UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA ELÉTRICA

Modelos Logísticos Quadráticos com Máxima Verossimilhança Penalizada para Previsão de Estrutura Secundária de Proteínas

RAUL NEDER PORRELLI

Orientador: Dr. Renato M. E. Sabbatini

Dissertação apresentada como parte dos requisitos para obtenção de título de Mestre em Engenharia Elétrica.

Porrelli, Raul Neder
P828B Modelos logísticos quadráticos com máxima verossimilhança penalizada para previsão de estrutura secundária de proteínas / Raul Neder Porrelli.--Campinas, SP: [s.n.], 1995.

Orientador: Renato M. E. Sabbatini.
Dissertação (mestrado) - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica.

1. Modelos log-lineares. 2. Sequência de aminoácidos. 3. Proteínas - Análise. I. Sabbatini, Renato M. E. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica. III. Título.

Agradecimentos

Ao Dr. Peter Munson, chefe da *Analytical Biostatiscs Section, Laboratoy of Structural Biology, Department of Computer Resources and Technology, National Institutes of Health,* E.U.A. (NIH), a quem pertencem muitas das idéias contidas aqui. Dr. Munson não somente colaborou na adptação do presente trabalho, realizado durante meu estágio de *fellowship*, para que se tornasse uma dissertação de mestrado, como deu condições técnicas e apoio financeiro para fazê-lo (custos de utilização do Convex C3830, diárias nas visitas ao NIH e empréstimo de equipamento).

Ao *Department of Computer Resources and Technology, National Institutes of Health,* E.U.A., na pessoa de seu diretor Dr. David Rodbard, responsável pela minha ida ao NIH, pelo já citado apoio.

À Dra. Valentina Di Francesco, companheira de *fellowship* no NIH, responsável pela otimização das rotinas FORTRAN no Convex, por sua ajuda e companheirismo.

Ao Núcleo de Informática Biomédica (NIB), na pessoa do seu coordenador Dr. Renato Sabbatini, meu orientador, por ter acolhido a mim e a este projeto. O NIB tem sido meu lar desde 1983 e tem me apoiado, mesmo nos períodos em que estive temporariamente filiado a outras instituições.

Ao Departamento de Engenharia Biomédica, nas pessoas daqueles que foram seus chefes durante a execução deste trabalho, Dr. Alberto Cliquet Jr., Dr. Sérgio Santos Mühlen e Dr. Eduardo Tavares Costa, pelo apoio e incentivo à atividade interdisciplinar.

À Rede Nacional de Pesquisa (RNP), na pessoa de seu coordenador, Prof. Eduardo Tadao Takahashi, que nos deu acesso privilegiado à Internet, sem o que seria impossível o uso remoto dos computadores do NIH. Além disso, a RNP permitiu que coordenasse as atividades que ali exerço com a execução deste trabalho, como na ocasião em que visitei o NIH em outubro de 1994, estando a serviço da RNP nos E.U.A.

Aos membros da Banca Examinadora, Dr. João Meidanis e Dr. Reginaldo Palazzo Jr., pelas sugestões e correções incorporadas a esta versão final.

Agradecemos, também, o apoio da CAPES, de março a julho de 1993 (bolsa de mestrado pela Faculdade de Engenharia Elétrica de Campinas) e do CNPq (bolsa do programa de Desenvolvimento Tecnológico e Industrial), desde março de 1994, pela RNP).

Resumo

Apesar do grande número de algoritmos existentes para a previsão de estrutura secundária de proteínas, determinadas técnicas estatísticas ainda não haviam sido exploradas. Utilizamos a metodologia de funções discriminantes logísticas na tentativa de ultrapassar a acurácia obtida por métodos que usaram redes neurais e teoria da informação. O número de parâmetros foi limitado explorando-se a natureza periódica das alfa-hélices e placas pregueadas beta. Uma grande variedade de modelos foi pesquisada, usando abordagem semi-paramétrica (máxima verossimilhança com penalização) combinada com seleção gradual de parâmetros. Mostramos que os modelos mais bem sucedidos tem ao redor de 800 parâmetros "efetivos" para o conjunto de dados utilizado. Os 340 parâmetros lineares e parte dos 800 parâmetros quadráticos puderam ser interpretados do ponto de vista fisicoquímico, contrastando com outros métodos da literatura. Após otimização e validação cruzada, a acurácia foi de 65.9% para três estados estruturais, o que representa um resultado ligeiramente superior aos dos algoritmos já publicados. A maior acurácia de previsão está concentrada numa porção dos resíduos e a confiança da previsão pode ser facilmente calculada. Exploramos a possibilidade de usar estes resíduos, previstos com alta confiabilidade, para prever a estrutura completa da proteína, assim como muitos outros artifícios para aumentar a eficiência do método, com resultados limitados. Embora tenhamos obtido apenas uma modesta melhora da acurácia, a maneira como implementamos o modelo sugere que utilizamos toda a

informação estrutural contida em segmentos de até 17 aminoácidos, no nível de complexidade que a quantidade de dados permite.

