:  freq[, 1] and freq[, 2]
Kruskal-Wallis chi-squared = 5.213, df = 4, p-value = 0.2661
```

Figura 46 – Teste de Kruskal-Wallis

Como p-value obtido é superior a 0.05, não rejeito H_0 com 95% de confiança. Neste evento o teste de comparação de médias não indica que existam diferenças entre as médias de ocorrência das duas variantes de TTR.

4.3.2 Procura de alterações significativas por aminoácido.

O Evento procurado neste ponto foi a *existência de uma alteração significativa no comportamento de cada aminoácido* e foram analisados os 127 aminoácidos nas dez simulações. Neste caso não foram utilizados só os aminoácidos representativos porque este procedimento tem um baixo custo computacional.

O que se pretende é encontrar mudanças de comportamento dos aminoácidos que possam ser interessantes para análises futuras. Considerou-se que uma mudança de comportamento interessante seria a alteração rápida da distância do aminoácido ao centro de massa da proteína em poucas frames. Isto é, para uma janela de tamanho n (relativamente pequena) o módulo do declive da recta que une o valor máximo de distância ao centro de massa ao valor mínimo, é elevado.

Neste contexto encontraram-se eventos em 66 aminoácidos, para uma janela de 200 frames, que estão apresentados Tabela 30 – *Anexo III-B*. As frequências registadas são baixas, no entanto, no aminoácido PRO_86 verificam-se alterações significativas de comportamento nas 5 simulações da WT-TTR e em uma simulação da L55P-TTR (ver Figura 47 e Figura 48).

Tabela 13 - Comportamento do aminoácido PRO_86

Aminoácido	Frequência	Simulação	Início-janela (ps)	Fim-janela (ps)
PRO_86	55	WTRUN2	7536	7736
PRO_86	53	WTRUN3	305	505
PRO_86	16	WTRUN4	626	826
PRO_86	88	WTRUN6	2544	2744
PRO_86	20	WTRUN5	4197	4397
PRO_86	27	L55PRUN1	8328	8528

Importa ainda salientar que, no aminoácido GLY_83 foram registadas 4 alterações significativas em 4 simulações de L55P-TTR e no aminoácido GLY_57 foram registadas 4 alterações significativas do comportamento em 3 simulações da proteína L55P-TTR e numa simulação de WT-TTR.

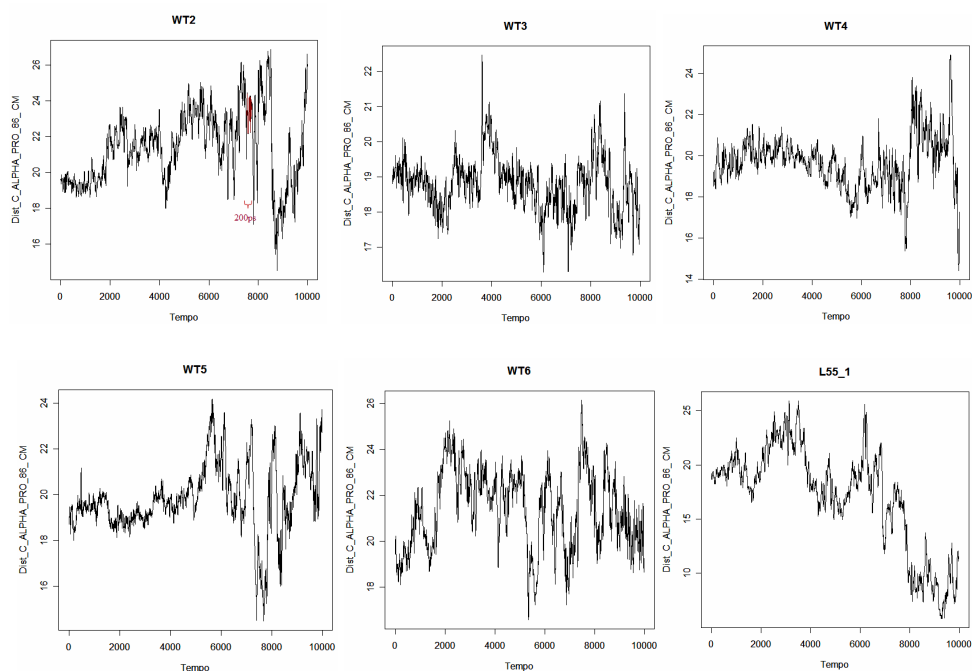


Figura 47 – Análise gráfica do evento “*existência de uma alteração significativa no comportamento do aminoácido PRO_86*” que se verifica nas corridas WT2, WT3, WT4, WT5, WT6, L55_1.

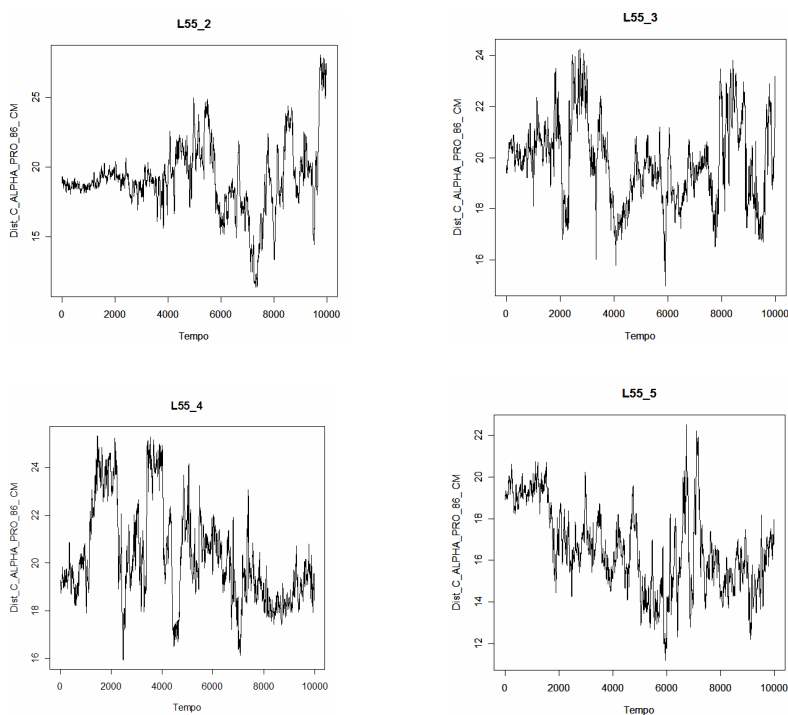


Figura 48 – Análise gráfica do aminoácido PRO_86 nas corridas onde não se verifica o evento.

Os aminoácidos que apresentam mais alterações bruscas de comportamento, ou seja, onde mais vezes o declive da recta é superior a 1, são apresentados na Tabela 14, na

qual a coluna *Frequência* apresenta o número de vezes que, numa *Simulação*, o aminoácido se aproxima ou afasta do centro de massa rapidamente.

Na Tabela 31- *Anexo III-B* são apresentados todos os aminoácidos onde se registaram alterações de comportamento, frequências dessas alterações, as simulações e o intervalo de tempo onde se verificam os valores elevados de declive.

Tabela 14 – Aminoácidos com comportamentos mais variáveis

Aminoácido	Frequência	Simulação
PRO_24	211	L55PRUN4
ALA_108	142	WTRUN3
GLY_47	135	WTRUN2
GLU_51	134	L55PRUN5
ALA_120	131	WTRUN2
VAL_30	130	WTRUN2
ASN_27	128	WTRUN3
GLY_57	124	L55PRUN2
SER_52	112	WTRUN5
THR_3	104	L55PRUN4

Para analisar se existe diferença significativa entre as frequências médias de ocorrência deste tipo de evento consoante o tipo de proteína foi efectuado o teste de Kruskal-Wallis. Considerem-se as seguintes hipóteses:

H_0 : As duas proteínas têm a mesma distribuição ou

vs

H_1 : Pelo menos uma das distribuições dá valores mais elevados do que os da outra

```
> kruskal.test(t[,1],t[,2])

Kruskal-Wallis rank sum test

data:  t[, 1] and t[, 2]
Kruskal-Wallis chi-squared = 23.9768, df = 3, p-value = 2.526e-05
```

Figura 49 – Teste Kruskal-Wallis

Como o p-value é inferior a 0.01 então Rejeito H_0 com 99% de confiança, isto é, não existe evidência estatística para aceitar a hipótese de que as duas amostras independentes possuem distribuição idêntica. Logo existe uma diferença significativa na frequência dos eventos consoante o tipo de proteína.

Uma abordagem, que poderia ser desenvolvida futuramente, seria considerar dois tipos de alterações significativas, isto é, considerar “*aproximação rápida ao centro de massa*” e “*afastamento rápido do centro de massa*”. E numa análise isolada, dos aminoácidos com comportamento mais variável, registar os tempos de ocorrência desses comportamentos e estudar relações possíveis entre as ocorrências vs tempo e o tipo de proteína.

Podem ser definidos outros eventos que se considerem interessantes para analisar a desnaturação da TTR. Neste trabalho foram procurados eventos simples para ser cumprida e definida uma linha de raciocínio. No futuro, pode, eventualmente, partir-se para abordagens mais complexas, por exemplo, incluir o factor tempo nos eventos.

Com os eventos obtidos podem usar-se regras para descrever relações entre os eventos e o tipo de proteína. Pode obter-se relações do tipo: “*se o aminoácido i “troca” com o aminoácido j, e o aminoácido k “troca” com o aminoácido l, então a proteína em estudo é a WT-TTR com Y% de probabilidade*”.

Podem ainda ser estudadas relações entre eventos e pares de aminoácidos. Por exemplo, “*se o aminoácido i está a uma distância do centro de massa entre 10 e 20, então o aminoácido j está a uma distância do centro de massa entre 40 e 45, com X% de probabilidade*” (um estudo semelhante foi desenvolvido na análise dos valores de SASA em Azevedo et. al., (2005)).

Os eventos registados neste capítulo permitiram a procura de regras de associação que distingam as duas variantes de TTR. No capítulo seguinte são apresentadas as regras de associação encontradas que relacionam eventos e variante de TTR.

4.4 Procura de Regras de Associação

Nesta secção são apresentadas as regras de associação obtidas, que relacionam evento e variante de TTR. Para obter as regras foi utilizado o programa CAREN e procuraram-se regras com valores de suporte e confiança elevados. O programa foi aplicado para três tabelas, cada uma das quais continha na primeira coluna o nome dos eventos e na segunda coluna o nome de cada corrida. A cada tabela foram ainda acrescentadas 10

linhas com nome da variante (na primeira coluna - WT ou L55P) e nome da corrida (na segunda coluna – WT2,...,WT6, L55P1,..., L55P5).

- Evento Nº 1 - “*Verifica-se uma troca de posição entre dois aminoácidos*”

Para este tipo de evento obteve-se a regra:

Regra nº1: ASN_98_spw_SER_77 & ASN_124_spw_GLU_63 → L55P Suporte=0.5; Confiança=1; Lift=2; Conviction= ∞
--

A regra obtida indica que, nas simulações em estudo, a probabilidade de estarmos perante a variante L55P-TTR sabendo que:

- o aminoácido ASN_98 está afastado do centro de massa enquanto SER_77 está próximo (ao longo de 300 frames) trocando depois de posição mantendo o comportamento ao longo de 300 frames;
- e do aminoácido ASN_124 estar mais afastado centro de massa que GLU_63 (ao longo de 300 frames) trocando depois de posição, relativamente ao centro de massa, ao longo de 300 frames;

é de 100%.

Pode ainda concluir-se que o antecedente e o consequente são positivamente dependentes porque o valor da medida de interesse *lift* é superior a 1. O valor de *conviction* também é superior a 1, logo segundo esta medida, a regra obtida é interessante. Com os gráficos do Anexo IV-A é possível verificar a ocorrência dos eventos da regra nº1.

- Evento Nº 2 - “*Verificam-se duas trocas entre os aminoácidos*”

Para este tipo de evento a única regra encontrada foi:

Regra nº2: GLU_42_troca_2x_A_ALA_91 → L55 Suporte=0.4; Confiança=1; Lift=2; Conviction= ∞

A regra obtida indica que, a probabilidade de estarmos perante a proteína L55P-TTR sabendo que o aminoácido GLU_42 está mais afastado do centro de massa enquanto ALA_91 está mais próximo, ao longo de 200 frames da primeira secção e nas secções seguintes alternam o comportamento, é de 100%. O valor de *lift* e de

conviction é mais uma vez superior a 1, o que indica que a regra obtida é interessante.

- Evento por aminoácido - *existência de uma alteração significativa no comportamento do aminoácido*

Para este evento as duas regras obtidas indicam duas situações já mencionadas no capítulo anterior, não acrescentam informação nva. No entanto o *lift* destas regras é superior a 1, o que nos indica que quanto mais o aminoácido GLY_83 apresentar um comportamento irregular maior é a propensão de este pertencer à variante L55P-TTR. Por outro lado o valor de *conviction* é superior a um, o que indica que estas regras são interessantes.

```
Regra nº3: GLY_83 → L55
          Sup = 0.4  Conf = 1; Lift=2; Conviction=∞

Regra nº4: PRO_86 → WT
          Sup = 0.5  Conf = 0.83; Lift=1.66; Conviction=30
```

Nesta fase do trabalho surgiram algumas questões. Qual o valor destas regras? Será que o número de simulações em análise é suficiente para obtermos conclusões seguras do comportamento da TTR? Qual o risco de falsas descobertas? Na secção seguinte são abordadas e discutidas estas questões.

4.5 Valor das regras obtidas

Nesta secção pretende-se analisar o valor das regras obtidas e o risco de falsas descobertas. Uma vez que só estão em estudo 10 simulações, será que as regras obtidas acontecem por acaso? Poderia ter-se obtido o mesmo resultado por acaso?

4.5.1 Como controlar o risco de falsas descobertas?

Segundo Webb (2006) as regras de associação só são interessantes se representam correlações não-triviais entre todos os items. No seu trabalho “*Discovering Significant rules*” são analisadas técnicas, para controlar o risco de falsas descobertas na obtenção de regras de associação. Estas técnicas baseiam-se no uso de testes estatísticos para seleccionar só as regras interessantes.

Uma das técnicas apresentadas por Webb (2006), designada por “Within-search approach”, consiste em aplicar testes estatísticos às regras e ajustando o nível de significância, através do “ajustamento de Bonferroni”, é definido o tamanho do espaço de procura de onde as regras devem ser extraídas. Através deste ajustamento é possível controlar, com mais rigor, o risco de falsas descobertas. Segundo Webb (2006) este método é útil quando o espaço de procura é pequeno ou o número de regras candidatas é grande.

Trabalhos recentes mostram o elevado interesse dos analistas no estudo de métodos para controlar o risco de falsas descobertas. Podem ser encontradas diferentes abordagens nos trabalhos de Benjamini (2001), Bruzese (2003) e Webb (2006).

Um método útil para validar regras é o “*método holdout*”. Este método consiste em partir os dados em dois conjuntos, um conjunto de exploração para obter as regras candidatas, e um conjunto de teste que é utilizado para escolher as regras finais (aplicando, por exemplo, uma avaliação estatística (Webb, 2006)). Nesta tese colocou-se a hipótese de se aplicar o método *holdout*, mas não foi utilizado porque só estão em estudo 10 simulações da desnaturação da TTR. Com mais simulações o poder estatístico aumentaria e o erro diminuiria. Mas aqui surge uma outra questão quantas simulações são necessárias?

Os baixos orçamentos obrigam a que sejam efectuadas poucas simulações dos fenómenos em estudo, o que aumenta o risco de falsas descobertas. Para tentar resolver esta questão alguns autores têm desenvolvido modelos para calcular o número de corridas necessárias para obter resultados reais.

No trabalho de Wei et. al., (2004) são analisados factores que influenciam na escolha da dimensão do exemplo, tais como, a variabilidade da população (que deve ser fixada consoante o estudo), o valor de significância (para aplicar os testes-t) ou o rácio de falsos positivos/negativos. O objectivo de Wei et. al. era fornecer uma indicação de como calcular um número aproximado de corridas. Segundo estes autores a variabilidade de dados Humanos é geralmente elevada, logo são sempre necessárias muitas corridas para um estudo.

No trabalho de Warnes (2006) é proposto um modelo de estimação do tamanho do exemplo, o qual foi implementado no package `ssize` do R. Todas estas abordagens são muito recentes e susceptíveis de muitas dúvidas uma vez que os parâmetros a otimizar em cada modelo dependem do contexto do problema em estudo.

4.5.2 Como medir a aleatoriedade das regras obtidas?

Segundo Witten I. (1999) a “aleatoriedade” de uma regra, ou seja, a probabilidade de uma regra aleatória dar o mesmo resultado (ou melhor) do que uma dada regra (R) é dada por:

$$m(R) = \sum_{i=p}^{\min(t,P)} P(i \text{ acertos em } t \text{ casos}) = \sum_{i=p}^{\min(t,P)} \frac{\binom{P}{i} \binom{T-P}{t-i}}{\binom{T}{t}}$$

Valores baixos de $m(R)$ indicam boas regras porque se pretende que a probabilidade de uma regra aleatória, com o mesmo suporte da regra R, dar o mesmo resultado, ou melhor, que a regra R seja baixa.

Em que:

P - nº total de casos da classe nos dados.

T - nº total de casos nos dados.

p - número de casos da classe que a regra selecciona.

t - nº total de casos que a regra selecciona.

Para analisar esta medida de aleatoriedade foi efectuado o seu cálculo para as duas últimas regras apresentadas no capítulo anterior.

Cálculo da aleatoriedade das seguintes regras:

