Universidade Técnica de Lisboa

Instituto Superior Técnico

Algoritmos para a predição da estrutura terciária de proteínas

Alexandre Paulo Lourenço Francisco

Licenciado

Dissertação para a obtenção do Grau de Mestre em Matemática Aplicada

Orientador: Doutor Arlindo Manuel Limede de Oliveira

Co-orientador: Doutor Amílcar dos Santos Costa Sernadas

Júri

Presidente: Doutor Amílcar dos Santos Costa Sernadas

Vogais: Doutor Pedro Manuel Corrêa Calvente Barahona

Doutor Arlindo Manuel Limede de Oliveira

Doutor Paulo Alexandre Carreira Mateus

Dezembro de 2004

Algoritmos para a predição da estrutura terciária de proteínas

Nome: Alexandre Paulo Lourenço Francisco
Curso de Mestrado em: Matemática Aplicada
Orientador: Doutor Arlindo Manuel Limede de Oliveira
Co-orientador: Doutor Amílcar dos Santos Costa Sernadas
Provas concluídas em:

Resumo: A predição da estrutura terciária de proteínas é um dos principais problemas da biologia molecular que permanece por resolver. Nesta dissertação introduzem-se os conceitos e resultados necessários ao entendimento deste problema. Os principais resultados de complexidade computacional acerca deste problema bem como a dificuldade computacional inerente são discutidos. São apresentados os métodos e algoritmos utilizados actualmente e levantam-se diversas questões em aberto. As técnicas de classificação são a abordagem com melhores resultados, nomeadamente através da utilização de Máquinas de Vectores de Suporte (SVM). Deste modo introduzem-se as ferramentas da aprendizagem estatística e as SVM's. Apresenta-se uma variante baseada em SVM's. E discutem-se os resultados obtidos e possíveis melhorias quer no conjunto de treino quer nos classificadores, nomeadamente através da introdução de classificadores mais sofisticados. Por fim são apresentadas as conclusões do trabalho e possíveis direcções para trabalho futuro.

Palavras chave: proteína; estrutura terciária; algoritmo; complexidade computacional; classificador; mapa de contacto.

Algorithms for the prediction of the tertiary structure of proteins

Abstract: The prediction of the tertiary structure of proteins is one of the main unsolved problems in molecular biology. This dissertation presents the concepts and results needed to understand this problem. The main computational complexity results about this problem, as well it hardness, are discussed. A survey on current methods and algorithms is given, leaving many open questions. Classification technics are the approach with the best known results, namely by using Support Vector Machines (SVM). Therefore, an introduction to statistical learning tools and to SVM's is given. An alternative view with SVM's follows. The results are discussed and some improvements in the training set and in the classifiers are pointed out, namely by the introduction of classifiers more sophisticated. Finally the conclusions of this work are presented, as well as possible directions for future work.

Keywords: protein; tertiary structure; algorithm; computational complexity; classifier; contact map.

Agradecimentos

Escrever esta secção não é uma tarefa fácil, foram inúmeras as pessoas que contribuíram para que esta dissertação chegasse a este ponto. Em primeiro lugar quero agradecer ao Professor Arlindo Oliveira pelo apoio e incentivo dados durante a realização deste trabalho. Quero também agradecer ao Professor Amílcar Sernadas pela apreciação crítica e pelas suas sugestões.

Quero também deixar aqui uma palavra de agradecimento ao grupo ALGOS do Inesc-ID, em particular no que diz respeito aos recursos informáticos facultados, e ao CLC. O meu obrigado a todos aqueles que contribuíram com sugestões e críticas.

Por último, mas não menos importante, agradeço à minha esposa Cátia pelo apoio e amor sempre demonstrados, obrigado.

Conteúdo

Capítulo 1

Introdução

A predição da estrutura terciária de proteínas é um dos principais problemas da biologia molecular que permanece por resolver. Este problema é usualmente designado por *Protein Folding Problem* e consiste no seguinte: dada a sequência de aminoácidos de uma proteína, qual a conformação espacial da mesma? Dado que a funcionalidade de uma proteína é determinada pela sua estrutura terciária e pelos grupos químicos à superfície, este é de facto um problema central para a bioquímica.

O sistema dinâmico inerente à formação das proteínas é bastante complexo. As equações que o descrevem envolvem inúmeras variáveis, pois é necessário ter em conta todo o meio envolvente e não apenas os átomos que compõem a proteína. Por outro lado, tal como se irá descrever adiante, existe uma componente probabilística neste sistema dinâmico. Portanto, a simulação exacta deste, face às técnicas e meios computacionais disponíveis, apenas é possível para proteínas com poucas dezenas de aminoácidos. Na natureza, mesmo para proteínas com centenas de aminoácidos, a formação e o seu enrolamento ocorre em segundos.

As abordagens conhecidas a este problema podem dividir-se em duas classes: aquelas que envolvem a utilização de funções de energia efectivas físicas e as que envolvem a utilização de funções de energia estatísticas. No primeiro caso pretende-se determinar a conformação que corresponde ao mínimo de uma dada função de energia que descreve a energia potencial da

molécula. No segundo caso parte-se da utilização dos dados conhecidos, e.g., bases de dados com informação acerca de proteínas com estrutura conhecida ou com informação acerca das propriedades dos aminoácidos, e pretende-se determinar quais os factores que contribuem para a definição da estrutura terciária das proteínas. Neste caso, o objectivo último é a obtenção de um modelo estatístico preditivo que permita inferir a estrutura para novas proteínas.

Esta dissertação divide-se em quatro partes: preliminares; panorâmica das abordagens à predição; predição com máquinas de vectores de suporte; e notas finais.

Na primeira parte introduzem-se os conceitos e os resultados da teoria da complexidade computacional necessários ao estudo dos algoritmos e problemas apresentados no seguimento. São também introduzidos os conceitos químicos e físicos inerentes à constituição e formação das proteínas. Por último, apresenta-se a formalização geral do problema da predição da estrutura terciária das proteínas.

Face à dificuldade na simulação da dinâmica do processo de formação das proteínas, às dificuldades em determinar as conformações com energia mínima, tarefas estas computacionalmente difíceis, têm vindo a ser propostos modelos e técnicas alternativos para o estudo deste problema. Deste modo, na segunda parte desta dissertação discutir-se-ão quais as dificuldades encontradas e apresentar-se-ão alguns dos modelos relacionados com a determinação da conformação de energia mínima. Em particular será discutida a complexidade do problema de determinar a estrutura da proteína à luz de cada um dos modelos. Serão também analisadas as consequências dos resultados de complexidade computacional apresentados no estudo do problema tal como se apresenta na natureza e, em particular, abordar-se-á o *paradoxo de Levinthal*. Dado que os resultados da análise da complexidade computacional do problema apontam para que este seja um problema computacionalmente difícil, apresentar-se-ão algumas técnicas de aproximação.

Por último é apresentada a representação da estrutura terciária das proteínas através do seu mapa de contactos que, embora com popularidade recente, revela propriedades interessantes, nomeadamente a independência em relação ao referencial.

Na óptica dos métodos de aproximação e da análise de proteínas com estrutura conhecida, na última parte é introduzida a predição através da utilização de máquinas de vectores de suporte. Introduzir-se-á em primeiro lugar a teoria inerente às máquinas de vectores de suporte e as propriedades relevantes face às técnicas de classificação comuns. Em segundo lugar, serão discutidos os resultados actualmente conseguidos com as máquinas de vectores de suportes (os melhores conseguidos até ao momento) e serão apresentadas algumas sugestões para melhorar os resultados, nomeadamente através da utilização das capacidades especiais das máquinas de vectores de suporte.

Concluir-se-á esta dissertação com algumas notas finais, onde serão apresentadas algumas conclusões acerca da panorâmica fornecida e em que se darão possíveis direcções para trabalho futuro.

Capítulo 2

Preliminares

Neste capítulo introduzir-se-ão os conceitos e resultados necessários ao estudo dos modelos, algoritmos e técnicas empregues na predição da estrutura terciária das proteínas. Em particular, introduzem-se os conceitos e os resultados da teoria da complexidade computacional importantes para a análise do problema à luz de cada modelo.

Apresenta-se também na segunda secção uma breve introdução aos conceitos bioquímicos inerentes à constituição e às propriedades físico-químicas das proteínas.

Na última secção discute-se a formalização do problema da predição da estrutura terciária das proteínas.

2.1 Complexidade computacional

Nesta secção introduzem-se os conceitos e as técnicas necessárias à análise e estudo da complexidade de um dado problema. A teoria aqui exposta tem por base as referências [GJ79, Pap94].

Surge então uma primeira questão: o que é um *problema*? Um *problema* é uma questão geral para a qual se pretende uma resposta. Em geral, um *problema* tem vários *parâmetros* cujos valores não são especificados. Um *problema* é descrito através da definição dos seus *parâmetros* e da especificação das propriedades que se pretendem na resposta, i.e., na *solução*. Uma *instância* de um problema é obtida através da especificação de valores

particulares para cada um dos *parâmetros*.

Usualmente para resolver um dado problema é necessário propor um modelo matemático para a representação do problema. Dado um modelo matemático e a representação de um problema sobre esse mesmo modelo, o objectivo principal passa então por desenvolver um *algoritmo* para resolver o problema. Um *algoritmo* é um *procedimento*, ou seja, um método bem definido e detalhado passo-a-passo para resolver um dado problema. Nota-se no entanto que, ao contrário do que é exigido a um *procedimento*, dada a descrição de uma instância de um problema, o *algoritmo* fornece sempre uma resposta, i.e., termina sempre.

É importante observar que um algoritmo está associado com um problema particular, sujeito a todas as restrições associadas e tem de receber como entrada uma instância do problema compatível.

Um algoritmo diz-se *correcto para uma dada instância de um problema* se devolve a resposta certa quando fornecida uma descrição da instância. Um algoritmo é *correcto para um dado problema* quando é correcto para todas as instâncias desse problema. Um algoritmo diz-se *correcto* quando é correcto para o problema para o qual foi desenvolvido.

Embora o objectivo de um algoritmo seja obter a resposta para o problema que foi proposto (i.e., um algoritmo deve, antes de tudo, ser correcto), é importante analisar a sua complexidade. Nota-se que, quando se fala em *complexidade computacional*, é importante distinguir entre *complexidade de um algoritmo* e *complexidade de um problema*.

Complexidade algorítmica

Para medir a complexidade de um algoritmo considerar-se-á o tempo[1] de execução do mesmo. É importante que a medida da complexidade traduza o desempenho do algoritmo sobre todas as instâncias do problema e que não

[1] O espaço, i.e., a memória, utilizado por um algoritmo na sua execução também pode ser considerado como medida de complexidade. Porém, o tempo revela-se, em muitos casos, mais importante.

reflicta pormenores de implementação. Deve assim satisfazer o seguinte:

- expressar a complexidade como função do tamanho da instância do problema;
- considerar apenas propriedades acerca do tempo de execução inerentes a todas as instâncias do problema;
- focar a dependência funcional do tempo de execução no tamanho da instância do problema, e não em medidas absolutas do tempo de execução.

Deste modo, utiliza-se a noção de *complexidade assimptótica para o pior caso* e, por forma a facilitar o estudo desta medida de complexidade, introduz-se de seguida a notação O.

Notação 2.1.1 Considerem-se as funções $f, g : \mathbb{N} \to \mathbb{N}$. Diz-se que $f(n)$ é $O(g(n))$ se existem constantes positivas $c \in \mathbb{R}$ e $n_0 \in \mathbb{N}$ tal que $f(n) \leq c \cdot g(n)$, para qualquer $n \geq n_0$.

Um algoritmo diz-se *eficiente* para um dado problema se devolver sempre resposta para qualquer instância do problema num período de tempo limitado polinomialmente em função do tamanho da instância, i.e., num período de tempo $O(f(n))$, em que $f(n)$ é um polinómio no tamanho n da instância.

Complexidade de um problema

Dado que a complexidade algorítmica permite comparar diversas formas de resolver um problema, pode-se agora definir a complexidade de um problema. O estudo da complexidade de problemas envolve apenas problemas de decidibilidade, i.e., cuja resposta é *sim* ou *não*. O estudo da complexidade dos problemas de optimização concretiza-se usualmente estabelecendo a correspondência entre estes e os problemas de decidibilidade.

Dado um problema, a sua *complexidade* define-se como a complexidade do algoritmo mais eficiente que resolve o problema.

Note-se que, aparentemente, medir a complexidade de um problema envolve determinar a complexidade de todos os algoritmos que resolvem o problema. Dado que esta é uma tarefa bastante difícil, ou impossível, dependendo do problema em causa, na próxima secção ver-se-á uma abordagem alternativa para determinar a complexidade de um problema.

Utilizando as noções anteriores, define-se *problema intratável*. Um problema diz-se *intratável* se não existe um algoritmo correcto e eficiente que o resolva. Ainda nesta secção ver-se-á em que classe de complexidade se incluem este tipo de problemas.

Redutibilidade

Como se referiu acima, determinar a complexidade de um problema nem sempre é uma tarefa fácil. Existem inúmeros problemas práticos e interessantes para os quais se desconhece a complexidade. Contudo, existem outras formas de medir a complexidade computacional de um problema, e.g., por *redução*.

Dados dois problemas de decidibilidade A e B, o problema A diz-se *redutível polinomialmente* ao problema B se existe uma transformação $\mathcal{R}$ que, para qualquer instância I do problema A, produz uma instância *equivalente* $\mathcal{R}(I)$ do problema B em tempo polinomial. Por *equivalente* entende-se que, uma resposta para $\mathcal{R}(I)$ enquanto instância de B, é uma resposta correcta para I enquanto instância de A.

Através de uma construção por redução polinomial é possível constatar que B é um problema com pelo menos o mesmo grau de dificuldade que o problema A. De facto, se existir um algoritmo eficiente que resolva B, então existe um algoritmo eficiente que resolve A.

O conceito de redutibilidade em tempo polinomial permite deste modo relacionar a dificuldade de problemas diferentes e conseguir assim discutir de forma rigorosa a intratabilidade.

Classes de complexidade

Os problemas são classificados em várias classes de acordo com a sua complexidade computacional. Nota-se que as classes apresentadas de seguida apenas incluem problemas de decidibilidade, i.e., cuja resposta é *sim* ou *não*. Na prática existem alguns problemas de optimização que se dizem estar numa destas classes dado que é possível estabelecer uma correspondência com a sua versão decidível.

A classe dos problemas cuja complexidade é polinomial, i.e., para os quais existe um algoritmo correcto e eficiente que os resolve, denota-se por P. Estes são muitas vezes referidos como *problemas tratáveis*.

A classe NP é uma outra classe bastante importante. Um problema de decidibilidade A diz-se estar na classe NP se existe um algoritmo que verifica as seguintes condições:

- se I é uma instância positiva de A, i.e., uma instância cuja resposta correcta é *sim*, então existe uma testemunha, ou seja, uma prova de que essa é a resposta correcta, tal que o algoritmo termina em tempo polinomial com resposta afirmativa;
- se I é uma instância negativa de A, ou seja, uma instância cuja resposta correcta é *não*, então qualquer que seja a testemunha fornecida o algoritmo termina em tempo polinomial com resposta negativa.

De facto, dado um problema em NP, não se sabe se existe um algoritmo correcto e eficiente que o resolva. No entanto, dada uma proposta de solução, é possível verificar se esta é ou não solução em tempo polinomial. Trivialmente, tem-se que $P \subseteq NP$.

Dada a noção de redutibilidade, define-se uma subclasse de NP, a classe dos problemas *NP-completos*. Um problema A pertence à classe dos *NP-completos* se qualquer problema na classe NP é redutível polinomialmente a A. Em particular, os problemas *NP-completos* são redutíveis entre si em tempo polinomial e são considerados como provavelmente intratáveis. O

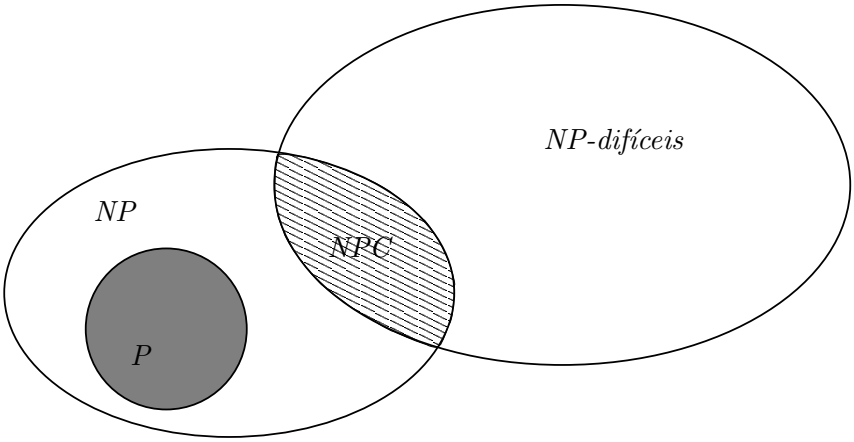

Figura 2.1: Dado o estado da arte, estas são as relações conhecidas entre as classes *P*, *NP*, *NP-completos* (*NPC*) e *NP-difíceis*.

primeiro problema que se demonstrou fazer parte desta classe foi o SAT[2] [Coo71]. Nota-se que, se existir um algoritmo correcto e eficiente para um problema *NP-completo*, quer a classe dos problemas *NP-completos* quer a classe *NP* são subclasses de *P*, i.e., *P=NP*.

Por vezes, dado um problema consegue-se apenas verificar que existe um problema *NP-completo* que se reduz polinomialmente a ele ou que qualquer problema *NP* se reduz polinomialmente a ele, não sendo possível verificar se o mesmo pertence à classe *NP*. Os problemas que verificam esta propriedade constituem a classe dos problemas *NP-difíceis*. Note-se que *NP* ∩ *NP-difíceis* = *NP-completos* e que, se existir um algoritmo correcto e eficiente para um problema *NP-difícil*, então *P=NP*.

Os problemas de optimização são normalmente estudados através a correspondência com a sua versão de problema de decisão. Observa-se no entanto que existem alguns para os quais a correspondência não é possível, ou pelo menos não é trivial. Um grande número destes problemas estão incluídos na classe dos problemas *NP-difíceis*.

[2]Problema da satisfação em lógica proposicional.

2.2 Bioquímica

As proteínas são as moléculas biológicas mais importantes, sendo responsáveis pelo funcionamento de todos os seres vivos. Cada proteína é construída a partir de blocos construtores simples, os aminoácidos. Existem vinte aminoácidos diferentes. Cada um deles consiste num átomo de carbono central - carbono α - ligado a um grupo amino (NH2), a um grupo carboxílico (COOH) e a uma cadeia lateral. Tal como se constata na tabela 2.1, a cadeia lateral é diferente em cada aminoácido e é responsável pelas características próprias de cada aminoácido, e.g., forma, tamanho e polaridade. A estrutura e propriedades de uma proteína dependem assim dos elementos básicos de cada aminoácido.

Nome	**Símbolos**		**Fórmula estrutural linear**
Alanina	Ala	A	CH3-CH(NH2)-COOH
Arginina	Arg	R	HN=C(NH2)-NH-(CH2)3-CH(NH2)-COOH
Asparagina	Asn	N	H2N-CO-CH2-CH(NH2)-COOH
Aspartato	Asp	D	HOOC-CH2-CH(NH2)-COOH
Cisteína	Cys	C	HS-CH2-CH(NH2)-COOH
Glutamina	Gln	Q	H2N-CO-(CH2)2-CH(NH2)-COOH
Glutamato	Glu	E	HOOC-(CH2)2-CH(NH2)-COOH
Glicina	Gly	G	NH2-CH2-COOH
Histidina	His	H	NH-CH=N-CH=C-CH2-CH(NH2)-COOH |_____|
Isoleucina	Ile	I	CH3-CH2-CH(CH3)-CH(NH2)-COOH
Leucina	Leu	L	(CH3)2-CH-CH2-CH(NH2)-COOH
Lisina	Lys	K	H2N-(CH2)4-CH(NH2)-COOH
Metionina	Met	M	CH3-S-(CH2)2-CH(NH2)-COOH
Fenilalanina	Phe	F	Ph-CH2-CH(NH2)-COOH
Prolina	Pro	P	NH-(CH2)3-CH-COOH |_____|
Serina	Ser	S	HO-CH2-CH(NH2)-COOH
Treonina	Thr	T	CH3-CH(OH)-CH(NH2)-COOH
Triptófano	Trp	W	Ph-NH-CH=C-CH2-CH(NH2)-COOH |_____|
Tirosina	Tyr	Y	HO-p-Ph-CH2-CH(NH2)-COOH
Valina	Val	V	(CH3)2-CH-CH(NH2)-COOH

Tabela 2.1: Aminoácidos

Numa proteína, os aminoácidos unem-se entre si através de *ligações peptídicas* que resultam da reacção do grupo amino de um aminoácido com o grupo carboxílico de outro aminoácido. A cadeia principal, figura 2.2, é restringida pelas limitações impostas aos ângulos de ligação entre cada um dos átomos, figura 2.3.