Sumário

Abreviaturas, Símbolos e Convenções

$\mathbf{0}$	Vetor nulo.
$\mathbf{1}$	Vetor unidade.
3_{10}	Alfa-hélice 1-3.
$\mathbf{A}$	Matriz alfa da forma quadrática do modelo logístico quadrático.
$\mathbf{A}$	Bloco 20x20 cuja repetição, ponderada por $\mathbf{W}$, forma a matriz $\mathbf{A}$.
$\mathbf{a}$	Vetor com os parâmetros lineares alfa.
a_0	Parâmetro linear aditivo. Primeiro elemento de θ_a.
a_{ij}	Parâmetro linear alfa, elemento de $\mathbf{a}$.
arg max{f(x)}	Valor de f para o maior x.
ASCII	*American Standard Code for Information Interchange.*
AxB	Interação entre os aminoácidos A e B, nesta ordem.
$\mathbf{B}$	Matriz beta da forma quadrática do modelo logístico quadrático.
$\mathbf{B}$	Bloco 20x20 cuja repetição, ponderada por $\mathbf{W}$, forma a matriz $\mathbf{B}$.
BNL	*Brookhaven National Laboratory.*
C_α	Carbono alfa.
C1	Constante de decisão beta.
C2	Constante de decisão alfa.
cos(x)	Cosseno de x.
$\mathbf{DC_H}$	Constante de decisão alfa de GGR.
$\mathbf{DC_E}$	Constante de decisão beta de GGR.

DSSP	Dicionary of Secondary Structure of Proteins (KABSCH & SANDER, 1983a).
e	Constante de Euler.
EMV	Estimativas de máxima verossimilhança.
fa,i	Variável binária sobre o estado alfa na i-ésima janela local.
fb,i	Variável binária sobre o estado beta na i-ésima janela local.
GCV	*General Cross-Validation index* (EUBANK, 1988).
GGR	Método preditivo de GIBRAT *et al.* (1987).
GOR	Método preditivo de GARNIER *et al.* (1978).
$\mathbf{h}$	Vetor contendo tabela de propriedade de aminoácidos.
h_i	Coeficiente de propriedade (e.g. hidofobicidade) do aminoácido i.
$\boldsymbol{I}$	Quantidade de informação.
$\boldsymbol{I_D}$	Quantidade de informação associada a observações fictícias.
$\boldsymbol{I_C}$	Quantidade de informação *coil*.
$\boldsymbol{I_E}$	Quantidade de informação beta.
$\boldsymbol{I_H}$	Quantidade de informação alfa.
$\boldsymbol{I_o}$	Quantidade de informação observada (em contraste com $\boldsymbol{I_D}$)
$\mathbf{L}$	Matriz quadrada de transformação linear.
L_{310}	Alfa-hélice 1-3 levógira (ao espelho).
L_a	Alfa-hélice levógira (ao espelho).
ln(x)	Logaritmo natural.
log(x)	Logaritmo natural (mantida a notação de GGR).
LNV	Logverossimilhança negativa.
$\mathbf{M}$	Número de observações fictícias.
MV	Máxima Verossimilhança.
max{x}	Máximo valor de x.
N	Número de observações. ou tamanho da cadeia protéica.
npar	Número efetivo de parâmetros em modelos penalizados.
nqua	Número de parâmetros quadráticos selecionados.
P	Nível de confiança (e.g. P<0.05).
$\mathbf{P_{max}}$	Nível de confiança para seleção de variáveis em GGR.
$\hat{p}$	Confiança das estimativas de um modelo.
$\mathbf{p_a}$	Vetor com as probabilidades de alfa-hélices.
$p_{a,i}$	Probabilidade do i-ésimo resíduo estar numa alfa-hélice.
$\mathbf{p_b}$	Vetor com as probabilidades de cordões beta.
$p_{b,i}$	Probabilidade do i-ésimo resíduo estar numa placa beta.
$\mathbf{p_c}$	Vetor com as probabilidades de *random coils*..
$p_{c,i}$	Probabilidade do i-ésimo resíduo não ter estrutura regular.