```
R3: GLY_83 → L55
      Sup = 0.40000  Conf = 1.00000
```

$$m(R_3) = \sum_{i=4}^{\min(4,5)} P(i \text{ acertos em } t \text{ casos}) = \sum_{i=4}^{\min(4,5)} \frac{\binom{5}{i} \binom{5}{5-i}}{\binom{10}{5}} = 0.02381$$

R4: PRO_86 → WT

Sup = 0.50000 Conf = 0.83333

$$m(R_4) = \sum_{i=5}^{\min(5,5)} P(i \text{ acertos em } t \text{ casos}) = \sum_{i=5}^{\min(5,5)} \frac{\binom{5}{i} \binom{5}{5-i}}{\binom{10}{5}} = 0.019841$$

A probabilidade de se obterem aleatoriamente regras tão boas ou melhores que as obtidas é relativamente baixa. Significa que se gerarmos aleatoriamente eventos 100 vezes, cerca de uma dessas vezes resultará numa regra semelhante. O valor de 2%, embora baixo, não deixa de apresentar alguns riscos de obtermos uma regras tão boa ou melhor que a obtida ao acaso.

5 Conclusão

Nesta dissertação foi desenvolvido um processo de extracção de conhecimento dos dados obtidos por simulação de desnaturações da proteína TTR para atingir um objectivo, obter mais conhecimento sobre os mecanismos de desnaturação da TTR.

Inicialmente foi efectuada uma análise unidimensional aos dados para a compreensão dos dados em estudo, foram encontradas diferenças interessantes dos valores mínimos e máximos da distância ao centro de massa consoante a variante de TTR. Depois partiu-se para uma análise multidimensional onde são aplicadas técnicas de Data Mining fundamentais para a extracção de conhecimento dos dados.

Uma vez que estavam em análise duas variantes de TTR, obtiveram-se duas partições dos 127 aminoácidos da proteína. Uma primeira partição que considera a informação das 5 simulações da variante WT-TTR, e uma segunda partição que considera a informação das 5 simulações da variante L55P-TTR. Verificou-se que as duas partições obtidas são muito diferentes, no entanto existem aminoácidos que ficam sempre no mesmo *cluster*, o que pode ser interessante para análise do ponto de vista bioquímico.

Com o objectivo de escolher os aminoácidos que representam o comportamento geral da proteína, obteve-se uma partição dos 126 aminoácidos (sem o aminoácido da posição 55), que considera toda a informação das 10 simulações, isto é, obteve-se uma partição consenso. A obtenção desta partição consenso permitiu a aplicação de um método muito interessante e útil, porque partindo de n quadros de dados (ou seja, um cubo de informação) é possível resumir a informação toda numa partição dos indivíduos.

Com esta partição seleccionaram-se os elementos representativos de cada classe e o aminoácido da posição 55, isto é, passou-se de um conjunto de 127 aminoácidos para um conjunto mais pequeno, de 15 aminoácidos. A partir dos 15 aminoácidos representativos foram procurados dois eventos que comparam os aminoácidos dois a dois relativamente ao centro de massa da TTR. Após a obtenção destes eventos, obtiveram-se regras de associação que relacionam o evento com a variante de TTR que tem mais propensão para a ocorrência desse evento. Estas regras podem vir a tornar-se muito úteis porque se se obtiver um elevado grau de certeza da regra, poderá vir a identificar-se o tipo de variante de TTR se ocorrer o antecedente da regra.

Para complementar este estudo foi ainda procurado um terceiro evento por aminoácido, para os 127 aminoácidos da TTR, e obtiveram-se regras que relacionam o evento “alterações bruscas de comportamento do aminoácido” com a variante de TTR que tem mais propensão para verificar esta ocorrência no aminoácido.

Várias poderão ser as questões levantadas ao longo deste trabalho, no Capítulo 4.5 foram abordados alguns pontos, dos quais se pode concluir que existe ainda muito trabalho a fazer para se poderem obter conclusões seguras quanto às características e mecanismos de desnaturação de TTR que possam ajudar na descoberta de conhecimento sobre a Doença dos Pezinhos.

O processo aqui apresentado pode ser aplicado, por exemplo, em valores de SASA obtidos por simulação da desnaturação da TTR, seria interessante comparar os resultados obtidos com os valores de SASA, com os resultados obtidos nesta tese. O mesmo processo pode ainda ser aplicado noutros contextos.

O processo desenvolvido apresenta algumas características que o torna interessante para trabalhos futuros, porque pode ser aplicado a um número grande de simulações, e cada etapa deste processo pode ser desenvolvida e reformulada. Por exemplo, podem ser implementadas outros algoritmos de obtenção de clusters, os clusters podem ser avaliados com outros critérios, os aminoácidos representativos podem ser obtidos de outras formas, podem ser procurados outro tipo de eventos, pode incluir-se o factor tempo aos eventos (desenvolver, por exemplo, um estudo de Data Mining Geo-Temporal), pode ser aplicada uma análise sequencial e podem ainda ser implementadas outras técnicas de Data Mining.

No final desta tese são imensos os pontos positivos, fica a sensação de que se poderá fazer muito mais e de que este trabalho contribuiu de forma positiva para o estudo da TTR.

Bibliografia

- Agrawal, R., Imielinski, T., Swami, A. "Mining association rules between sets of items in large data bases." *SIGMOD Conference On Management of Data*, Maio 1993: 207-216.
- Agrawal, R., Srikant, R. "Fast Algorithms for Mining Association Rules." *International Conference On Very Large Databases, Santiago, Chile*, 1994: 487-499.
- Agrawal, R., Srikant, R. "Mining Sequential Patterns." *Eleventh International Conf. on Data Engineering*, 1995: 3-14.
- Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., Watson, J. *Molecular Biology of The Cell*. Garland Publishing 3ª Edição, 1994.
- Andrade C. "A peculiar form of peripheral neuropathy: familial atypical generalized amyloidosis with special involvement of the peripheral nerves." *Brain*, 1952: 75:408-427.
- Azevedo, P. J. "CAREN A java based Apriori implementation for classification purposes." (*Personal Communication*), 2003.
- Azevedo, P., Cândida Silva, Rodrigues, J., Ferreira, N., Brito, R. "Detection of Hydrophobic Clusters in Molecular Dynamics Protein Simulations Using Association Rules." *ISBMDA* , 2005: 329-337.
- Azevedo, P., Jorge, A. "Comparing Rule Measures for Predictive Association Rules." *Springer Berlin / Heidelberg- European Conference on Machine Learning* , 2007: 510-517.
- Bayardo, R., Agrawal, R. "Mining the Most Interesting Rules." *5th International Conference on Knowledge Discovery and Data Mining, KDD - 99* , 1999: 145-153.
- Berendsen HJC, Van der Spoel D, Van Drunen R. "GROMACS: a messagepassing parallel molecular dynamics implementation." *Comp. Phys. Commum.*, 1995: 91:43-56.
- Berry, M J. A., Linoff, Gordon S. *Data Mining Techniques*. New York: Jonh Wiley & Sons, Inc., 2004.
- Bertrand P., Mufti, G. "Loevinger's measures of rule quality for assessing cluster stability." *Computacional Statistics and Data Analysis* 50, 2005: 992-1015.
- Bertrand, P., Mulfi, G, Moubarki, L. "Determining the number of groups from measures of cluster stability." *Conf. Int. Symp. on Applied Shochastic Models and Data Analysis* , 2005.

- Bittner, T. "Rough sets in spatio-temporal data mining ." *Proceedings of International Workshop on Temporal, Spatial and Spatio-Temporal Data Mining* 89--104.
- Branco, J. *Uma Introdução à Análise de Clusters*. Évora : Congresso Anual de Sociedade Portuguesa de Estatística, 2004.
- Brin, S., Motwani, R., Silverstein, C. "Beyond market baskets: generalizing association rules to correlations." *ACM SIGMOD/PODS'97* , 1997: 265-276 (a).
- Brin, S., Motwani, R., Ullman, J.D., Tsur, S. "Dynamic itemset counting and implication rules for market basket data." *ACM SIGMOD 1997 Int. Conf. on Management of Data*, 1997: 255-264 (b).
- Brito, R., Dubitzky, Rodrigues J. "Protein Folding and Unfolding Simulations: A New Challenge for Data Mining." *A Journal of Integrative Biology* , 2004: Vol 8, Nº 2.
- Brock, G., Pihur, V., Datta, S., Datta, S. "clValid: Validation of Clustering Results." *Contributed Packages*. 27 de 1 de 2008. <http://www.louisville.edu/~g0broc01/research> (acedido em 30 de 1 de 2008).
- Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. "CHARMM: a program for macromolecular energy, minimization, and dynamics calculations." *J. Comput. Chem.*, 1983: 4:187-217.
- Bruzzese D., Davino C. *Significant Knowledge Extraction from Association Rules*. 2006. http://www.iasc-isi.org/Proceedings/2006/COMPSTAT_Satellites/KNEMO/Lavori/Papers%20CD/Davino%20Bruzzese.pdf (acedido em 8 de 2007).
- Bruzzese D., Davino C. "Visual Post- Analisis of Association Rules." *Journal of Visual Languages and Computing*, 2003: 621-635.
- Caiado, J., Crato, N., Peña, D. "A periodogram-based metric for time series classification." *Computacional Statistics & Data Analysis*, 2006: 2668-2684.
- Calinski, R., Harabasz, J. "A dendrite method for clustering analysis." *Communications in Statistics* 3, 1974: 1-27.
- Cejtin, H., Edler, J., Gottlieb, A., Helling, R., Li, H., Philbin, J., Wingreen, N., Tang, C. "Fast tree search for enumeration of a lattice model of protein folding." *The Journal of chemical physics Vol. 116, Nº1*, 2002: 352-359.