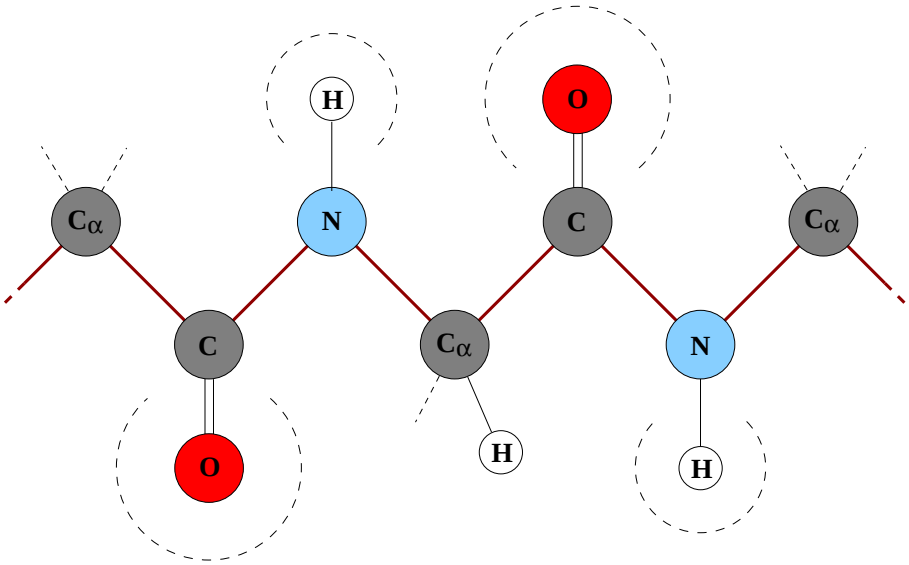

Figura 2.2: Ligação peptídica

Tal como se observa na tabela 2.1, a prolina difere de todos os outros aminoácidos pelo facto de a ligação entre o carbono central, i.e., carbono α, e o átomo de azoto `N` envolvido na ligação peptídica estar integrada num anel. Deste modo as posições relativas destes dois átomos estão fixas. Nos outros aminoácidos não se verifica esta situação. Deste modo e de forma simplificada, temos dois grupos de aminoácidos: a prolina e os restantes. Usualmente a ligação peptídica entre dois aminoácidos em que nenhum é prolina é uma ligação parcialmente dupla e portanto relativamente rígida, o que significa que as ligações simples entre o átomo de carbono α e os átomos de azoto e carbono vizinhos permitem a rotação. Os ângulos de rotação são limitados pelas nuvens electrónicas dos átomos de oxigénio do grupo carboxílico e dos átomos de hidrogénio ligados ao átomo de azoto envolvido na ligação peptídica.

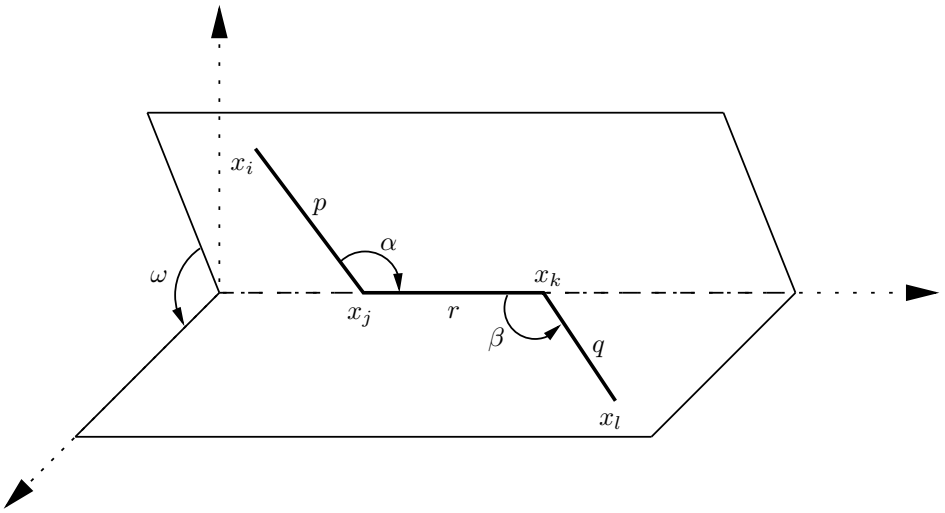

Figura 2.3: Ângulos das ligações entre átomos.

As proteínas assumem conformações espaciais variadas, mas verifica-se usualmente que consistem na combinação de certas estruturas designadas por *estruturas secundárias*. As estruturas secundárias são de dois tipos: *hélice-alpha* e *folha-beta*. A estrutura final de uma proteína pode ser vista como uma combinação de estruturas secundárias ligadas de forma específica.

Para capturar matematicamente a geometria de uma proteína atribui-se a cada átomo i um ponto no espaço tridimensional, i.e., o ponto $x_i^T = (x_{i1}, x_{i2}, x_{i3})$. Se dois átomos k e j estão ligados entre si, o vector de ligação r é dado da forma usual:

$$r = x_k - x_j,$$

o comprimento da ligação é dado por:

$$\|r\| = \sqrt{\langle r, r \rangle},$$

em que $\langle \_, \_ \rangle$ é o produto interno usual em $\mathbb{R}^3$. Os vectores de ligação p e q calculam-se de forma idêntica ao vector r. De acordo com a notação presente na figura 2.3, os ângulos de ligação α e β e o ângulo diedral ω determinam-se a partir das seguintes expressões:

$$\cos\alpha = \frac{\langle p,r\rangle}{\|p\|\|r\|},\ \sin\alpha = \frac{\|p\times r\|}{\|p\|\|r\|},\ \cos\beta = \frac{\langle q,r\rangle}{\|q\|\|r\|},\ \sin\beta = \frac{\|q\times r\|}{\|q\|\|r\|},$$
$$\cos\omega = \frac{\langle p\times r, r\times q\rangle}{\|p\times r\|\|r\times q\|} \text{ e } \sin\omega = \frac{\langle q\times p, r\rangle\|r\|}{\|p\times r\|\|r\times q\|},$$

onde $\_ \times \_$ é o produto externo usual em $\mathbb{R}^3$. O conjunto dos comprimentos de ligação, dos ângulos de ligação e dos ângulos diedrais fixa a geometria da molécula. Sob condições biológicas normais, o comprimento das ligações e os ângulos de ligação são bastante rígidos, sendo portanto os ângulos diedrais ao longo da ligação peptídica a determinar as principais características da geometria final da proteína.

No meio biológico uma proteína não é uma estrutura estática. Os átomos movimentam-se face às condições do ambiente, nomeadamente face à temperatura. Analisa-se de seguida os princípios físicos que regem o sistema dinâmico constituído pelos átomos da proteína. Por forma a simplificar a exposição, utilizar-se-á um vector x com as coordenadas de todos os N átomos da molécula:

$$x^T = (x_1, ..., x_N) = (x_{11}, x_{12}, x_{13}, ..., x_{N1}, x_{N2}, x_{N3}).$$

Dado que x contém três coordenadas para cada átomo, no caso de uma proteína real este vector terá uma dimensão entre 1500 e 30000.

O balanço das forças presentes na molécula e a dinâmica resultante pode ser aproximada através de uma equação diferencial estocástica:

$$M\ddot{x} + C\dot{x} + \nabla V(x) = D\dot{W}(t).$$

O primeiro termo descreve a energia cinética e consiste no produto da matriz diagonal de massa M pela aceleração $\ddot{x}$. Em cada posição da diagonal da matriz M está a massa do átomo correspondente a essa coordenada. O segundo termo descreve o excesso de energia dissipado e absorvido pelo ambiente envolvente e resulta do produto entre a matriz de amortecimento definida positiva C e a velocidade $\dot{x}$. O terceiro termo traduz a alteração da energia potencial, é expresso como o gradiente de uma função de potencial V característica da molécula. Por último, o termo do lado direito representa

um factor aleatório correspondente à flutuação das forças de colisão entre a molécula e o meio envolvente que dissipam energia; este termo é representado pelo produto do ruído branco normalizado $\dot{W}(t)$ por uma matriz D adequada, em que $W(t)$ é o processo de Wiener[3].

De facto, a interacção com o meio envolvente é bastante mais complexa. Os termos de amortecimento e de flutuação são apenas uma descrição simplificada. No entanto, em vez de melhorar a representação destes termos, representações mais exactas passam por incluir átomos do meio envolvente relevantes no vector de átomos, nomeadamente moléculas de água circundantes. A função potencial é por sua vez também extendida por forma a incluir as interacções destes novos átomos entre si e com a molécula.

Assuma-se que a temperatura é próxima do zero absoluto. Nesse caso toda a energia é potencial e portanto a molécula fica num estado estacionário que corresponde a um mínimo local da função de energia potencial. Sob condições ambientes normais, existe a presença de energia cinética e portanto a molécula descreve oscilações aleatórias em torno do mínimo local. De facto, caso a energia cinética seja elevada, a molécula pode mesmo aproximar-se de outro mínimo local e portanto assumir uma conformação espacial diferente. Ao contrário do que acontece com moléculas razoavelmente rígidas, caracterizadas pela função de energia potencial possuir apenas um mínimo global, a superfície da função de energia potencial de uma proteína é bastante irregular e, portanto, exibe um grande número de mínimos locais. Deste modo, quando a temperatura é suficientemente elevada e as forças de interacção aumentam, existe uma grande probabilidade das oscilações fazerem com que a molécula mude de geometria passando para outro mínimo local.

A frequência destas transições depende da temperatura e das barreiras de energia existentes num caminho favorável à transição sobre a superfície da função de energia potencial. Dado que as transições de níveis de energia

[3]Processo estocástico em tempo contínuo, $W(t)$ para $t \geq 0$, tal que o incremento $W(t) - W(s)$ é Gaussiano com média 0 e variância $t - s$, para qualquer $0 \leq s < t$. Os incrementos para intervalos de tempo cuja intersecção é vazia são independentes.

mais altos para níveis de energia mais baixos são muito mais frequentes, tem-se que a molécula permanece a maior parte do tempo com uma geometria que corresponde ao mínimo global da função de energia potencial. Esta é razão pela qual a maior parte dos investigadores assume que a conformação que correspondente ao mínimo global de energia potencial é de facto a conformação nativa (hipótese termodinâmica). Nomeadamente Anfinsen *et al* constatou na década de 1950 que muitas proteínas simples têm uma única conformação nativa [AHSW61, Afi73]. Posteriormente este facto foi verificado para um grande número de proteínas. Observa-se no entanto que, se as barreiras de energia forem suficientemente elevadas, a molécula pode passar um período de tempo biologicamente significativo com uma geometria correspondente a um mínimo local da função de energia potencial.

Tal como se verifica pela teoria envolvida e pela complexidade do sistema dinâmico em causa, determinar a estrutura das proteínas é um problema bastante complexo. Nota-se no entanto que as proteínas assumem a sua conformação nativa rapidamente, ainda que o espaço de conformações possíveis seja extremamente grande. A forma como a natureza determina a conformação ideal permanece mal compreendida.

2.3 Problema PF

O problema da predição da estrutura terciária das proteínas tem uma formulação diferente para cada um dos modelos propostos para o seu estudo. Dado que no seguimento deste trabalho serão apresentados diferentes modelos, para cada um deles formalizar-se-á o problema em detalhe. A formalização geral é apresentada de seguida. Veja-se a ilustração na figura 2.4.

Problema 2.3.1 (PF) ***Instância do Problema:*** *Uma sequência s correspondente aos aminoácidos constituintes de uma dada proteína, i.e.,*

$$s \in \{A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V\}^n$$

com $n \in \mathbb{N}$. ***Objectivo:*** *Determinar a estrutura terciária da proteína, i.e., determinar a conformação espacial nativa da cadeia de aminoácidos.*

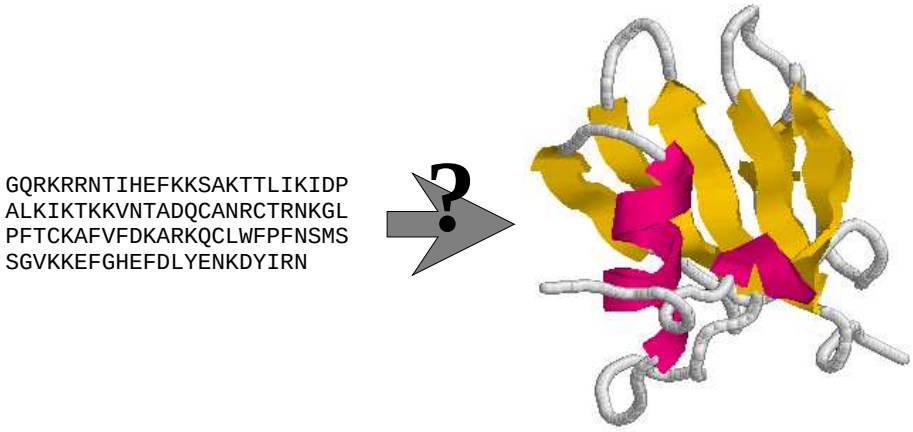

Figura 2.4: Esta é a questão a que se pretende responder no problema da predição da estrutura terciária das proteínas. Dada uma cadeia de aminoácidos, à esquerda, pretende-se determinar qual a conformação espacial da proteína, à direita.

Tal como indicado acima, o problema será referido no seguimento como *PF* dado que este é conhecido na literatura por *Protein Folding Problem*.

Nota-se que para além da cadeia de aminoácidos, são também conhecidas as propriedades químicas e físicas de cada aminoácido. Este conhecimento bem como dados acerca do comportamento do sistema dinâmico em causa são muitas vezes utilizados nas tentativas de resolução do problema. De facto cada modelo selecciona um determinado conjunto de prioridades a ter em conta e impõe algumas restrições por forma a simplificar o problema.

Do ponto de vista computacional, levantam-se algumas questões: Existirá um algoritmo correcto para este problema? Qual será a complexidade computacional deste problema? Que técnicas de aproximação poderão ser úteis? Poder-se-á utilizar algum método de classificação sobre este problema? Nos próximos dois capítulos tentar-se-á dar respostas a estas questões.

Capítulo 3

Panorâmica das abordagens à predição

Ao longo dos últimos anos têm sido propostas várias abordagens ao problema da predição da estrutura terciária das proteínas, i.e., ao problema *PF*. Embora os resultadas ainda sejam pouco expressivos, este problema têm vindo a ser fonte de um grande número de problemas interessantes para a modelação matemática, análise numérica e ciência da computação, nomeadamente no que diz respeito a tópicos como equações diferenciais, equações diferenciais estocásticas, aproximação de funções multidimensionais, classificação de dados, desenvolvimento de algoritmos e complexidade computacional.

Neste capítulo ir-se-ão expor alguns dos modelos propostos para a representação do problema *PF*, estudar a sua complexidade e referir algumas das técnicas de aproximação utilizadas. Deste modo, na primeira secção analisar-se-ão alguns dos modelos relacionados com a minimização de funções de energia efectivas físicas, sendo discutidos alguns modelos simplificados e estudada a complexidade da formalização do problema *PF* sobre os mesmos. Isto permitirá extrair alguma informação acerca da complexidade do problema que será exposta na segunda secção, onde será abordado também o chamado *paradoxo de Levinthal*. Na terceira secção será apresentado o estado da arte no que diz respeito às técnicas de aproximação. Na quarta e última secção conclui-se este capítulo com a apresentação de

uma representação da estrutura terciária das proteínas que tem vindo a ganhar popularidade e que será importante para o estudo efectuado no terceiro capítulo.

3.1 Minimização de energia

O sistema dinâmico inerente a uma proteína, tal como se viu no capítulo anterior, pode ser descrito pela equação diferencial estocástica

$$M\ddot{x} + C\dot{x} + \nabla V(x) = D\dot{W}(t).$$

Esta equação é de facto uma simplificação baseada na aproximação de Born-Oppenheimer. Nota-se contudo que para se obter uma descrição exacta da energia em função da estrutura dever-se-ia de utilizar a equação de Schrödinger. Na secção 2.2, viu-se que na maior parte dos casos a conformação nativa de uma proteína é a de menor energia potencial, i.e., corresponde ao mínimo global da função de energia potencial $V(x)$. Nesta secção apresentar-se-ão modelos simplificados do problema de encontrar a conformação de energia potencial mínima e estudar-se-á a sua complexidade computacional.

A definição da função de energia potencial $V(x)$ inclui vários termos, nomeadamente termos associados ao comprimento das ligações, aos ângulos de ligação, aos ângulos de torsão e às forças de atracção e repulsão entre átomos não ligados (forças electroestáticas e forças de *van der Waals*).

Apresentam-se de seguida três modelos diferentes, por ordem decrescente de complexidade, nos quais se tenta minimizar uma dada função de energia. Nota-se que, algumas vezes, nas funções de energia apenas são considerados alguns termos dos referidos acima.

3.1.1 Modelo PSP

Um dos principais objectivos da biologia molecular é determinar a configuração de energia miníma de uma proteína dada a cadeia de aminoácidos. Até ao momento não foi encontrado nenhum algoritmo que resolva este pro-

blema eficientemente, pelo que se suspeita que este seja um problema intratável. É claro que a procura exaustiva, ainda que permita encontrar a solução, está fora de questão pois o espaço inerente às conformações possíveis é exponencial no tamanho da molécula e excede largamente os limites impostos pelo hardware actual.

Os factos anteriores levam a que se acredite que o problema em causa é intratável. No entanto este facto nunca foi provado. Para se estudar estas questões em detalhe recorre-se à teoria da complexidade computacional. Nesta secção apresentam-se alguns resultados obtidos por Ngo e Marks [NM92] acerca do problema *PF*. Importa referir que se adoptará um determinado modelo matemático e que se estuda a formulação do problema *PF* sobre esse modelo. De facto, os resultados obtidos não têm relevância para a questão levantada por Levinthal em 1968, que sugere que uma proteína não pode percorrer todas as conformações possíveis aquando da sua formação.

A estratégia é a usual quando se estuda a completude *NP*. A parte não trivial passa por mostrar a redução polinomial de um problema *NP-completo* ao problema em questão. A redução será demonstrada em vários passos: em primeiro lugar define-se um problema computacional discreto, *DLP* (*Diamond Lattice Path*) e mostra-se que é *NP-completo* por redução a partir do problema *NP-completo PARTITION* [GJ79]; em segundo lugar mostra-se que o problema *PF*, formalizado sobre o modelo *PSP*, é *NP-completo* por redução a partir do problema *DLP*.

O modelo associado ao problema abstracto *DLP* é inspirado na geometria das cadeias alifáticas; veja-se a relação com os alcanos. Um alcano é uma cadeia de n carbonos da forma $\mathtt{CH_3(CH_2)_{n-2}CH_3}$. Cada carbono tem hibridação sp^3, i.e., liga-se a quatro átomos vizinhos com geometria tetraédrica. A torsão de cada ligação está relativamente desimpedida a temperaturas normais. No entanto o ângulo diedral assume em geral um de três valores, $\{60°, 180°, -60°\}$.

O problema *DLP* consiste na predição da estrutura de uma cadeia de n

carbonos na qual a geometria tetraédrica das ligações é exacta, e a escolha de quais as ligações que favorecem os ângulos diedrais de 180° e as que assumem energia mínima com qualquer um dos ângulos $\{60^\circ, 180^\circ, -60^\circ\}$ é arbitrária. Dadas as condições anteriores, o comprimento e os ângulos das ligações são precisamente os que encontramos num cristal de diamante, o qual é inteiramente constituído por carbonos sp^3. Desta forma, o problema discreto *DLP* pode ser analisado sobre um reticulado.

O *reticulado de diamante* D é o conjunto infinito de pontos no espaço tridimensional que são ocupados por átomos de carbono num cristal de diamante. Escolhendo a orientação e unidade de comprimento adequadas, $D \subseteq \mathbb{R}^n$ é o conjunto de pontos x tais que

$$x = x' + 4 \cdot (i, j, k) + l \cdot (1, 1, 1),$$

em que $x' \in \{(0,0,0), (0,2,2), (2,0,2), (2,2,0)\}$, $i, j, k \in \mathbb{Z}$ e $l \in \{0, 1\}$. Dois pontos x e y tais que $x - y \in \{\pm(1,1,1), \pm(-1,1,1), \pm(1,-1,1), \pm(1,1,-1)\}$ dizem-se vizinhos. Um *caminho no reticulado de diamante* D é uma sequência de pontos em D tal que qualquer par de pontos consecutivos na sequência corresponde a um par de vizinhos em D.