PDB	*Protein Data Bank.*
Pot(x)	Função potência de Fourier.
RMS	Raiz do erro quadrático médio.
$\mathbf{r}$	Vetor cosseno, para criar periodicidade em $\mathbf{W}$.
R_i	i-ésimo resíduo da janela.
$\mathbf{s}$	Vetor seno, para criar periodicidade em $\mathbf{W}$.
$\hat{S}$	Máxima verossimilhança de um modelo.
$S_{\theta'_x}$	Desvio padrão dos elementos de θ'.
S_i	Estado estrutural do i-ésimo resíduo da janela.
sen(x)	Seno de x.
T	Como expoente (e.g. $\mathbf{A}^T$) representa a matriz transposta.
traço(A)	Soma dos elementos da diagonal principal de **A**.
$t_{x,i}$	Valor t de Student para o i-ésimo parâmetro do estado x.
$\mathcal{V}$	Verossimilhança das estimativas de um modelo.
ν	Contribuição individual de um resíduo para a verossimilhança.
w_{ij}	Elemento de $\mathbf{W}$
$\mathbf{w}_{\|i-j\|}$	Vetor contendo os elementos de uma diagonal de $\mathbf{W}$.
$\mathbf{W}_X^n$	Matriz com os coeficientes para a montagem de $\mathbf{A}$ a partir de **A**.
$\tilde{\mathbf{X}}$	Variável independente, estendida para conter interacões.
$\mathbf{x}$	Variável independente, binária, com a sequência.
x_{ij}	Aminoácido do tipi i, na j-ésima posição da janela. Elemento de $\mathbf{x}$.
y_i	i-ésimo valor obtido pelo modelo, e.g. p_x.
$\hat{y}_i$	i-ésimo valor observado nos dados, e.g. f_x.
$\mathbf{z}(\mathbf{x})$	Vetor $\mathbf{z}$, expoente na função vetorial logística.
z_a	Variável intermediária alfa, expoente na função logística.
z_b	Variável intermediária beta, expoente na função logística.
$\frac{\partial y}{\partial x}$	Derivada parcial de y com relação a x.
ϕ	Primeiro ângulo diédrico.
Λ	Matriz binária que determina os parâmetros penalizados.
λ	Penalidade aplicada à verossimilhança.
Λ_{ij}	Elemento de Λ.
π	3.1415...
$\prod_{i=1}^{N} x_i$	$(x_1.x_2. \ldots .x_{i-1}.x_i)$
θ'_x	Vetor contendo somente parâmetros quadráticos do estado x.
θ_a	Vetor combinando a_0, **a**, e **A**.

θ_b	Vetor combinando b_0, **b**, e **B**.
$\sum_{i=1}^{N} x_i$	$(x_1+x_2+\ldots+x_{i-1}+x_i)$
τ	Ângulo plano entre N-C_α e C_α-C.
χ^2	Teste de *chi* quadrado.
ω	Terceiro ângulo diédrico.
ψ	Segundo ângulo diédrico.

As abreviaturas dos aminoácidos se encontram na Fig. 2.

Seguimos os padrões da Associação Brasileira de Normas Técnicas para dissertações e teses.

Adotamos as modificações do novo acordo ortográfico internacional. Daí a ortografia de *aminoácido*, *alfa-hélice*, *físicoquímica* e *superrepresentação*. A palavra *protéica* foi acentuada, contrariando o modo como é pronunciada no Estado de São Paulo, pois assim foi encontrada nos dicionários.

1. Introdução

Proteínas são polipeptídeos, isto é, cadeias lineares de aminoácidos ligados covalentemente entre si. Com algumas exceções, 20 diferentes tipos de aminoácidos entram na composição de uma proteína. A Figura 1 mostra a estrutura de um aminoácido genérico (STRYER, 1981). Cada um deles é formado por um átomo de carbono central (C_α) ao qual estão ligados um átomo de hidrogênio, um grupo amina (NH_3) e um grupo carboxila (COOH). O que distingue cada monômero é a cadeia lateral, também ligada ao átomo C_α. A Figura 2 mostra os 20 tipos de aminoácidos mais frequentes na composição das proteínas. Ali podemos ver as ligações peptídicas, entre o grupo carboxila de um amino ácido e o grupo amina do outro (DOOLITTLE, 1985).

Uma proteína pode ser formada por uma ou mais cadeias polipeptídicas, contendo ou não elementos adicionais tais como grupos prostéticos, metais coordenados e cadeias de carbohidratos (STRYER, 1981). O comportamento químico e, consequentemente, a função de uma proteína dependem de sua estrutura tridimensional (BRANDEN & TOOZE, 1991). Além da conformação da cadeia principal, a orientação de cada cadeia lateral no espaço define estas propriedades. A Figura 3 ilustra estes conceitos. Na Fig. 3a podemos ver a cadeia principal de uma proteína, na Fig. 3b uma representação mais realista, com todos os átomos das cadeias laterais, mostrando o aspecto globular que, à primeira vista, tem pouca relação com a forma da cadeia principal.