- Coenen, F., G. Goulbourne and P. Leng. "Tree Structures for Mining Association Rules ." In *Data Mining and Knowledge Discovery*, 25-51. Kluwer Academic Publishers, 2004.

Costa, J., Bação. "O Papel do Data Mining Geo-Espacial nos Location Based Services." Porto: 3ª Conferência da Associação Portuguesa de Sistemas de Informação Porto, de 15 a 17 Outubro, 2003.

Dafas, P., Garcez, A. "Applied temporal rule mining to time series." *School of Informatics London*. 2006.
http://www.soi.city.ac.uk/project/DOC_TechReport/TR_2006_DOC_01.pdf (acedido em 2007).

Das, G., Lin, K.I., Mannila, H., Renganathan, G., Smyth, P. "Rule Discovery from time series." *Fourth International Conference on Knowledge Discovery and Data Mining*, 1998: 16-22.

Dill, K. A. & Chan, H. S. "From Levinthal to pathways to funnels." *Nature Structure Biology* N°4, 1997: 10-19.

Ewing CS, Thacher TS, Hagler AT. "Derivation of Class II force fields 7. Nonbonded force field parameters for organic compounds." *Journal Physic Chemistry B* 103, 1999: 6998-7014.

Faceli, K., Carvalho, A., Souto, M. "Validação de Algoritmos de Agrupamento." 2005.
ftp://ftp.icmc.usp.br/pub/BIBLIOTECA/rel_tec/RT_254.pdf (acedido em 8 de 2007).

Faísca, P. "O Mistério das Proteínas." *Gazeta da Física, Sociedade portuguesa da Física*, Art. 07, Vol 29_3,, 2004: 33-39.

Fayyad (2), U., Piatetsky-Shapiro, Padhraic S. "From Data Mining to Knowledge Discovery in Databases." *Advances in Knowledge Discovery and Data Mining* (Advances in Knowledge Discovery and Data Mining, AAAI Press), 1996: 1-34.

Fayyad, U., Piatetsky-Shapiro G. "The KDD Process for Extracting Useful knowledge from Volumes of Data." *Communications of the ACM* Vol.39, N°11, November 1996: 27-34.

Feller, S. E., Zhang, Y. H., Pastor, R. W., et al. "Constant-pressure molecular-dynamics simulation - the Langevin piston method." *Journal of Chemistry and Physics* N°103, 1995: 4613-4621.

Ferreira, P., Silva, C., Brito, R., Azevedo, P. "A Closer Look on Protein Unfolding Simulations through Hierarchical Clustering." *Computational Intelligence and Bioinformatics and Computational Biology*, 2007: 461 - 468.

Ganascia, J., G. *Deriving the learning bias from rule properties*. Clarendon Press New York, NY, USA, 1991.

Gomes, C. M. "Implicações do (des)rolamento proteico em patologias humanas." 2005.
<http://www.itqb.unl.pt/~gomes/articles/pat.pdf> (acedido em 5 de 2007).

- Gonçalves, E. “Regras de Associação e suas Medidas de Interesse Objectivas e Subjectivas.” *Journal of Computer Science Vol 6*, 2005.
- Gordon, A. D. *Classification 2nd Edition*. Chapman & Hall/CRC, 1999.
- Grubmuller, Helmut. *Solvate 1.0*. 1996.
www.mpibpc.mog.de/groups/grubmueller/start/software/solvate/docu.html (acedido em 2007).
- Guralnik V., Srivastava J. “Event Detection from Series Data.” *Fifth ACM SIGKDD international conference on Knowledge discovery and data mining* , 1999: 33-42.
- Hahsler, M., Hornik, K. “New Probabilistic Interest Measures for Association Rules.” *Mitarbeiter: Michael Hahsler* . Março de 2007. http://www.ai.wu-wien.ac.at/~hahsler/research/hyperConfidence_IDA2007/hyperConfidence.pdf (acedido em 2007).
- Halkidi, M., Batistakis, Y., Vazirgiannis, M. “On Clustering Validation Techniques.” Montagem por Kluwer Academic Publishers. *Journal of Intelligent Information Systems*, 17:2/3, 2001: 107-145.
- Hartigan, J. In *Clustering Algorithms*, Cap.2;9. John Wiley & Sons, 1975.
- Hennig, C. *Package fpc Version 1.1-1*. 22 de 4 de 2004. <http://cran.r-project.org/web/packages/fpc/index.html> (acedido em 2007).
- Honda, R., Takimito, H., Konoshi, O., “Semantic Indexing and Temporal Rule Discovery for Time-series Satellite Images.” *International Workshop on Multimedia Data Mining*. Boston, 2000. 82-90.
- Honig, B. “Protein Folding: From the Levinthal Paradox to Structure Prediction.” *Journal Molecular Biology*, 1999: 283-293.
- Huang, Y. F., C. M. Wu. “Mining Association Rules Using Pruning Techniques.” *International Conference in Data Mining*, 2002: 227-234.
- Imielinski, T., Mannila, H. “A Database Perspective on Knowledge on Discovery.” *Communications of the ACM N°11 39* (1996): 58-64.
- Jain, A., Dubes, R. *Algorithms for Clustering Data*. Prentice Hall, 1988.
- Jakel, J., Nollenburg, M., “Validation in the Cluster Analysis of Gene Expression Data.” 2004. <http://i11www.iti.uni-karlsruhe.de/members/noelle/pub/jn-vcage-04.pdf> (acedido em 7 de 2007).

Kalé, L., Skeel, R., BBhandarkar, M., Brunner, R., Gursoy, A., Krawetz, N., Philips, J., Shinozaki, A., Varadarajan, K., Schulten, K. "NAMD2: Greater Scability for Parallel Molecular Dynamics." *Journal of Computational Physics*, 1999: 151, 283-312.

Kauzmann, W. "The three dimensional Structures of Proteins." *Biophysical Journal* 4 (1964): 43-54.

Kovacs, F., Legány, C., Babos, A. "Cluster Validity Measurement Techniques." *6th International Symposium of Hungarian Researchers on Computational Intelligence*. Budapest, 2005.

Krzanowski, W., Lai, Y. "A criterion for determining the number of groups in a dataset using sum of squares clustering." *Biometrics* 44 (1985): 23-34.

Lenca, P., Vaillant B., Meyer P., Lallich S. *Association Rule interestingness measures: experimental and theoretical studies*. Vol. 1. Lyon: PM-PP-06-06 , 2006.

Levinthal, C. "Are there pathways for protein folding." *Journal de Chemie Physique* 65, 1968: 44.

Liu, B., Hsu W., Chen, S., Ma, Y. *Analysing the subjective Interestingness of Association Rules*. IEEE, Intelligent Systems, 2000.

Liu, B., Hsu, W., Chen, S. "Analysing the subjective Interestingness of Association Rules." *IEEE Inteligent Systems* 15 (2000): 47-55.

Liu, B., Pednault, E., Smyth, P. "Business Applications of Data Mining." *Communications of the ACM* 45 (2002): 49-53.

Luis, M. "Polineuropatia Amiloidótica Familiar do Tipo Português: do artigo original ao futuro." *Sinapse -Sociedade Portuguesa de Neurologia* 6 (2006): 40-42.

Mannila H., Toivonen H. "Discovering generalized episodes using minimal occurrences." *Knowledge Discovery and Data Mining*, 1996: 146-151.

Mannila, H., Toivonen H., Verkamo A. "Discovery of Frequent Episodes in Event Sequences." *Kluwer Academic Publishers*, 1997: 259-289.

Mannila, H., Toivonen, H., Verkamo A. "Discovering frequent episodes in sequences." *1st Int'l Conference on Knowledge and Data Mining*, 1995: 210-215.

Maroco, J. *Análise Estatística com utilização do SPSS*. Edições Silabo, 2003.

Mennis, J., Liu, J. W. "Mining Association Rules in Spatio- Temporal Data." *7th International Conference on GeoComputation*. 2003.

Milligan, G., Cooper, M. "An examination of procedures for determining the number of clusters in a data set." *Pchychometrika* 50, 1985: 159-179.

Morrison, R., Boyd R. *Química Orgânica*. Fundação Calouste Gulbenkian 6ª Edição Cap. 36, 1992.

Murteira, Bento José Ferreira. *Probabilidades e Estatística*. McGraw-Hill, 1990.

Nunes, A., Barroso, A., Pacheco, J., Tavares, M., Mimoso, J., Gama, M., Faísca, P., Ferreira, P., Patrício, P., Álvares, R., Konotop, V. *PRISMA à Luz da Física*. <http://cftc.cii.fc.ul.pt/PRISMA/capitulos/capitulo4/modulo4> (acedido em 10 de 2007).

Omiecinski, E. R. "Alternative interest measures for mining associations in databases." *IEEE Transaction on Knowledge and Data Engineering* 15, n.º 1 (2003): 57-69.

Pace C. N., Shidley, B. A., Macnutt, M., Gajiwala, K. "Forces contributing to the conformational stability of Proteins." *The FASEB Journal* 10 (1996): 75-83.

Pal, N. R.; Bezdek, J. C. "On Cluster validity for fuzzy c-means model." *IEEE Transactions on Fuzzy Systems* 3 (1995): 370-379.

Pande V., Baker I., Chapman J., Elmer S. P., Khaliq S., Larson S.M., Rhee Y.M., Shirts M.R., Snow C.D., Sorin E.J., Zagrovic B. "Atomistic protein Folding Simulations on the Submillisecond Time Scale Using Worldwide Distributed Computing." *Biopolymers* 68 (2003): 91-109.

Pande, V., Grosberg, A., Tanaka, T., Rokhsar, D. "Pathways for protein folding: is a "new view" needed?" *Currente Opinion in Structural Biology* 8 (1998): 68-79.

Pearlman DA, Case DA, Caldwell JW, Ross SW, Cheatham TE III, et al. "AMBER, a package of computer programs for applying molecular mechanics, normal-mode analysis, molecular-dynamics and free-energy calculations to simulate the structural and energetic properties of molecules." *Comput. Phys. Commum*, 1995: 91:1-41.

Quintas, A., Saraiva, M., Brito, R. "The Tetrameric Protein Transthyretin Dissociates to a Non-native Monomer in Solution." *Biological Chemistry* 274 , n.º 46 (1999): 32943-32949.

Quintas, A., Vaz, D., Cardoso, I., Saraiva, M., Brito, R. "Tetramer Dissociation and Monomer Partial Unfolding Precedes Protofibril Formation in Amyloidogenic Transthyretin Variants." *Biological Chemistry* 276, Nº29, n.º 29 (2001).

R., Pereira M. *Modelo de Predicção da Conformação Tridimensional de Proteínas Globulares a partir do método AB INITIO utilizando Rede Neural com uma função de base radial*. Tese Doutorado, Florianópolis: Universidade Federal Santa Catarina Florianópolis, 2006.