Dado um caminho $(m_0, m_1, ..., m_N)$ em D, o vector $\delta_i = m_i - m_{i-1}$ será designado por ligação i.

Problema 3.1.1 (DLP [NM92]) ***Instância do Problema:*** *Um inteiro positivo N; caminhos (P_0, P_1, P_2) e (Q_0, Q_1, Q_2) em D; e um conjunto $I \subseteq \{2, 3, ..., N-1\}$.* ***Objectivo:*** *Determinar se existe um caminho em D, $(m_0, m_1, ..., m_N)$, tal que $(m_0, m_1, m_2) = (P_0, P_1, P_2)$, $(m_{N-2}, m_{N-1}, m_N) = (Q_2, Q_1, Q_0)$ e $\delta_{i-1} = \delta_{i+1}$ para qualquer $i \in \{2, 3, ..., N-1\} \setminus I$.*

Cada δ_i, i.e., ligação i, com índices em I poderá ter um ângulo diedral com valor em $\{60^\circ, 180^\circ, -60^\circ\}$. As restantes ligações terão ângulos diedrais com valor 180°. O tamanho de uma instância de *DLP* é linear em $|I|$ e não em N.

A prova de que *DLP* é *NP-completo* consiste em verificar que *DLP* está em *NP* e que um problema *NP-completo* é redutível polinomialmente a *DLP*. O facto de que $DLP \in NP$ é evidente, pois dada uma solução para uma dada instância, i.e., um caminho, basta considerar o caminho $(m_0, m_1, m_2) = (P_0, P_1, P_2)$ e o conjunto de valores para os ângulos diedrais que podem oscilar. Com estes dados é possível calcular os pontos $\{m_{N-2}, m_{N-1}, m_N\}$ em tempo polinomial em $|I|$ e verificar a condição $(m_{N-2}, m_{N-1}, m_N) = (Q_2, Q_1, Q_0)$. O candidato a problema *NP-completo* redutível polinomialmente a *DLP* é o problema *PARTITION*.

Problema 3.1.2 (PARTITION) ***Instância do Problema:*** *Um conjunto finito* $A \subset \mathbb{N}_0$. ***Objectivo:*** *Determinar se existe* $A' \subseteq A$ *tal que* $\sum_{n \in A'} n = \sum_{n \in A \setminus A'} n$.

Para se verificar que o problema *PARTITION* é redutível polinomialmente ao problema *DLP* basta encontrar um procedimento para converter qualquer instância de *PARTITION* numa instância de *DLP* em tempo polinomial. Na proposição seguinte prova-se que esse procedimento existe, i.e., que *DLP* é *NP-completo*, seguindo-se a prova original de Ngo e Marks em 1992 a menos de notação.

Proposição 3.1.3 ([NM92]) *O problema DLP é NP-completo.*

Prova: Seja $x = (x_1, x_2, x_3) \in D$. Escolhendo-se a unidade de distância e a orientação convenientes, tem-se que os pontos vizinhos de x são $x - \vec{u}$, $x + \vec{v}_0$, $x + \vec{v}_1$ e $x + \vec{v}_2$, em que $\vec{u} = (1, 1, 1)$, $\vec{v}_0 = (-1, 1, 1)$, $\vec{v}_1 = (1, -1, 1)$ e $\vec{v}_2 = (1, 1, -1)$. Dada uma instância de *PARTITION*, i.e., um conjunto finito $A = \{n_1, ..., n_k\} \subset \mathbb{N}_0$, define-se uma instância de *DLP* da seguinte forma:

$$
\begin{array}{lcllcl}
B & = & \frac{1}{2}\sum_{n\in A} n & \Delta_0 & = & (2\vec{u}+\vec{v}_0+\vec{v}_1)B \\
 & & & & & \\
P_0 & = & 0 & Q_2 & = & P_1+\Delta_0 \\
P_1 & = & P_0+\vec{v}_0 & Q_1 & = & P_1+\Delta_0+\vec{u} \\
P_2 & = & P_0+\vec{v}_0+\vec{u} & Q_0 & = & P_1+\Delta_0+\vec{u}+\vec{v}_0 \\
 & & & & & \\
N & = & 4B+3 & & & \\
I & = & \{2, 2n_1+2, 2n_1+2n_2+2, ..., 4B+2\} & & &
\end{array}
$$

Observa-se que B é um número inteiro, porque a soma $\sum_{n\in A} n$ é um número par (caso contrário não existiria trivialmente uma partição de A). O tempo de construção da instância de *DLP* é claramente polinomial no tamanho da instância de *PARTITION*. Falta agora mostrar que a instância de *DLP* obtida tem resposta positiva se e só se a instância do problema *PARTITION* tem resposta afirmativa. Sabe-se que cada ligação pode ter uma entre oito possíveis orientações, i.e., $\delta_i \in \{\pm\vec{u}, \pm\vec{v}_0, \pm\vec{v}_1, \pm\vec{v}_2\}$. Portanto, fixando-se m_0, m_1 e m_2 em P_0, P_1 e P_2, respectivamente, e tomando-se $\Delta = m_{N-2} - m_1$, pretende-se verificar em que condições se tem $\Delta = \Delta_0$. Dado que o deslocamento pretendido na direcção de $\vec{u}$ é dado por:

$$\Delta_0 \cdot u = (2\vec{u}\cdot\vec{u} + \vec{v}_0\cdot\vec{u} + \vec{v}_1\cdot\vec{u})B$$

e que não existem duas ligações seguidas na direcção $\vec{u}$, tem-se que este é o deslocamento máximo possível. Como as projecções dos elementos de $\{\pm\vec{u}, \pm\vec{v}_0, \pm\vec{v}_1, \pm\vec{v}_2\}$ na direcção $\vec{u}$ são dadas por:

$$
\begin{array}{cc}
\vec{u}\cdot\vec{u} & 1 \\
\vec{v}_0\cdot\vec{u} = \vec{v}_1\cdot\vec{u} = \vec{v}_2\cdot\vec{u} & \frac{1}{3} \\
-\vec{v}_0\cdot\vec{u} = -\vec{v}_1\cdot\vec{u} = -\vec{v}_2\cdot\vec{u} & -\frac{1}{3} \\
-\vec{u}\cdot\vec{u} & -1
\end{array}
$$

verifica-se ainda que o deslocamento máximo é atingido quando $\delta_i = \vec{u}$, para $i \in \{2, 4, 6, ..., 4B\}$, e quando $\delta_i \in \{\vec{v}_0, \vec{v}_1, \vec{v}_2\}$, para $i \in \{3, 5, 7, ...4B+1\}$, e.g. figura 3.1. Sejam V_0, V_1 e V_2 os conjuntos das ligações, i.e., dos δ_i's, entre m_1 e m_{N-2} nas direcções $\vec{v}_0$, $\vec{v}_1$ e $\vec{v}_2$, respectivamente. O número de ligações nestas direcções é $2B$, i.e., $|V_0|+|V_1|+|V_2| = 2B$. Seja $\vec{w}$ a projecção de $\vec{v}_0 + \vec{v}_1$ no plano ortogonal a $\vec{u}$, tem-se que:

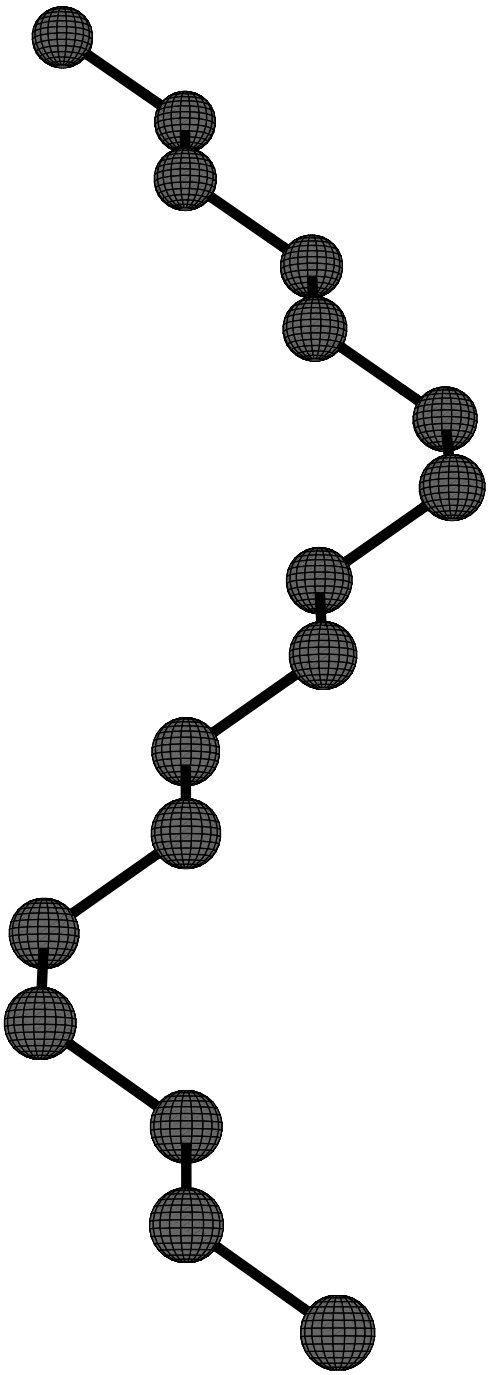

Figura 3.1: Construção para a prova da proposição 3.1.3, em que a instância de *PARTITION* é dada por $A = \{1, 3, 2\}$ e tem resposta afirmativa. Apenas estão representados os átomos de carbono. Admitindo a numeração de 0 a 15 dos átomos de carbono de baixo para cima, tem-se que $I = \{2, 4, 10, 14\}$. Portanto apenas as ligações 2, 4, 10 e 14 podem ter um ângulo diedral diferente 180°.

$$\Delta \cdot \vec{w} = |V_0|\vec{v}_0 \cdot \vec{w} + |V_1|\vec{v}_1 \cdot \vec{w} + |V_2|\vec{v}_2 \cdot \vec{w}$$

e portanto $|V_2| = 0$, pois o deslocamento pretendido na direcção $\vec{w}$ é

$$\Delta_0 \cdot \vec{w} = (\vec{v}_0 \cdot \vec{w} + \vec{v}_1 \cdot \vec{w})B$$

e, por simetria, $\vec{v}_0 \cdot \vec{w} = \vec{v}_1 \cdot \vec{w} > \vec{v}_2 \cdot \vec{w}$. Portanto tem-se que

$$\Delta = 2B\vec{u} + |V_0|\vec{v}_0 + |V_1|\vec{v}_1.$$

Para que $\Delta = \Delta_0$, é necessário que $|V_0| = |V_1| = B$. Contudo não é permitida uma escolha arbitrária das ligações para V_0 e V_1 dado que a direcção de δ_{i-1} e de δ_{i+1} apenas podem diferir quando $i \in I$. Como os valores em I subdividem a sequência de ligações em $V_0 \cup V_1$ em subsequências de comprimento n_1, n_2, ..., n_k, cujas ligações em $V_0 \cup V_1$ tem a mesma direccção, i.e., v_0 ou v_1, tem-se que existe um caminho em D que satisfaz as restrições impostas se e só se a instância do problema *PARTITION* tem resposta afirmativa. Logo, o problema *DLP* é *NP-completo.* ∎

Considerem-se os aminoácidos presentes numa dada proteína, todos os átomos da molécula e as respectivas ligações atómicas são conhecidos. Note-se que um átomo pode ocupar qualquer ponto do espaço e que existem determinadas restrições em cada ligação entre os átomos. Neste modelo, o problema *PF* consiste em determinar as posições dos átomos no espaço por forma a que a energia potencial da proteína seja mínima.

De facto, o problema *PF* formalizado como acima corresponde a um problema mais geral: o problema de determinar a conformação espacial de um dado polímero, quando são conhecidos os átomos que o constituem e as ligações entre estes, através da minimização da função de energia potencial da molécula, *PSP* (*Polymer Structure Prediction*). O problema *PSP* é um problema de optimização e ir-se-á reescrever o mesmo sob a forma de um problema de decisão, no qual o objectivo é determinar se a função de energia potencial assume algum valor abaixo de um determinado limite. Um problema de decisão derivado de um problema de optimização é sempre trivialmente redutível a este.

Problema 3.1.4 (PSP [NM92]) ***Instância do Problema:*** *Um tuplo de* $(N+1)$ *atómos,* $(m_0, m_1, ..., m_N)$*, cada um dos quais poderá ocupar uma posição qualquer no espaço tridimensional; um comprimento de ligação de equilíbrio* l_0^b *e um coeficiente positivo* K_b^{lig} *para cada par de átomos adjacentes,* $b = (m_i, m_{i+1})$*; um ângulo de equilíbrio* θ_a^0 *e um coeficiente* K_a^{ang} *para cada triplo de átomos adjacentes,* $a = (m_i, m_{i+1}, m_{i+2})$*; um ângulo diedral de equilíbrio* ϕ_d^0*, um inteiro positivo* n_d *e um coeficiente positivo* K_d^{die} *para cada quadruplo de átomos adjacentes,* $d = (m_i, m_{i+1}, m_{i+2}, m_{i+3})$*; um coeficiente positivo* K_{ij}^{nloc} *e um raio de equilíbrio* r_{ij}^0 *para cada par de átomos* $i > j$ *no tuplo; e um limite de energia,* U_0*.* ***Objectivo:*** *Determinar se existem valores para as variáveis* l_b*,* θ_a *e* ϕ_d*, tais que a função de energia potencial*

$$\begin{aligned} V(x) &= \sum_b K_b^{lig}(l_b - l_b^0)^2 \\ &+ \sum_a K_a^{ang}(\theta_a - \theta_a^0)^2 \\ &+ \sum_d K_d^{die}(1 - \cos[n_d\phi_d - \phi_d^0]) \\ &+ \sum_{i>j} K_{ij}^{nloc} f(r_{ij}/r_{ij}^0) \end{aligned}$$

tenha um valor que não exceda U^0*. (A variável* r_{ij} *é a distância entre os átomos* m_i *e* m_j*. Consideram-se instâncias de PSP aquelas em que a função* f *tem um mínimo global em* 1*.)*

Importa referir que o problema *PSP* pode não estar em *NP*. De facto, para um algoritmo verificar uma solução candidata em tempo polinomial teria de utilizar precisão infinita e, portanto, para verificar se este problema está em *NP* seria necessário fazer uma sofisticada análise do erro. Por outro lado, ainda que o problema de decisão esteja em *NP*, o problema de optimização correspondente poderá não estar em *NP*.

Verifica-se então de seguida que o problema *PSP* é *NP-difícil*. Para

tal, ir-se-á estabelecer uma redução polinomial do problema *DLP* para o problema *PSP*. De facto considerar-se-á o problema *DLP'* que tem como instâncias as de *DLP* que se obtêm por redução a partir de instâncias do problema *PARTITION*. É óbvio que *DLP'* é *NP-completo* pela proposição 3.1.3.

Proposição 3.1.5 ([NM92]) *O problema PSP é NP-difícil.*

Prova: Antes de mostrar a redução polinomial a partir do problema *DLP'*, nota-se que para que a redução seja eficiente é necessário que o tamanho da instância de *PSP* seja polinomial no tamanho da instância *DLP'* correspondente. Dado que a primeira tem tamanho linear em N e a segunda tem tamanho linear em $|I|$, isto pode não acontecer. Este facto pode ser corrigido através de uma codificação da instância de *PSP* em que são atribuídos valores por omissão às variáveis K^{lig}, K^{ang}, K^{die}, l^0, θ^0, ϕ^0 e n, sendo apenas guardada uma lista de pares ordenados de comprimento $|I|$ para cada conjunto de variáveis em que o primeiro elemento do par corresponde ao índice da variável que deve ser alterada e o segundo elemento corresponde ao seu novo valor. Desta forma, a codificação da instância de *PSP* obtida é linear em $|I|$. Nota-se ainda que as instâncias de *PSP* obtidas por redução a partir de instâncias de *DLP'* devem possuir parâmetros que possam ser guardados em memória finita. Assumir-se-á, portanto, que a unidade dos comprimentos das ligações é $\sqrt{3}$ e a unidade dos ângulos é $\arcsin(-\frac{1}{3})$.

Dada uma instância de *DLP'*, a cadeia de átomos obtida por redução na instância de *PSP* possui duas regiões: a região *variável* e a região de *suporte*. A região *variável* consiste em ν átomos, $(m_0, ..., m_\nu)$, em que ν é igual ao parâmetro N da instância de *DLP'*. Os parâmetros associados à região *variável* são os seguintes: l_b^0 toma o valor $1(\times\sqrt{3})$ para todo o b; θ_a^0 toma o valor $1(\times \arcsin(-\frac{1}{3}))$ para todo o a; ϕ_d^0 toma o valor $180°$ para todo o d; dado $d = (m_i, m_{i+1}, m_{i+2}, m_{i+3})$, n_d toma o valor 3 sempre que $i + 2 \in I$ e 1 caso contrário. Os coeficientes K_b^{lig}, K_a^{ang} e K_d^{die} podem

assumir qualquer valor positivo. Da redução sabe-se também que a cadeia de átomos associada à região *variável* pode ser orientada de forma a ocupar um caminho em D. Sem perda de generalidade assume-se que a cadeia está orientada de modo a que os átomos $m_{\nu-2}$, $m_{\nu-1}$ e m_ν ocupem as posições Q_2, Q_1 e Q_0, respectivamente.

Falta garantir que, caso a instância de DLP' tenha resposta afirmativa, os átomos m_0, m_1 e m_2 ocupem as posições P_0, P_1 e P_2, respectivamente. Para isso utilizar-se-á a região de *suporte* e os termos não locais. Qualquer que seja d, n_d toma o valor 1 na região de *suporte* e, portanto, ao colocarem-se os restantes parâmetros geométricos, a região de *suporte* tem apenas uma configuração de energia mínima igual a 0. À excepção de alguns coeficientes K_{ij}^{nloc} que tomam um valor positivo arbitrário, todos tomam o valor 0. Desta forma, os termos não locais assumem o mínimo quando $r_{ij} = r_{ij}^0$. Introduzem-se quatro restrições de distância para cada um dos átomos $\{m_0, m_1, m_2\}$: três para remover os três graus de liberdade espaciais e uma para remover a simetria. Os parâmetros geométricos da região de *suporte* são escolhidos de forma a que, na conformação de energia mínima igual a 0 da região de suporte, os átomos m_N, m_{N-1}, m_{N-2} e m_{N-3} ocupem as posições determinadas pelos pontos P_0', P_1', P_2' e P_3' não coplanares, respectivamente. As distâncias r_{ij}^0 para os doze pares $\{m_0, m_1, m_2\} \times \{m_N, m_{N-1}, m_{N-2}, m_{N-3}\}$ tomam os valores das distâncias determinadas pelos doze pares de pontos em $\{P_0, P_1, P_2\} \times \{P_0', P_1', P_2', P_3'\}$, respectivamente. A região de *suporte* pode não corresponder a um caminho em D.

O limite U_0 toma o valor que resulta de somar os mínimos de cada termo. Portanto, a instância de *PSP* obtida por redução a partir de uma instância de *DLP'*, tem resposta afirmativa se e só se a instância de *DLP'* tem resposta afirmativa, i.e., *PSP* é *NP-difícil* dado que *DLP'* é *NP-completo.* ∎

3.1.2 Modelo LPE

Em 1993, Unger e Moult [UM93] sugerem um outro modelo para capturar a essência do processo de enrolamento das proteínas durante a sua formação. Nesta secção ir-se-ão apresentar o modelo e os resultados obtidos relativamente à complexidade computacional de se encontrar a conformação de energia mínima. As implicações decorrentes desta análise serão discutidas na última secção deste capítulo. Tal como no estudo apresentado em [NaMK94], o problema resultante da interpretação do problema *PF* sobre este modelo será designado por *LPE* (*Lattice Polymer Embedding*).

Ao contrário do que acontece com o problema *PSP*, em que se pretende determinar vários parâmetros geométricos de forma a que os termos de energia potencial associados ao comprimento e aos ângulos das ligações entre os átomos sejam mínimos, o problema *LPE* consiste em determinar a configuração espacial óptima de uma cadeia de elementos de forma a que determinadas propensões dos mesmos sejam satisfeitas. O modelo consiste num reticulado finito 3D, i.e., numa grelha tridimensional cujos pontos são coordenadas no espaço com coeficientes inteiros, tal que, dada uma proteína, ou seja, uma cadeia de aminoácidos, cada aminoácido ocupa um ponto do reticulado, dois aminoácidos vizinhos na proteína ocupam posições adjacentes no reticulado e dois aminoácidos não ocupam simultaneamente o mesmo ponto do reticulado.