1.1. Estrutura Protéica

A estrutura protéica tem diferentes níveis de complexidade, como ilustrado na Figura 4. Cada cadeia polipeptídica pode ser caracterizada pela sequência de aminoácidos que a compõe. Esta lista de monômeros chama-se *estrutura primária* de uma proteína. Existem alguns elementos estruturais que se repetem, tais como alfa-hélices e placas pregueadas beta, cujo reconhecimento pode simplificar a visualização da estrutura tridimensional. A lista destes elementos, arranjados na ordem em que ocorrem na sequência de aminoácidos, denomina-se *estrutura secundária*. A estrutura tridimensional de uma cadeia protéica, definida pela posição de cada um de seus átomos, incluindo aqueles dos radicais laterais, é a *estrutura terciária*. Como algumas proteínas são compostas por mais de uma cadeia, formando complexos, define-se também *estrutura quaternária*, como o arranjo espacial de mais de uma cadeia (STRYER, 1981).

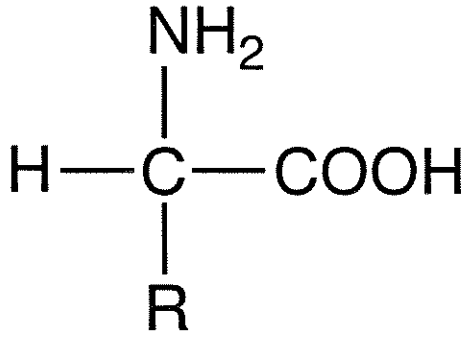


Figura 1 - Fórmula estrutural de um amino ácido genérico. O "C" no centro é o carbono alfa, que se liga a quatro radicais diferentes (menos no caso da glicina onde "R" é um átomo de hidrogênio, veja Fig. 2), "NH2" é o grupo amina, "COOH" o grupo ácido carboxílico e "R" representa a cadeia lateral, diferente para cada amino ácido. Baseada em STRYER (1981).

Os dois tipos de estrutura secundária são as alfa-hélices (Fig. 5) e as placas pregueadas beta (Fig. 6). O primeiro tipo surge de interação local entre os aminoácidos de um segmento. Cada um estabelece pontes de hidrogênio com outro, situado quatro posições adiante. Esta é uma estrutura compacta. Nas placas pregueadas, as porções estendidas da cadeia interagem lado a lado, podendo vir de regiões bastante distintas. As pontes de hidrogênio se estabelecem entre os átomos de cada cadeia justaposta, formando um plano ou placa, que é então chamada de