- Robertis, E., Robertis, E. Jr. *Biologia Celular e Molecular*. 13. Fundação Caloustre Gulbenkian, 1987.
- Roddick, J., Spiliopoulou M. "A survey of Temporal Knowledge Discovery Paradigms and Methods." *IEEE Transactions on Knowledge and Data Engineering*. 14 (2002).
- Rousseeuw, P., J. "Sillouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of Computational and Applied Mathematics* 20, 1987: 53-65.
- Saporta, G. *Probabilités analyse des données et statistique*. Gulf Publishing, 1990.
- Sarker, B. K., Mori, T., Uehara, K. "Parallel algorithms for mining association rules in time series data." *Technical Report 2745/2003* (2003): 1-13.
- Sarle, W. S., Kuo, A.H. "The MODECLUS procedure." *Technical Report Cary* (NC:SAS Institute Inc.), 1993: 256.
- Savasere, A., Omiecinski, E., Navathe, S. "An efficient algorithm for mining association rules in large databases." *21th International Conference On Very Large Databases*, 1995: 432-444.
- Shearer, C. "The CRISP-DM Model: The New Blueprint for Data Mining." *Journal of Data Warehousing* 5 N° 4 (2000): 4:19.
- Shirley, B. A., Stanssens, P., Hahn, U. & Pace, C. N. "Contribution of hydrogen bonding to the conformational stability of ribonuclease T1." *Biochemistry* 31 (1992): 725-732.
- Silberschatz, A., Tuzhilin, A. "What Makes Patterns Interesting in Knowledge Discovery Systems." *IEEE Trans. Knowledge and Data Engineering* 8 (1996): 970-974.
- Silva, H. "Métodos de Partição e Validação em Análise Classificatória baseados em Teoria de Grafos." Tese Doutorado, FCUP, 2005.
- Sousa, F. "Novas Metodologias e Validação em Classificação Hierárquica Ascendente." Tese Doutorado, Universidade Nova de Lisboa, 2000.
- Tan, P-N, Kumar, V., Srivastava, J. "Selecting the Right Interestingness Measure for Association Patterns." *Fourth International Conference on Knowledge Discovery and Data Mining*, 2002: 121-127.
- Tan, P-N., Kumar V. "Interesting Measures for Association Pattern: A Perspective." *Workshop on Postprocessing in Machine Learning and Data Mining*, 2000.
- Toivinen, H. "Sampling Large Databases for Association Rules." *22nd Very Large Data Base Conference*, 1996: 134-145 (a).

- Toivonen, H. "Discovery of Pattenrs im Large Data Collections." *Dep.of Computer Science, Seria A, Report A-1996-5*, 1996.
- Veloso, M. "Regras de Associação Aplicadas a um Método de Apoio ao Planeamento de Recursos Humanos." Dissertação de Mestrado em Análise de Dados e Sistemas de Apoio à Decisão, Faculdade Economia Porto, Porto, 2003.
- Vonesh, F.E., Chinchilli, V.M. *Linear and Nonlinear Models for the analysis of repeated measures*. New York: Marcel Dekker, 1997.
- Ward, J. H. "Hierarchical grouping to optimize an objective function." *Journal of the American Statistical Association* 58 (1963): 236-244.
- Warnes, G. "Sample size estimation for microarray experiments." 2006.
http://www.urmc.edu/smd/biostat/people/faculty/manuscripts/0606_warnes_liu.pdf
 (acedido em 9 de 2007).
- Webb, Geoffrey I. "Discovering significant rules." *International Conference on Knowledge Discovery and Data Mining*, 2006: 434-443.
- Wei, C., Li, J., Bumgarner, R. "Sample size for detecting differentially expressed genes microarray experiments." *BMC Genomics*, 2004.
- Wei, G., Sun, Y., Zhou, Y., Feig, M. "Molucular Multiresolution Surfaces." 2005.
http://arxiv.org/PS_cache/math-ph/pdf/0511/0511001v1.pdf (acedido em 2007).
- Witten I., Frank E. *Data Mining: practical machine learning tools abd techniques with Java implementatons*. Morgan Kaufman Publishers, 1999.
- Wolynes, P. G., Onuchic, J. N. & Thirumalai, D. "Navigating the folding routes." *Science* 267 (1995): 1619-1620.
- Zhang, L., Hermans, J. "Hydrophilicity of cavities in proteins." *Proteins: Structure, Function and Genetics* 24, 1996: 433-438.
- Zien, A., Fluck, J., Zimmer, R., Lengauer, T. "Microarrays: How many do you need?" *Journal of Computational Biology*, 2003: 653-667.

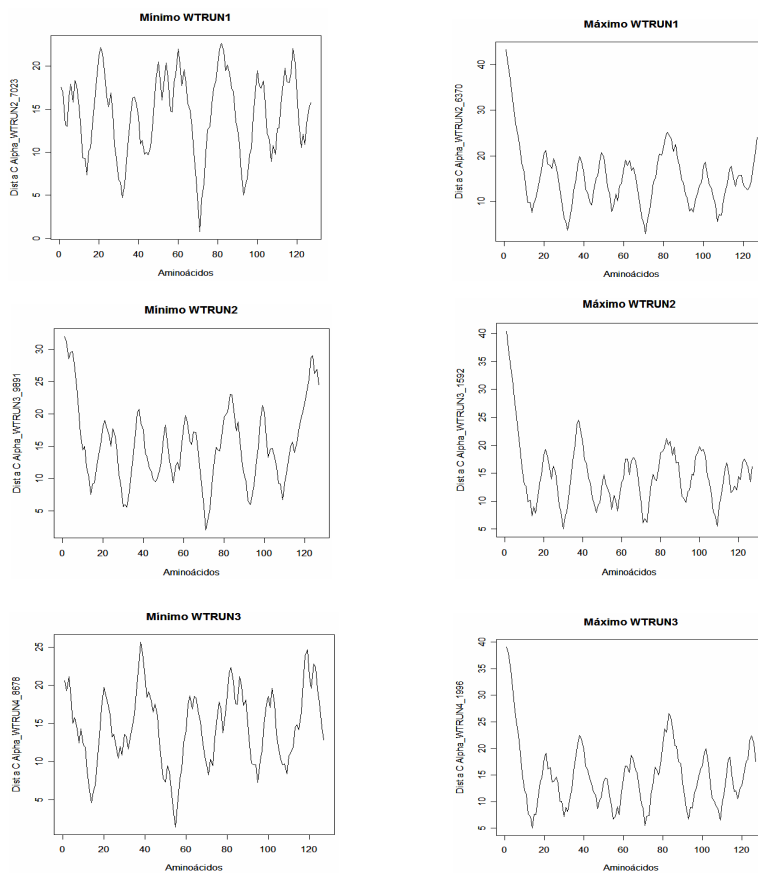
Anexos

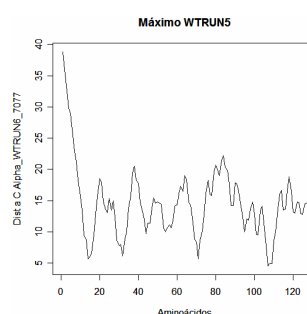
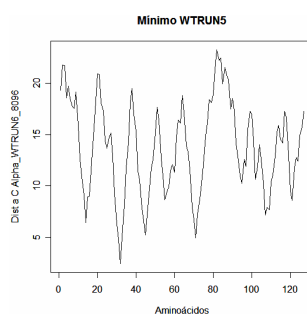
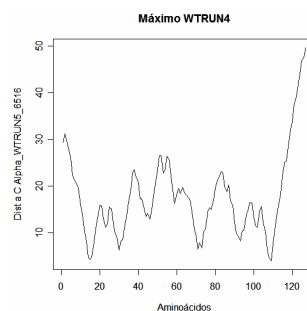
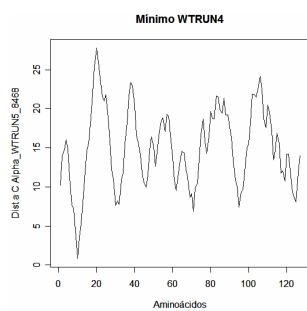
Anexo I – Análise Univariada

A. Valores máximos e mínimos atingidos em cada corrida

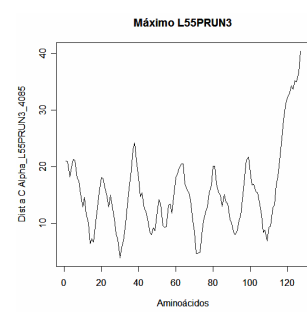
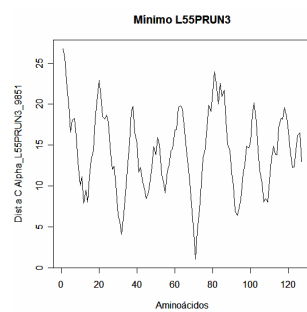
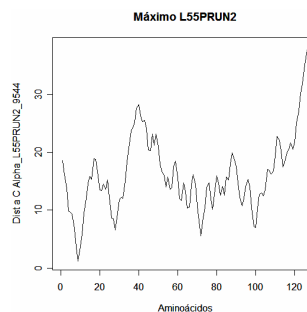
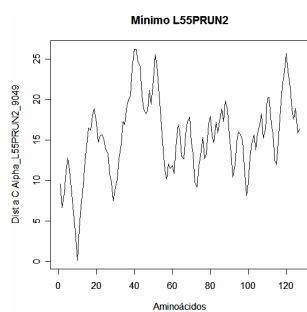
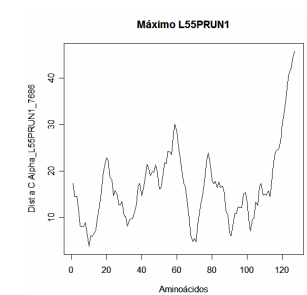
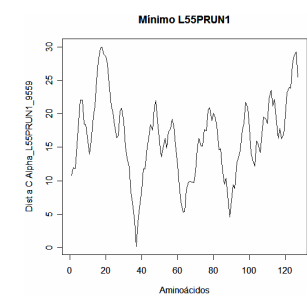
Cada gráfico representa a posição dos 127 aminoácidos no instante onde é atingido o valor máximo de distância ao centro de massa, e o valor mínimo ao longo da simulação da desnaturação da proteína TTR.

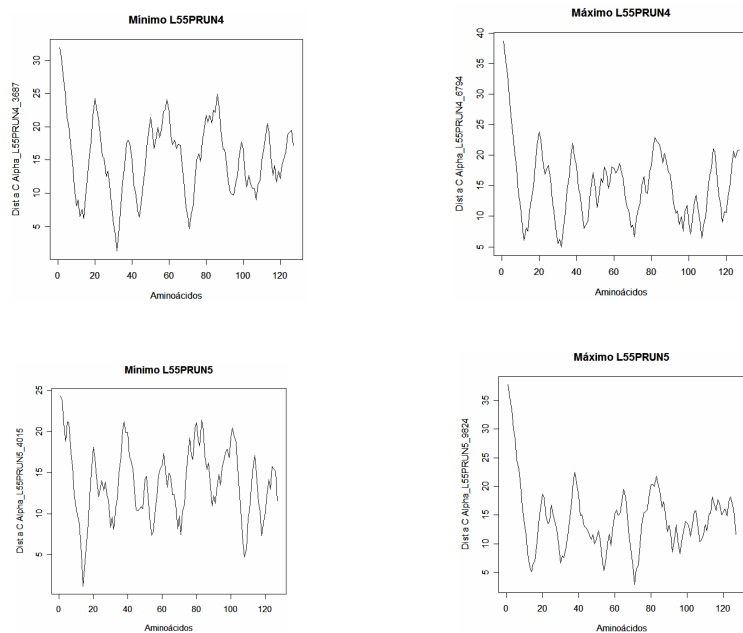
Para a variante WT-TTR temos 10 gráficos, dois para cada corrida. Por exemplo, na corrida WTRUN2 o valor mínimo da distância ao centro de massa é atingido aos 7023 ps pelo aminoácido que fica entre a posição 60 e 80, isto é, pela VAL_71.





Para a variante L55P-TTR são apresentados os 10 gráficos seguintes:





B. Gráficos do aminoácido da posição 55 ao longo das 10 simulações

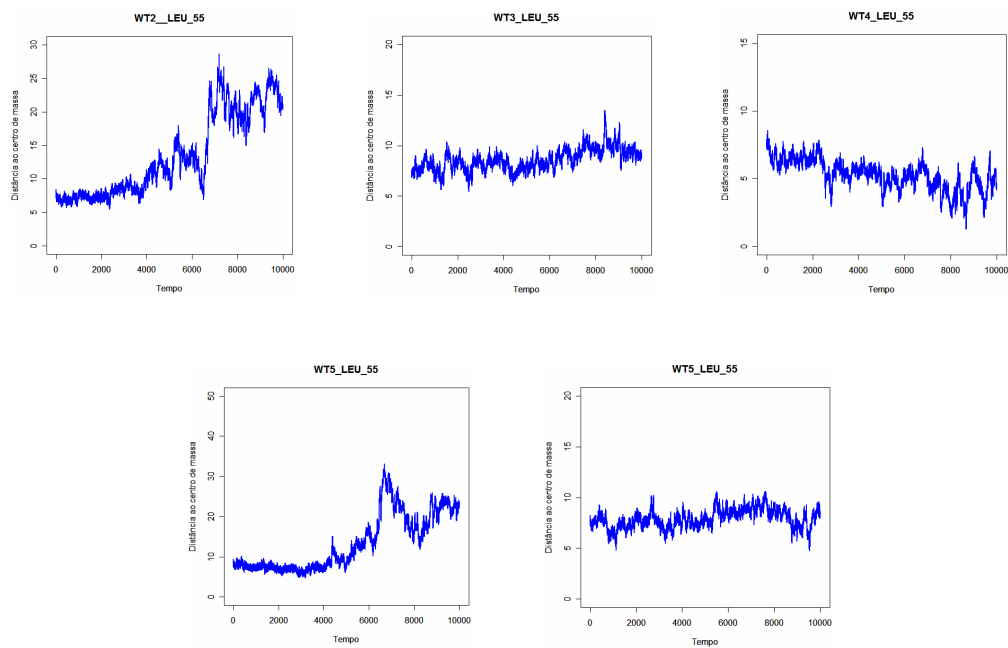


Figura 50 – Evolução da distância ao centro de massa do aminoácido LEU_55 da variante WT-TTR.

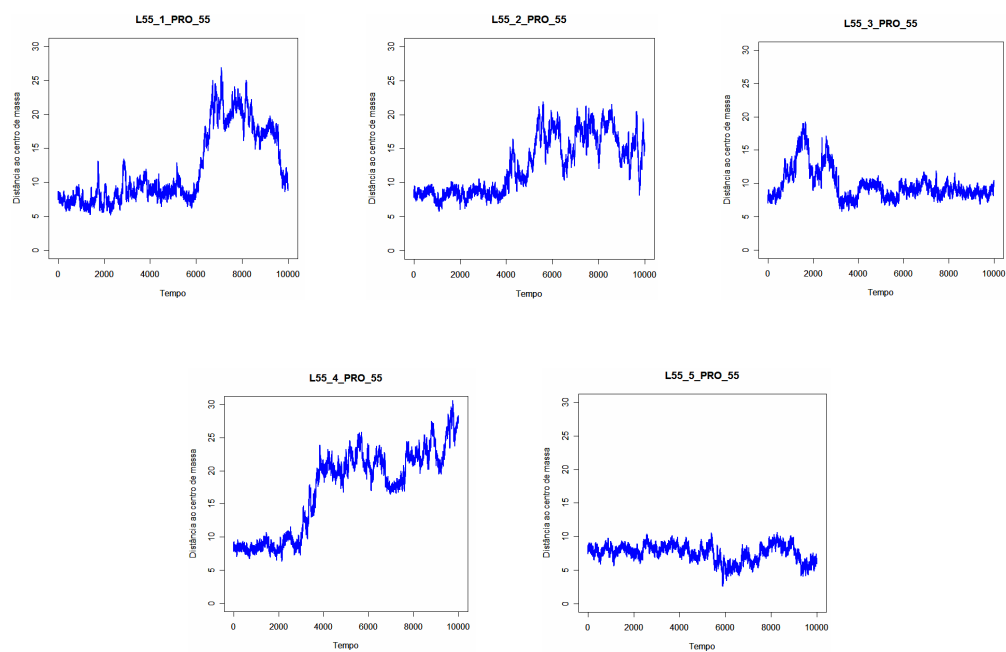


Figura 51 – Evolução da distância ao centro de massa do aminoácido LEU_55 da variante WT-TTR.

Anexo II – Classificação hierárquica - Resultados

A. Clustering da WT-TTR

A1- Optimização dos Índices de Milligan e Cooper

Na simulação WTRUN2, pela análise gráfica do índice G2 (que deve ser maximizado), verifica-se que a quebra mais acentuada ocorre de 14 para 15, isto é, existe uma mudança significativa no valor dos índices para 14 *clusters*. No índice G3 (que deve ser minimizado) existe uma queda acentuada de 13 para 14 classes. O número de classes escolhido foi 14.

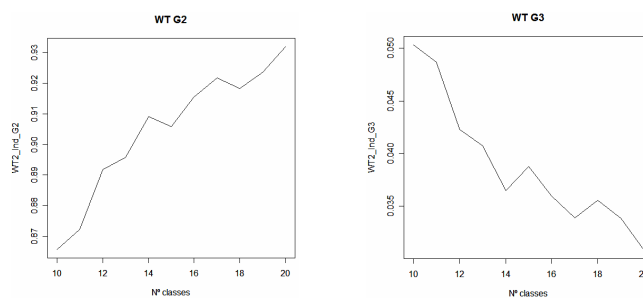


Figura 52 - Índices de Milligan e Cooper – G2 e G3 para a simulação WTRUN2

Para o quadro de dados WTRUN3 verifica-se que o número de classes que otimiza os índices é 13.

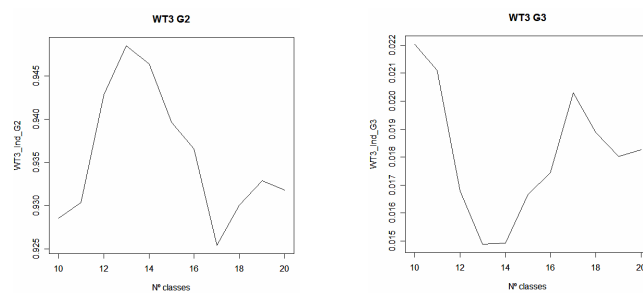


Figura 53 - Índices de Milligan e Cooper – G2 e G3 para a simulação WTRUN3

Quanto às restantes simulações o critério de escolha do número de classes é análogo. Na corrida WTRUN4 o número de classes onde se verificam alterações significativas dos índices é 14. O mesmo acontece na corrida WTRUN5. Quanto à corrida WTRUN6 os gráficos obtidos não apresentam grandes alterações, no entanto para 13 classes apresentam uma pequena perturbação.

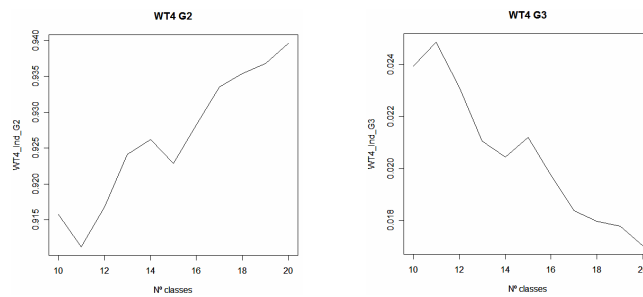


Figura 54 - Índices de Milligan e Cooper – G2 e G3 para a simulação WTRUN4

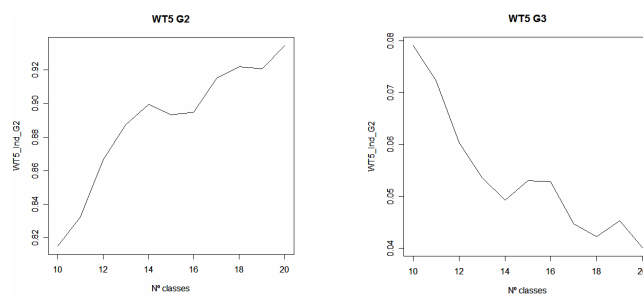


Figura 55 - Índices de Milligan e Cooper – G2 e G3 para a simulação WTRUN5

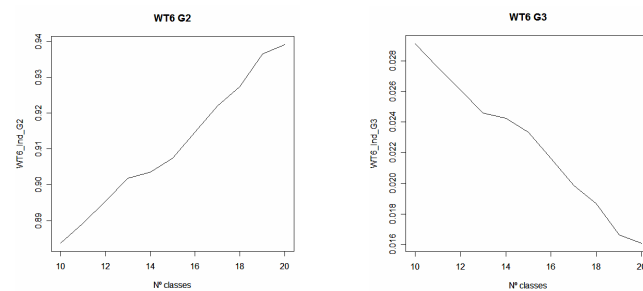


Figura 56 - Índices de Milligan e Cooper – G2 e G3 para a simulação WTRUN6

A2- Custering de cada corrida da WT-TTR

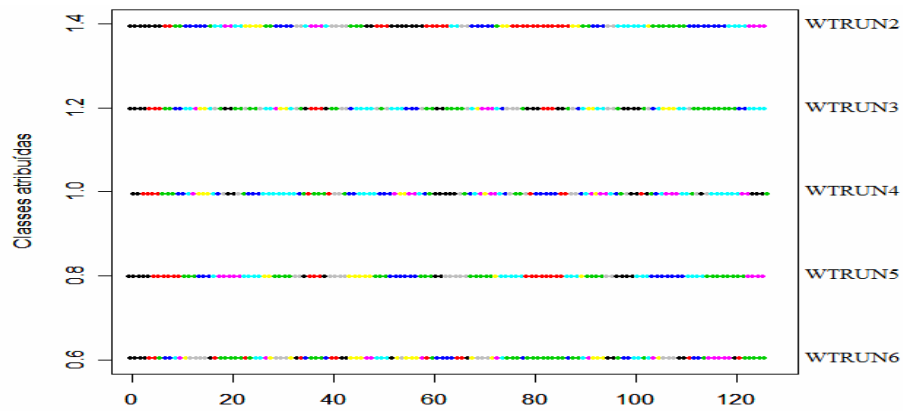
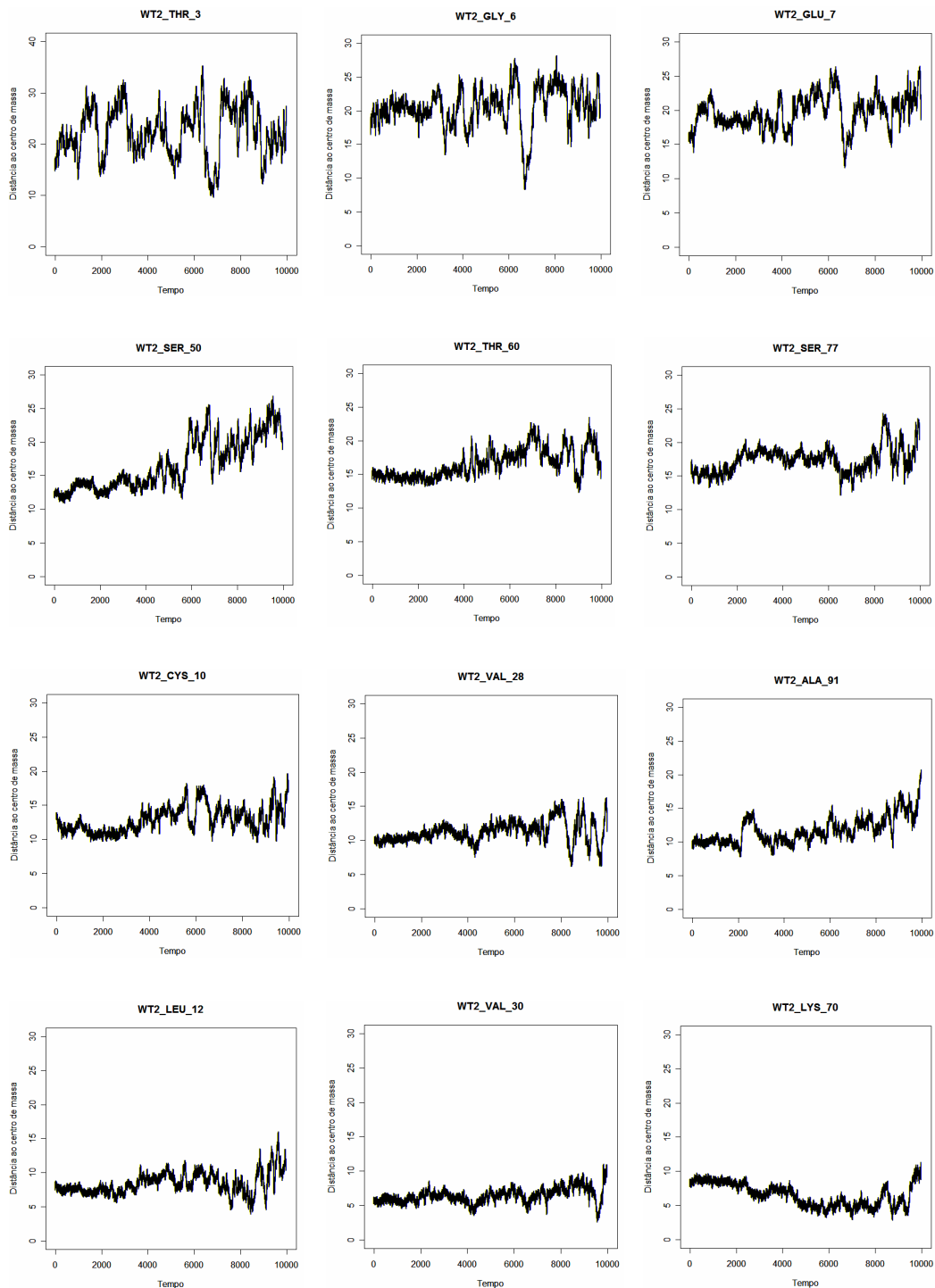


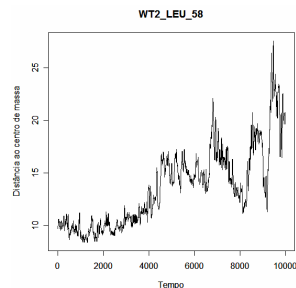
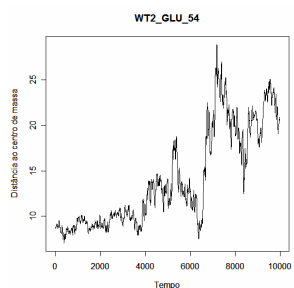
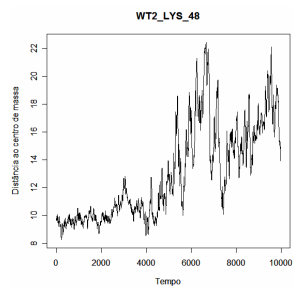
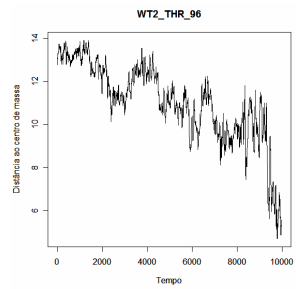
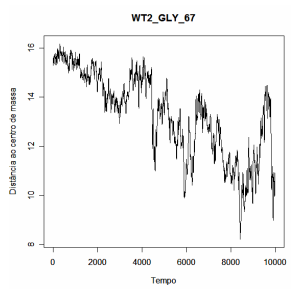
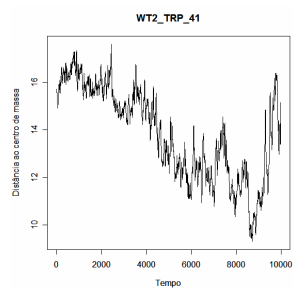
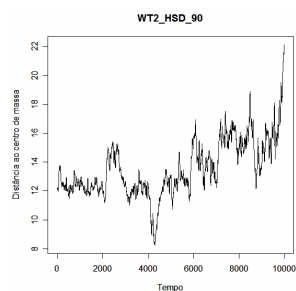
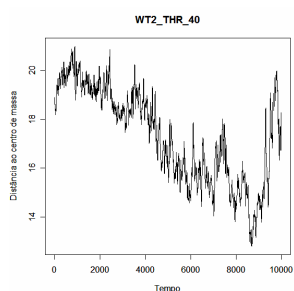
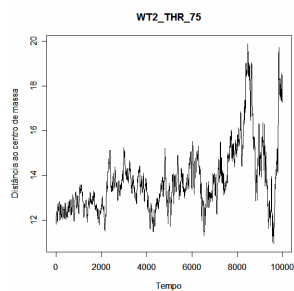
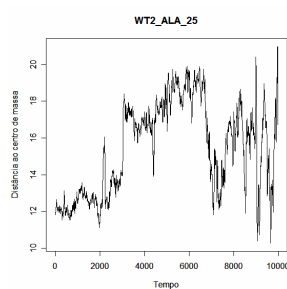
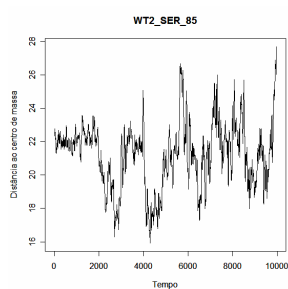
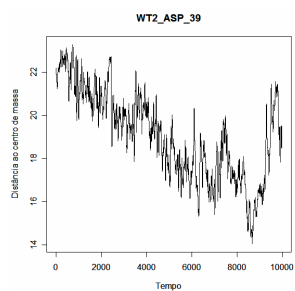
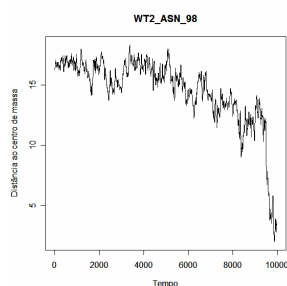
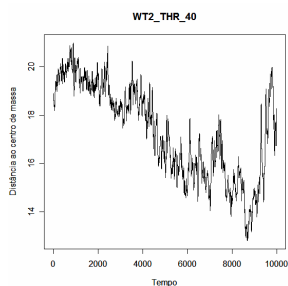
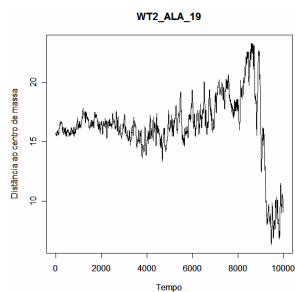
Figura 57 - Clustering de cada corrida da WT-TTR

Tabela 15 - WTRUN2- Classificação hierárquica: elementos das 14 classes

1	2	3	4	5	6	7	8	9	10	11	12	13	14
GLY_1	SER_8	CYS_10	LEU_12	ASP_18	VAL_20	PRO_24	ARG_34	LYS_48	ALA_81	ILE_107	SER_112	ALA_120	ASN_124
PRO_2	LYS_9	PRO_11	MET_13	ALA_19	ARG_21	ALA_25	LYS_35	THR_49	LEU_82	ALA_108	PRO_113	VAL_121	PRO_125
THR_3	SER_50	LEU_17	VAL_14	GLY_22	ALA_37	ILE_26	TRP_41	GLY_53	GLY_83	ALA_109	TYR_114	VAL_122	LYS_126
GLY_4	GLU_51	VAL_28	LYS_15	SER_23	ASP_38	ASN_27	GLU_42	GLU_54	ILE_84	LEU_110	SER_115	THR_123	GLU_127
THR_5	SER_52	ALA_29	VAL_16	ALA_36	ASP_39	THR_75	PRO_43	LEU_55	SER_85	LEU_111	TYR_116		
GLY_6	THR_60	ALA_45	VAL_30	THR_40		LYS_76	PHE_44	HSD_56	PRO_86		SER_117		
GLU_7	GLU_61	SER_46	HSD_31	VAL_65		GLU_89	GLY_67	GLY_57	PHE_87		THR_118		
	GLU_62	GLY_47	VAL_32	GLU_66		HSD_90	ILE_68	LEU_58			THR_119		
	GLU_63	ASP_74	PHE_33	ASN_98		ARG_104	THR_96	THR_59					
	PHE_64	ALA_91	TYR_69	ASP_99			ALA_97						
	SER_77	GLU_92	LYS_70	SER_100									
	TYR_78	TYR_105	VAL_71	GLY_101									
	TRP_79	THR_106	GLU_72	PRO_102									
	LYS_80		ILE_73	ARG_103									
	HSD_88		VAL_93										
			VAL_94										
			PHE_95										

Gráficos para visualização do comportamento de alguns aminoácidos da simulação WTRUN2.