Modelos deste tipo são normalmente utilizados dado que as simulações sobre os mesmos têm um custo computacional muito menor do que as simulações em modelos mais próximos da realidade, i.e., do que em modelos em que os pontos são quaisquer. Estes modelos discretos capturam ainda assim propriedades bastante importantes das proteínas: dois elementos vizinhos na cadeia têm de permanecer ligados no reticulado; e dois elementos não podem estar simultaneamente a ocupar a mesma posição. Estas propriedades são de facto responsáveis pelas restrições no movimento das cadeias durante o processo real de enrolamento das proteínas.

Outras propriedades das proteínas, tais como as propensões de cada aminoácido, são integradas no modelo através da definição de uma função de energia. Existem várias hipóteses para definir a função de energia: dada informação acerca da estrutura secundária da cadeia ou de parte da cadeia, a função de energia pode ser definida de acordo; também se pode definir a função de energia através das propriedades químicas de cada aminoácido, separando-se os aminoácidos em diferentes grupos químicos e determinando-se os coeficientes de afinidade entre os mesmos.

Problema 3.1.6 (LPE [UM93]) ***Instância do Problema:*** *Uma cadeia de elementos* $S = \{s_1, s_2, ..., s_{|S|}\}$ *para ser representada num reticulado 3D* $L = \{1, 2, ..., |S|\}^3$*, um coeficiente* $c(s_i, s_j)$ *para cada par de elementos* $s_i, s_j \in S$ *e uma função* $f : \{1, ..., |S|\}^3 \rightarrow \mathbb{R}$*.* ***Objectivo:*** *Encontrar o mínimo global de energia* U^{min} *da cadeia. A cadeia tem de ser representada no reticulado* L *de forma a que dois vizinhos na cadeia ocupem posições vizinhas no reticulado (também se consideram os pontos da diagonal como vizinhos) e que dois elementos da cadeia não ocupem o mesmo ponto do reticulado. A função de energia* U *define-se da seguinte forma:*

$$U = \sum_{i<j} c(s_i, s_j) f(|x(s_i) - x(s_j)|, |y(s_i) - y(s_j)|, |z(s_i) - z(s_j)|)$$

em que $(x(s_i), y(s_i), z(s_i)) \in L$ *é o ponto em que o elemento* s_i *está posicionado.*

Unger e Moult mostraram que o problema *LPE*, em que no modelo subjacente a função de energia é definida de acordo com a segunda hipótese indicada acima, é *NP-difícil*. A prova consiste em encontrar uma redução eficiente do problema *OLA* (*Optimal Linear Arrangement*) [GJ79] para o problema *LPE*.

Problema 3.1.7 (OLA) ***Instância do Problema:*** *Uma lista de N pessoas, uma tabela que indica que pares são amigos e uma distância máxima* D_{max}*.* ***Objectivo:*** *Determinar se as pessoas podem ser sentadas num fila*

de N cadeiras de forma a que a soma D das distâncias entre amigos não exceda D_{max}. A distância entre dois amigos é 1 se estes estiverem sentados lado a lado, 2 se existir uma cadeira entre eles, etc. A soma D é realizada tendo em conta todos os pares de amigos. Duas pessoas não podem ocupar a mesma cadeira.

Apresenta-se de seguida a prova de que *LPE* é *NP-difícil*, importando referir que alguns detalhes da prova original foram omitidos.

Proposição 3.1.8 *([UM93]) O problema LPE é NP-difícil.*

Prova: Pretende-se encontrar uma redução do problema *OLA* para o problema *LPE*. Dada uma instância do problema *OLA*, i.e., N pessoas entre as quais existem amigos e uma distância máxima D_{max}, pretende-se saber se é possível sentá-las em N cadeiras de modo a que a soma das distâncias entre amigos não ultrapasse D_{max}. Considere-se o seguinte procedimento para transformar uma instância de *OLA* numa instância de *LPE*:

1. Construir uma cadeia com elementos de dois tipos: um conjunto de elementos que correspondem às pessoas da instância de *OLA*; um conjunto de elementos extra para dar flexibilidade à cadeia. Os elementos do primeiro conjunto são colocados na cadeia em qualquer ordem separados por k elementos do segundo conjunto.

2. Definir os coeficientes $c(s_i, s_j)$ para cada par de elementos da seguinte forma: se s_i e s_j corresponderem a um par de amigos na instância de *OLA*, $c(s_i, s_j)$ toma o valor 1; em todos os outro casos $c(s_i, s_j)$ toma o valor 0. Com esta escolha a função de energia irá depender apenas dos pares de amigos.

3. Definir a função f da seguinte forma:

$$f(x, y, z) = \begin{cases} x & \text{se } y = z = 0 \\ \frac{N^3}{6} & \text{c.c.} \end{cases}$$

Nota-se que esta escolha leva a que a configuração de energia mínima ocorra quando todos os elementos envolvidos na relação de amizade estiverem alinhados sobre uma recta com direcção do vector $(1, 0, 0)$. Suponha-se que tal não se verifica, então existe um par de amigos s_i e s_j que não está sobre uma recta nesta direcção e logo $y(s_i) \neq y(s_j)$ e $z(s_i) \neq z(s_j)$. Portanto $f(|x(s_i) - x(s_j)|, |y(s_i) - y(s_j)|, |z(s_i) - z(s_j)|) = \frac{N^3}{6}$ e, uma vez que todos os termos da função de energia são não negativos, o valor da energia da cadeia é no mínimo $\frac{N^3}{6}$. Esta não pode ser a configuração de energia mínima, pois a energia de uma configuração em que todos os elementos correspondentes a pessoas da instância de *OLA* estão sobre uma recta com a direccção do vector $(1, 0, 0)$ em qualquer ordem, é no máximo $\frac{N^3-N}{6}$.

4. Escolher um valor inteiro suficientemente grande para k de modo a que as sequências de elementos extra sejam suficientemente flexíveis para permitir todas as permutações dos elementos correspondentes a pessoas da instância de *OLA* sobre a recta indicada acima.

Verifica-se facilmente que o procedimento descrito é polinomial no tamanho da instância de *OLA*. Portanto a redução em causa é polinomial. De acordo com a construção da instância de *LPE* apresentada, tem-se que todos os elementos envolvidos na relação de amizade inerente à instância de *OLA* ficam posicionados sobre uma recta na direcção $(1, 0, 0)$ na configuração de energia mínima e, o valor deste mínimo é precisamente o menor valor que se consegue obter na soma da instância de *OLA* original. Logo, se existir um algoritmo eficiente que resolva o problema *LPE*, existe também um algoritmo eficiente que resolve o problema *OLA*, i.e., *LPE* é *NP-difícil*. ∎

3.1.3 Modelo H-P

Para testar as várias hipóteses relativas ao processo de formação das proteínas, foram propostos inúmero modelos matemáticos abstractos que capturam determinados aspectos. O modelo *H-P* (*hydrophilic-hydrophobic model*)

é um dos mais estudados. Em particular o modelo *H-P* bidimensional proposto por Dill [Dil90] tem vindo a ter bastante sucesso nas experiências realizadas. Este modelo é de facto um caso especial do modelo *LPE* discutido na seccção anterior.

Neste modelo, as proteínas são representadas por sequências de 0's e 1's, em que 0 corresponde a um aminoácido hidrófilo e 1 corresponde a um aminoácido hidrófobo. O objectivo consiste em representar as sequências num reticulado bidimensional sem que ocorram sobreposições e tal que a energia seja miníma. Dada uma configuração sobre o reticulado, atribui-se uma pontuação igual ao número de pares de 1's adjacentes no reticulado que não sejam adjacentes na sequência. Por exemplo, na figura 3.2, a pontuação é 5. A pontuação definida desta forma captura um modelo simples de minimização de energia, no qual os 1's hidrófobos tendem a permanecer próximos evitando a exposição ao meio envolvente, enquanto que os 0's hidrófilos são neutrais. Assume-se assim que a configuração que maximiza a pontuação é a que corresponde à configuração nativa, i.e., à configuração de menor energia.

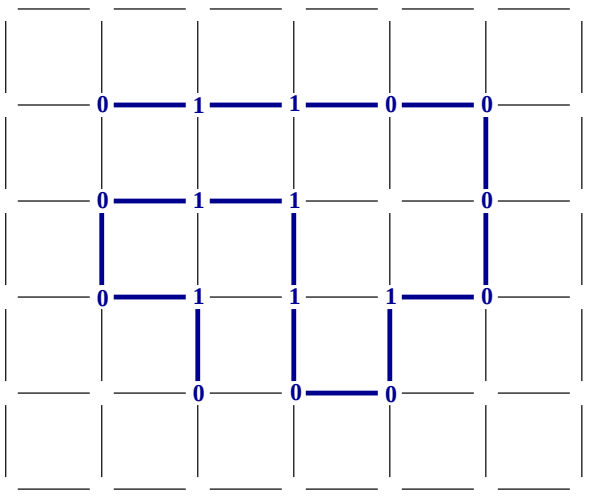

Figura 3.2: Exemplo da configuração de uma sequência.

Portanto, um problema bastante interessante consiste em, dada uma sequência de 0's e 1's, determinar qual a configuração que maximiza a pon-

tuação. Nesta seccção ir-se-ão apresentar alguns resultados acerca da complexidade deste problema, o qual se designará por problema *H-P*.

Problema 3.1.9 (H-P) ***Instância do Problema:*** *Uma sequência de* 0*'s e* 1*'s.* ***Objectivo:*** *Determinar a configuração sobre um reticulado bidimensional com maior pontuação.*

São várias as referências que se encontram na literatura a respeito da complexidade do problema 3.1.9. Nesta secção apresenta-se o resultado obtido por Crescenzi, Goldman, Papadimitriou, Piccolboni e Yannakakis [CGP$^+$98] que mostraram em 1998 que o problema 3.1.9 é *NP-completo* e que envolve uma redução do problema *HC* (*Hamilton Cycle*).

Problema 3.1.10 (HC) ***Instância do Problema:*** *Um grafo não dirigido* $G = (V, E)$*, em que* V *é o conjunto de vértices e* E *é o conjunto de arestas.* ***Objectivo:*** *Determinar se existe um caminho hamiltoniano, ou seja, um ciclo simples em* G *que visita todos os vértices* V*.*

Em primeiro lugar mostra-se que a versão do problema 3.1.9 para um conjunto de sequências finito, problema 3.1.11, é *NP-completo*. Neste contexto, o reticulado bidimensional é o grafo $R = (\mathbb{Z}^2, L)$, em que $\mathbb{Z}^2$ é o conjunto de vértices e $L = \{((x, y), (x', y')) : |x - x'| + |y - y'| = 1\}$ é o conjunto de arcos. Considere-se o conjunto de sequências $S = \{s_1, ..., s_m\}$ sobre o alfabeto $\{0, 1\}$, uma conformação das sequências de S sobre o reticulado R é um mapa bijectivo $f : \{(i, j) : 1 \leq i \leq m, 1 \leq j \leq |s_i|\} \rightarrow \mathbb{Z}^2$ tal que $(f(i, j), f(i, j + 1)) \in L$, para todo o i e j tais que $1 \leq i \leq m,\ 1 \leq j \leq |s_i|$. Dada uma conformação f, $f(i, j)$ e $f(i, j + 1)$ dizem-se vizinhos. No seguimento dir-se-á que ocorre uma *perda* quando dois pontos adjacentes em L não são vizinhos e um e um só deles é imagem de f a partir de um par (i, j) tal que $s_i[j] = 1$.

Problema 3.1.11 ([CGP$^+$98]) ***Instância do Problema:*** *Um conjunto de sequências* $S = \{s_1, ..., s_m\} \subset \{0, 1\}^*$ *e um inteiro* E*.* ***Objectivo:*** *De-*

terminar se existe uma configuração de S sobre o reticulado bidimensional com E perdas no máximo.

Proposição 3.1.12 ([CGP$^+$98]) *O problema 3.1.11 é NP-completo.*

Prova: Face à extensão desta prova, apenas se indicam os pontos principais. Para mais detalhes ver [CGP$^+$98]. Este resultado mostra-se por redução a partir do problema *HC*. Dada uma instância de *HC*, i.e., dado um grafo G, é necessário construir um conjunto de sequências finito S e determinar um inteiro E tal que existe um caminho hamiltoniano em G sse existe uma configuração de S sobre R com E perdas no máximo.

A redução utiliza um mapa proposto por Trevisan [Tre97] denotado por T. O mapa T mapeia os vértices do grafo G em $\{0,1\}^{8n}$ de forma a que $d_H(T(u),T(v)) = \frac{7n}{2}$, se (u,v) é um arco, e $d_H(T(u),T(v)) = 4n$, se (u,v) não são adjacentes em G. A distância d_H refere-se à distância de Hamming entre duas sequências de bits e n refere-se ao número de vértices de G. Deste modo, o conjunto de sequências S contém L sequências, em que L é polinomial em n e tal que existem n sequências diferentes, i.e., existem L/n cópias de cada sequência. Cada sequência s_i, com $i \neq 1, n$, contém $T(v_i)$ como subsequência no meio. De forma a garantir a equivalência indicada atrás, o meio de cada sequência é constituído na sua maioria por 0's com alguns 1's escolhidos cuidadosamente. A primeira e última sequências são ligeiramente diferentes: possuem 1's adicionais entre cada par de 0's. As sequências diferem praticamente apenas na estrutura dos $T(v)$'s, sendo a restante estrutura semelhante.

A configuração sobre R é obtida colocando-se todas as sequências, à excepção da primeira e da última, lado a lado com todos os pontos iniciais alinhados horizontalmente. A primeira e última sequências colocam-se de forma a que os seus 0's fiquem alinhados horizontalmente com os 0's das restantes sequências. As perdas ocorrem apenas entre subsequências correspondentes a $T(v)$'s e, portanto, são limitadas pela propriedade do mapa. Logo, esta configuração garante no máximo $7n$ perdas, i.e., $E = 7n$.

O recíproco sai observando-se que qualquer configuração com no máximo E perdas terá uma estrutura semalhante à descrita. O primeiro passo consiste em verificar a existência de uma região densa rectangular, com área aproximadamente L^2 e circunscrita e inscrita por rectângulos com lados de comprimento próximo de L. Então mostra-se que a maioria das sequências cortam o rectângulo inscrito de alto a baixo.

Dado que o problema *HC* é *NP-completo*, pela redução anterior tem-se que o problema 3.1.11 é *NP-completo.* ∎

A proposição 3.1.13 resulta como consequência da proposição 3.1.12 e do facto de que problema *HC* permanece na classe dos problemas *NP-completos* ainda que restrito a grafos planares especiais[1].

Proposição 3.1.13 ([CGP$^+$98]) *O problema 3.1.9 é NP-completo.*

Tal como é referido em [CGP$^+$98], a redução anterior pode generalizar-se para o caso 3D, e portanto a versão tridimensional do problema 3.1.9 é *NP-completo.* Em 1998, Berger e Leighton [BL98] mostraram também, independentemente dos primeiros, que uma versão tridimensional do problema 3.1.9, problema 3.1.14, é *NP-completo*, proposição 3.1.15.

Problema 3.1.14 ([BL98]) ***Instância do Problema:*** *Um inteiro m e uma sequência de 0's e 1's.* ***Objectivo:*** *Determinar se existe uma configuração da sequência sobre o reticulado tridimensional com pontuação igual a m.*

Proposição 3.1.15 ([BL98]) *O problema 3.1.14 é NP-completo.*

Apenas por curiosidade, nota-se que a prova deste teorema segue por redução do problema *BP* (*Bin Packing*) ao problema 3.1.14.

Problema 3.1.16 (BP) ***Instância do Problema:*** *Um conjunto finito de $U \subset \mathbb{N}$ e uma capacidade $B \in \mathbb{N}$.* ***Objectivo:*** *Determinar se existe uma*

[1]Um grafo planar especial possui regiões com vértices de grau 3, ligadas entre si por caminhos de comprimento 2 e, quando se colapsam os vértices de grau 2, fica triligado.

partição U_1, ..., U_k de U tal que a soma dos elementos de cada U_i seja inferior a B.

3.2 Paradoxo de Levinthal

Se o tamanho do espaço de configurações possíveis de uma proteína é exponencial no tamanho da mesma, como pode esta determinar a configuração óptima em poucos segundos? Considere-se a seguinte ilustração, a ligação entre dois aminoácidos pode assumir diferentes estados (considerem-se 3 estados), portanto se uma proteína tiver 101 aminoácidos podem existir 3^{101} configurações; mesmo que uma proteína pudesse analizar 10^3 configurações por segundo, precisaria de 10^{27} segundos para analizar todas. Com base neste facto, Levinthal concluiu que uma procura aleatória não poderia levar à configuração óptima. Ainda assim as proteínas atingem a configuração óptima em poucos segundos.

No estudo do problema *PSP* utiliza-se um modelo contínuo e uma função de energia que se pretende minimizar, i.e., captura-se a essência envolvida na mecânica molecular. Por outro lado, o facto de se estudar o problema sobre um modelo contínuo fornece informações acerca da dificuldade do mesmo que não se adquirem quando se considera um modelo discreto. Embora seja comum assumir-se que os problema contínuos são pelo menos tão difíceis como as versões discretas correspondentes, o que é verdade na maior parte dos casos, é possível que um problema discreto seja intratável enquanto que a sua versão contínua seja resolvida eficientemente[2].

O problema *LPE* distingue-se de outros problemas formalizados sobre reticulados em dois detalhes: a função de energia faz parte da instância; na redução apresentada, uma das direcções é privilegiada. O facto da função de energia fazer parte da instância leva a que um algoritmo desenvolvido para resolver *LPE* desconheça a função de energia e portanto terá de ser suficientemente geral, o que poderá contribuir para uma maior complexi-

[2]Veja-se o exemplo da programação linear, ainda que existam algoritmos eficientes para o caso geral, o problema posto para variáveis inteiras é *NP-completo*.

dade do problema. Como se viu quer no problema *PSP* quer no problema *H-P* a forma da função é conhecida e portanto pode ser explorada pelo algoritmo. Ao privilegiar-se uma direcção na redução a partir do problema *OLA*, faz-se com que o polímero em causa tenha três valores de energia diferentes mediante a orientação. Este facto não é comum nos modelos sobre reticulados convencionais nem nas proteínas e poderá contribuir também com complexidade adicional.

O problema *H-P* assenta provavelmente no modelo mais simples alguma vez considerado, mas ainda próximo da realidade. A função de energia é conhecida à partida e captura uma das propriedades mais bem conhecidas das proteínas, nomeadamente a propriedade de afinidade com a água por parte dos aminoácidos. Verifica-se que, mesmo definido sobre um reticulado bidimensional, é um problema *NP-completo*.

Indicados alguns detalhes no estudo de cada um dos problemas, nota-se que o problema *PSP* é muito mais geral que o problema *PF*, o que significa que existem instâncias de *PSP* que não são de forma alguma instâncias de *PF*. Deste modo, dado que o resultado de *PSP* ser *NP-difícil* estabele o comportamento no pior caso, não se pode conluir que o problema *PF* seja *NP-difícil*. No entanto este resultado revela alguns dados importantes. Os algoritmos desenvolvidos para resolver problemas de optimização gerais não são eficientemente aplicáveis ao problema *PF*, porque não consideram informações específicas acerca do domínio de aplicação, são bastante lentos e não têm utilidade prática. Logo, pode-se inferir que um algoritmo correcto e eficiente que resolva o problema *PF* terá de ter em conta informação específica acerca das proteínas. O mesmo é verdade para os algoritmos de aproximação.

É também importante referir que, se for encontrado um algoritmo correcto e eficiente para o problema *PSP*, este servirá também para resolver o problema *PF* e, tendo em conta o estudo da secção 3.1.1, servirá para resolver qualquer problema *NP-completo*. A existência de um tal algoritmo

implicaria que $P = NP$. O resultado acerca da complexidade não compromete no entanto a existência de um algoritmo não correcto e eficiente para o problema *PSP* que permita resolver instâncias de *PF* correctamente.

Logo, como se viu acima e tal como se discute em [NaMK94], para que a questão de Levinthal seja resolvida uma das seguintes afirmações tem de ser verdadeira:

1. $P = NP$. Esta possibilidade é considerada pouco provável, pois como se sabe implicaria que todos os problemas NP-*completos* serião resolvidos eficientemente.