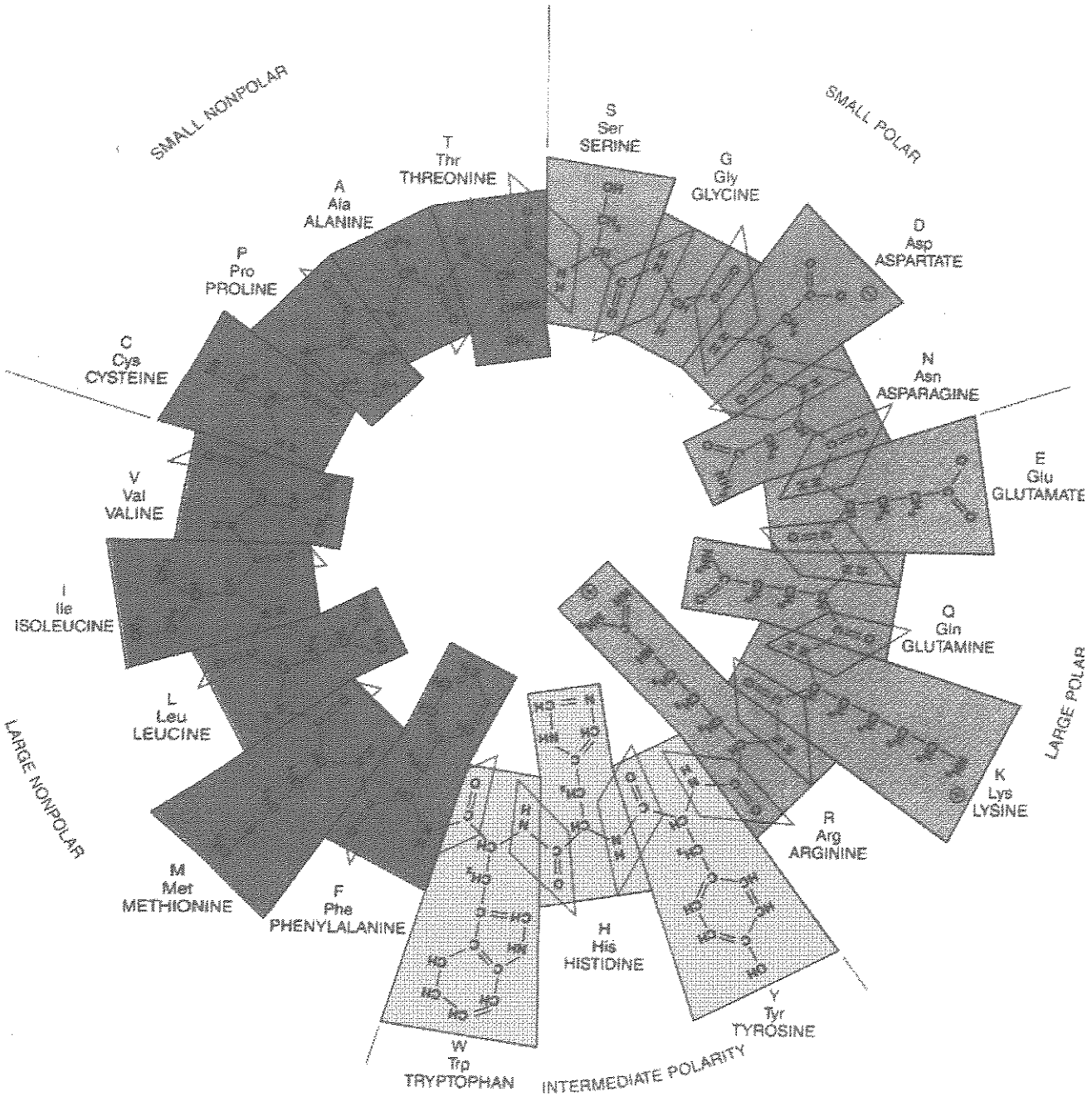
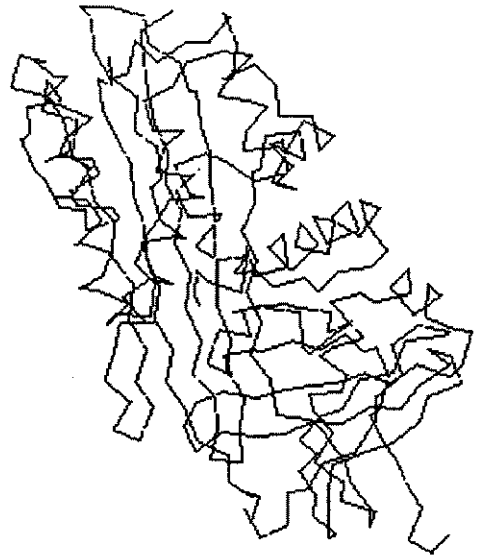
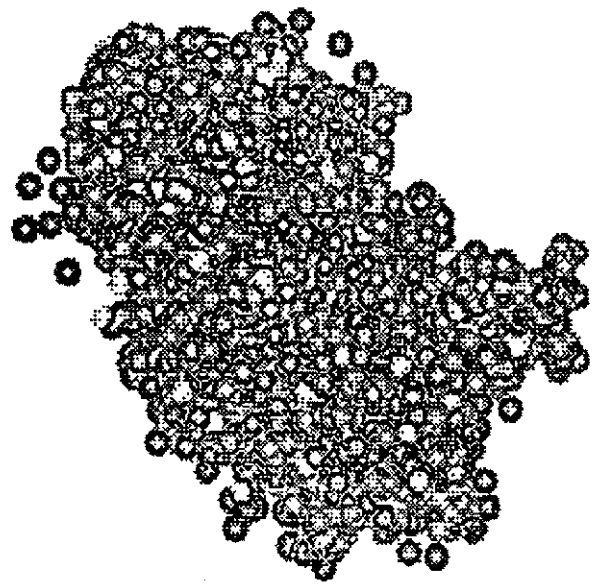


Figura 2 - Fórmulas estruturais dos 20 aminoácidos mais frequentemente encontrados nas proteínas naturais. Aqui eles se encontram covalentemente associados em estrutura circular para ilustrar a ligação peptídica entre o grupo carboxila de um amino ácido com o grupo amina do subsequente. Os aminoácidos foram agrupados de acordo com sua hidrofobicidade. Cada um pode ser codificado por uma ou três letras. (Adaptado de DOOLITTLE, 1985).