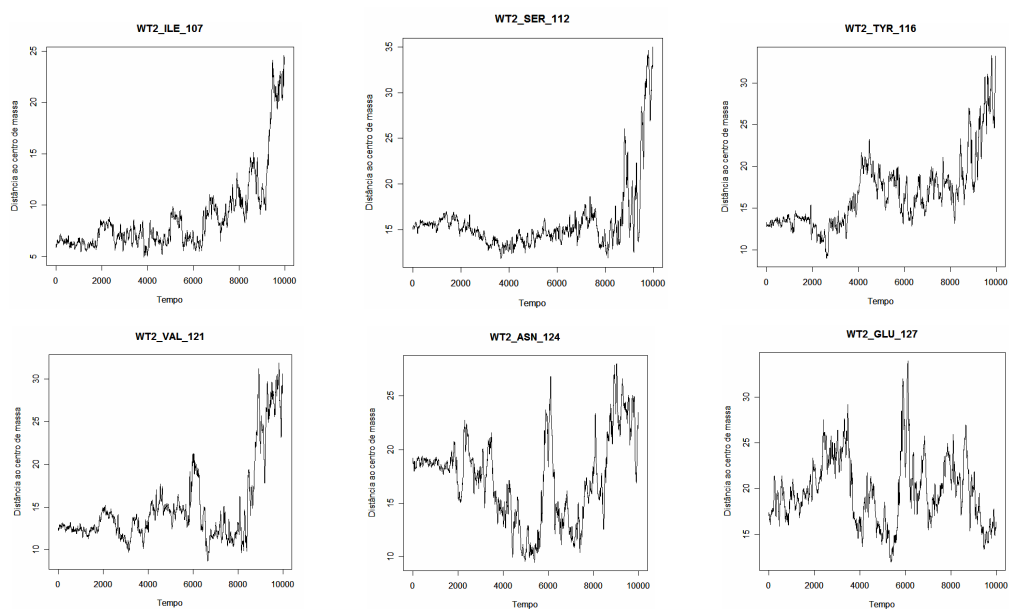
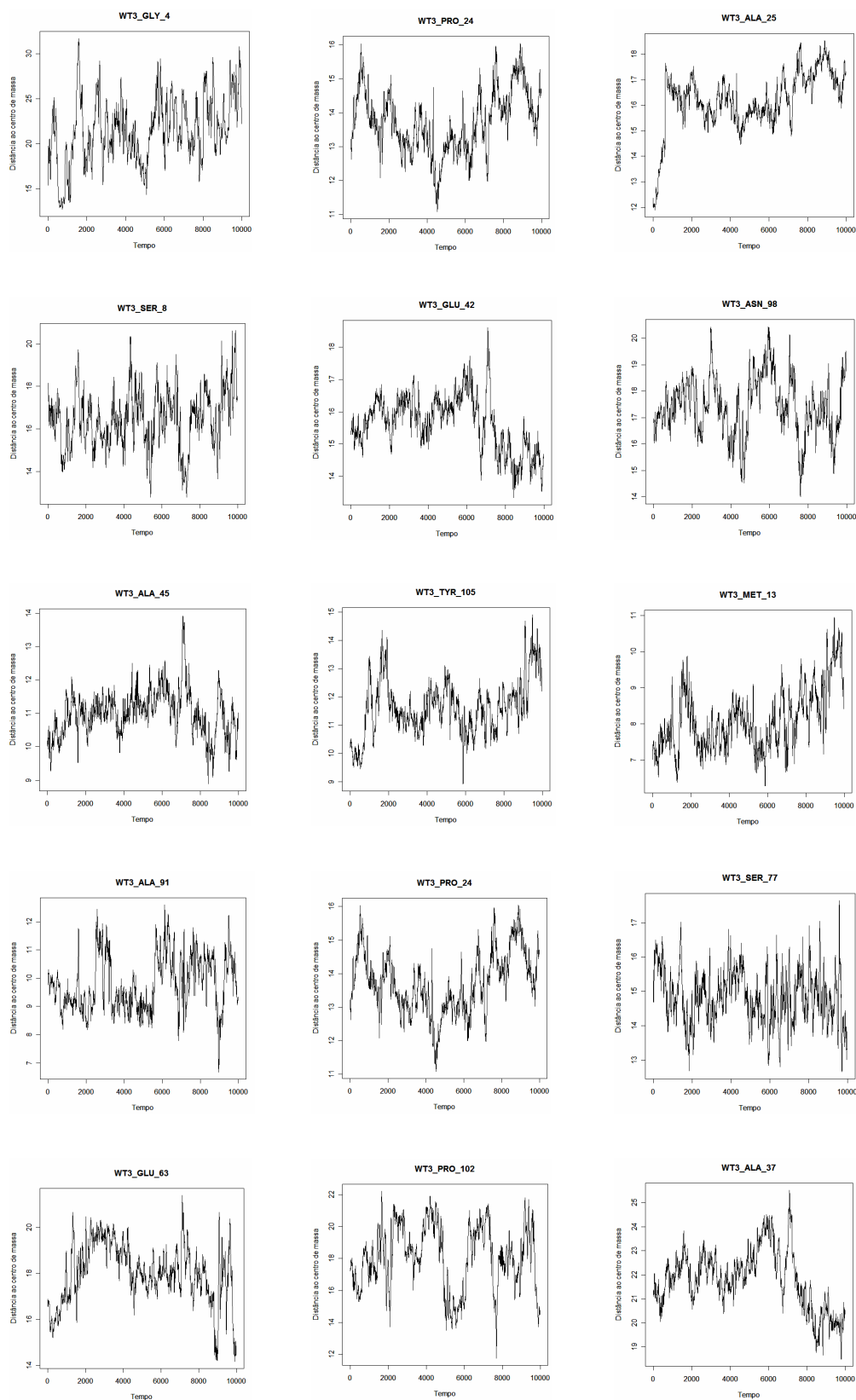


Tabela 16 - WTRUN3- Classificação hierárquica: elementos das 13 classes

1	2	3	4	5	6	7	8	9	10	11	12	13
GLY_1	THR_5	SER_8	CYS_10	LEU_12	VAL_14	LYS_15	ASP_18	VAL_20	ALA_37	TYR_116	VAL_122	ASN_124
PRO_2	GLY_6	LYS_9	PRO_11	MET_13	VAL_30	VAL_16	PRO_24	ARG_21	ASP_38	SER_117	THR_123	PRO_125
THR_3	GLU_7	ALA_19	ALA_45	LEU_17	VAL_71	HSD_31	ASN_27	ALA_36	ASP_39	THR_118		LYS_126
GLY_4		GLY_22	HSD_56	VAL_28	GLU_72	VAL_32	ARG_34	THR_40	GLY_83	THR_119		GLU_127
		SER_23	GLY_57	ALA_29	ILE_73	LYS_70	PRO_43	GLU_62	ILE_84	ALA_120		
		ALA_25	LEU_58	PHE_33		GLU_92	PHE_44	GLU_63	SER_85	VAL_121		
		ILE_26	THR_75	SER_46		VAL_93	SER_50	LYS_80				
		LYS_35	HSD_90	GLY_47		ILE_107	SER_52	ALA_81				
		TRP_41	TYR_105	LYS_48		ALA_108	THR_59	LEU_82				
		GLU_42		THR_49		ALA_109	ILE_68	PRO_86				
		GLU_51		GLY_53			LYS_76	PHE_87				
		THR_60		GLU_54			SER_77	ASP_99				
		GLU_61		LEU_55			TYR_78	SER_100				
		PHE_64		TYR_69			GLU_89	GLY_101				
		VAL_65		ASP_74			THR_96	PRO_102				
		GLU_66		ALA_91			ALA_97					
		GLY_67		VAL_94			ARG_104					
		TRP_79		PHE_95			SER_112					
		HSD_88		THR_106								
		ASN_98		LEU_110								
		ARG_103		LEU_111								
		PRO_113										
		TYR_114										
		SER_115										

Gráficos para visualização do comportamento de alguns aminoácidos da simulação WTRUN3.



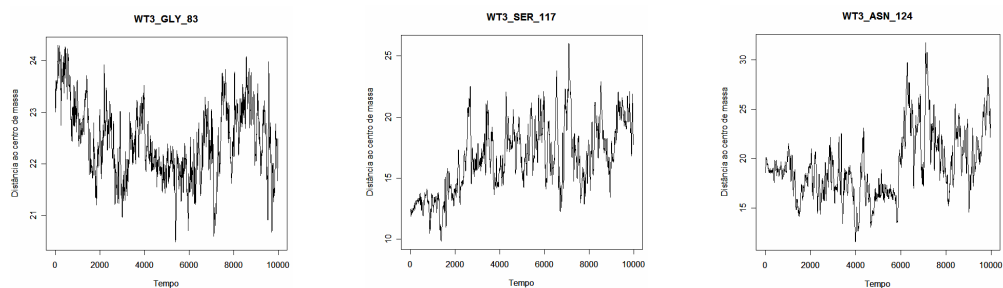
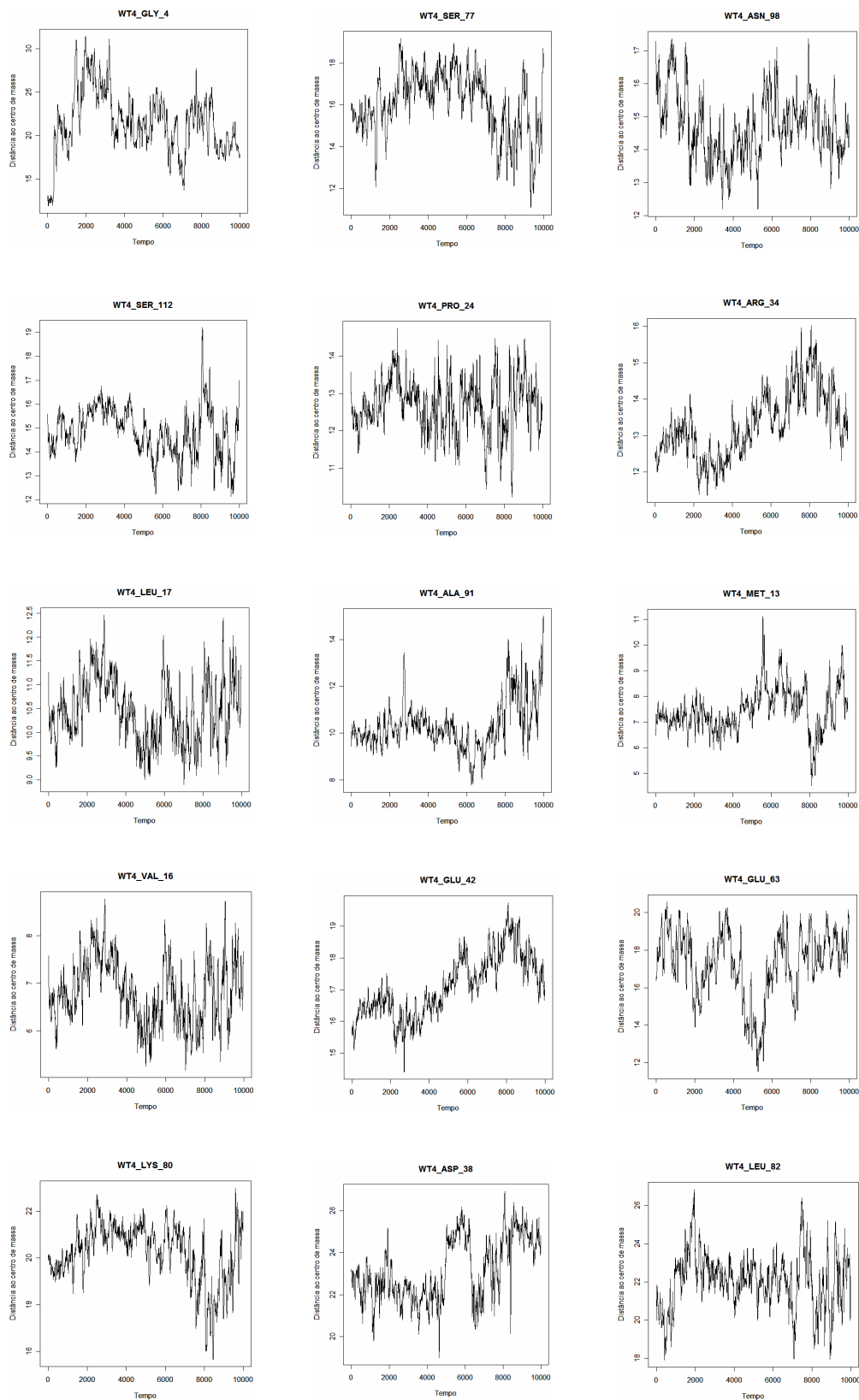


Tabela 17 - WTRUN4- Classificação hierárquica: elementos das 14 classes

1	2	3	4	5	6	7	8	9	10	11	12	13	14
GLY_1	THR_3	GLU_7	CYS_10	LEU_12	MET_13	VAL_14	ALA_19	VAL_20	ALA_36	ALA_37	ALA_81	TYR_116	VAL_122
PRO_2	GLY_4	SER_8	PRO_11	LEU_17	GLY_53	LYS_15	GLY_22	ARG_21	THR_40	ASP_38	LEU_82	SER_117	THR_123
	THR_5	LYS_9	ASP_18	ASN_27	HSD_56	VAL_16	TRP_41	GLU_61	LYS_80	ASP_39	GLY_83	THR_118	
	GLY_6	SER_23	PRO_24	VAL_28	GLY_57	GLU_54	GLU_42	GLU_62	PRO_86		ILE_84	THR_119	
		LYS_35	ALA_25	ALA_29	LYS_70	LEU_55	GLU_66	GLU_63	PHE_87		SER_85	ALA_120	
		PRO_43	ILE_26	VAL_30	GLU_72	VAL_71	LYS_76	PHE_64	PRO_102			VAL_121	
		THR_60	ARG_34	HSD_31	ILE_73	VAL_93	TRP_79	VAL_65					
		GLY_67	PHE_44	VAL_32	GLU_92		HSD_88	SER_100					
		SER_77	ALA_45	PHE_33	VAL_94		GLU_89	GLY_101					
		TYR_78	SER_50	SER_46	PHE_95		ASP_99	ARG_103					
		ASN_98	GLU_51	GLY_47	ILE_107		PRO_113	TYR_114					
		ARG_104	SER_52	LYS_48	ALA_108		SER_115	ASN_124					
		SER_112	THR_59	THR_49	ALA_109			PRO_125					
		GLU_127	ILE_68	LEU_58				LYS_126					
			THR_75	TYR_69									
			HSD_90	ASP_74									
			ALA_97	ALA_91									
			TYR_105	THR_96									
				THR_106									
				LEU_110									
				LEU_111									

Gráficos para visualização do comportamento de alguns aminoácidos da simulação WTRUN4.



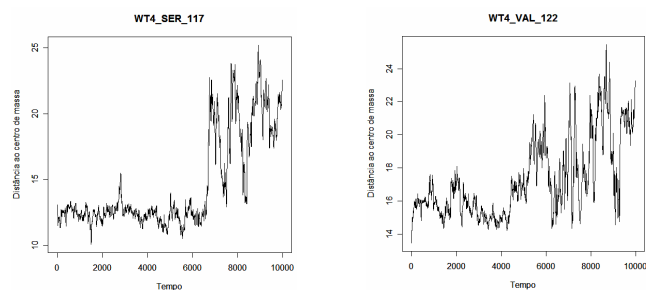
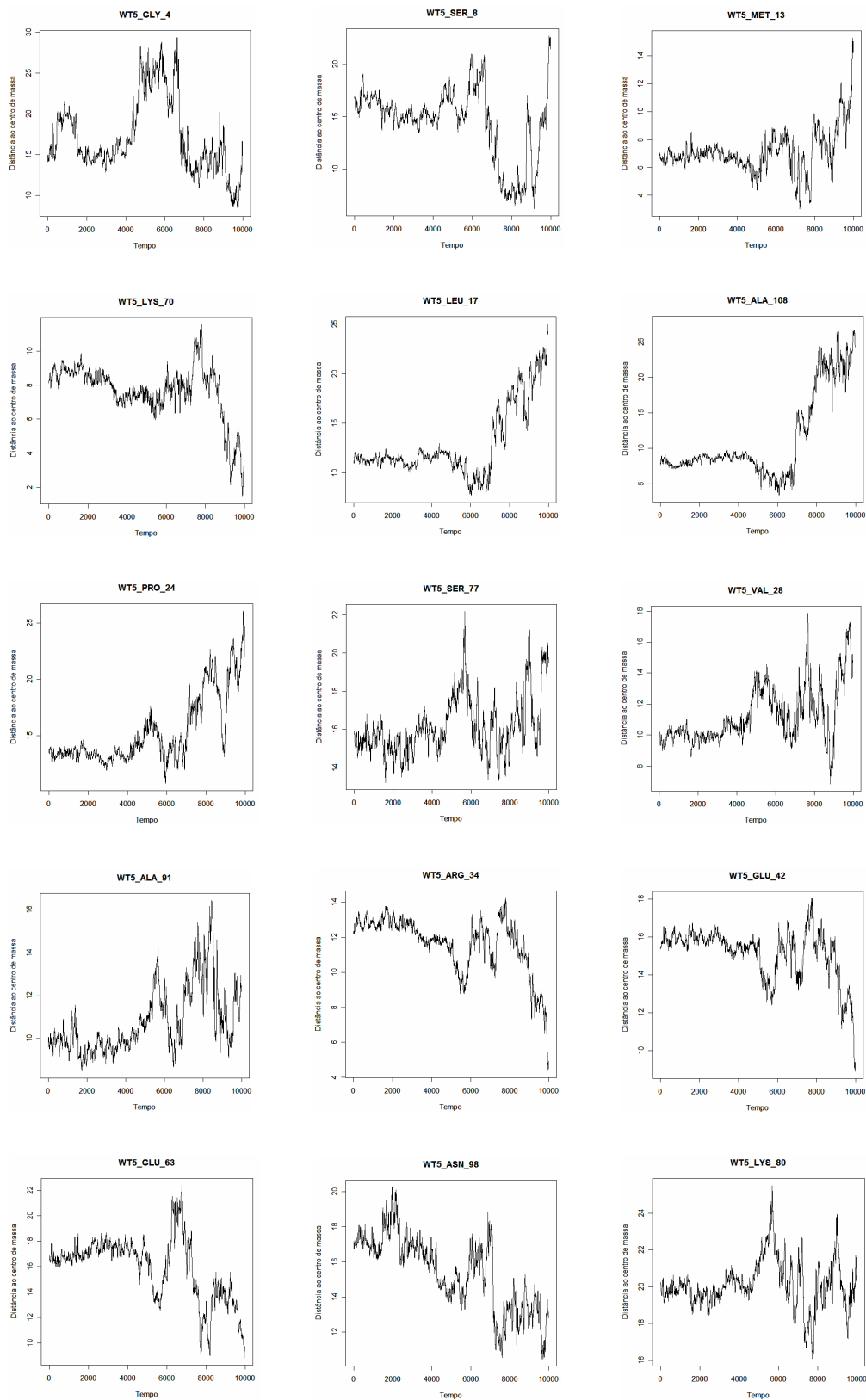


Tabela 18 - WTRUN5- Classificação hierárquica: elementos das 14 classes

1	2	3	4	5	6	7	8	9	10	11	12	13	14
GLY_1	GLY_6	LEU_12	LYS_15	ASP_18	ALA_19	VAL_28	ARG_34	ALA_36	ALA_37	SER_50	GLY_53	SER_112	ASN_124
PRO_2	GLU_7	MET_13	VAL_16	PRO_24	VAL_20	ALA_29	LYS_35	THR_40	ASP_38	GLU_51	GLU_54	PRO_113	PRO_125
THR_3	SER_8	VAL_14	LEU_17	ALA_25	ARG_21	ALA_45	TRP_41	GLU_62	ASP_39	SER_52	LEU_55	TYR_114	LYS_126
GLY_4	LYS_9	VAL_30	TYR_105	ILE_26	GLY_22	SER_46	GLU_42	GLU_63	LYS_80	THR_59	HSD_56	SER_115	GLU_127
THR_5	CYS_10	HSD_31	THR_106	ASN_27	SER_23	GLY_47	PRO_43	ASN_98	ALA_81	THR_60	GLY_57		
	PRO_11	VAL_32	ILE_107	THR_75		LYS_48	PHE_44	ASP_99	LEU_82	GLU_61	LEU_58		
		PHE_33	ALA_108	LYS_76		THR_49	PHE_64	SER_100	GLY_83	TYR_116			
		TYR_69	ALA_109	SER_77		ASP_74	VAL_65	GLY_101	ILE_84	SER_117			
		LYS_70	LEU_110	TYR_78		ALA_91	GLU_66		SER_85	THR_118			
		VAL_71	LEU_111	TRP_79			GLY_67		PRO_86	THR_119			
		GLU_72		HSD_88			ILE_68		PHE_87	ALA_120			
		ILE_73		GLU_89			THR_96			VAL_121			
		GLU_92		HSD_90			ALA_97			VAL_122			
		VAL_93		PRO_102						THR_123			
		VAL_94		ARG_103									
		PHE_95		ARG_104									

Gráficos para visualização do comportamento de alguns aminoácidos da simulação WTRUN5.



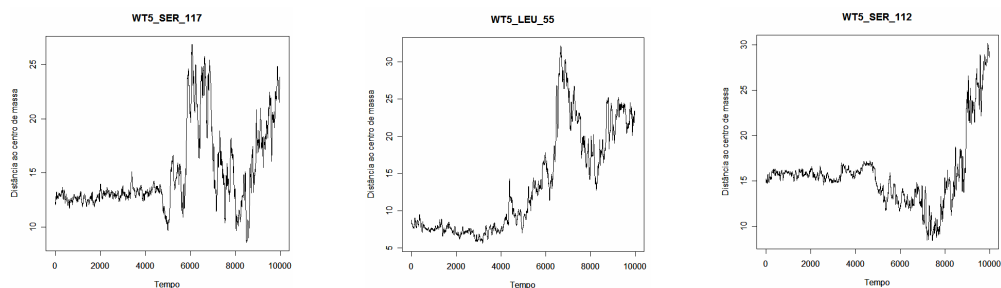
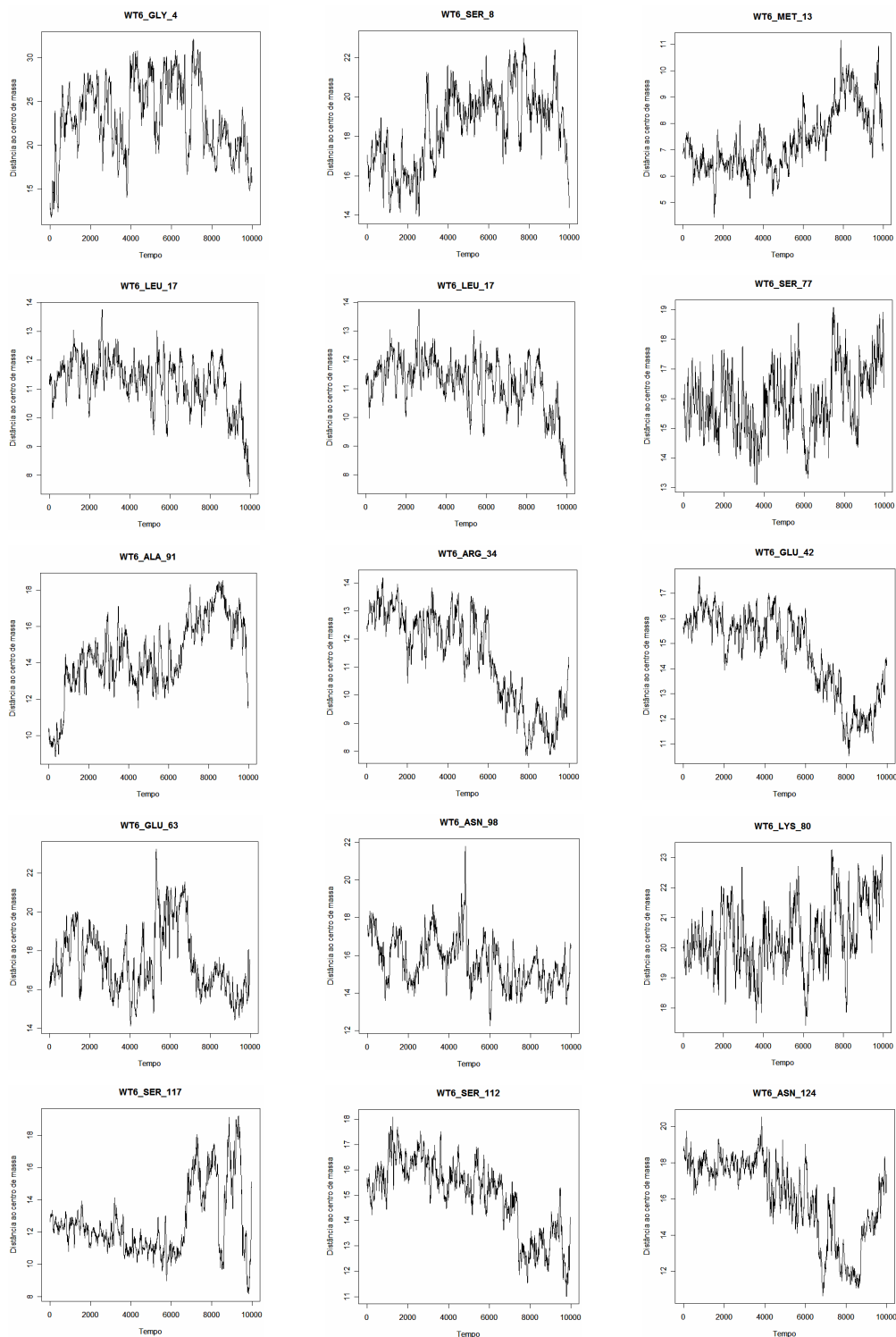


Tabela 19 - WTRUN6- Classificação hierárquica: elementos das 13 classes

1	2	3	4	5	6	7	8	9	10	11	12	13
GLY_1	THR_5	GLU_7	SER_8	CYS_10	PRO_11	LEU_12	MET_13	LEU_17	ASP_18	ALA_19	ALA_36	GLY_101
PRO_2	GLY_6	VAL_20	LYS_9	ILE_26	VAL_28	ALA_29	VAL_14	ARG_34	PRO_24	GLY_22	THR_40	PRO_102
THR_3		ARG_21	GLU_62	ASN_27	LYS_48	PHE_33	LYS_15	PRO_43	LYS_35	SER_23	ASP_99	
GLY_4		ALA_37	GLU_63	SER_50	THR_49	ALA_45	VAL_16	PHE_44	TRP_41	ALA_25	SER_100	
		ASP_38	PHE_64	GLU_51	THR_59	SER_46	VAL_30	GLY_53	GLU_42	GLU_61	PRO_113	
		ASP_39	VAL_65	SER_52	ASP_74	GLY_47	HSD_31	ILE_68	THR_60	LYS_76	TYR_114	
		LYS_80	TRP_79	THR_75	GLU_92	GLU_54	VAL_32	THR_96	GLU_66	SER_77		
		ALA_81	HSD_88	ALA_91	TYR_105	HSD_56	LEU_55	ALA_97	GLY_67	TYR_78		
		LEU_82		ARG_104	TYR_116	GLY_57	LYS_70	LEU_110	SER_112	GLU_89		
		GLY_83			SER_117	LEU_58	VAL_71	LEU_111	VAL_122	HSD_90		
		ILE_84			THR_118	TYR_69	GLU_72	VAL_121		ASN_98		
		SER_85			THR_119	ILE_73	ILE_107			ARG_103		
		PRO_86			ALA_120	VAL_93	ALA_108			SER_115		
		PHE_87				VAL_94	ALA_109			THR_123		
						PHE_95				ASN_124		
						THR_106				PRO_125		
										LYS_126		
										GLU_127		

Gráficos para visualização do comportamento de alguns aminoácidos da simulação WTRUN6.



A3 - Partição consenso da WT-TTR

Clustering da proteína WT-TTR usando os 5 clusterings anteriores de cada simulação.

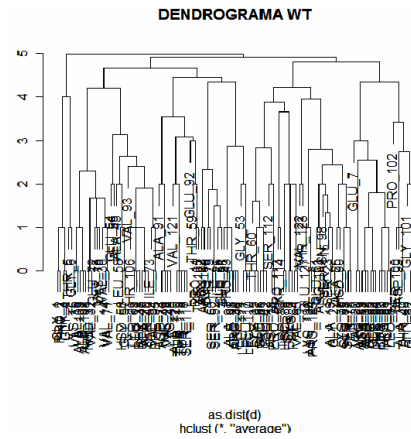


Figura 58 - Dendrograma WT-TTR

Corte do dendrograma em h=4.

Tabela 20 – Clustering da WT (Partição I)

1	2	3	4	5	6	7	8	9	10	11
GLY_1	SER_8	CYS_10	LEU_12	ASP_18	VAL_20	ARG_34	ALA_36	GLY_53	TYR_116	ASN_124
PRO_2	LYS_9	PRO_11	MET_13	PRO_24	ARG_21	LYS_35	THR_40	GLU_54	SER_117	PRO_125
THR_3	ALA_19	LEU_17	VAL_14	ALA_25	ALA_37	TRP_41	ASP_99	LEU_55	THR_118	LYS_126
GLY_4	GLY_22	VAL_28	LYS_15	ILE_26	ASP_38	GLU_42	SER_100	HSD_56	THR_119	GLU_127
THR_5	SER_23	ALA_29	VAL_16	ASN_27	ASP_39	PRO_43	GLY_101	GLY_57	ALA_120	
GLY_6	THR_60	ALA_45	VAL_30	SER_50	LYS_80	PHE_44	PRO_102	LEU_58	VAL_121	
GLU_7	GLU_61	SER_46	HSD_31	GLU_51	ALA_81	GLU_66			VAL_122	
	GLU_62	GLY_47	VAL_32	SER_52	LEU_82	GLY_67			THR_123	
	GLU_63	LYS_48	PHE_33	THR_59	GLY_83	ILE_68				
	PHE_64	THR_49	TYR_69	THR_75	ILE_84	THR_96				
	VAL_65	ASP_74	LYS_70	LYS_76	SER_85	ALA_97				
	TRP_79	ALA_91	VAL_71	SER_77	PRO_86					
	HSD_88	TYR_105	GLU_72	TYR_78	PHE_87					
	ASN_98	THR_106	ILE_73	GLU_89						
	ARG_103	LEU_110	GLU_92	HSD_90						
	SER_112	LEU_111	VAL_93	ARG_104						
	PRO_113		VAL_94							
	TYR_114		PHE_95							
	SER_115		ILE_107							
			ALA_108							
			ALA_109							

B. Clustering da L55P-TTR

B1 – Optimização dos Índices de Milligan e Cooper

Quanto à escolha do número de classes para a proteína L55P-TTR o raciocínio é o mesmo. Desta forma, na primeira corrida, o número de classes escolhido é 15 como se pode observar na figura seguinte:

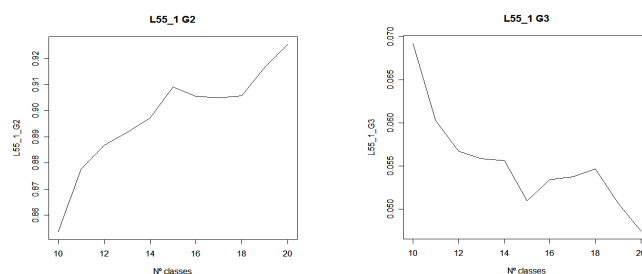


Figura 59 - Índices de Milligan e Cooper – G2 e G3 para a simulação L55PRUN1

Na corrida L55PRUN2 a escolha do número de *clusters* não é fácil uma vez que os gráficos se aproximam de rectas. No entanto, para 19 classes, o valor do Índice G2 apresenta uma pequena flutuação, e o valor do Índice G3 decresce bastante de 18 para 17 classes.

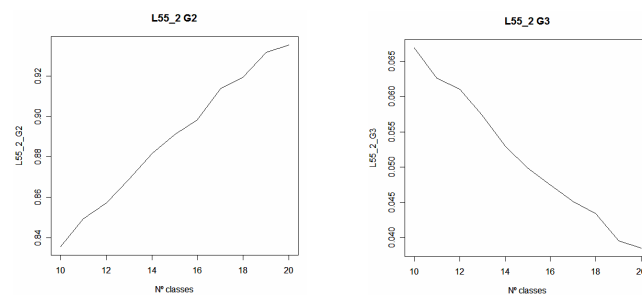


Figura 60 - Índices de Milligan e Cooper – G2 e G3 para a simulação L55PRUN2

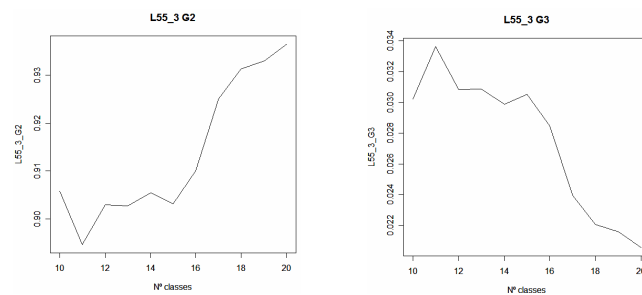


Figura 61 - Índices de Milligan e Cooper – G2 e G3 para a simulação L55PRUN3

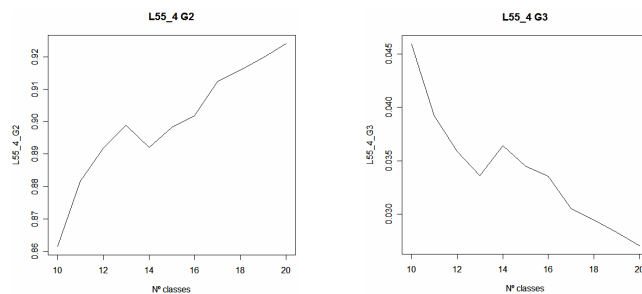


Figura 62 - Índices de Milligan e Cooper – G2 e G3 para a simulação L55PRUN4

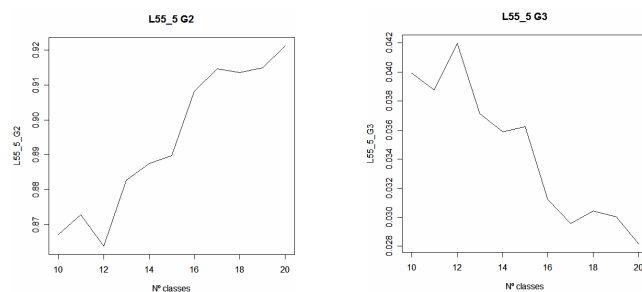


Figura 63 - Índices de Milligan e Cooper – G2 e G3 para a simulação L55PRUN5

B2 - Clustering de cada corrida da L55P-TTR

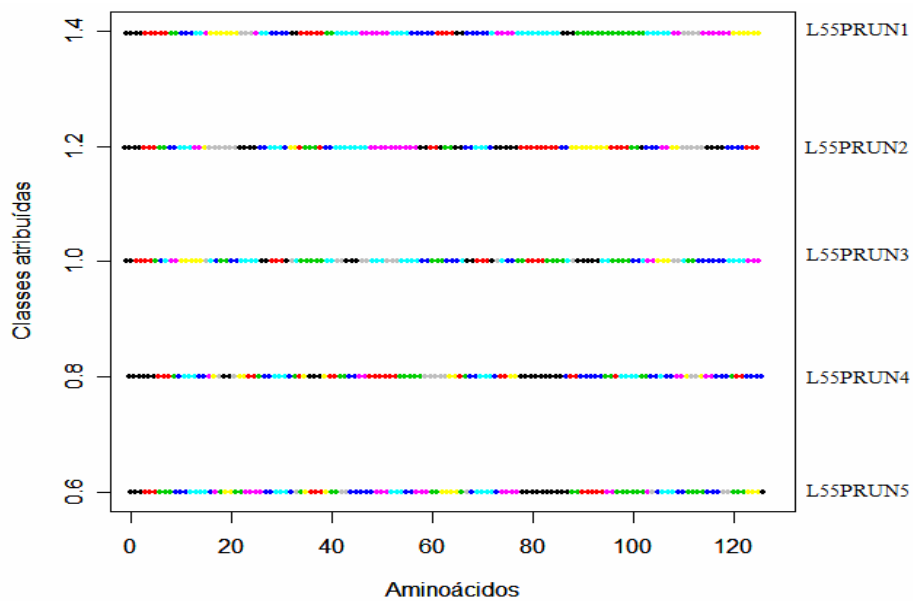
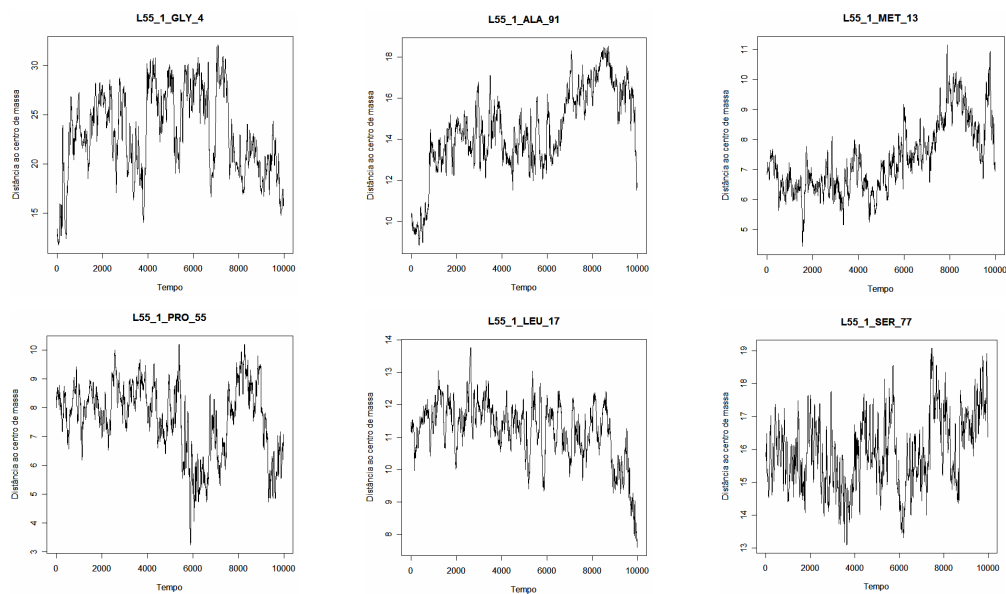


Figura 64 - Clustering de cada corrida da L55P-TTR

Tabela 21 – Clustering da L55PRUN1: 15 classes

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
GLY_1	THR_5	CYS_10	LEU_12	LYS_15	LEU_17	ASP_18	PRO_24	ARG_34	ALA_36	TRP_41	LEU_58	TRP_79	TYR_116	VAL_122
PRO_2	GLY_6	PRO_11	MET_13	VAL_16	ASN_27	ALA_19	ALA_25	LYS_35	ALA_37	GLU_42	THR_59	LYS_80	SER_117	THR_123
THR_3	GLU_7	ALA_91	VAL_14	VAL_28	LYS_48	VAL_20	ILE_26	GLY_67	ASP_38	THR_96	THR_60	ALA_81	THR_118	ASN_124
GLY_4	SER_8	GLU_92	VAL_30	ALA_29	THR_49	ARG_21	SER_112	ILE_68	ASP_39	ALA_97	GLU_61	LEU_82	THR_119	PRO_125
	LYS_9	VAL_93	HSD_31	PRO_43	SER_50	GLY_22	PRO_113	HSD_88	THR_40	ASN_98	GLU_62	GLY_83	ALA_120	LYS_126
		VAL_94	VAL_32	PHE_44	GLU_51	SER_23	TYR_114	GLU_89	GLU_63	ASP_99		ILE_84	VAL_121	GLU_127
		PHE_95	PHE_33	ALA_45	SER_52		SER_115	HSD_90	PHE_64	SER_100		SER_85		
			TYR_69	SER_46	GLY_53				VAL_65	GLY_101		PRO_86		
			LYS_70	GLY_47	THR_75				GLU_66	PRO_102		PHE_87		
			VAL_71	GLU_54	LYS_76					ARG_103				
			GLU_72	PRO_55	SER_77					ARG_104				
			ILE_73	HSD_56	TYR_78									
				GLY_57	LEU_110									