2. *O problema PSP é intratável porque* $P \neq NP$*, mas toda a proteína que atinge a configuração óptima em tempo polinomial corresponde a uma instância de PSP que pode ser resolvida correcta e eficientemente. O conjunto de tais instâncias constituem um forma restrita do problema PSP que é tratável.* De facto poderão existir algoritmos que levem em conta determinadas propriedades das proteínas para além da função de energia potencial. Pode também acontecer que a função de energia potencial apresentada na definição do problema *PSP* não contenha determinados termos de energia existentes no caso das proteínas.

3. *Não é garantido que uma proteína consiga a configuração óptima em tempo polinomial.* Quer do ponto de vista matemático, quer do ponto de vista físico, existe alguma garantia de que um sistema físico faça o quer que seja dado que os princípios da mecânica-quântica são probabilísticos? A análise levada a cabo até agora baseia-se nos limites da física clássica em que os resultados deterministas são possíveis. Ao considerarem-se algoritmos probabilistícas os resultados de complexidade aqui apresentados não são violados e podem existir algoritmos que resolvam o problema *PF* eficientemente.

4. *A configuração nativa de uma proteína pode ser determinada a partir da função de energia potencial através de um algoritmo que não involva a minimização global e cuja complexidade computacional é ainda desconhecida.*

Esta hipótese tem vindo a ser apontada por vários investigadores. Do ponto de vista experimental nem todas as proteínas parecem atingir a configuração nativa com o mínimo global de energia, mas sim com um minímo local. No entanto, apenas aquelas em que a configuração nativa coincide com o minímo global de energia parecem sobreviver à selecção natural.

3.3 Técnicas de aproximação

Face aos resultados de complexidade computacional para os diferentes modelos apresentados na secção 3.1, é natural que se tente desenvolver algoritmos e aplicar técnicas de aproximação que permitam obter a conformação espacial cuja energia seja o mais próximo possível da energia mínima.

O sistema dinâmico que resulta do conjunto de forças presentes na molécula e da interacção desta com o meio envolvente é representado rigorosamente através da equação de *Schrödinger*. No entanto, dada a complexidade associada à resolução da mesma, adopta-se normalmente uma descrição aproximada do sistema dinâmico segundo uma perspectiva física mais clássica. Tal como foi referido na secção 2.2, este sistema dinâmico pode ser aproximado matematicamente pela equação diferencial estocástica

$$M\ddot{x} + C\dot{x} + \nabla V(x) = D\dot{W}(t).$$

Admita-se que os átomos do solvente, i.e., os átomos das moléculas do meio envolvente, relevantes estão incluídas no vector x e que não existem oscilações devidas à energia cinética do sistema. Neste caso o objectivo é determinar o mínimo global da função de energia potencial $V(x)$ de acordo com a discussão apresentada na secção 2.2.

Normalmente a aproximação de *Born-Oppenheimer* é adequada para a função de energia potencial $V(x)$. Fixando-se as posições x dos átomos, o valor da energia é obtido como o menor valor próprio associado ao Hamiltoniano do sistema, um operador diferencial parcial associado à função de onda dos electrões. A aproximação dos valores próprios obtém-se através

dos métodos *ab-initio*. Contudo, os cálculos da mecânica quântica inerentes a estes métodos não são realizáveis para moléculas complexas face aos recursos computacionais actuais. Assim, utilizam-se descrições das moléculas em termos de ligações e interacções atómicas efectivas. Os cálculos teóricos da mecânica quântica apenas são utilizados em pequenas partes constituintes das moléculas e, os modelos são construídos com base nos dados assim obtidos e nos dados obtidos experimentalmente.

Estes modelos, também conhecidos por campos de forças, são expressos através da soma de vários termos cuja expressão matemática pretende capturar a natureza física desse termo. Em geral, a expressão para $V(x)$ deverá incluir termos associados ao comprimento das ligações, aos ângulos das ligações, aos ângulos de torsão e às interacções não locais, i.e.,

$$\begin{aligned} V(x) &= \sum_b K_b^{lig}(l_b - l_b^0)^2 \\ &+ \sum_a K_a^{ang}(\theta_a - \theta_a^0)^2 \\ &+ \sum_d K_d^{die}(1 - \cos[n_d\phi_d - \phi_d^0]) \\ &+ \sum_{i>j} K_{ij}^{nloc} f(r_{ij}/r_{ij}^0). \end{aligned}$$

Os primeiros termos envolvem os comprimentos das ligações e os valores dos ângulos destas, a sua representação baseia-se numa aproximação harmónica de acordo com a lei de Hooke e, pretendem impor uma penalização por forma a que os comprimentos l_b e os ângulos de ligação θ_a permaneçam o mais próximo possível dos valores de referência l_b^0 e θ_a^0 que dependem dos átomos em questão. O terceiro conjunto termos descreve a energia de rotação associada aos ângulos de torsão e, geralmente, obtêm-se a partir da expansão de uma série de Fourier. O último conjunto de termos descreve a energia associada às interacções não locais. Neste caso a função f corresponde ao potencial de Lennard-Jones em que cada valor r_{ij} corresponde à distância entre o átomo i e o átomo j e os valores r_{ij}^0 correspondem às distâncias de

equilíbrio. Para além destes podem ainda ser adicionados outros termos, nomeadamente termos associados à energia de solvatação. Nota-se que no problema 3.1.4 se utilizou uma expressão deste tipo para descrever a energia potencial da molécula.

Na predição da estrutura terciária das proteínas pretende-se minimizar globalmente a função de energia potencial $V(x)$. Os métodos usuais de minimização permitem identificar eficientemente mínimos locais. No entanto não se conhecem métodos eficientes para determinar mínimos globais. A exploração total da complexa hipersuperfície da função de energia potencial $V(x)$ apenas é exequível para sistemas pequenos. Por exemplo, em [LS88] foi estimado que, para um dado péptido com apenas cinco aminoácidos, o número de mínimos locais da função $V(x)$ ronda os 10^{11}. Verifica-se assim a necessidade de desenvolver métodos eficientes para determinar o mínimo global destas superfícies.

Métodos combinatórios

Alguns dos métodos propostos consistem na decomposição do problema da predição da estrutura terciária das proteínas em subproblemas. Em primeiro lugar determinam-se as estruturas terciárias para subsequências da cadeia de aminoácidos original. Normalmente a estrutura encontrada corresponde a um mínimo da energia potencial para a subsequência em causa. Em segundo lugar procede-se à combinação das estruturas obtidas por forma a obter-se a estrutura terciária da proteína original. Este processo de combinação está naturalmente sujeito a uma minimização de energia.

Uma implementação básica deste processo designa-se por método construtivo [GS87, VS85]. Neste caso a ideia baseia-se na combinação sucessiva de fragmentos menores, retendo-se as conformações de menor energia para cada um deles. Ainda que simples de implementar, o número de conformações cresce exponencialmente sendo necessário impor um limite apertado no valor de energia aquando da selecção das estruturas a guardar para

cada fragmento.

Com base no princípio da programação dinâmica, foi proposta uma modificação do método construtivo [VD90, VJSD93]. Importa referir que o princípio de optimalidade utilizado na programação dinâmica não se aplica directamente ao caso das proteínas, i.e., o facto da conformação de um dada sequência corresponder a um mínimo de energia não implica que todas as subsequências possuam uma conformação de energia mínima.

É comum nestes métodos considerar como fragmentos subsequências de aminoácidos para as quais se conhece a estrutura secundária, i.e., subsequências que correspondem a hélices-alfa ou folhas-beta. Passa-se assim a trabalhar com um nível de abstracção superior dado que, em vez de átomos ou aminoácidos, os elementos construtores básicos são as estruturas secundárias.

Algoritmos genéticos

Os algoritmos genéticos utilizam técnicas de optimização baseadas nos princípios da teoria evolucionária [Hol73]. Estes têm vindo também a ser aplicados ao problema da predição da estrutura terciária das proteínas para determinar o mínimo global da função de energia potencial $V(x)$. As variáveis do problema, e.g., comprimentos e ângulos das ligações, correspondem à sequência de genes de um cromossoma. O processo de optimização tem por base a manipulação de uma população de cromossomas através de processos de selecção, recombinação e mutação.

Em geral, a população inicial é constituída por um conjunto de cromossomas finito em que o valor de cada gene, i.e., o valor de cada variável, é seleccionado aleatoriamente. A aptidão de cada cromossoma é determinada através de uma função objectivo, e.g., a função de energia potencial $V(x)$. Dada uma população de cromossomas, obtém-se uma nova população através da selecção de um subconjunto finito de cromossomas da população original e da aplicação dos processos de recombinação e mutação sobre os

seleccionados. O processo de selecção consiste em escolher os elementos com maior aptidão; no caso de se utilizar a função $V(x)$ como função objectivo, são seleccionados os elementos que correspondem a um valor menor de $V(x)$. A recombinação realiza-se através da escolha de dois indivíduos e à troca de informação entre estes, i.e., constitui-se um novo indivíduo cujos genes são seleccionados entre os dois indivíduos escolhidos. O processo de mutação ocorre com baixa probabilidade e consiste na alteração aleatória de genes, atribuindo-se valores aleatórios a algumas variáveis num dado cromossoma. O processo de geração de novas populações termina quando um dado nível de optimalidade for atingido.

As aplicações dos algoritmos genéticos à predição da estrutura terciária das proteínas recorrem por vezes a bases de dados de conformações nativas, i.e., conformações que correspondem a mínimos de energia potencial, de forma a acelerar a convergência para a solução óptima. De facto, conformações aparentemente nativas foram identificadas para um número razoável de proteínas [TEHL91] e a identificação de conformações correspondentes a mínimos globais para proteínas com estruturas secundárias especiais têm também tido algum sucesso [LM93]. Foram também propostas versões modificadas de algoritmos genéticos. Uma das versões modificadas avalia a aptidão com base em mínimos locais e mostrou-se que esta garante um melhor desempenho [JCM$^+$92].

Nota-se que, quando as variáveis são fortemente correlacionadas, o que acontece no problema da predição da estrutura terciária das proteínas, a especificação da recombinação e da mutação torna-se bastante importante. Por outro lado as diferenças entre a formulação da função objectivo e a sua implementação exacta influência a capacidade de predição das conformações óptimas. Desta forma, a aplicação dos algoritmos genéticos estará em princípio restrita a classes específicas de problemas.

Arrefecimento simulado

O processo de arrefecimento simulado foi introduzido em 1983 por Kirkpatrick [KGV83] e tem como inspiração o facto de que o aquecimento do metal seguido de um arrefecimento lento torna-o num estado cristalino mais uniforme. O papel da temperatura neste processo é bastante importante, pois o aumento da mesma permite que as conformações atinjam estados de energia mais altos, com probabilidade dada pela lei de Boltzmann, possibilitando assim que as barreiras de energia sejam vencidas. Desta forma, as conformações deixam de estar restritas a um mínimo global e é possível explorar toda a superfície da função de energia potencial $V(x)$.

Antes de descrever esta metodologia em detalhe, é importante analisar o método de Metropolis Monte Carlo. A contribuição de Metropolis em relação ao método de Monte Carlo clássico revela-se com a utilização da distribuição de Boltzmann na orientação no espaço de procura. Esta abordagem consiste na escolha arbitrária de uma conformação inicial e em determinar a sua energia potencial, E_i. A cada passo, uma nova conformação é obtida através da alteração aleatória de algumas variáveis. Determina-se a energia potencial, E_n, de cada nova conformação e esta será aceite se se verificar a condição

$$P_B(T, E_n) \geq RND.$$

Nesta condição, $P_B(T, E)$ é um factor de probabilidade e $0 \leq RND \leq 1$ é gerado aleatoriamente. Na técnica de Metropolis o factor de probabilidade é obtido através da distribuição de Boltzmann, i.e.,

$$P_B(T, E_n) = \exp(\tfrac{-(E_n - E_i)}{k_B T}),$$

em que k_B é a constante de Boltzmann e T é a temperatura. Nota-se que todas as conformações com energia $E_n \leq E_i$ serão aceites visto que, nesse caso, $P_B(T, E_n) \geq 1$. Deste modo e a cada passo a procura é dirigida no sentido de uma menor energia, no entanto o facto de conformações com maior

energia que a conformação actual serem aceites aleatoriamente permitem a procura junto de outro mínimo local.

É óbvio que o sucesso de uma procura deste tipo depende directamente da escolha da temperatura. Se a temperatura for elevada serão aceites mais conformações e explorar-se-á uma maior região. Caso a temperatura seja baixa a procura ficará restrita aos mínimos locais mais próximos da conformações inicial. Para ultrapassar este inconveniente introduz-se então o conceito de arrefecimento [KGV83], i.e., toma-se uma temperatura elevada para a qual todos as conformações são acessíveis e baixa-se gradualmente a temperatura até se obter uma conformação com energia mínima. Se a temperatura for diminuída correctamente e de forma suficientemente lenta, o mínimo global de energia pode ser identificado.

A eficiência desta técnica para a optimização global tem vindo a ser testada e têm-se verificado resultados divergentes. Por exemplo, no estudo [OKK92] verifica-se a convergência para o mínimo global de energia potencial de um dado péptido num número significativo de simulações. No entanto, no trabalho [NVS91] contesta-se a possibilidade de se determinar o mínimo para esse mesmo péptido através da técnica de arrefecimento simulado. Ainda que esta divergência possa ser atribuída a diferentes formas de fazer a modelação, e.g., escolha de variáveis, este facto demonstra a natureza estocástica do método. Mesmo assim esta abordagem tem sido aplicada a uma grande variedade de proteínas [MGR91, Sno92], tendo sido introduzidas algumas melhorias através da combinação desta com outras técnicas [RVS91].

Métodos por alisamento e deformação

Sugeridos pela primeira vez por Stillinger e Weber [Sti85, SW88], os métodos por alisamento e deformação tem como intuição o facto de que na natureza as propriedades macroscópicas são um efeito dos detalhes microscópicos. Deste modo alisando a superfície de potencial, espera-se que a maioria dos mínimos locais desapareça e que as restantes propriedades da superfície ape-

nas denotem um único mínimo. Através da adição sucessiva de detalhes, as aproximações feitas pelo alisamento são recuperadas e obtém-se o mínimo global da superfície.

Embora seja verdade que um método com este possa falhar o mínimo global, uma implementação correcta desta ideia permitirá determinar um mínimo local muito próximo do mínimo global com uma fracção das operações necessárias aos métodos baseados em algoritmos genéticos ou na técnica de arrefecimento simulado.

Uma técnica de alisamento bastante atractiva é o método da equação de difusão [PKS89, LS87]. Seja $V(x)$ a função de energia potencial, a solução $V(x,t)$ da equação de difusão

$$\frac{\partial^2 V(x,t)}{\partial x^2} = \frac{\partial V(x,t)}{\partial t}$$

com condição inicial $V(x,0) = V(x)$, que pode ser obtida explicitamente se $V(x)$ for uma combinação linear de Gaussianas, fica cada vez mais regular à medida que o t aumenta. Portanto, $V(x,t)$ pode ser minimizada através de métodos locais quando t é suficientemente grande. Escolhe-se um mínimo local para um dado t como ponto inicial e, através da aplicação de métodos locais, é possível obter o mínimo para um outro valor de $t' < t$. Decrementando sucessivamente t e calculando uma sequência de mínimos locais de $V(x,t')$ para cada $t' < t$, obtém-se para $t = 0$ um mínimo para a função de energia potencial $V(x)$. A escolha do mínimo inicial e da sequência decrescente de t's determinará se o mínimo para $t = 0$ corresponde ou não o mínimo global de $V(x)$.

Infelizmente, para funções gerais, este é um processo pesado do ponto de vista computacional e, actualmente, estes métodos são principalmente úteis quando a função de energia potencial é uma soma de funções univariadas. A motivação para a aplicação destes métodos à predição da estrutura terciária das proteínas deve-se ao facto de as funções de energia potencial utilizadas serem bem aproximadas por somas deste tipo.

Métodos por partição e limite[3]

Tradicionalmente os métodos por partição e limite, introduzidos por Land e Doig em 1960 [LD60], são os métodos de eleição para resolver problemas de optimização global de natureza combinatória, e.g. [NRT89]. Estes são formulados normalmente como programas lineares inteiros. Dado que o problema da predição da estrutura terciária das proteínas tem aspectos combinatórios, estes métodos surgem também como uma abordagem ao problema. Nota-se no entanto que neste problema as variáveis são contínuas e não inteiras e, portanto, estes método não podem ser aplicados imediatamente à minimização da energia potencial.

Ainda que a aplicação destes métodos se resuma a pequenas proteínas, estes permitem calcular o mínimo global com segurança e encontrar também todos os mínimos e pontos de sela numa vizinhança do mínimo global, i.e., determinar todos os zeros do sistema não linear $\nabla V(x) = 0$ sujeito a determinadas restrições. Portanto, dado que estes métodos permitem não só obter o mínimo global mas também um conjunto de mínimos locais, o seu desenvolvimento permitirá não só estudar a conformação nativa correspondente ao mínimo global, mas também analisar as conformações temporárias que poderão existir durante o processo de enrolamento.

Alguns resultados da aplicação destes métodos ao problema da predição da estrutura terciária das proteínas podem ser consultados em [AMF95].

Aproximação em reticulados

Têm vindo também a ser propostos algoritmos de aproximação para os modelos sobre reticulados, em particular para o modelo *H-P* apresentado na secção 3.1.3. Neste caso, dada uma sequência $s \in \{0,1\}^n$, pretende-se determinar a configuração sobre o reticulado que maximiza a pontuação, i.e., o número de pares de 1's adjacentes no reticulado que não são adjacentes na sequência. Quanto maior for a pontuação menor será a energia do sis-

[3] Do inglês *branch and bound.*

tema, pois existem menos aminoácidos hidrófobos em contacto com o meio envolvente.

Considere-se o caso bidimensional e uma sequência $s \in \{0,1\}^n$, com $n \in \mathbb{N}$. Dada uma qualquer conformação de s sobre o reticulado, tem-se que dois 1's podem ser adjacentes se e só se existir um número par de elementos entre si, i.e., s_i e s_j são adjacentes se e só se $|i-j|$ é ímpar. Deste modo pode-se dividir o conjunto de 1's em s em duas classes de equivalência, a classe x e a classe y, de tal modo que a diferença entre os índices de dois elementos na mesma classe é par. Logo, um 1 da classe x só pode ser adjacente a 1 da classe y e, portanto, o número de contactos possíveis ou a pontuação máxima é

$$P_{max}(s) = \min(X + Y) + T,$$

em que X é número de elementos na classe x, Y é o número de elementos na classe y e T é o número de 1's nas extremidades de s. Em termos de energia, dir-se-á que o valor óptimo é $E_{OPT}(s) = -P_{max}(s)$. Observa-se que no caso tridimensional pode-se definir um limite análogo.

Este é então um problema de minimização de energia e, dado um algoritmo de aproximação $\mathcal{A}$, é necessário determinar o desempenho do mesmo. O *factor de aproximação absoluto* α é dado por [GJ79]

$$\alpha = \sup\{r \leq 1 | \forall_{s \in \{0,1\}^n, n \in \mathbb{N}} \frac{E_{\mathcal{A}}(s)}{E_{OPT}(s)} \geq r\},$$

onde $E_{\mathcal{A}}(s) = -P(s)$ e $P(s)$ é a pontuação conseguida com o algoritmo de aproximação $\mathcal{A}$. Uma vez que $E_{OPT} \leq E_{\mathcal{A}}(s) \leq 0$, tem-se que $0 \leq \alpha \leq 1$. Portanto, à medida que α se aproxima de 1, melhor é o desempenho do algoritmo $\mathcal{A}$.

O primeiro algoritmo de aproximação foi proposto por Hart e Istrail em 1996 [HI96], com factores de aproximação de $\frac{1}{4}$ e de $\frac{3}{8}$ para os modelos *H-P* bidimensional e tridimensional, respectivamente. Em 2002, Newman [New02] obteve um algoritmo de aproximação para este modelo no caso

bidimensional com um factor de aproximação de $\frac{1}{3}$. Newman mostrou ainda que, no caso bidimensional e para qualquer algoritmo de aproximação, o factor de aproximação é limitado por $\frac{1}{2}$.

Classificadores

As abordagens que recorrem a técnicas de classificação baseadas na inferência de conhecimento a partir de bases de dados, e.g. [BWF+00], e na utilização de classificadores como as redes neuronais e as máquinas de vectores de suporte têm ganho popularidade [PB02, ZK03, BPFV03]. Nota-se que alguns dos métodos anteriores, tal como foi referido, recorrem inúmeras vezes ao conhecimento empírico disponível e, portanto, os métodos de classificação desempenham aí um papel bastante importante.