(a)

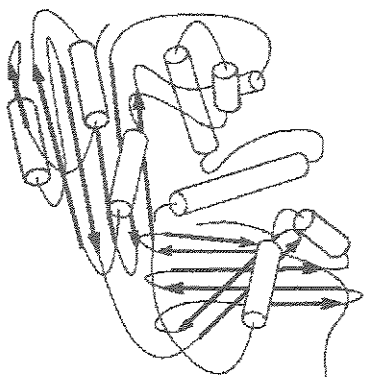


(b)

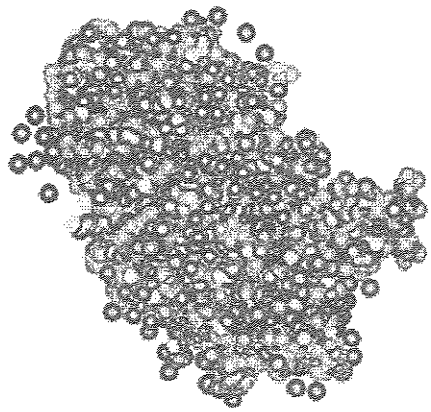
Figura 3 - Representações gráficas da estrutura da proteína "anti-tripsina alfa" (arquivo PDB, código 7API; ENGH *et al.*, 1989). (a) Traçado da cadeia principal (b) Modelo da mesma proteína, na mesma posição, mas substituindo cada átomo por uma esfera de tamanho proporcional. Ilustra como a superfície externa de uma proteína globular depende da orientação das cadeias laterais expostas. Sua relação com o formato da cadeia principal é complexa. (Figura feita pelo candidato com auxílio do programa MACIMDAD, 1993).

NTSHHDQDHPTFNKITPNLAEFAFSLYRQLAHQSN
STNIFFSPVSIATAFAMLSLGTKADTHDEILEGLNF
NLTEIPEAQIHEGFQELLRTLNQPDSQLQLTTGNG
LFLSEGLKLVDKFLEDVKKLYHSEAFTVNFGDTEE
AKKQINDYVEKGTQGKIVDLVKELDRDTVFALVNYI
FFKGKWERPFEVKDTEEEDFHVDQVTTVKVPMMKR
LGMFNIQHCKKLSSWVLLMKYLGNATAIFFLPDE
GKLQHLENELTHDIITKFLENEDRRSASLHLPKLSI
TGTYDLKSVLGQLGITKVFSNGADLSGVTEEAPLK
LSKAVHKAVLTIDEKGTEAAGAMFLEAIPM

(a)



(b)



(c)

Figura 4 - Tipos de estrutura protéica.(a) A sequência de aminoácidos da proteína "anti-tripsina alfa 1" (arquivo PDB, 7API; ENGH *et al.*, 1989) corresponde à sua *estrutura primária.* Os segmentos que correspondem a alfa-hélices foram marcados em vermelho e os que correspondem a cordões beta, em azul, representando a *estrutura secundária.* Desta forma o texto representa uma combinação dos dois tipos de estrutura. (b) Cadeia principal da mesma proteína (*estrutura terciária*) onde foram assinaladas as porções alfa ("cilindros") e beta ("flechas"), elementos de *estrutura secundária.* (Novamente foram combinados dois tipos de estrutura). (c) Modelo tridimensional da mesma proteína, representando sua *estrutura terciária.* Existem pequenas discrepâncias entre as estruturas secundárias de (a) e (b) uma vez que a primeira foi retirada do PDB e a segunda do programa DSSP de KABSCH & SANDER (1983a, veja o texto). Figura feita pelo candidato com auxílio do programa MACIMDAD (1993).

pregueada, devido ao efeito criado pelo aspecto de *zig-zag* das cadeias estendidas. O plano recebe o nome de placa pregueada beta (*beta pleated sheet*) e cada segmento justaposto, cordão beta (*beta strand*). Duas cadeias podem ter orientação paralela ou antiparalela, e as placas podem ser exclusivamente paralelas, antiparalelas ou mistas (STRYER, 1981; BRANDEN & TOOZE, 1991).