				ASP_74	LEU_111									
				TYR_105										
				THR_106										
				ILE_107										
				ALA_108										
				ALA_109										

Gráficos para visualização do comportamento de alguns aminoácidos da simulação L55PRUN1.



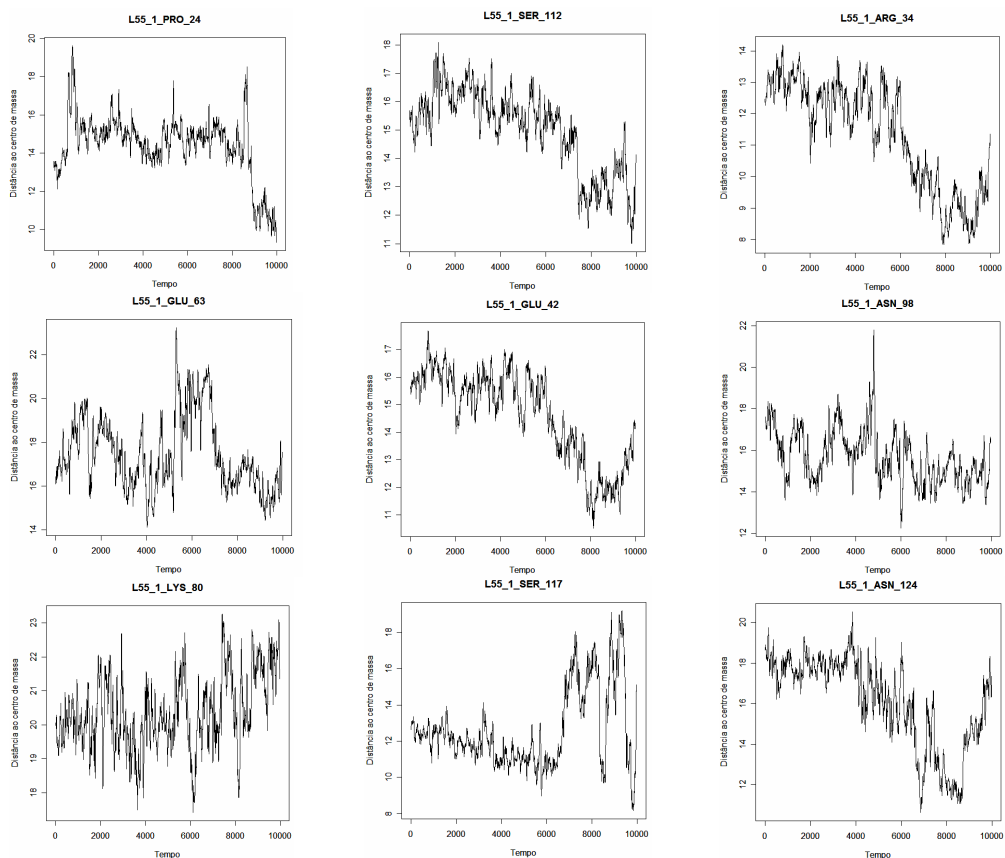


Tabela 22 - Clustering da L55PRUN2: 17 classes

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
GLY_1	THR_5	SER_8	CYS_10	LEU_12	LYS_15	LEU_17	ASP_18	PRO_24	ALA_36	ALA_37	TRP_41	PRO_43	SER_50	GLU_92	SER_112	SER_117
PRO_2	GLY_6	LYS_9	PRO_11	MET_13	VAL_16	ARG_34	ALA_19	ALA_25	THR_40	ASP_38	GLU_42	PHE_44	GLU_51	VAL_93	PRO_113	THR_118
THR_3	GLU_7	VAL_65	VAL_28	VAL_14	GLU_54	LYS_35	VAL_20	ILE_26	LYS_80	ASP_39	HSD_88	ALA_45	SER_52	VAL_94	TYR_114	THR_119
GLY_4	GLU_62	GLU_66	ALA_29	VAL_30	PRO_55	HSD_90	ARG_21	ASN_27	ALA_81		GLU_89	SER_46	GLY_53	PHE_95	SER_115	ALA_120
	GLU_63	PRO_102	PHE_33	HSD_31	HSD_56	ALA_91	GLY_22	THR_60	LEU_82		VAL_121	GLY_47			TYR_116	
	ASN_98	ARG_103	TYR_69	VAL_32	GLY_57	THR_96	SER_23	GLU_61	GLY_83		VAL_122	LYS_48				
	ASP_99		LYS_70	VAL_71	LEU_58	ALA_97		PHE_64	ILE_84		THR_123	THR_49				
	SER_100		ASP_74	GLU_72	THR_59	LEU_110		GLY_67	SER_85		ASN_124					
	GLY_101		TYR_105	ILE_73	ALA_108	LEU_111		ILE_68	PRO_86							
			THR_106		ALA_109			THR_75	PHE_87							
			ILE_107					LYS_76	PRO_125							
								SER_77	LYS_126							
								TYR_78	GLU_127							
								TRP_79								
								ARG_104								

Gráficos para visualização do comportamento de alguns aminoácidos da simulação L55PRUN2.

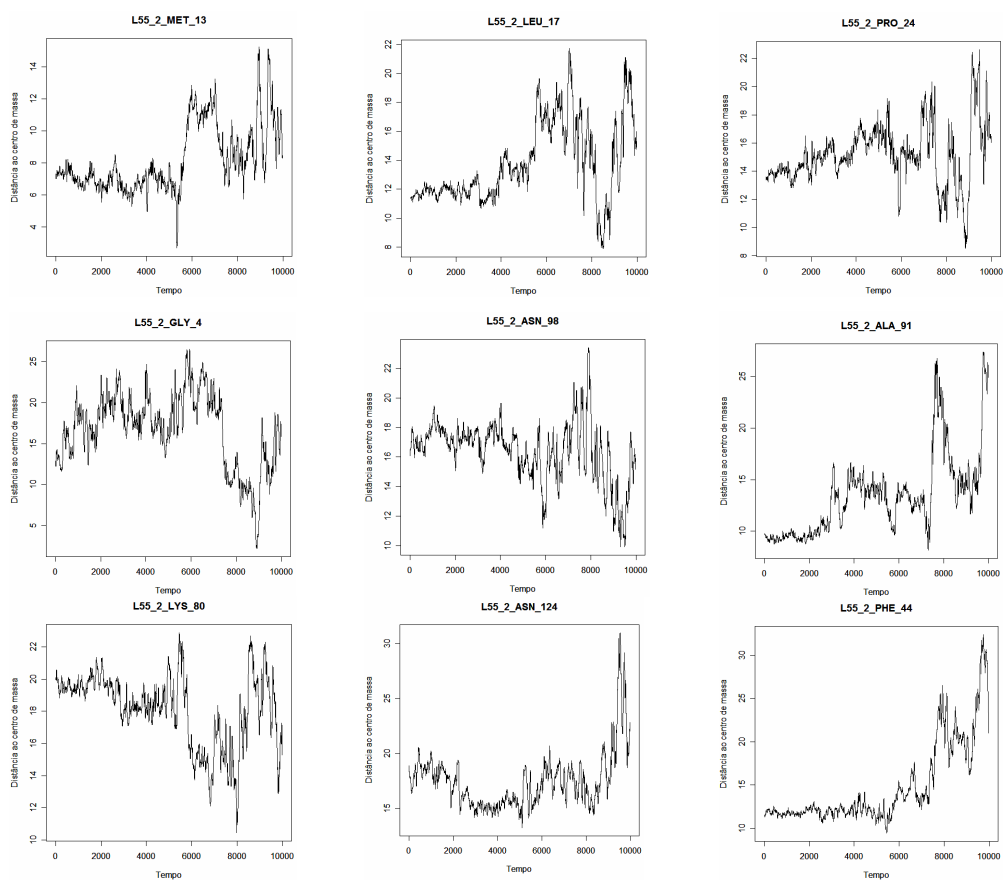
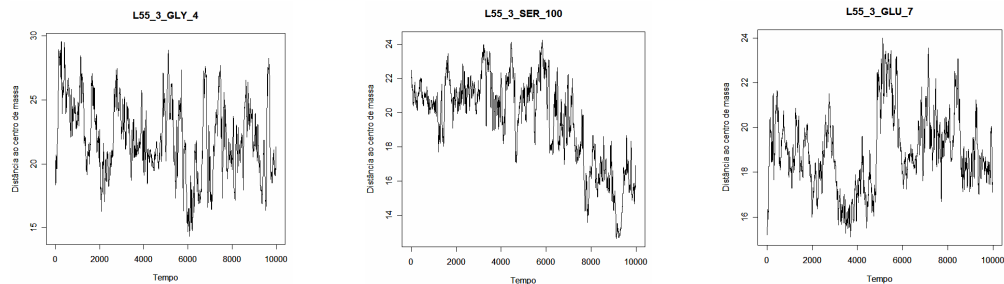


Tabela 23 - Clustering da L55PRUN3: 14 classes

1	2	3	4	5	6	7	8	9	10	11	12	13	14
GLY_1	THR_3	GLU_7	SER_8	LYS_9	CYS_10	ILE_107	GLU_54	VAL_28	VAL_30	ASN_98	TYR_116	VAL_121	PRO_125
PRO_2	GLY_4	GLU_62	VAL_65	HSD_56	PRO_11	LEU_12	PRO_55	ALA_29	HSD_31	ALA_36	SER_117	VAL_122	LYS_126
	THR_5	GLU_63	GLU_66	GLY_57	TYR_105	MET_13	LEU_17	PHE_33	VAL_32	THR_40	THR_118	THR_123	GLU_127
	GLY_6	ASP_99	PRO_102	LEU_58	THR_106	VAL_14	ARG_34	TYR_69	VAL_71	PHE_87	THR_119	ASN_124	
	ALA_81	SER_100	ARG_103	THR_59		LYS_15	HSD_90	LYS_70	GLU_72	HSD_88	ALA_120		
	LEU_82	GLY_101	ALA_19	LYS_35		VAL_16	LEU_110	ASP_74	ILE_73	PRO_113			
	GLY_83	VAL_20	GLY_22	THR_96		ALA_108	LEU_111	ALA_91		TYR_114			
	ILE_84	ARG_21	SER_23	ALA_97		ALA_109	THR_75	ALA_45					
		PHE_64	THR_60	ASP_18			PRO_43	SER_46					
		TRP_79	GLU_61	PRO_24			PHE_44	GLY_47					
		LYS_80	GLY_67	ALA_25			LYS_48	GLU_92					
		SER_85	SER_77	ILE_26			THR_49	VAL_93					
		PRO_86	TYR_78	ASN_27			GLY_53	VAL_94					
		ALA_37	SER_115	ILE_68				PHE_95					
		ASP_38		LYS_76									
		ASP_39		ARG_104									
				TRP_41									
				GLU_42									
				GLU_89									
				SER_50									
				GLU_51									
				SER_52									
				SER_112									

Gráficos para visualização do comportamento de alguns aminoácidos da simulação L55PRUN3.



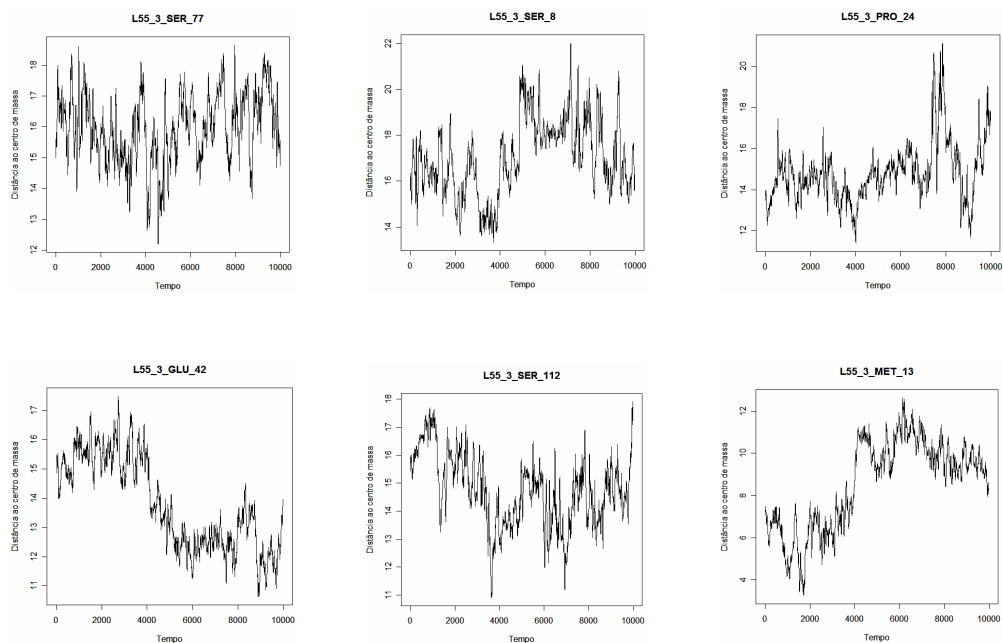
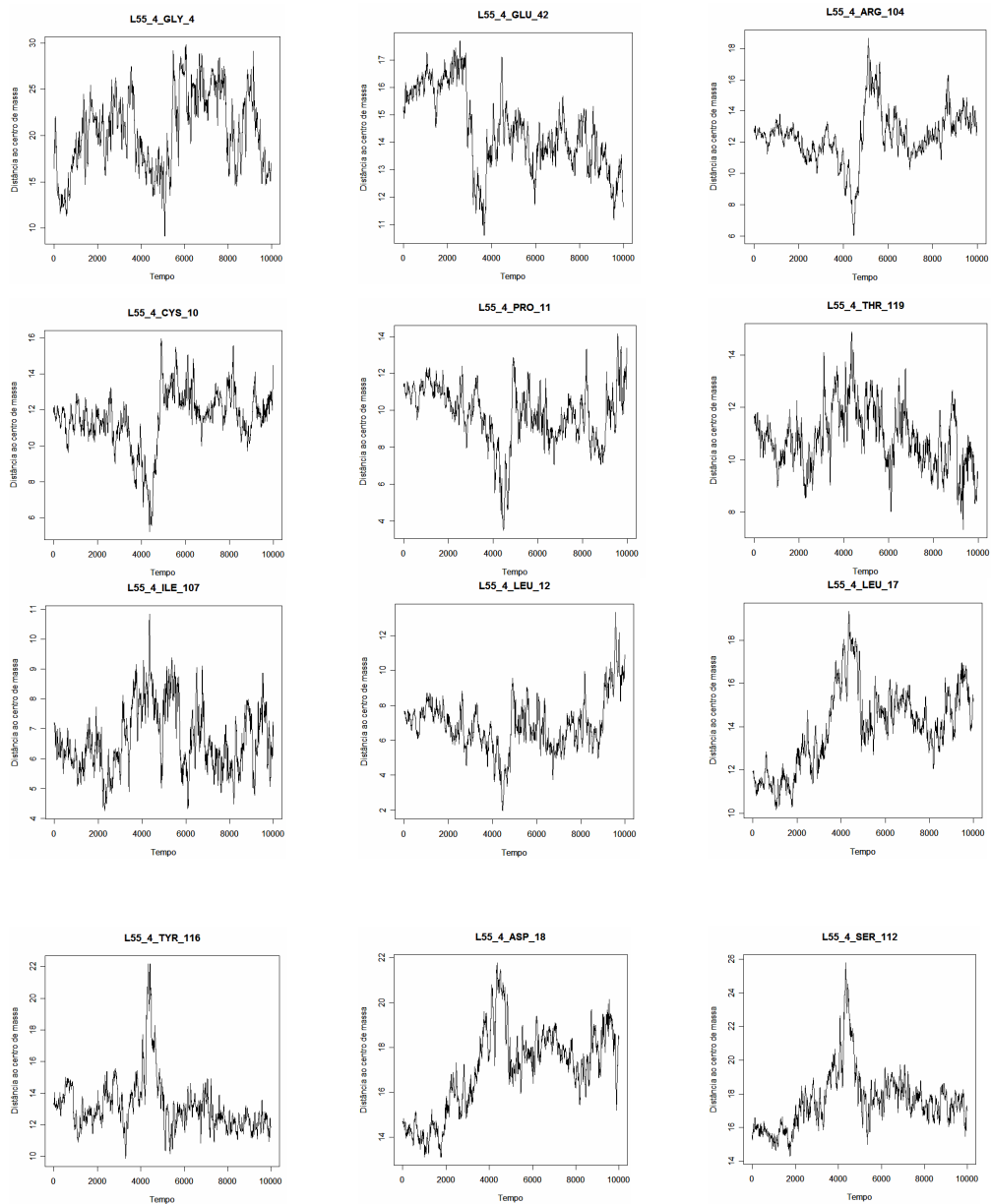


Tabela 24 - Clustering da L55PRUN4: 13 classes

1	2	3	4	5	6	7	8	9	10	11	12	13
GLY_1	GLU_7	ARG_103	PRO_11	ILE_107	LEU_17	VAL_65	GLU_62	ALA_81	SER_50	HSD_56	HSD_88	ASP_99
PRO_2	SER_8	THR_96	TYR_105	LEU_12	LEU_110	GLU_66	GLU_63	LEU_82	GLU_51	GLY_57	ASN_124	SER_100
THR_3	GLY_67	ALA_97	THR_106	MET_13	LEU_111	SER_23	PHE_64	GLY_83	SER_52	LEU_58	PRO_125	GLY_101
GLY_4	LYS_9	ASN_27	LYS_15	VAL_14	LYS_48	SER_77	ALA_19	ILE_84	GLU_54	THR_59	LYS_126	PRO_102
THR_5	LYS_35	ILE_68	VAL_16	VAL_30	GLY_47	TYR_78	GLY_22	VAL_20	THR_49	PRO_55	GLU_127	
GLY_6	ALA_25	ARG_104	ALA_108	HSD_31	TYR_116	SER_115	THR_60	ARG_21	GLY_53			
	ILE_26	CYS_10	ALA_109	VAL_32	SER_117	ASP_18	GLU_61	TRP_79				
	LYS_76	ARG_34	VAL_28	VAL_71		PRO_24	PRO_113	LYS_80				
	TRP_41	PRO_43	ALA_29	GLU_72		SER_112	TYR_114	SER_85				
	GLU_42	PHE_44	PHE_33	ILE_73		ALA_36		PRO_86				
	GLU_89	VAL_121	TYR_69			THR_40		ALA_37				
	HSD_90		LYS_70					ASP_38				
	THR_75		ASP_74					ASP_39				
	ASN_98		ALA_91					PHE_87				
	VAL_122		ALA_45									
	THR_123		SER_46									
			GLU_92									
			VAL_93									
			VAL_94									
			PHE_95									
			THR_118									
			THR_119									
			ALA_120									

Gráficos para visualização do comportamento de alguns aminoácidos da simulação L55PRUN3.



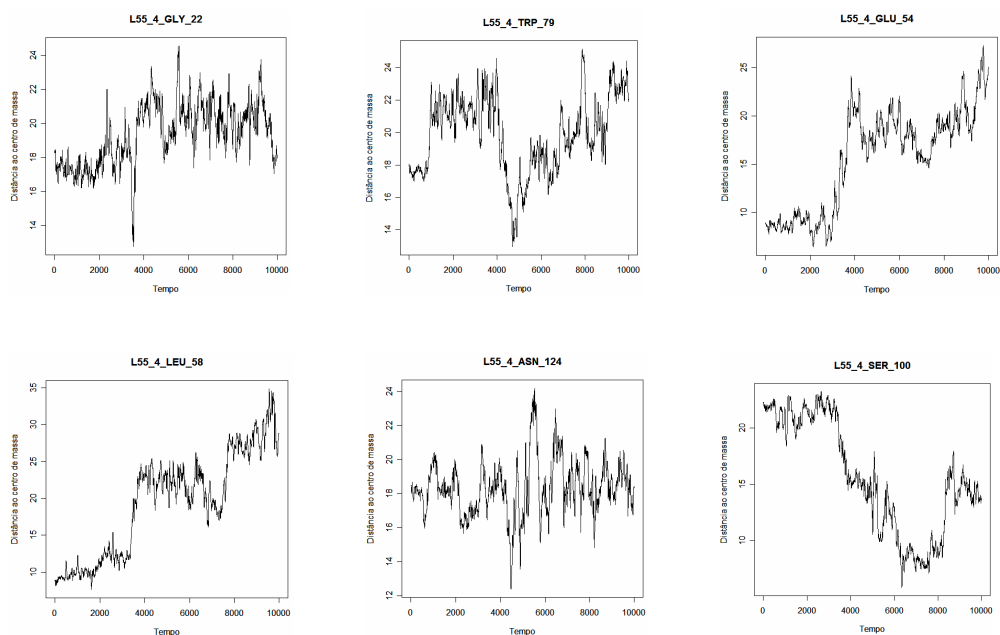
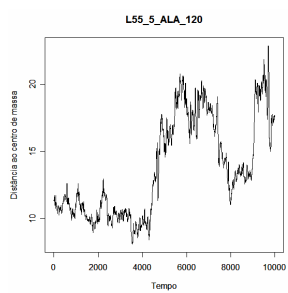
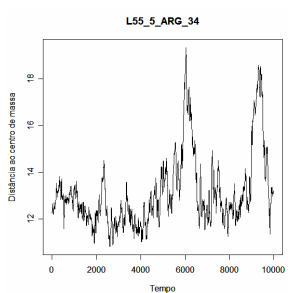
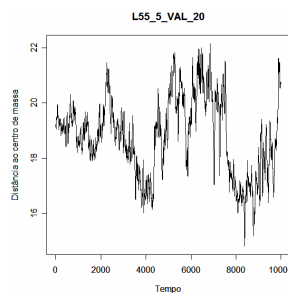
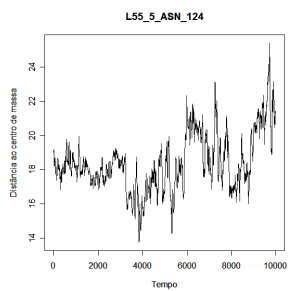
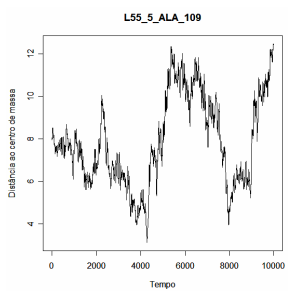
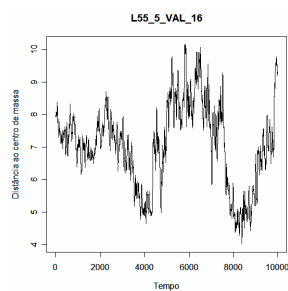
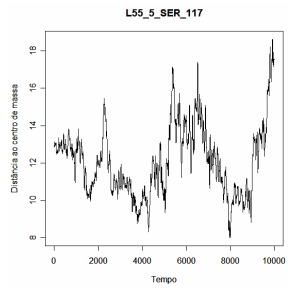
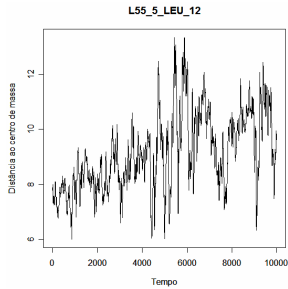
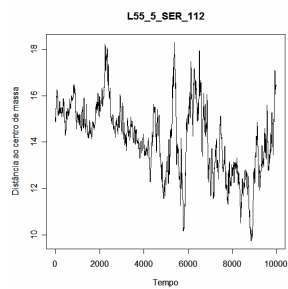
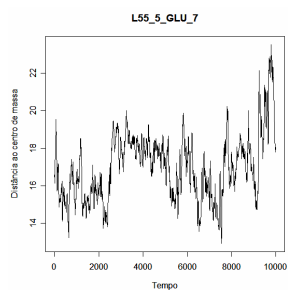
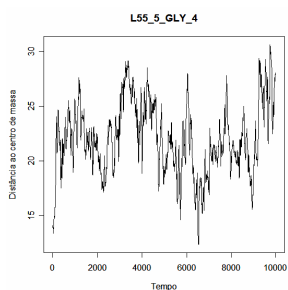
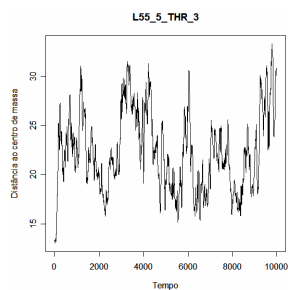


Tabela 25 - Clustering da L55PRUN5: 11 classes

1	2	3	4	5	6	7	8	9	10	11
GLY_1	GLY_4	GLU_7	CYS_10	MET_13	ASP_18	VAL_20	ARG_34	TRP_79	ALA_91	ASN_98
PRO_2	THR_5	SER_8	PRO_11	VAL_14	PRO_24	ARG_21	PRO_43	LYS_80	GLU_92	ASP_99
THR_3	GLY_6	LYS_9	LEU_12	LYS_15	ALA_25	ALA_36	PHE_44	ALA_81	VAL_93	SER_100
	ALA_37	ALA_19	LEU_17	VAL_16	ILE_26	THR_40	ILE_68	LEU_82	VAL_94	GLY_101
	ASP_38	GLY_22	VAL_28	VAL_30	ASN_27	GLU_63	TYR_105	GLY_83	PHE_95	PRO_102
	ASP_39	SER_23	ALA_29	HSD_31	SER_50	PHE_64	THR_119	ILE_84		ARG_103
		LYS_35	PHE_33	VAL_32	GLU_51	VAL_65	ALA_120	SER_85		
		TRP_41	ALA_45	GLY_53	SER_52	GLU_66		PRO_86		
		GLU_42	SER_46	GLU_54	LEU_58	ASN_124		PHE_87		
		GLU_61	GLY_47	PRO_55	THR_59	PRO_125		HSD_88		
		GLU_62	LYS_48	LYS_70	THR_60	LYS_126		GLU_127		
		GLY_67	THR_49	VAL_71	THR_75					
		GLU_89	HSD_56	GLU_72	LYS_76					
		HSD_90	GLY_57	ILE_73	SER_77					
		SER_112	TYR_69	ILE_107	TYR_78					
		PRO_113	ASP_74	ALA_108	THR_96					
		TYR_114	THR_106	ALA_109	ALA_97					
		SER_115	LEU_110		ARG_104					
		VAL_121	LEU_111							
		VAL_122	TYR_116							
		THR_123	SER_117							
			THR_118							



B3 - Partição consenso da L55P-TTR

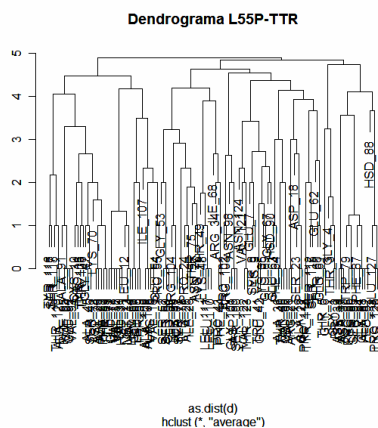


Figura 65 - Dendrograma L55P-TTR

Classificação obtida com um corte do dendrograma em $h=4$.

Tabela 26 – Classificação L55 (PartiçãoII)

[illegible]

C. Partição consenso - sem a posição 55

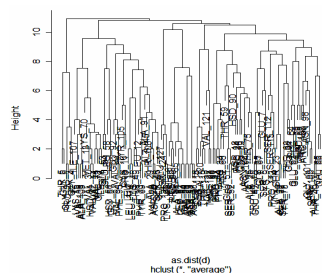


Figura 66 - Dendrograma TTR

D. Escolha dos aminoácidos Representativos (sem pos. 55)

Tabela 27 – Aminoácidos Representativos

Classe	WT_2	WT_3	WT_4	WT_5	WT_6	L55_1	L55_2	L55_3	L55_4	L55_5	Aa. Repr
1	GLY_4	GLY_4	GLY_4	THR_3	GLY_4	GLY_4	GLY_4	GLY_4	THR_3	GLY_4	GLY_4
2	GLU_89	GLU_42	GLU_89	GLU_66	GLU_66	GLU_42	LYS_35	LYS_35	GLU_89	GLU_42	GLU_42
3	TYR_105	LEU_17	LEU_17	THR_49	LEU_17	LEU_110	THR_105	CYS_10	LEU_17	LYS_48	LEU_17
4	LYS_15	MET_13	ALA_108	LYS_15	LYS_70	MET_13	LYS_70	TYR_69	THR_106	VAL_32	MET_13
5	PRO_24	PRO_24	ILE_26	ILE_26	SER_52	ASN_27	PRO_24	ILE_26	SER_50	SER_50	PRO_24
6	SER_23	SER_77	ALA_19	LYS_76	SER_77	SER_23	LYS_76	SER_77	GLY_22	LYS_76	SER_77
7	LYS_80	LEU_82	ALA_81	LYS_80	LYS_80	ILE_84	SER_85	ALA_37	ASP_38	LYS_80	LYS_80
8	GLU_92	ALA_29	ALA_91	ALA_91	VAL_28	VAL_94	VAL_94	ALA_29	VAL_28	ALA_45	ALA_91
9	PHE_44	ARG_34	ARG_34	ARG_34	ARG_34	THR_96	THR_96	ARG_34	ARG_34	PRO_43	ARG_34
10	VAL_65	ASN_98	SER_100	ALA_63	ASN_98	GLY_101	ASN_98	ALA_36	ASN_98	ASP_99	ASN_98
11	GLU_63	GLU_63	GLU_63	GLU_63	GLU_62	GLU_62	GLU_63	GLU_62	GLU_61	GLU_62	GLU_63
12	SER_112	PRO_113	PRO_112	SER_112	VAL_122	SER_112	SER_112	PRO_113	LEU_111	SER_112	SER_112
13	THR_117	SER_117	SER_117	SER_117	SER_117	SER_117	SER_117	SER_117	SER_117	SER_117	SER_117
14	ASN_124	ASN_124	ASN_124	ASN_124	ASN_124	PRO_125	ASN_124	ASN_124	ASN_124	ASN_124	ASN_124

Na classe 5 existe um caso de empate, isto é, 2 aminoácidos são representativos 3 vezes (PRO_24, ILE_26) uma vez que estes dois aminoácidos estão próximos na cadeia peptídica, em princípio um acompanha o comportamento do outro. Logo pode escolher-se qualquer um. O mesmo acontece na classe 4. Para as classes 2 e 8, como existem empates, dos aminoácidos repetidos escolheram-se aqueles que não estão próximos na cadeia polipeptídica aos aminoácidos representativos já seleccionados nas outras classe. O objectivo é ter um conjunto de aminoácidos uniformemente distribuidos pela cadeia, para que estes representem o comportamento de toda a cadeia.

Anexo III – Eventos

A. Comparação de aminoácidos dois a dois

Evento N° 1 - “troca” de posição entre dois aminoácidos

Para cada corrida foram procuradas as trocas entre aminoácidos. Nos ficheiros a designação dos dois aminoácidos LEU_55 e PRO_55 foi substituída pela designação POS_55 para ser considerada uma posição única.

Tabela 28- Eventos “troca” de posição entre dois aminoácidos

N°	Evento	Freq_WT	Freq_L55P	N°	Evento	Freq_WT	Freq_L55P
1	ALA_91 spw GLU_42	2	0	30	ASN_98 spw LEU_17	2	3
2	ALA_91 spw GLY_4	0	1	31	ASN_98 spw LYS_80	0	1
3	ALA_91 spw LEU_17	1	2	32	ASN_98 spw PRO_24	2	4
4	ALA_91 spw LYS_80	0	1	33	ASN_98 spw SER_77	2	5
5	ALA_91 spw MET_13	0	2	34	GLU_42 spw GLY_4	1	3
6	ALA_91 spw PRO_24	1	2	35	GLU_63 spw ALA_91	0	2
7	ALA_91 spw SER_77	1	3	36	GLU_63 spw ARG_34	0	2
8	ARG_34 spw ALA_91	3	4	37	GLU_63 spw ASN_98	0	3
9	ARG_34 spw GLY_4	0	1	38	GLU_63 spw GLU_42	2	2