Actualmente, a utilização das técnicas de classificação no âmbito da predição da estrutura terciária das proteínas incide em especial sobre a predição dos mapas de contacto. Na próxima secção ir-se-á analisar esta representação dos contactos entre os aminoácidos de uma proteína e de que forma se pode recuperar a estrutura terciária desta a partir do seu mapa de contactos. Os resultados obtidos para a predição dos mapas de contacto são ainda pouco exactos. Os melhores resultados, cerca de 20% de exactidão, foram conseguidos com redes neuronais [BPFV03] e com máquinas de vectores de suporte [ZK03].

É importante notar que, para a predição da estrutura terciária das proteínas ser possível a partir do seu mapa de contactos, os testes realizados até ao momento parecem indicar que é suficiente a predição de 50% dos contactos. Estes resultados foram apresentados informalmente durante o encontro CASP4 [LLH01].

3.4 Mapas de contacto

Na literatura física representa-se frequentemente a energia potencial da conformação espacial de uma proteína através dos potenciais de contacto entre

pares de aminoácidos [VND99]. Seja $S = \{s_1, s_2, ..., s_n\}$, com $n \in \mathbb{N}$, a sequência de aminoácidos de uma dada proteína. A energia potencial em função dos potenciais de contacto entre pares de aminoácidos representa-se através da expressão

$$\sum_{i<j}^{n} M_{ij} w(s_i, s_j),$$

em que $w(s_i, s_j)$ é o potencial de contacto entre s_i e s_j e o *mapa de contacto* M é uma matriz $n \times n$ cujas entradas são 0 ou 1: $S_{i,j} = 1$ se os aminoácidos s_i e s_j estão em *contacto*, $S_{i,j} = 0$ caso contrário. Usualmente, dado que se identifica a conformação espacial pela cadeia peptídica, dois aminoácidos s_i e s_j estão em contacto quando a distância entre os carbonos centrais ou carbonos α é inferior a um dado limite R_c. Existem outras definições alternativas de contacto. É comum dizer-se que dois aminoácidos também estão em contacto quando a distância entre dois átomos, um de cada aminoácido, é inferior a um dado valor limite ou quando a distância entre dois átomos pesados[4], um de cada aminoácido, é inferior a um dado valor limite. O valor limite R_c pode variar entre 6 e 12Å.

Dado que existem 20 aminoácidos, é necessário definir 210 potenciais de contacto diferentes. A escolha destes pode basear-se apenas na polaridade tal como acontece no modelo *H-P* estudado anteriormente, pode recorrer a outras propriedades fisico-químicas de cada aminoácido ou combinações das mesmas e pode também utilizar dados empíricos, nomeadamente utilizando informação extraída das bases de dados PDB [BWF$^+$00] e AAindex [KK99]. Por outro lado, é comum utilizarem-se processos de análise e aproximação, alguns desses discutidos na secção anterior, na determinação dos valores dos potenciais de contacto.

Uma vez que os mapas de contacto são independentes do referencial de coordenadas utilizado, estes são convenientes para a comparação estrutural das proteínas e para a procura de estruturas similares em bases de dados.

[4]Num aminoácido os átomos de `C`, `N` e `O` dizem-se pesados e os átomos de `H` dizem-se leves.

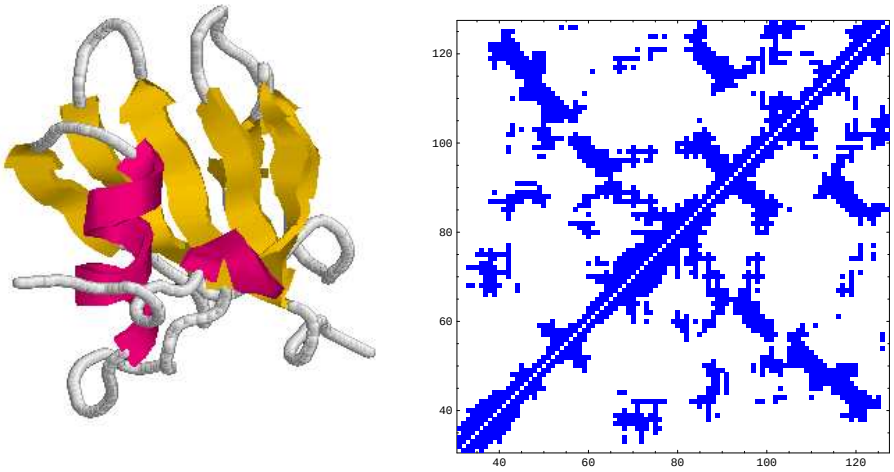

Figura 3.3: Uma proteína e o respectivo mapa de contacto, neste caso dois aminoácidos estão em contacto se existem dois átomos pesados, um de cada aminoácido, a uma distância inferior a 8Å.

Um objectivo mais aliciante consiste em utilizar a representação estrutural sob a forma de mapa de contacto de uma dada proteína para procurar no espaço dos mapas de contacto possíveis para a mesma aquele que corresponde à conformação nativa [VND99]. Do ponto de vista computacional, a vantagem em utilizar esta abordagem reside no facto de que é mais eficiente a manipulação da estrutura através do mapa de contacto. Em particular, pequenas alterações no mapa de contacto de uma proteína conduzem a alterações significativas na conformação espacial da cadeia peptídica.

No entanto existem algumas questões relativas à coerência das alterações introduzidas nos mapas de contacto, e, nomeadamente, é importante determinar se o mapa obtido é fisicamente realizável, i.e., se corresponde a uma conformação possível da proteína. Nota-se que alterações aleatórias num mapa de contacto conduzem com grande probabilidade a mapas não realizáveis fisicamente dado que, para uma proteína com N aminoácidos, o número de mapas de contacto possíveis é $O(2^{N^2})$, enquanto que o número de mapas realizáveis fisicamente é $O(2^N)$ [VD00]. A utilização de mapas de contacto na procura da conformação de menor energia para uma dada

proteína é uma técnica relativamente recente [VD00].

As estruturas secundárias de uma proteína são facilmente detectadas através do seu mapa de contacto. Tal como se pode ver na figura 3.3, as hélices-alpha detectam-se através de pequenas bandas ao longo da diagonal principal, o que faz sentido dado que as hélices-alpha envolvem o contacto de um dado aminoácido com os seus vizinhos na cadeia peptídica. Por sua vez, as folhas-beta revelam-se através de pequenas bandas paralelas ou perpendiculares à diagonal principal.

A propriedade fundamental dos mapas de contacto é a possibilidade de permitir a reconstrução da estrutura terciária das proteínas. O processo de predição pode ser dividido em três passos fundamentais [BPFV03]. O primeiro passo consiste em analisar a estrutura primária, i.e., a sequência de aminoácidos, possivelmente através do alinhamento de sequências, por forma a determinar algumas propriedades estruturais da sequência, por exemplo, identificar as estruturas secundárias. No segundo passo, o mais difícil, pretende-se determinar o mapa de contacto a partir da sequência de aminoácidos. No terceiro e último passo recuperam-se as coordenadas espaciais através do mapa de contacto e das estruturas identificadas no primeiro passo.

As técnicas a utilizar no primeiro passo são bem conhecidas (ver por exemplo [PBFC01, PPRB01]) e os processos para realizar o terceiro passo estão também documentados na literatura [VKD97, Ven03]. O segundo passo constitui de facto o desafio central do problema; actualmente existem alguns trabalhos neste sentido, e.g., [BPFV03, ZK03]. No entanto os resultados têm sido insatisfatórios quando é necessária no mínimo uma precisão de 50% [LLH01]. No último capítulo desta dissertação ir-se-á estudar a aplicação das máquinas de vectores de suporte à predição de mapas de contacto, sendo em particular discutida a abordagem tomada em [ZK03].

Capítulo 4

Predição com SVM's

No último capítulo apresentaram-se os mapas de contacto como uma possível representação da estrutura terciária das proteínas. Verificou-se que a predição dos mesmos a partir da estrutura primária das proteínas constitui um dos passos mais importantes e um dos maiores desafios. Alguns dos métodos mais utilizados na predição dos mapas de contacto consistem de facto em técnicas de classificação, nomeadamente através da utilização de redes neuronais ou de *máquinas de vectores de suporte* (SVM's). Face à utilização de outras técnicas, estas parecem ser as mais promissoras ainda que, actualmente, os resultados continuem a ser pouco expressivos.

Neste capítulo estudar-se-á a aplicação das SVM's na predição de mapas de contacto. Em primeiro lugar ir-se-ão introduzir as SVM's. De seguida será apresentado em detalhe o processo de aquisição de informação e de preparação dos conjuntos de treino, serão apresentados os modelos obtidos através da utilização de SVM's e discutir-se-ão os resultados de classificação conseguidos.

4.1 Máquinas de vectores de suporte

Um dos problemas fundamentais da teoria da aprendizagem consiste na classificação de objectos em classes. Suponha-se que são dadas duas classes de objectos devidamente rotulados. Dado um novo objecto pretende-se determinar a classe a que pertence. Normalmente designa-se este caso por *reco-*

nhecimento de padrões ou *classificação (binária)* e formaliza-se da seguinte forma, dado um conjunto de dados empíricos, i.e., o *conjunto de treino*,

$$(x_1, y_1), ..., (x_m, y_m) \in \mathcal{X} \times \{\pm 1\}, \tag{4.1}$$

onde $\mathcal{X}$ é um conjunto não vazio de *padrões* x_i (ou *instâncias*, ou *inputs*) designado por *domínio* e $y_i \in \{\pm 1\}$ são as *etiquetas* (ou *outputs*). Dividem-se os padrões $\{x_1, ..., x_m\}$ pelas respectivas classes através da atribuição de etiquetas. Nota-se que neste caso existem apenas duas classes de padrões, os que são etiquetados com -1 e os que são etiquetados com $+1$.

No entanto, o objectivo principal é classificar padrões novos. Veja-se o exemplo anterior: dado um novo padrão $x \in \mathcal{X}$ acerca do qual não existe informação empírica pretende-se prever a etiqueta $y \in \{\pm 1\}$ correspondente, i.e., escolher um $y \in \{\pm 1\}$ de forma a que (x, y) seja *similar* aos exemplos de treino 4.1. Embora a caracterização de similaridade seja simples no caso das etiquetas, pois duas etiquetas podem ser ou iguais ou diferentes, o mesmo não se passa com os padrões sendo esta a questão central da teoria da aprendizagem. Considere-se uma medida de similaridade da forma

$$\begin{aligned} k : \mathcal{X} \times \mathcal{X} &\longrightarrow \mathbb{R} \\ (x, x') &\longmapsto k(x, x'), \end{aligned}$$

ou seja, uma função que, dados dois padrões x e x', retorna um número real que caracteriza a sua similaridade. Se k for simétrica, i.e., $k(x, x') = k(x', x)$, diz-se que k é um *kernel*. Medidas de similaridade desta forma são difíceis de estudar. Deste modo ver-se-á em primeiro lugar um tipo simples de medida de similaridade que será objecto de generalização.

O *produto interno* usual em $\mathbb{R}^n$ é um exemplo de uma medida de similaridade simples. Este permite lidar com todas as construções geométricas que podem ser formuladas em termos de ângulos, comprimentos e distâncias, mas não é suficientemente geral para lidar com muitos problemas interessantes. De facto os padrões podem ser qualquer tipo de objecto e não têm que necessariamente pertencer a um espaço com produto interno. Nestes

casos, é comum utilizar-se um mapa $\Phi : \mathcal{X} \longrightarrow \mathcal{H}$ que represente os padrões em $\mathcal{X}$ como vectores num espaço com produto interno $\mathcal{H}$. E, mesmo que os padrões originais existam num espaço com produto interno, em determinados problemas utiliza-se um mapa para mudar a representação para uma mais adequada (Φ pode ser não linear). O espaço $\mathcal{H}$ obtido designa-se por *espaço principal*.

Deste modo, as vantagens do mapeamento dos dados num espaço principal são: pode-se definir uma medidade similaridade com base no produto interno; permite lidar com os dados de forma geomética e desta forma estudar algoritmos de aprendizagem baseados em álgebra linear e geometria analítica; a liberdade na escolha de Φ permite definir e estudar um maior número de medidas de similaridade e de algoritmos.

4.1.1 Teoria da aprendizagem estatística

No reconhecimento de padrões com duas classes, procura-se inferir uma função $f : \mathcal{X} \longrightarrow \{\pm 1\}$ a partir do conjunto de treino. Suponha-se que os dados são gerados independentemente a partir de uma distribuição probabilística desconhecida, mas fixa, $P(x, y)$. O objectivo é descobrir uma função f que classifique correctamente os exemplos desconhecidos e que valide o cojunto de treino, i.e., para as amostras (x, y), que também são geradas a partir de $P(x, y)$, $f(x) = y$. A correcção da classificação é medida em termos da *função de perda zero-um c* definida por

$$c(x, y, f(x)) = \frac{1}{2}|f(x) - y|.$$

A perda é zero se (x, y) é classificado correctamente, um caso contrário. No entanto, como apenas se têm os dados de treino, não existe forma de seleccionar entre duas funções qual é a preferível. Conclui-se que, minimizando a média do *erro de treino* (*risco empírico*)

$$R_{\text{emp}}[f] = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2}|f(x_i) - y_i|,$$

não existe garantia de um *erro de teste* pequeno sobre os exemplos de teste obtidos a partir da distribuição $P(x, y)$. O erro de teste designa-se por *risco* e pode ser definido para qualquer função de perda desde que o seguinte integral exista:

$$R[f] = \int \frac{1}{2}|f(x) - y| \; dP(x, y).$$

Neste caso em que se utiliza a *função de perda zero-um*, o *risco* é igual à probabilidade de uma instância x ser mal classificada.

A teoria da aprendizagem estatística, também conhecida por teoria VC (Vapnik-Chervonenkis), mostra que é necessário restringir o conjunto de funções a partir do qual f pode ser escolhida. Nota-se que o conjunto obtido deve ter uma *capacidade* adequada para o tamanho do conjunto de treino em causa. A teoria VC fornece limites para os erros do teste e a minimização desses limites, que depende não só do risco empírico mas também da capacidade da classe de funções, e leva ao princípio da *minimização do risco estrutural*.

Existem várias medidas de capacidade. No entanto a mais conhecida é a *dimensão de VC*. Cada função da classe separa os padrões de uma determinada forma e logo induz uma determinada etiquetação para os padrões. Como as etiquetas estão em $\{\pm 1\}$, existem 2^m diferentes formas de etiquetar m padrões. Uma classe de funções muito rica poderá realizar todas as separações 2^m e, nesse caso, diz-se *estilhaçar* os m pontos. É claro que nem todas as classes de funções conseguem separar os m pontos. A *dimensão de VC* para uma classe de funções é definida como o maior m tal que existe um conjunto de m pontos que a classe consegue estilhaçar e ∞ se este não existir. Para que seja possível obter predições não triviais, é importante que a classe de funções seja *restrita*, i.e., que a sua capacidade seja suficientemente pequena em relação ao conjunto de treino de forma a que um erro de treino pequeno garanta um erro de teste pequeno. Ao mesmo tempo, a capacidade deve ser grande o suficiente para fornecer funções que estejam aptas a modelar as dependências escondidas em $P(x, y)$. A escolha do conjunto de

funções é desta forma crucial para aprender a partir dos dados.

4.1.2 Classificação com hiperplanos

Descrever-se-á agora um algoritmo de classificação baseado em hiperplanos e que pode ser aplicado quando o conjunto de dados está representado num espaço com produto interno. Considere-se a classe dos hiperplanos num dado espaço com produto interno $\mathcal{H}$ e a função de decisão f para cada $w \in \mathcal{H}$ e $b \in \mathbb{R}$ dada por

$$f(x) = \text{sgn}(\langle w, x \rangle + b).$$

Nestas condições Vapnik propôs um algoritmo de aprendizagem para problemas que são separáveis por hiperplanos, i.e., *linearmente separáveis*. Este algoritmo baseia-se em duas propriedades. A primeira propriedade é a existência de um único hiperplano óptimo que corresponde à margem de separação maximal entre qualquer ponto do conjunto de treino e o hiperplano, o qual é dado pela solução de

$$\max_{w \in \mathcal{H}, b \in \mathbb{R}} \min\{\|x - x_i\| \mid x \in \mathcal{H}, \langle w, x \rangle + b = 0,\ i = 1, ..., m\}.$$

A segunda propriedade está relacionada com o facto de a capacidade da classe de hiperplanos separadores diminuir com o aumento da margem.

Para construir o hiperplano óptimo tem que se resolver a equação

$$\min_{w \in \mathcal{H}, b \in \mathbb{R}} \tau(w) = \frac{1}{2}\|w\|^2, \tag{4.2}$$

com as seguintes condições

$$y_i(\langle w, x_i \rangle + b) \geq 1, \tag{4.3}$$

para todo o $i = 1, ..., m$.

Nota-se que, se $\|w\| = 1$, então $y_i(\langle w, x_i \rangle + b) \geq 1$ será igual à distância de x_i ao hiperplano. De uma forma geral, tem que se dividir 4.3 por $\|w\|$ para se obter essa distância. Deste modo, se a equação 4.2 for satisfeita

para todo o $i = 1, ..., m$ com um $\|w\|$ de comprimento mínimo, a margem de separação será optimizada.

A função τ é designada por *função objectivo*, enquanto que as inequações 4.3 são designadas por *inequações de restrição*. Conjuntamente formam o *problema de optimização restrito*. Este tipo de problemas são resolvidos introduzindo os *multiplicadores de Lagrange* $\alpha_i \geq 0$ e o *Lagrangeano* L dado por

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i (y_i(\langle w, x_i \rangle + b) - 1),$$

em que $\alpha = (\alpha_1, ..., \alpha_m)$. É portanto necessário minimizar L em relação às *variáveis principais* w e b e maximizar relativamente às *variáveis duais* α_i, o que significa, terá que ser encontrado um *ponto de sela*.

Nota-se que se uma inequação de restrição não for verificada, então $y_i(\langle w, x_i \rangle + b) - 1) < 0$ e neste caso L poderá ser incrementada através do incremento do α_i correspondente. Por outro lado, para que L seja decrementada, w e b terão que ser alterados. De forma a que $\alpha_i(y_i(\langle w, x_i \rangle + b) - 1) < 0$ não seja um número negativo demasiado grande, a alteração de w e de b vai garantir que, desde que o problema seja separável, a restrição seja eventualmente satisfeita. Para o caso das restrições da forma $y_i(\langle w, x_i \rangle + b) - 1 > 0$, o α_i correspondente terá que ser zero visto este ser o valor que maximiza L.

O facto de que, no ponto de sela, as derivadas de L em relação às variáveis principais terem de ser zero faz com que

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \text{ e } w = \sum_{i=1}^{m} \alpha_i y_i x_i.$$

Deste modo, o vector solução depende de um subconjunto dos padrões de treino, nomeadamente os padrões com α_i diferentes de zero, que se designam por *vectores de suporte*. Conclui-se assim que o hiperplano é completamente determinado pelos padrões mais próximos dele.

Na prática o problema que se resolve, designado por problema de *opti-*

mização dual, é a maximização de $W(\alpha)$, em que $\alpha \in \mathbb{R}^m$ e

$$W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle,$$

atendendo a que $\alpha_i \geq 0$, para todo o $i = 1, ..., m$, e que $\sum_{i=1}^{m} \alpha_i y_i = 0$. Portanto, a função de decisão do hiperplano é dada por

$$f(x) = \text{sgn} \left(\sum_{i=1}^{m} y_i \alpha_i \langle x, x_i \rangle + b \right),$$

em que b é obtido através da exploração das condições

$$\alpha_i [y_i (\langle x_i, w \rangle + b) - 1] = 0,$$

para todo o $i = 1, ..., m$.

4.1.3 Classificação com vectores de suporte

As *máquinas de vectores de suporte* (*SVM*s) constituem um método de aprendizagem supervisionada introduzido por Vapnik [Vap95]. Estas são normalmente utilizadas quando as instâncias pertencem a um espaço $\mathcal{X}$ em que não são separáveis linearmente ou quando o espaço $\mathcal{X}$ não possui produto interno. Nota-se que até agora tudo foi formalizado num espaço $\mathcal{H}$ com produto interno.

Para expressar as fórmulas anteriores em termos de padrões num espaço arbitrário $\mathcal{X}$, é necessário aplicar um *kernel* k, i.e.,

$$\begin{aligned} k : \mathcal{X} \times \mathcal{X} &\longrightarrow \mathcal{H} \\ (x, x') &\longmapsto \langle \Phi(x), \Phi(x') \rangle. \end{aligned}$$

k mapeia os objectos de $\mathcal{X}$ num espaço $\mathcal{H}$ com produto interno através de Φ e utiliza como medida de similaridade o produto interno em $\mathcal{H}$. Obtém-se assim a função de decisão f dada por

$$f(x) = \text{sgn} \left(\sum_{i=1}^{m} y_i \alpha_i \langle \Phi(x), \Phi(x_i) \rangle + b \rangle \right),$$

e o programa quadrático

$$\text{maximizar } W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j k(x_i, x_j), \tag{4.4}$$

em que $\alpha \in \mathbb{R}^m$, com as condições $\alpha_i \geq 0$, para todo o $i = 1, ..., m$, e $\sum_{i=1}^{m} \alpha_i y_i = 0$.