O reconhecimento dos elementos de estrutura secundária (placas pregueadas beta e alfa-hélices) foi inicialmente baseado na visão subjetiva do cristalógrafo (BERNSTEIN *et al.*, 1977). A identificação destes elementos era publicada juntamente com as coordenadas dos átomos. Uma forma quantitativa de identificar estes elementos é através dos *ângulos diédricos* ϕ e ψ, ilustrados na Figura 7. Por eles pode-se inferir a "trajetória" da cadeia principal, se estendida (estrutura beta) ou compacta (RICHARDSON & RICHARDSON, 1989). Um instrumento útil para a representação destes ângulos é o *gráfico de Ramachandran* (RAMACHANDRAN & SASSEKHARAN, 1968, Fig. 8) onde ϕ e ψ ocupam respectivamente os eixos das ordenadas e das abscissas. O uso deste método tende a superestimar a presença de estruturas regulares, principalmente do tipo beta (NISHIKAWA & NOGUCHI, 1991). Isto se deve à existência de porções estendidas da cadeia principal que não formam placas pregueadas. No caso das alfa-hélices, são necessários pelo menos quatro aminoácidos com ângulos na região α da Fig.8 para que tenhamos um passo de hélice, de forma que o método as vezes identifica segmentos alfa inválidos, mais curtos (LAMBERT & SCHERAGA, 1989).

Outro método que identifica elementos de estrutura secundária são os padrões característicos de pontes de hidrogênio. Nas estruturas alfa estas ligações ocorrem entre os aminoácidos separados por um passo de hélice (cada resíduo de posição *i* se liga ao de posição *i+4*). Nas placas beta as pontes se formam entre os aminoácidos de diferentes cordões. KABSCH & SANDER (1983) desenvolveram um método quantitativo para detectar elementos de estrutura secundária em proteínas de estrutura terciária determinada, através da estimativa da "força" das pontes de hidrogênio baseada em modelo energético. A estrutura secundária é atribuída aos segmentos que satisfazem determinados padrões. Estes autores reconhecem 8 tipos de estrutura, não necessariamente exclusivas, em contraposição às 3 já apresentadas (alfa, beta e ausência de estrutura regular, também chamada de *random coil*). A maioria dos autores simplesmente ignora os outros cinco tipos de estruturas, interpretando todos como *coil*. Alguns autores (e.g. GARNIER *et al.*, 1978) propuseram