10	ARG_34 spw LEU_17	3	4	39	GLU_63 spw GLY_4	4	4
11	ARG_34 spw LYS_80	0	1	40	GLU_63 spw LEU_17	1	2
12	ARG_34 spw MET_13	1	2	41	GLU_63 spw LYS_80	0	4
13	ARG_34 spw PRO_24	1	2	42	GLU_63 spw MET_13	0	1
14	ARG_34 spw SER_77	0	2	43	GLU_63 spw PRO_24	1	2
15	ASN_124 spw ALA_91	1	0	44	GLU_63 spw SER_77	3	2
16	ASN_124 spw ASN_98	3	2	45	LEU_17 spw GLU_42	1	3
17	ASN_124 spw GLU_42	0	1	46	LEU_17 spw GLY_4	1	2
18	ASN_124 spw GLU_63	2	5	47	LYS_80 spw GLU_42	0	2
19	ASN_124 spw GLY_4	3	4	48	LYS_80 spw GLY_4	4	3
20	ASN_124 spw LEU_17	2	2	49	LYS_80 spw LEU_17	0	2
21	ASN_124 spw LYS_80	4	4	50	LYS_80 spw PRO_24	1	1
22	ASN_124 spw PRO_24	3	2	51	LYS_80 spw SER_77	0	1
23	ASN_124 spw SER_112	2	2	52	MET_13 spw GLU_42	0	1
24	ASN_124 spw SER_117	4	4	53	MET_13 spw GLY_4	0	1
25	ASN_124 spw SER_77	3	1	54	POS_55 spw ALA_91	2	4
26	ASN_98 spw ALA_91	2	3	55	POS_55 spw ARG_34	2	4
27	ASN_98 spw ARG_34	1	2	56	POS_55 spw ASN_124	2	2
28	ASN_98 spw GLU_42	2	4	57	POS_55 spw ASN_98	2	3
29	ASN_98 spw GLY_4	3	3	58	POS_55 spw GLU_42	2	3
59	POS_55 spw GLU_63	2	3	78	SER_112 spw LYS_80	2	2
60	POS_55 spw GLY_4	2	3	79	SER_112 spw PRO_24	2	1

61	POS_55 spw LEU_17	2	4	80	SER_112 spw SER_77	2	3
62	POS_55 spw LYS_80	2	3	81	SER_117 spw ALA_91	1	3
63	POS_55 spw MET_13	0	3	82	SER_117 spw ARG_34	2	3
64	POS_55 spw PRO_24	2	3	83	SER_117 spw ASN_98	4	4
65	POS_55 spw SER_112	2	2	84	SER_117 spw GLU_42	5	3
66	POS_55 spw SER_117	2	3	85	SER_117 spw GLU_63	3	2
67	POS_55 spw SER_77	2	3	86	SER_117 spw GLY_4	3	2
68	PRO_24 spw GLU_42	3	3	87	SER_117 spw LEU_17	1	3
69	PRO_24 spw GLY_4	1	3	88	SER_117 spw LYS_80	1	2
70	PRO_24 spw LEU_17	0	2	89	SER_117 spw PRO_24	2	4
71	SER_112 spw ALA_91	1	1	90	SER_117 spw SER_112	5	3
72	SER_112 spw ARG_34	1	1	91	SER_117 spw SER_77	4	3
73	SER_112 spw ASN_98	3	4	92	SER_77 spw GLU_42	0	1
74	SER_112 spw GLU_42	2	3	93	SER_77 spw GLY_4	3	3
75	SER_112 spw GLU_63	2	2	94	SER_77 spw LEU_17	1	2
76	SER_112 spw GLY_4	1	3	95	SER_77 spw PRO_24	1	3
77	SER_112 spw LEU_17	1	2				

Evento Nº 2 - “Existem duas trocas de comportamento de dois aminoácidos em relação ao centro de massa da proteína”

Tabela 29- Eventos “troca” de posição entre dois aminoácidos

Nº	Eventos	L55P	WT				
1	ALA_91_troca_2x_A_POS_55	2	0	18	GLU_42_troca_2x_A_LEU_17	1	0
2	ALA_91_troca_2x_B_ARG_34	2	0	19	GLU_42_troca_2x_A_POS_55	1	0
3	ALA_91_troca_2x_B_ASN_124	1	0	20	GLU_42_troca_2x_A_PRO_24	0	1
4	ALA_91_troca_2x_B_ASN_98	2	0	21	GLU_42_troca_2x_A_SER_117	2	2
5	ALA_91_troca_2x_B_SER_112	2	1	22	GLU_42_troca_2x_B_ASN_124	0	3
6	ALA_91_troca_2x_B_SER_117	3	0	23	GLU_42_troca_2x_B_ASN_98	2	0
7	ARG_34_troca_2x_A_SER_117	1	0	24	GLU_42_troca_2x_B_LYS_80	2	0
8	ARG_34_troca_2x_B_GLU_63	1	0	25	GLU_42_troca_2x_B_SER_112	1	0
9	ARG_34_troca_2x_B_SER_112	1	1	26	GLU_42_troca_2x_B_SER_77	1	0
10	ASN_124_troca_2x_A_POS_55	0	1	27	GLU_63_troca_2x_A_ASN_124	1	1
11	ASN_98_troca_2x_A_GLU_63	0	1	28	GLU_63_troca_2x_A_POS_55	2	0
12	ASN_98_troca_2x_A_POS_55	2	0	29	GLU_63_troca_2x_A_SER_117	1	0
13	ASN_98_troca_2x_A_SER_112	1	1	30	GLU_63_troca_2x_B_ASN_124	1	3
14	ASN_98_troca_2x_A_SER_117	2	0	31	GLY_4_troca_2x_A_ASN_124	1	2
15	ASN_98_troca_2x_B_ASN_124	3	2	32	GLY_4_troca_2x_A_ASN_98	0	1
16	ASN_98_troca_2x_B_GLU_63	2	0	33	GLY_4_troca_2x_A_GLU_42	1	0
17	GLU_42_troca_2x_A_ALA_91	4	0	34	GLY_4_troca_2x_A_GLU_63	3	2
				35	GLY_4_troca_2x_A_LEU_17	1	1

36	GLY_4_troca_2x_A_LYS_80	0	1
37	GLY_4_troca_2x_A_POS_55	0	1
38	GLY_4_troca_2x_A_PRO_24	0	1
39	GLY_4_troca_2x_A_SER_112	1	1
40	GLY_4_troca_2x_A_SER_117	0	1
41	GLY_4_troca_2x_A_SER_77	1	1
42	GLY_4_troca_2x_B_ASN_124	1	1
43	GLY_4_troca_2x_B_ASN_98	2	1
44	GLY_4_troca_2x_B_GLU_42	1	0
45	GLY_4_troca_2x_B_GLU_63	2	1
46	GLY_4_troca_2x_B_LYS_80	3	2
47	GLY_4_troca_2x_B_SER_112	1	0
48	GLY_4_troca_2x_B_SER_77	1	0
49	LEU_17_troca_2x_A_ALA_91	1	1
50	LEU_17_troca_2x_A_POS_55	3	0
51	LEU_17_troca_2x_B_ASN_124	2	1
52	LEU_17_troca_2x_B_ASN_98	2	0
53	LEU_17_troca_2x_B_GLU_63	1	0
54	LEU_17_troca_2x_B_LYS_80	1	0
55	LEU_17_troca_2x_B_PRO_24	3	0
56	LEU_17_troca_2x_B_SER_112	2	1
57	LEU_17_troca_2x_B_SER_117	2	0
58	LEU_17_troca_2x_B_SER_77	3	0
59	LYS_80_troca_2x_A_ALA_91	2	0
60	LYS_80_troca_2x_A_ARG_34	1	0
61	LYS_80_troca_2x_A_ASN_124	2	1
62	LYS_80_troca_2x_A_ASN_98	2	0

63	LYS_80_troca_2x_A_GLU_63	2	0
64	LYS_80_troca_2x_A_POS_55	2	1
65	LYS_80_troca_2x_A_SER_112	2	0
66	LYS_80_troca_2x_A_SER_117	2	2
67	PRO_24_troca_2x_A_ALA_91	2	0
68	PRO_24_troca_2x_A_ARG_34	1	1
69	PRO_24_troca_2x_A_POS_55	2	1
70	PRO_24_troca_2x_A_SER_117	2	1
71	PRO_24_troca_2x_B_ASN_124	0	2
72	PRO_24_troca_2x_B_ASN_98	1	0
73	PRO_24_troca_2x_B_GLU_63	1	0
74	PRO_24_troca_2x_B_SER_112	1	3
75	PRO_24_troca_2x_B_SER_77	1	1
76	SER_112_troca_2x_A_POS_55	2	2
77	SER_112_troca_2x_A_SER_117	1	2
78	SER_112_troca_2x_B_ASN_124	1	2
79	SER_117_troca_2x_A_POS_55	2	1
80	SER_117_troca_2x_B_ASN_124	1	2
81	SER_77_troca_2x_A_ALA_91	3	0
82	SER_77_troca_2x_A_ARG_34	2	0
83	SER_77_troca_2x_A_ASN_124	1	0
84	SER_77_troca_2x_A_ASN_98	1	0
85	SER_77_troca_2x_A_POS_55	1	1
86	SER_77_troca_2x_A_SER_117	1	2
87	SER_77_troca_2x_B_ASN_124	1	1
88	SER_77_troca_2x_B_ASN_98	1	0
89	SER_77_troca_2x_B_GLU_63	1	1

B. Procura de um evento por aminoácido

Tabela 30- Aminoácidos onde se verificou o evento “*alteração significativa no comportamento do aminoácido*”.

Evento	Freq_WT	Freq_L55P	Evento	Freq_WT	Freq_L55P
GLY_4	0	1	ALA_91	1	0
LYS_9	0	1	PHE_95	1	0
ARG_21	0	1	ALA_108	1	0
ASP_38	0	1	LEU_110	1	0
PRO_43	0	1	LEU_111	1	0
SER_46	0	1	SER_112	1	0
VAL_65	0	1	SER_117	1	0
SER_77	0	1	THR_119	1	0
HSD_90	0	1	VAL_121	1	0
ALA_97	0	1	ASN_124	1	0
ASN_98	0	1	PRO_2	1	1
SER_100	0	1	THR_3	1	1
THR_106	0	1	GLY_6	1	1
GLY_22	0	2	SER_8	1	1
PRO_24	0	2	SER_23	1	1
ILE_84	0	2	LYS_35	1	1
GLY_83	0	4	GLY_47	1	1
VAL_16	1	0	GLU_51	1	1
ASP_18	1	0	VAL_94	1	1
ILE_26	1	0	PRO_102	1	1
ASN_27	1	0	THR_118	1	1
VAL_30	1	0	THR_123	1	1
ARG_34	1	0	GLY_57	1	3
ALA_37	1	0	ALA_29	2	0
ALA_45	1	0	HSD_31	2	0
SER_52	1	0	LYS_48	2	0
GLY_53	1	0	THR_59	2	0
LEU_55	1	0	ALA_120	2	0
LEU_58	1	0	GLY_1	2	1
GLU_66	1	0	GLU_92	2	1
GLY_67	1	0	GLY_101	2	1
LEU_82	1	0	TYR_116	2	1
GLU_89	1	0	PRO_86	5	1

Tabela 31 - Frequência dos eventos e janela onde ocorre a alteração mais significativa (com maior declive).

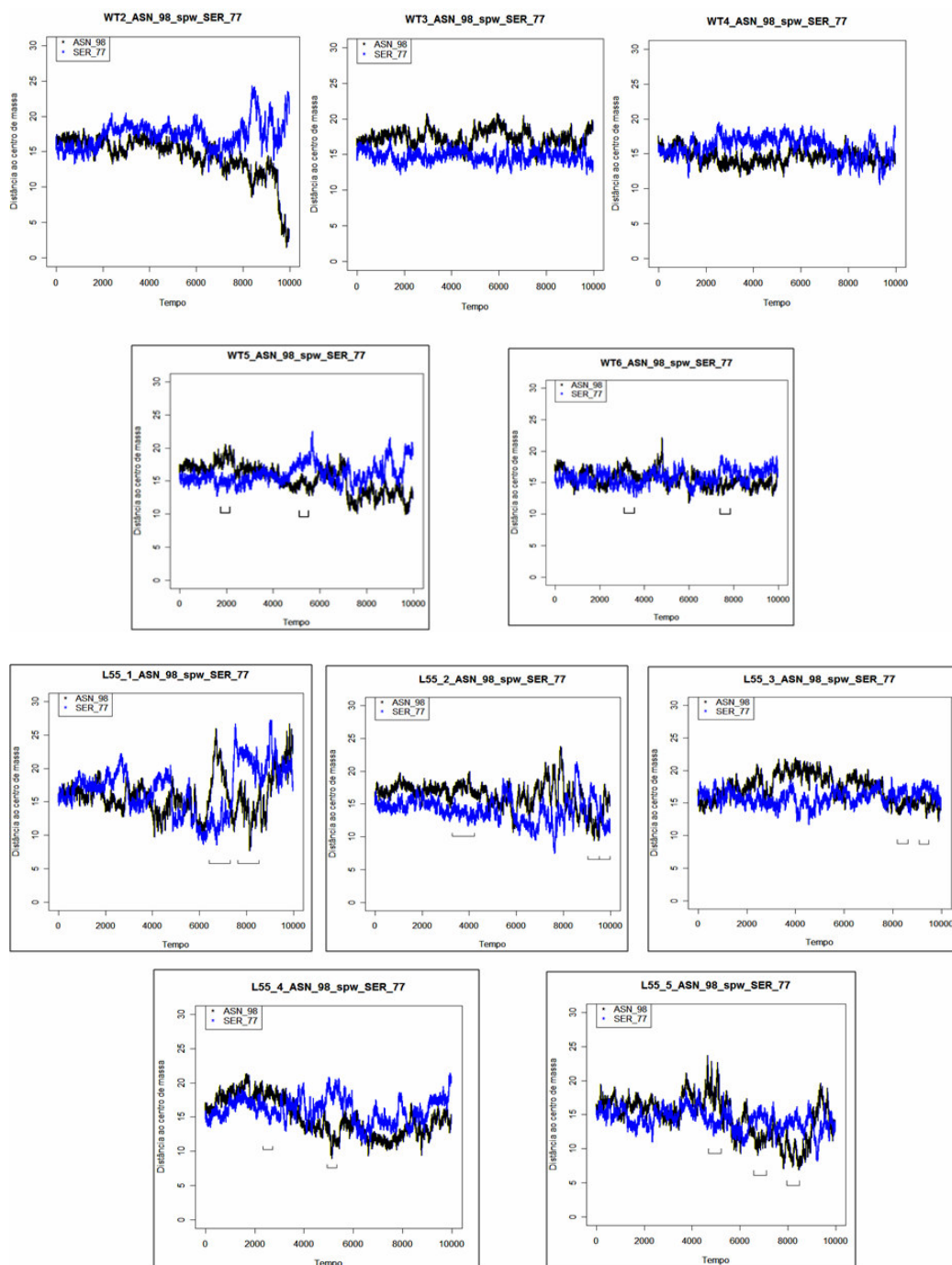
Aa	Freq	Simulação	I-janela	F-janela	Aa	Freq	Simulação	I-janela	F-janela
ASP_38	15	L55_1	9658	9858	GLY_83	7	L55_5	766	966
GLY_22	16	L55_1	2398	2598	ILE_84	7	L55_5	2218	2418
GLY_47	3	L55_1	8328	8528	LYS_9	26	L55_5	5049	5249
GLY_6	39	L55_1	2034	2234	THR_123	3	L55_5	4607	4807
GLY_83	31	L55_1	6652	6852	ALA_120	131	WT2	431	631
HSD_90	35	L55_1	81	281	ALA_29	14	WT2	8946	9146
PRO_102	3	L55_1	3633	3833	ASP_18	28	WT2	3196	3396
PRO_86	27	L55_1	8328	8528	GLU_89	62	WT2	7932	8132
SER_46	2	L55_1	641	841	GLY_1	83	WT2	342	542
THR_118	18	L55_1	8386	8586	GLY_101	49	WT2	124	324
GLU_92	32	L55_2	6777	6977	GLY_47	135	WT2	2630	2830
GLY_57	124	L55_2	1191	1391	GLY_53	2	WT2	3520	3720
PRO_24	77	L55_2	7909	8109	GLY_67	46	WT2	2837	3037
SER_8	38	L55_2	8223	8423	HSD_31	2	WT2	6527	6727
TYR_116	1	L55_2	9405	9605	LEU_111	21	WT2	1264	1464
ALA_97	44	L55_3	4518	4718	PHE_95	64	WT2	4130	4330
GLY_1	6	L55_3	4213	4413	PRO_86	55	WT2	7536	7736
GLY_101	6	L55_3	2105	2305	VAL_30	130	WT2	7698	7898
GLY_4	13	L55_3	8249	8449	ALA_108	142	WT3	198	398
GLY_57	3	L55_3	4426	4626	ALA_91	18	WT3	1	201
GLY_83	11	L55_3	1737	1937	ASN_27	128	WT3	3520	3720
PRO_2	1	L55_3	3452	3652	GLU_66	77	WT3	4722	4922
PRO_43	39	L55_3	4455	4655	GLU_92	83	WT3	256	456
SER_23	1	L55_3	5230	5430	GLY_6	14	WT3	1439	1639
VAL_65	26	L55_3	2185	2385	HSD_31	62	WT3	9520	9720
ARG_21	94	L55_4	9764	9964	ILE_26	33	WT3	526	726
ASN_98	9	L55_4	4785	4985	LEU_110	43	WT3	550	750
GLY_83	19	L55_4	5247	5447	LYS_48	19	WT3	9222	9422
ILE_84	5	L55_4	4420	4620	PRO_86	53	WT3	305	505
LYS_35	19	L55_4	3538	3738	THR_118	54	WT3	2716	2916
PRO_24	211	L55_4	3771	3971	THR_123	47	WT3	4968	5168
SER_100	12	L55_4	6007	6207	THR_59	6	WT3	7510	7710
SER_77	3	L55_4	2925	3125	ASN_124	79	WT4	124	324
THR_106	16	L55_4	4457	4657	GLY_57	24	WT4	218	418
THR_3	104	L55_4	535	735	LYS_48	41	WT4	6299	6499
VAL_94	4	L55_4	8527	8727	PRO_86	16	WT4	626	826
GLU_51	134	L55_5	2941	3141	ALA_120	18	WT5	1794	1994
GLY_22	59	L55_5	3864	4064	ALA_37	55	WT5	5773	5973

GLY_57	1	L55_5	8516	8716	GLU_51	14	WT5	8818	9018
GLY_1	24	WT5	1834	2034	VAL_16	14	WT5	3248	3448
LEU_55	13	WT5	4962	5162	ALA_29	41	WT6	1104	1304
PRO_102	5	WT5	9177	9377	ALA_45	2	WT6	5693	5893
PRO_2	4	WT5	3862	4062	ARG_34	29	WT6	594	794
PRO_86	20	WT5	4197	4397	GLU_92	6	WT6	426	626
SER_112	12	WT5	456	656	GLY_101	1	WT6	6170	6370
SER_117	5	WT5	5554	5754	LEU_58	12	WT6	1807	2007
SER_23	1	WT5	1154	1354	LEU_82	2	WT6	3945	4145
SER_52	112	WT5	3750	3950	LYS_35	29	WT6	9692	9892
THR_119	39	WT5	8464	8664	PRO_86	88	WT6	2544	2744
THR_3	2	WT5	5064	5264	SER_8	19	WT6	1511	1711
THR_59	4	WT5	1076	1276	TYR_116	67	WT6	8194	8394
TYR_116	1	WT5	4532	4732	VAL_94	84	WT6	719	919
VAL_121	30	WT5	273	473					

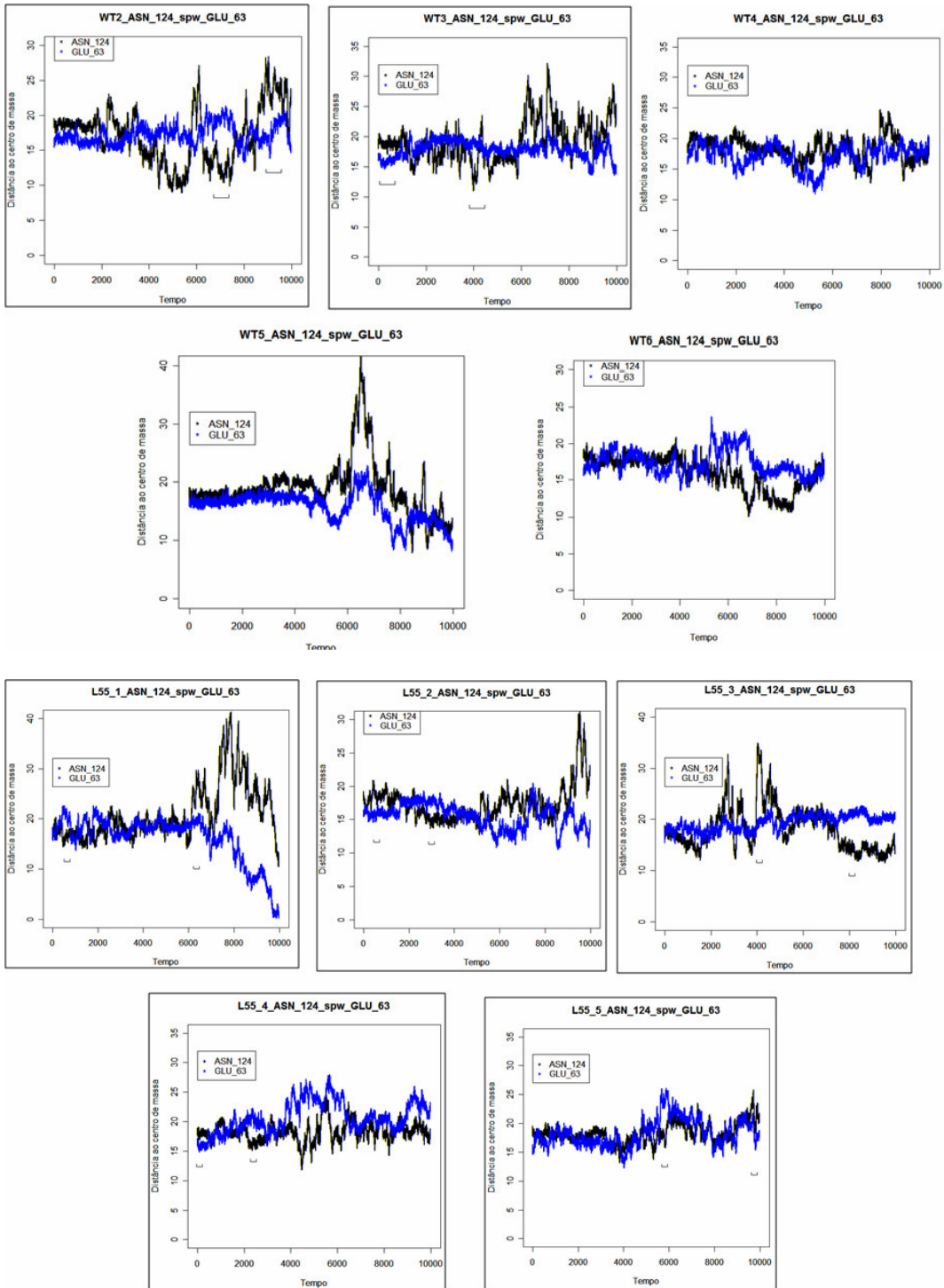
Anexo IV- Regras de Associação

A. Análise Gráfica da regra n°1

Regra n°1: **ASN_98_spw_SER_77** & ASN_124_spw_GLU_63 → L55P



Regra nº1: ASN_98_spw_SER_77 & **ASN_124_spw_GLU_63** → L55P



Anexo V- Código em R