Na prática, um hiperplano separador pode não existir. Para permitir a possibilidade de existirem exemplos que não verifiquem 4.3, são introduzidas variáveis de diferença

$$\xi_i \geq 0, \tag{4.5}$$

para todo o $i = 1, ..., m$, de forma a enfraquecer as restrições 4.3. Obtêm-se assim as restrições

$$y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i, \tag{4.6}$$

para todo o $i = 1, ...m$.

Logo, um classificador em que a capacidade e a soma dos ξ_i com $1 \leq i \leq m$ são controladas, generaliza bem o problema. Uma realização possível de um tal *classificador marginal leve* é obtida minimizando a função objectivo

$$\tau(\mathbf{w}, \xi) = \frac{1}{2}\|w\| + C \sum_{i=1}^{m} \xi_i,$$

com $C > 0$ e em relação às restrições 4.5 e 4.6. Reescrevendo em termos dos multiplicadores de Lagrange, fica-se com o problema de maximizar 4.4 em relação às restrições $\alpha_i \geq 0$, para todo o $i = 1, ..., m$, e $\sum_{i=1}^{m} \alpha_i y_i = 0$. A constante C determina a relação entre a maximização marginal e a minimização do erro de treino. Pode-se calcular b explorando o facto de que, para todos os vectores de suporte x_i com $\alpha_i \leq C$, $\xi_i = 0$ e logo

$$\sum_{i=1}^{m} \alpha_j y_j k(x_i, x_j) + b = y_i.$$

Outra possível realização de uma variante marginal leve do hiperplano óptimo usa ν-parametrização. Neste caso, em vez de um parâmetro C,

tem-se um parâmetro $\nu \in]\,0,1\,]$ que dá um limite inferior para os vectores de suporte e um limite superior para os padrões de treino com variáveis de diferença que não sejam 0. O erro da função objectivo principal é expressável por $(\frac{1}{\nu m}\sum_i \xi_i) - \rho$ e as restrições de separação são

$$y_i(\langle w, x_i\rangle + b) \geq \rho - \xi_i,$$

para todo o $i = 1, ...m$. O dual consiste na maximização da parte quadrática de 4.4 em relação a $0 \leq \alpha_i \leq \frac{1}{\nu m}$, $\sum_i \alpha_i y_i = 0$ e à restrição adicional $\sum_i \alpha_i = 1$.

Para mais detalhes acerca das SVM's ver, por exemplo, [CS00, SS02]. Na continuação deste trabalho utilizar-se-á uma implementação das SVM's por Joachims [Joa99], $\texttt{SVM}^{light}$.

4.2 Predição de mapas de contacto

A predição dos mapas de contacto constitui um dos passos fulcrais no processo de predição da estrutura terciária das proteínas, tal como se viu na secção 3.4. Recentemente, Zhao e Karypis [ZK03] utilizaram as SVM's para predizer mapas de contacto a partir da estrutura primária das proteínas. Os resultados alcançados foram os melhores até agora conseguidos, com uma exactidão média de 22.4%. No entanto, estes resultados ficam longe dos 50% de exactidão mínima necessários para garantir a predição da estrutura terciária [LLH01].

Nesta secção analizar-se-á o trabalho de Zhao e Karypis [ZK03]. Simultanemente serão feitos alguns testes alternativos e, em especial, serão dadas algumas sugestões para melhorar os resultados. Nota-se que, infelizmente, os resultados dos testes realizados não reproduzem a exactidão conseguida por Zhao e Karypis. Este facto deve-se a factores que ficarão claros no decurso deste trabalho.

4.2.1 Formulação do problema

A predição de mapas de contacto pode ser tratada como um problema de classificação. Considere-se o conjunto das proteínas: todos os pares de aminoácidos em cada proteína podem ser classificados como estando ou não em contacto, i.e., pode-se separar os pares de aminoácidos em instâncias positivas e negativas.

Considere-se a sequência de aminoácidos $\{a_1, a_2, ..., a_n\}$ de uma dada proteína. No seguimento assumir-se-á que dois aminoácidos a_i e a_j estão em contacto se a distância entre dois átomos pesados, um de cada aminoácido, for inferior a 8Å. Nota-se também que o trabalho aqui apresentado foca os contactos não locais, i.e., apenas serão considerados pares de aminoácidos a_i e a_j tais que a distância entre si na sequência é maior do 6, i.e., $|i - j| \geq 6$.

Para cada par de aminoácidos a_i e a_j de uma determinada proteína é criada uma instância ou padrão com várias características, e.g., propriedades físico-químicas do ambiente, informação acerca da estrutura secundária e correlações evolucionárias. Neste caso concreto, o conjunto de padrões $\mathcal{X}$

Índice	Descrição	Domínio
1	posição na sequência do aminoácido a_i	$\mathbb{N}$
2	posição na sequência do aminoácido a_j	$\mathbb{N}$
3	estrutura secundária	$\{0, 1\}$
...		
20	estrutura secundária	$\{0, 1\}$
21	separação da sequência	$\mathbb{N}$
22	conservação da sequência	$\mathbb{R}$
23	conservação da sequência	$\mathbb{R}$
24	CMA	$\mathbb{R}$
...		
63	CMA	$\mathbb{R}$
64	perfil da sequência	$\mathbb{R}$
...		
135	perfil da sequência	$\mathbb{R}$
136	perfil da sequência	$\mathbb{N}$
...		
2135	perfil da sequência	$\mathbb{N}$

Tabela 4.1: Descrição de cada componente dos padrões.

possui todas as instâncias possíveis descritas em detalhe na tabela 4.1. As etiquetas $y \in \{\pm 1\}$ são atribuídas a cada par de aminoácidos da forma esperada, i.e., $+1$ caso os aminoácidos estejam em contacto, -1 caso contrário.

Tal como se discutiu na secção 4.1, em primeiro lugar é necessário um conjunto de vectores de treino. Deste modo, seleccionar-se-á um conjunto de proteínas cuja estrutura é conhecida, determinar-se-ão as instâncias para cada par de aminoácidos em cada proteína, i.e., os vectores de treino, e as etiquetas para cada um dos vectores de treino. A partir deste conjunto de vectores de treino produzir-se-á um classificador que permitirá classificar instâncias cujas etiquetas são desconhecidas, i.e., predizer o mapa de contacto para proteínas desconhecidas.

4.2.2 Extracção das características

O conjunto de características a ter em conta é vasto. É necessário seleccionar um subconjunto destas de acordo com a sua relevância, evitando a selecção de características que forneçam informação redundante. Na selecção consideram-se normalmente características que forneçam informação acerca da estrutura primária, da estrutura secundária, das propriedades químicas e físicas, e dos dados biológicos e evolucionários. Portanto, a selecção de características é de facto um problema complexo, sendo necessário extremo cuidado para que o conjunto de informação seleccionado contenha todos os aspectos relevantes e seja tratável em tempo útil.

Descreve-se agora o método utilizado para extrair as características para cada par de aminoácidos. As características terão assim por base as estruturas primárias das proteínas, i.e., as sequências de aminoácidos, o alinhamento múltiplo destas, as estruturas secundárias e as propriedades físico-químicas dos aminoácidos.

A análise de mutações correlacionadas (CMA[1]) tem vindo a ser proposta para a predição de contactos não locais, e.g. [ORV99, PTG99]. Na evolução

[1]Do inglês *Correlated Mutations Analysis*.

biológica, observa-se que os pares de aminoácidos distantes na estrutura primária e próximos na estrutura terciária parecem ser conservados ou sofrer mutações de forma correlacionada, i.e., as frequências de ocorrência de um dado aminoácido numa determinada posição dependem de um ou mais aminoácidos em outras posições da sequência. Deste modo, espera-se que posições com forte correlação correspondam a posições próximas na estrutura terciária.

Na expressão 4.7, para o cálculo dos valores de correlação, em vez de se considerarem apenas algumas propriedades físico-químicas tais como o volume ou a hidrofobia, consideram-se as 10 primeiras componentes resultantes da análise de componentes principais sobre os 494 índices da base de dados AAindex1 [KK99, TK96, NKK88]. Nesta base de dados, um índice para aminoácidos representa uma das propriedades biológicas ou físico-químicas e consiste num conjunto de 20 valores, i.e., um valor para cada aminoácido. Na extracção das componentes principais apenas 19 componentes se revelam importantes, uma vez que todas as outras têm valores próximos de 0. Para a análise de componentes principais foi utilizada a ferramenta `Mathematica5.0`. Nota-se que, no trabalho desenvolvido em [ZK03], apenas foram considerados 142 índices da base de dados AAindex1.

Dado um alinhamento múltiplo de sequências para uma proteína, para cada par de aminoácidos i e j da proteína calcula-se a seguinte correlação

$$r_{ij} = \frac{1}{{N_{MSA}}^2} \sum_{l=1}^{N_{MSA}} \frac{(q_i^l - m_i)(q_j^l - m_j)}{s_i s_j}, \tag{4.7}$$

em que N_{MSA} é o número de sequências no alinhamento multiplo, q_i^l e q_j^l são os valores de uma das 10 componentes principais indicadas acima para a sequência l nas posição i e j, respectivamente, e m_i, m_j, s_i, s_j são a média e o desvio padrão da componente principal em causa para cada posição.

A partir do alinhamento múltiplo de sequências calcula-se ainda a conservação para cada posição da sequência. Este cálculo consiste em determinar o valor da entropia dos aminoácidos que aparecem numa dada posição

no alinhamento múltiplo de sequências como se segue:

$$cons(i) = -\sum_{k=1}^{20} p(a_k|i) \log(p(a_k|i)) \tag{4.8}$$

em que a_k é um dos 20 aminoácidos e $p(a_k|i)$ é o quociente entre o número de sequências que contêm a_k na posição i e o número de sequências no alinhamento múltiplo.

Tendo em conta a análise das mutações correlacionadas e o cálculo da conservação indicados acima, descreve-se agora as diferentes características recolhidas para cada par de aminoácidos i e j numa dada proteína.

Conservação da sequência: dado o alinhamento múltiplo de sequências, calculam-se os valores de conservação de acordo com a expressão 4.8 para i e j.

Separação da sequência: calcula-se a distância entre os dois aminoácidos na sequência, i.e., $|i - j|$.

CMA: calcula-se em primeiro lugar o valor da expressão 4.7 para i, j e cada uma das 10 primeiras componentes principais referidas acima; atendendo aos 6 agrupamentos de propriedades disponíveis na base de dados AAindex1, seleccionam-se as 5 primeiras componentes principais em cada um dos grupos e calcula-se o valor da expressão 4.7 para i, j e cada uma das componentes principais em cada grupo.

Estruturas secundárias: para cada aminoácido utilizam-se 3 valores para indicar se o aminoácido em causa faz parte de uma hélice-alpha, de uma folha-beta ou de um enrolamento arbitrário não tipificado; se um aminoácido pertence a uma das estruturas secundárias o respectivo parâmetro toma valor 1, caso contrário toma valor 0; calculam-se estes valores para as posições $i-1$, i, $i+1$, $j-1$, j e $j+1$, i.e., calculam-se 18 valores para cada par de aminoácidos i e j.

Perfis da sequência: para as posições i e j no alinhamento múltiplo de sequências, calculam-se as frequências de ocorrência de todos os 210 pares de aminoácidos possíveis; para além do par (i, j), calculam-se as frequências para os pares $(i-1, j-1)$, $(i+1, j+1)$, $(i-1, j+1)$ e $(i+1, j-1)$; calculam-se ainda para todas as posições envolvidas nestes pares a média de cada uma das 12 primeiras componentes principais referidas acima.

A utilização dos perfis das sequências tem introduzido melhorias na predição dos mapas de contacto [FOVC01].

O conjunto de treino foi obtido a partir de 100 proteínas seleccionadas da base de dados PDB [BWF$^+$00], procedendo-se ao alinhamento múltiplo das sequências e à extracção das características acima indicadas. O alinhamento múltiplo de sequências foi obtido com o ClustalW [THG94]. Dado que o conjunto de padrões obtidos continha um grande número de casos negativos, i.e., pares que não estavam em contacto, realizou-se uma amostragem de forma a que o conjunto de treino ficasse com o mesmo número de padrões negativos e positivos. Importa referir que no trabalho desenvolvido em [ZK03] foram consideradas no alinhamento múltiplo sequências homologas extraídas através do PSI-BLAST para cada uma das proteínas. Nesse caso, foi ainda utilizado um método preditivo para determinar as estruturas secundárias das sequências homologas. No presente caso utilizaram-se apenas as sequências das proteínas seleccionadas da base de dados PDB e, portanto, as estruturas secundárias são conhecidas à priori. Nota-se que, face aos recursos computacionais disponíveis, os vectores de treino apenas incluem os pares (i, j) tais que $6 \leq |i - j| \leq 30$ e que, no decurso desta dissertação, foi desenvolvida uma aplicação em `C` de forma a permitir a extracção automática dos vectores de treino a partir dos ficheiros `pdb` obtidos da base de dados PDB [BWF$^+$00].

4.2.3 Classificador com kernel linear

O primeiro classificador foi obtido a partir do conjunto de treino atrás especificado com um *kernel* linear. Dado que $\mathcal{X} \subseteq \mathbb{R}^{2135}$, utilizou-se o produto interno usual como medida de similaridade. Os parâmetros de optimização da $\texttt{SVM}^{light}$ assumiram os valores por omissão.

Os resultados obtidos com o classificador linear estão resumidos na tabela 4.2 e na figura 4.1. Nota-se que as instâncias associadas às proteínas `2HGF` e `4JDW` faziam parte do conjunto de treino e que apenas as instâncias da proteína `5RXN` são exemplos de teste, estas representam os resultados conseguidos para todos os testes realizados. Embora se tenha consigo predizer

pdbID	**#Amin.**	**#Cont.**	**#Contactos preditos**		
			Total	**Correctos**	**Precisão**
`2HGF`	127	2194	2528	1002	45.7%
`4JDW`	423	9566	9532	4024	42.2%
`5RXN`	54	1010	1224	652	53.3%

Tabela 4.2: Resultados da classificação com *kernel* linear.

em média cerca de 50% dos contactos, a precisão é muito baixa. A figura 4.1 testemunha este facto, os contactos preditos correspondem aos contactos sobre a diagonal e, portanto, não acrescentam conhecimento relevante. De facto, nem os mapas de contacto das proteínas utilizadas na obtenção do conjunto de treino se conseguem recuperar, e.g., proteínas `2HGF` e `4JDW`.

Deste modo, os resultados aqui alcançados são qualitativamente inferiores aos conseguidos em [ZK03], onde se consegue uma predição média de 22.4% com uma precisão mais elevada, i.e., a precentagem de contactos incorrectos é menor. Nesse trabalho as proteínas foram divididas em classes de acordo com a estrutura secundária e o número de padrões de treino era bastante superior, estes dois factores contribuem para um aumento na especificidade do classificador numa dada classe de proteínas e para a universalidade do classificador, respectivamente.

Uma das correcções a fazer neste trabalho consiste em, seguindo a ideia

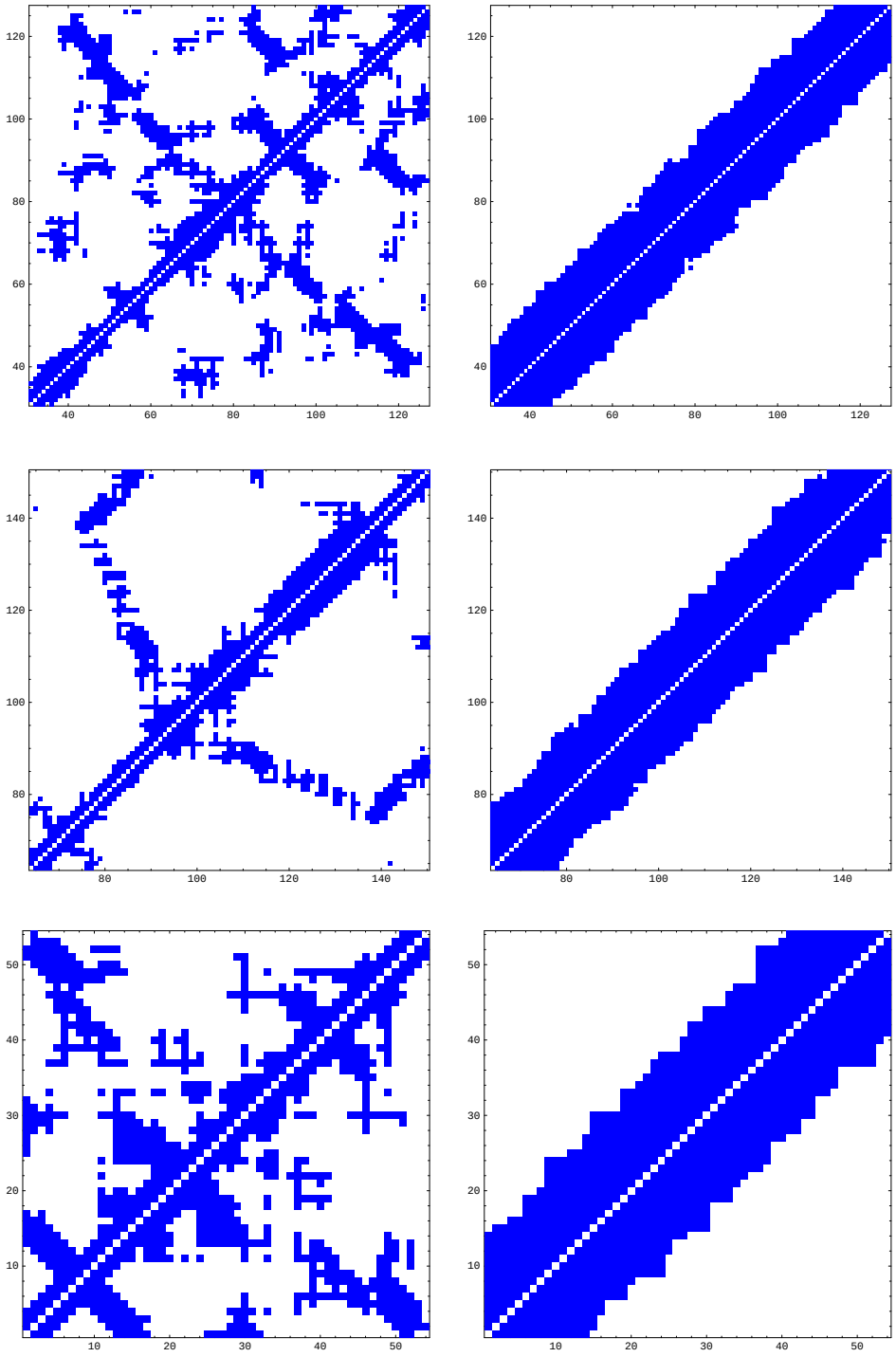

Figura 4.1: Resultados da classificação obtidos neste trabalho com *kernel* linear (mapa original e mapa obtido) para as proteínas `2HGF`, `4JDW` e `5RXN`, respectivamente.

apresentada em [ZK03], considerar subconjuntos de características de forma a verificar-se quais têm um papel relevante na predição dos mapas de contacto. Em particular determinar se, para dadas classes de proteínas, algumas características são mais relevantes do que outras.

4.2.4 Classificador com kernel RBF

Neste caso utilizou-se o *kernel RBF*[2] e o mesmo conjunto de treino indicado anteriormente. Observa-se que este kernel não foi utilizado no trabalho de [ZK03]. O *kernel RBF* k é definido por

$$k(x, x') = \exp(-\frac{\|x - x'\|^2}{2\sigma^2}),$$

em que σ é um parâmetro a especificar. O classificador foi obtido com o valor de σ dado por omissão pela $\texttt{SVM}^{light}$. Nota-se que, com a utilização do *kernel RBF*, está-se de facto a mapear $\mathcal{X}$ para um espaço $\mathcal{H}$ com dimensão infinita e, portanto, o mapa Φ não se define trivialmente. No entanto, como se viu na secção 4.1, o conhecimento de Φ não é importante.

Na tabela 4.3 e na figura 4.2 estão resumidos os resultados obtidos com o classificador *RBF*. Ao contrário do que aconteceu no caso linear, em que o valor 0 serviu de fronteira na distinção entre instâncias positivas e negativas, aqui utilizou-se o valor -0.5. Já em [ZK03] chama-se a atenção para o facto de que o número de contactos parece crescer linearmente com o tamanho da proteína e, nesse caso concreto, consideram-se como instâncias positivas as primeiras L_p instâncias, em que se assume a ordenação das instâncias por ordem decrescente das etiquetas e L_p é o comprimento da sequência de aminoácidos.