```
#Read data
```

```
dad<-read.csv(choose.files(),sep='\t')
attach(dad)
```

```
#Moving Average Filter
```

```
d<-VAL_71
m<-length(d)
n<-15
apply(as.array((n+1):m),1,function(x)mean(d[(x-n):x]))->s
plot(s,type = "l",ylab='Dist_C_ALPHA_VAL_71_ CM',xlab='Tempo')
```

```
##Statistics
```

```
mean(WTRUN2[-1])>medias
summary(WTRUN2[-1])>sumario
cor(WTRUN2[, -1], method = "pearson", use = "complete")>corr
write.table(corr,'correlações.txt')
```

```
##Plots
```

```
d<-VAL_71
m<-length(d)
n<-15
apply(as.array((n+1):m),1,function(x)mean(d[(x-n):x]))->s
op <- par(mfrow = c(1,2))
plot(s,type = "l",ylab='Dist_C_ALPHA_VAL_71_ CM',xlab='Tempo',ylim =
c(0,40))
title(main = "VAL_71", font.main = 4)
d<-GLY_1
m<-length(d)
n<-15
apply(as.array((n+1):m),1,function(x)mean(d[(x-n):x]))->s
plot(s,type = "l",ylab='Dist_C_ALPHA_GLY_1_ CM',xlab='Tempo',ylim =
c(0,40))
title(main = "GLY_1", font.main = 4)
```

```
d<-SER_50
m<-length(d)
n<-20
apply(as.array((n+1):m),1,function(x)mean(d[(x-n):x]))->F
matplot(c(1, 10002), c(0,40), type= "n", xlab = "Tempo", ylab =
"Distância ao centro de massa",main = "WT_SER_23_spw_SER_50")
matpoints(s, pch = "l")
matpoints(F, pch = "l",col = "blue")
legend(1, 32, c("SER_23", "SER_50"), pch = "***",col =
c('black','blue'))
```

```
##Find number of clusters - Millican & Cooper Index
```

```
m<-t(run[,-1][,-55])
ind<-NULL
for (i in 10:20){
ds<- dist(m,method='euclidean')
tr<- hclust(ds,method='ward')
ind<-
c(ind,cluster.stats(ds,cutree(tr,i),silhouette=TRUE,G2=TRUE,G3=TRUE)$g
3)
ind}
L55_5_G3<-ind
```

```
tudoWT_s55<-
rbind(WT2_G2,WT2_G3,WT3_G2,WT3_G3,WT4_G2,WT4_G3,WT5_G2,WT5_G3,WT6_G2,W
T6_G3)
tudoL55_s55<-
rbind(L55_1_G2,L55_1_G3,L55_2_G2,L55_2_G3,L55_3_G2,L55_3_G3,L55_4_G2,L
55_4_G3,L55_5_G2,L55_5_G3)
```

##Hierarchical Clustering

```
hclust(dist(t(WTRUN2[,-1])), method = "euclidean"), method = "ward")-
>tree
plclust(tree, hang = 0.1, unit = TRUE, level = FALSE, main =
"Dendrograma",hmin =0, xlab = NULL, ylab = "Distância")
cutree(tree,14)->aWTRUN2
write.table(aWTRUN2, 'cl_WTRUN2.txt')
plot(rep(1,127),col=aWTRUN2,pch=20,xlab='Aminoácidos', ylab='Classes
atribuídas')
```

##Dissimilarities

```
d <- matrix(0,127,127)
m<-clWT
rownames(d)<-rownames(m)
for(i in 1:(nrow(m)-1)) for(j in (i+1):nrow(m)) d[i,j]<-d[j,i]<-
sum(as.numeric((m[i,]!=m[j,])))
rownames(d)<-rownames(m)
colnames(d)<-rownames(m)
d
plclust(hclust(as.dist(d),method = "average"))
cutree(hclust(as.dist(d),method = "average"),h=4)->cl_finais_WT
write.table(cl_finais_WT, 'cl_finais_WT.txt')
plot(rep(1,127),col=cl_finais_WT,pch=20,xlab='Aminoácidos',
ylab='Classes atribuídas', main='Classificação_WT')
```

##Representatives Aa.

```
aa<-NULL
for (i in 1:14){
  t(run)[-1,][classes==i,]->t
  m<-apply(t,2,mean)
  rbind(m,t)->d
  dist(d)->t1
  as.matrix(t1)->t1
  which.min(t1[-1,1])->aal
  aa<-c(aa,aal)
  aa
}
```

##Find events

```
# EV1
n=300
run<-WTRUN2
attach(run)
res<-
c('GLY_4','GLU_42','LEU_17','MET_13','PRO_24','SER_77','LYS_80','ALA_9
1','ARG_34','ASN_98','GLU_63','SER_112','SER_117','ASN_124','PRO_55')
evento<-NULL
```

```

for(i in 2:length(res)){
  for(j in 1:(i-1)){
    s <- as.numeric(run[res[i]]>run[res[j]])
    soma<-vector(mode = "numeric", length =(length(s)-n+1) )
    for(l in 1:length(soma)){
      soma[l]<-sum(s[l:(l+n-1)])
    }
    a<-length(which(soma==0))

    b<-length(which(soma==n))

    if (a>0 & b>0)
      evento<-rbind(evento,paste(res[i], 'spw', res[j]))
    print(evento)
  }
}

```

```

# EV2
n=200
for(i in 1:(length(res)-1)){
  for(j in (1+i):(length(res))){
    print(i)
    print(j)
    s <- as.numeric(run[,res[i]]>run[,res[j]])
    soma<-vector(mode = "numeric", length =(length(s)-n+1))
    for(l in 1:length(soma)){
      soma[l]<-sum(s[l:(l+n-1)])
    }
    #plot(soma,type='l')

    if (length(which(soma[1:2000]==n))>0 &
length(which(soma[2001:(10002-n-2000)]==0))>0
    & length(which(soma[(10002-n-2000+1):(10002-n+1)]==n))>0)

      evento<-rbind(evento,paste(res[i], 'troca_2x_A', res[j]))

    if (length(which(soma[1:2000]==0))>0 &
length(which(soma[2001:(10002-n-2000)]==n))>0
    & length(which(soma[(10002-n-2000+1):(10002-n+1)]==0))>0)
      evento<-rbind(evento,paste(res[i], 'troca_2x_B', res[j]))

    }
    print(evento)
  }

eventoF<-rbind(eventoF, cbind(evento, rep('WT6', length(evento))))

```

```

#EV_P_AA
WTRUN2[, -1] -> run
attach(run)
m<-NULL
evento<-NULL
n<-200
a<-NULL
Freq<-NULL
time<-NULL
for(j in 1:127){
  print(j)
  v<-run[, j]
  for(i in 1:(length(v)-n)){
    Max<-max(v[i:(i+n-1)])
    Min<-min(v[i:(i+n-1)])
    x1<-which.max(v[i:(i+n-1)])
    x2<-which.min(v[i:(i+n-1)])
    m[i]<-abs((Max-Min)/(x1-x2))
  }
  f<-length(which(m>1))
  M<-which.max(m)
  if (f>=1)
  a<-cbind(a, j)
  if (f>=1)
  Freq<-cbind(Freq, f)
  if (f>=1)
  time<-cbind(time, M)
}
t(rbind(a, Freq, time))

```

```

#Find frequent events
freq<-table(ev[,1])
rownames(freq)<-c(1:length(freq))
barplot(sort(freq, decreasing=T)[1:10], col=c('lightblue'), horiz=T)
title(main = "Events frequencies", font.main = 4)
freq1<-table(ev[1:156,1])
rownames(freq1)<-c(1:length(freq))
barplot(sort(freq1, decreasing=T)[1:10], col=c('lightblue'), horiz=T)
title(main = "Eventos mais frequentes na WT-TTR", font.main = 4)
freq2<-table(ev[156:391,1])
rownames(freq2)<-c(1:length(freq2))
barplot(sort(freq2, decreasing=T)[1:10], col=c('lightcyan'), horiz=T)
title(main = "Eventos mais frequentes na L55P-TTR", font.main = 4)

```