Os valores constantes na tabela 4.3 dizem apenas respeito a padrões de treino. Foram realizados testes com proteínas não seleccionadas para o conjunto de treino. No entanto o classificador não conseguiu predizer nenhum contacto. Suspeita-se que este facto se deve principalmente ao número de

[2] *Gaussian Radial Basis Function.*

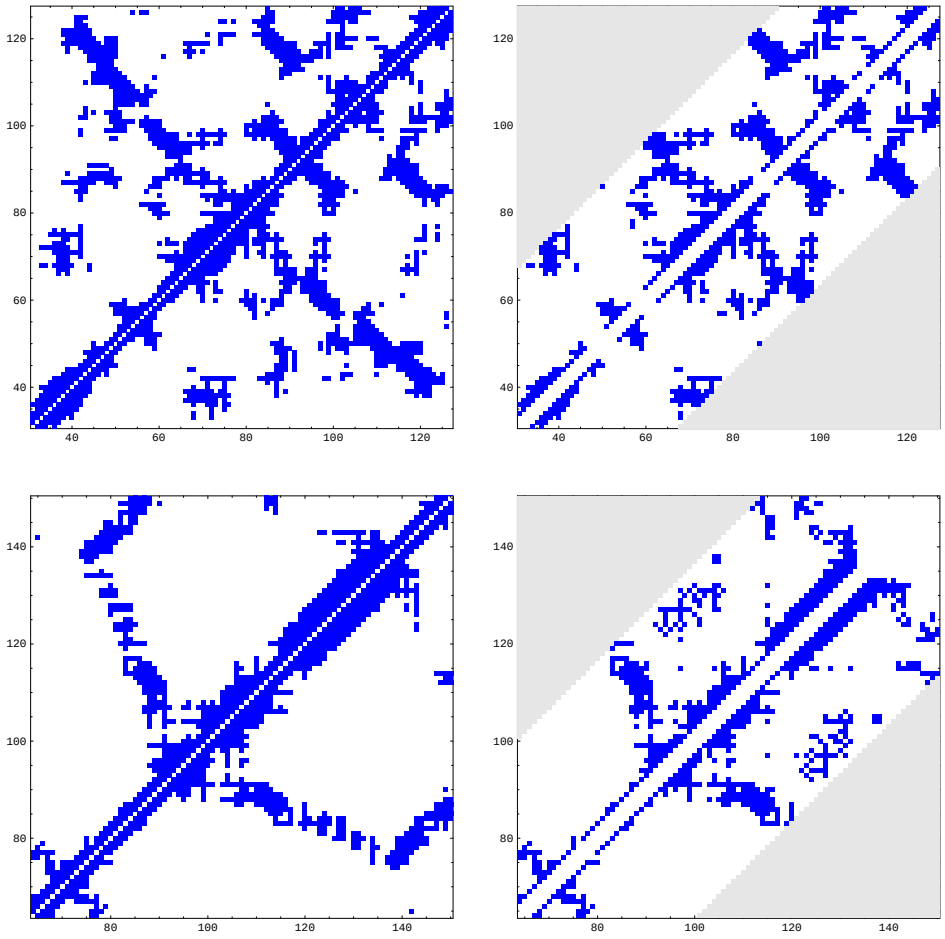

Figura 4.2: Resultados da classificação obtidos neste trabalho com *kernel RBF* (mapa original e mapa obtido) para as proteínas `2HGF` e `4JDW`, respectivamente.

pdbID	#Amin.	#Cont.	#Contactos preditos		
			Total	Correctos	Precisão
2HGF	127	2194	1212	1208	99.7%
4JDW	423	9566	4752	3956	83.2%

Tabela 4.3: Resultados da classificação com *kernel RBF*.

proteínas seleccionadas e à diversidade das mesmas. Na continuação futura deste trabalho ir-se-á considerar um maior número de proteínas no conjunto de treino e, tendo em vista os recursos computacionais, apenas proteínas de uma dada classe, nomeadamente da mesma classe estrutural e da mesma ordem de grandeza.

Porém, é importante observar que a recuperação dos contactos no conjunto de treino é conseguida com uma precisão média de 90%. E, atendendo a que no conjunto de treino apenas se consideraram pares cuja distância entre si na sequência está entre 6 e 30, a recuperação dos padrões de treino é da ordem dos 100%, ver a figura 4.2. Logo, dada a precisão conseguida, é importante explorar a utilização do *kernel RBF* por forma a melhorar os resultados de predição.

Capítulo 5

Conclusões e trabalho futuro

O principal objectivo desta dissertação consistiu em fornecer uma panorâmica das contribuições da Matemática e da Ciência da Computação para a resolução e entendimento do problema da predição da estrutura terciária das proteínas. Foram estudados os algoritmos e as técnicas desenvolvidos na abordagem ao problema, os resultados acerca da complexidade computacional do mesmo e, em especial, foi dada uma visão alternativa baseada em SVM's.

5.1 Conclusões

Os resultados apresentados na secção 3.1 não garantem a intratabilidade do problema da predição da estrutura terciária das proteínas. No entanto os modelos propostos actualmente levam a formulações do problema intratáveis do ponto de vista computacional. Mesmo nos modelos mais simples, por exemplo o modelo *H-P*, prova-se que a formulação do problema *PF* está na classe dos problemas *NP-completos*. Porém, o resultado acerca do problema *PSP* é o mais importante, pois este reflecte o problema no espaço real tridimensional.

Dada a dificuldade do problema e tal como tem vindo a acontecer recentemente, a forma de resolver o problema passa por encontrar técnicas de optimização global aplicáveis à função de energia potencial $V(x)$. Algumas das abordagens foram referidas sumariamente na secção 3.3, tendo sido

dadas referências para os trabalhos relevantes.

A abordagem com melhores resultados conseguidos passa pela utilização de técnicas de classificação ou aprendizagem estatística, algumas vezes combinadas com outras técnicas que, embora correctas, não se podem aplicar face aos recursos computacionais necessários. A divisão do processo de predição em três passos fundamentais, tal como se viu na secção 3.4, tem garantido melhores resultados. O passo mais difícil consiste na predição dos mapas de contacto das proteínas. Estes foram introduzidos na secção 3.4 tendo sido discutidas as propriedades dos mesmos, nomeadamente a capacidade de permitir recuperar totalmente a estrutura terciária. Das várias técnicas aplicadas na predição dos mapas de contacto, os melhores resultados foram conseguidos com máquinas de vectores de suporte.

No capítulo 4 utilizaram-se as máquinas de vectores de suporte para a predição dos mapas de contacto, note-se em particular a utilização do *kernel RBF* nesta dissertação. Os resultados conseguidos com o *kernel* linear aproximam-se do mínimo necessário para garantir a recuperação da estrutura terciária. No entanto a precisão da predição é muito baixa e o classificador não parece ter apreendido de facto informação relevante. Conclui-se assim que a aplicação do *kernel* linear ao universo global não é viável, o conjunto de treino terá de ser aumentado e dever-se-á restringir a classes de proteínas bem definidas. No trabalho de Zhao e Karypis [ZK03] os resultados são qualitativamente superiores, facto esse que confirma as conclusões anteriores dado que aí as proteínas foram divididas em classes estruturais e o conjunto de treino era maior.

A utilização do *kernel RBF* garantiu uma maior precisão e permitiu recuperar a 100% os padrões de treino. Contudo os resultados para padrões de teste foram maus. Portanto, ter-se-á que refazer o conjunto de características e construir um conjunto de treino mais cuidado. Em particular é importante combinar o *kernel RBF* com a informação adquirida na comparação do trabalho aqui desenvolvido com o trabalho desenvolvido em [ZK03], permitirá

sem dúvida melhorar os resultados.

Dado que foram utilizadas sequências homólogas em [ZK03] e no presente trabalho não, estas parecem ser relevantes e deverão ser incorporadas no conjunto de características.

5.2 Trabalho Futuro

Dada a panorâmica apresentada é fácil de constatar um grande número de questões em aberto e de problemas por resolver. Actualmente este problema é investigado por pessoas das mais variadas áreas. No que diz respeito aos classificadores e às máquinas de vectores de suporte é importante referir trabalhos como os [PBVF03, HSS+02, HLL04], o último dos quais apresentado em Junho de 2004 e que reafirma algumas das conclusões indicadas nesta dissertação. Existe um grande número de abordagens ao problema e é de facto necessário perceber e indicar quais os pontos fortes de cada abordagem.

Note-se a importância de novos métodos de optimização global eficientes. Actualmente existem excelentes técnicas de optimização local. Todavia a sua aplicação à optimização global revela-se ineficiente mesmo em casos mais simples. Esta não é uma área restrita ao problema da predição da estrutura terciária das proteínas, embora o seu desenvolvimento contribua decididamente para este problema.

Os mapas de contacto constituem uma das representações a estudar, de facto a facilidade com que se manipulam pode ter importância para a determinação da conformação nativa recorrendo à minimização de energia potencial. Em particular, é necessário o desenvolvimento de algoritmos eficientes que permitam inferir se um mapa de contacto é ou não realizável fisicamente.

A utilização de técnicas de aprendizagem estatística constitui uma das áreas a investigar, em particular as máquinas de vectores de suporte. Deste modo, não pondo de lado outros caminhos a seguir, será importante analisar de que forma se podem melhorar os resultados conseguidos. Após o

estudo aqui apresentado existem várias pontos a ter em conta, e.g.: quais as características relevantes para a predição; explorar a utilização do *kernel RBF*; construir conjuntos de treino mais cuidados; desenvolver *kernéis* mais sofisticados que consigam analisar a sequência de aminoácidos; e determinar a importância da utilização das sequências homólogas.

Dado o estado da arte, a sua importância e a relevância das questões em aberto, esta é sem dúvida uma área da ciência a estudar no futuro próximo.

Bibliografia

[Afi73] C. B. Afinsen. Principles that govern the folding of protein chains. *Science*, 183(4096):223–230, 1973.

[AHSW61] C. B. Afinsen, E. Haber, M. Sela, and F. H. White. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. In *Proceedings of the National Academy of Sciences*, volume 47, pages 1309–1314, USA, 1961.

[AMF95] I. P. Androulakis, C. D. Maranas, and C. A. Floudas. Global minimum potential energy conformations of oligopeptides. Manuscript, 1995.

[BL98] B. Berger and T. Leighton. Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. *Journal of Computational Biology*, 5(1):27–40, 1998.

[BPFV03] P. Baldi, G. Pollastri, P. Frasconi, and A. Vullo. New Machine Learning Methods for the Prediction of Protein Topologies. In P. Frasconi and R. Shamir, editors, *Artificial Intelligence and Heuristic Methods for Bioinformatics*. IOS Press, 2003.

[BWF+00] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissing I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235–242, 2000. http://www.rcsb.org/pdb/.

[CGP+98] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, and M. Yannakakis. On the complexity of protein folding. *Journal of Computational Biology*, 5(2), 1998.

[Coo71] S. A. Cook. The complexity of theorem proving procedures. In *Proceedings of the 3rd ACM Symposium on Theory of Computing*, pages 151–158, 1971.

[CS00] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods.* Cambridge University Press, 2000. Reimpresso com correcções em 2000 e 2001.

[Dil90] K. A. Dill. Dominant forces in protein folding. *Biochemistry*, 29(31):7133–7155, August 1990.

[FOVC01] P. Fariselli, O. Olmea, A. Valencia, and R. Casadio. Prediction of contact maps with neural networks and correlated mutations. *Protein Engineering*, 14(11):835–843, 2001.

[GJ79] M. R. Garey and D. S. Johnson. *Computers and Intractability, A Guide to the Theory of NP-Completeness.* W. H. Freeman and Company, 1979.

[GS87] K. D. Gibson and H. A. Scheraga. Revised algorithms for the build-up procedure for predicting protein conformations by energy minimization. *Journal of Ccomputational Chemistry*, 8:826–834, 1987.

[HI96] W. E. Hart and S. Istrail. Fast protein folding in the hydrophobic-hydrophilic model within three-eights of optimal. *Journal of Computational Biology*, 3:53–96, 1996.

[HLL04] C. Hu, X. Li, and J. Liang. Developing optimal nonlinear scoring function for protein design. *Bioinformatics*, Junho 2004. Em impressão.

[Hol73] J. Holland. Genetic algorithms and the optimal allocation of trials. *SIAM Journal of Computing*, 2:88–105, 1973.

[HSS+02] J. Hu, X. Shen, Y. Shao, C. Bystroff, and M. Zaki. Mining protein contact maps. In *2nd BIOKDD Workshop on Data Mining in Bioinformatics*, pages 3–10, 2002.

[JCM+92] R. S. Judson, M. E. Colvin, J. C. Meza, A. Huffer, and D. Gutierrez. Do intelligent configuration search techniques outperform random search for large molecules? *International Journal of Quantum Chemistry*, 44:270–290, 1992.

[Joa99] T. Joachims. Making large-scale SVM learning pratical. In B. Schlkopf, C. Burges, and A. Smola, editors, *Advances in Kernels Methods - Support Vector Learning*. MIT-Press, 1999.

[KGV83] S. Kirkpatrick, C. D. Geddat, Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671.680, 1983.

[KK99] S. Kawashima and M. Kanehisa. AAindex: amino acid index database. *Nucleic Acids Research*, 27(1):368–369, January 1999. http://www.genome.ad.jp/dbget/aaindex.html.

[LD60] A. H. Land and A. G. Doig. An Automatic Method for Solving Discrete Programming Problems. *Econometrica*, 28:497–520, 1960.

[LLH01] A. M. Lesk, L. Lo Conte, and T. J. P. Hubbard. Assessment of novel fold targets in CASP4: predictions of three-dimensional structures, secondary structures, function and genetics. *Proteins*, 45(5):98–118, 2001.

[LM93] S. M. Le Grand and K. M. Merz, Jr. The application of genetic algorithm to the minimization of potential energy functions. *Journal of Global Optimization*, 3:49–66, 1993.

[LS87] Z. Li and H. A. Scheraga. Monte Carlo approach to the multiple-minima problem in protein folding. In *Proceedings of the National Academy of Sciences*, volume 84, pages 15–29, 1987.

[LS88] Z. Li and H. A. Scheraga. Structure and free energy of complex thermodynamic systems. *Journal of Molecular Structures*, 179:333–352, 1988.

[MGR91] L. B. Morales, R. Garduno-Juarez, and D. Romero. Applications of simulated annealing to the multiple-minima problem in small peptides. *Journal of Biomolecular Structures & Dynamics*, 8(4):721–735, Fevereiro 1991.

[NaMK94] J. Ngo and J. Marks ant M. Karplus. Computational Complexity, Protein Structure Prediction, and the Levinthal Paradox. In K. Merz, Jr. and S. LeGrand, editors, *The Protein Folding Problem and Tertiary Structure Prediction*. Birkhäuser, 1994.

[New02] A. Newman. A new algorithm for protein folding in the HP model. In *Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms*, volume 1018, pages 876–884, San Francisco, California, January 2002.

[NKK88] K. Nakai, A. Kidera, and M. Kanehisa. Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Engineering*, 2(2):93–100, July 1988.

[NM92] J. Ngo and J. Marks. Computational Complexity of a Problem in Molecular-Structure Prediction. *Protein Engineering*, 5(4):313–321, 1992.

[NRT89] G. L. Nemhauser, A. H. G. Rinnooy Kan, and M. J. Todd, editors. *Optimization*, chapter Integer Programming, pages 447–527. North-Holland, Amsterdam, 1989.

[NVS91] A. Nayeem, J. Vila, and H. A. Scheraga. A comparative study of the simulated-annealing and Monte Carlo-with-minimization approaches to the minimum-energy structures of polypeptides: met-enkephalin. *Journal of Ccomputational Chemistry*, 12:594–605, 1991.

[OKK92] Y. Okamoto, T. Kikuchi, and H. Kawai. Prediction of low-energy structures of met-enkephalin by Monte Carlo simulated annealing. *Chemistry Letters*, 7:1275–1278, 1992.

[ORV99] O. Olmea, B. Rost, and A. Valencia. Effective use of sequence correlation and conservation in fold recognition. *Journal of Molecular Biology*, 293:1221–1239, 1999.

[Pap94] C. H. Papadimitriou. *Computational Complexity*. Addison Wesley, 1994. Reimpresso com correcções em Agosto de 1995.

[PB02] G. Pollastri and P. Baldi. Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics*, 18(1):S62–S70, 2002.

[PBFC01] G. Pollastri, P. Baldi, P. Fariselli, and R. Casadio. Prediction of coordination number and relative solvent accessibility in proteins. *Proteins*, 47:142–153, 2001.

[PBVF03] G. Pollastri, P. Baldi, A. Vullo, and P. Frasconi. Prediction of protein topologies using GIOHMMs and GRNNs. *Advances in Neural Information Processing Systems*, 15, 2003.

[PKS89] L. Piela, J. Kostrowicki, and H. A. Scheraga. The multiple-minima problem in the conformational analysis of molecules. Deformation of the potential energy hypersurface by the diffusion equation method. *Journal of Physical Chemistry*, 493:3339–3346, 1989.

[PPRB01] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*, 47:228–235, 2001.

[PTG99] D. D. Pollock, W. R. Taylor, and N. Goldman. Coevolving protein residues: maximum likelihood identification and relationship to structure. *Journal of Molecular Biology*, 287:187–198, 1999.

[RVS91] D. R. Ripoll, M. J. Vasquez, and H. A. Scheraga. The electrostatically driven Monte Carlo method: application to conformational analysis of decaglycine. *Biopolymers*, 31(3):319–330, Fevereiro 1991.

[Sno92] M. E. Snow. Powerful simulated annealing algorithm locates global minimum of protein folding potentials from multiple starting conformations. *Journal of Computational Chemistry*, 13(5):579–584, Junho 1992.

[SS02] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.

[Sti85] F. H. Stillinger. Role of potential-energy scaling in the lowtemperature relaxation behavior of amorphous materials. *Physical Review B*, 32:3134–3141, 1985.

[SW88] F. H. Stillinger and T. A. Weber. Nonlinear optimization simplified by hypersurface deformation. *Journal of Statistical Physics*, 52:1429–1445, 1988.

[TEHL91] P. Tuffery, C. Etchebest, S. Hazout, and R. Lavery. A new approach to the rapid determination of protein side chain conformations. *Journal of Biomolecular Structures & Dynamics*, 8:1267–1289, 1991.

[THG94] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensivity of progressive multiple sequence alignment throught sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673–480, 1994.

[TK96] K. Tomii and M. Kanehisa. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Engineering*, 9(1):27–36, January 1996.

[Tre97] L. Trevisan. When Hamming meets Euclid: the approximability of geometric TSP and MST. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, pages 21–29, 1997.

[UM93] R. Unger and J. Moult. Finding the lowest free energy conformation of a protein is an NP-hard problem: proof and implications. *Bulletin of Mathematical Biology*, 55(6):1183–1198, November 1993.

[Vap95] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[VD90] S. Vajda and C. De Lisi. Determining minimum energy conformations of polypeptides by dynamic programming. *Biopolymers*, 29:1755–1772, 1990.

[VD00] M. Vendruscolo and E. Domany. Protein folding using contact maps. *Vitamins and Hormones*, 58:171–212, 2000.

[Ven03] M. Vendruscolo. Protein folding using contact maps and contact vectors. In P. Frasconi and R. Shamir, editors, *Artificial Intelligence and Heuristic Methods for Bioinformatics*. IOS Press, 2003.

[VJSD93] S. Vajda, M. S. Jafri, U. Sezerman, and C. De Lisi. Necessary conditions for avoiding incorrect polypeptide folds in conformational search by energy minimizations. *Biopolymers*, 33:173–192, 1993.

[VKD97] M. Vendruscolo, E. Kussel, and E. Domany. Recovery of protein structure from contact maps. *Folding and Design*, 2:295–306, 1997.

[VND99] M. Vendruscolo, R. Najmanovich, and E. Domany. Protein folding in contact map space. *Physical Review Letters*, 82(3):656–659, 1999.

[VS85] M. Vásquez and H. A. Scheraga. Use buildup and energy-minimization procedures to compute low-energy stuctures of the backbone of enkephalin. *Biopolymers*, 24:1437–1447, 1985.

[ZK03] Y. Zhao and G. Karypis. Prediction of Contact Maps Using Support Vector Machines. In *Third IEEE Symposium on Bioinformatics and Bioengineering (BIBE'03)*, pages 26–33, March 2003.

Lista de Figuras

Lista de Tabelas

Índice