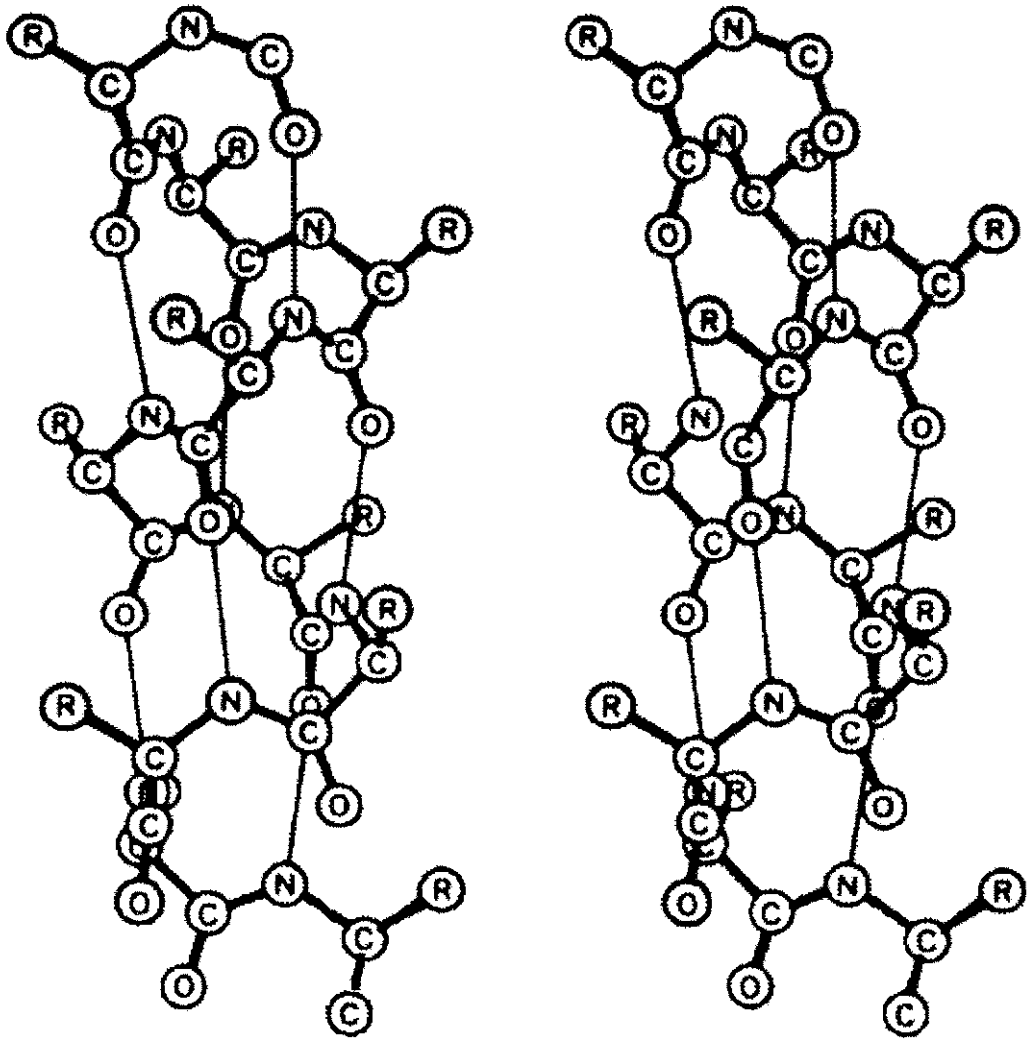


Figura 5 - Par estereográfico representando um exemplo de alfa-hélice. Com o auxílio de dispositivo apropriado pode ser vista a conformação helicoidal da cadeia peptídica. A porção N-terminal fica em cima e a C-terminal, em baixo. Os oxigênios do grupo carboxila fazem ponte de hidrogênio com os hidrogênios do grupo amina do aminoácido situado quadro posições adiante (C-terminal). O passo desta hélice é 3.6 resíduos por volta. (Adaptado de PIMENTEL & SPRATLEY, 1974, p.611).

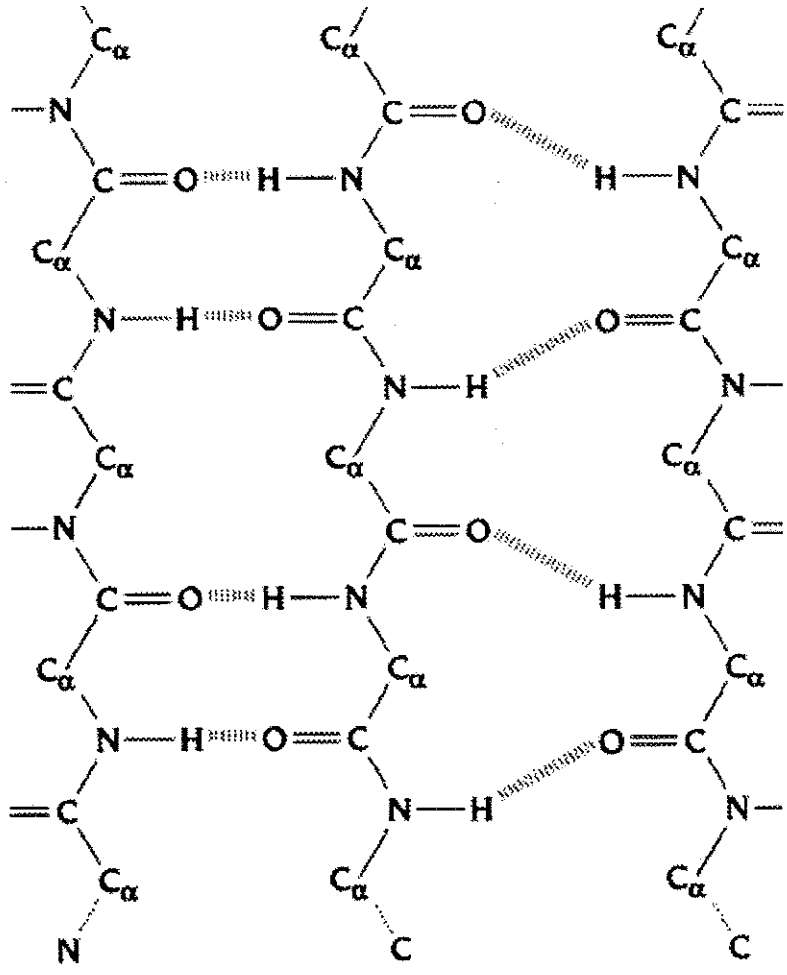


Figura 6 - Exemplo de placa pregueada beta. Três segmentos estendidos da cadeia peptídica se associam lado a lado, formando pontes de hidrogênio entre átomos da cadeia principal. Esta placa tem uma porção anti-paralela, onde as cadeias tem orientações opostas (a do meio "desce" e a da esquerda "sobe") e uma porção paralela, onde a orientação é a mesma (cadeias do meio e da direita "descem"). A cadeias laterais ligadas aos carbonos alfa (C_α) tem orientação perpendicular ao plano da placa (veja Fig. 11). Adaptado de BRANDEN & TOOZE (1991), p.18.

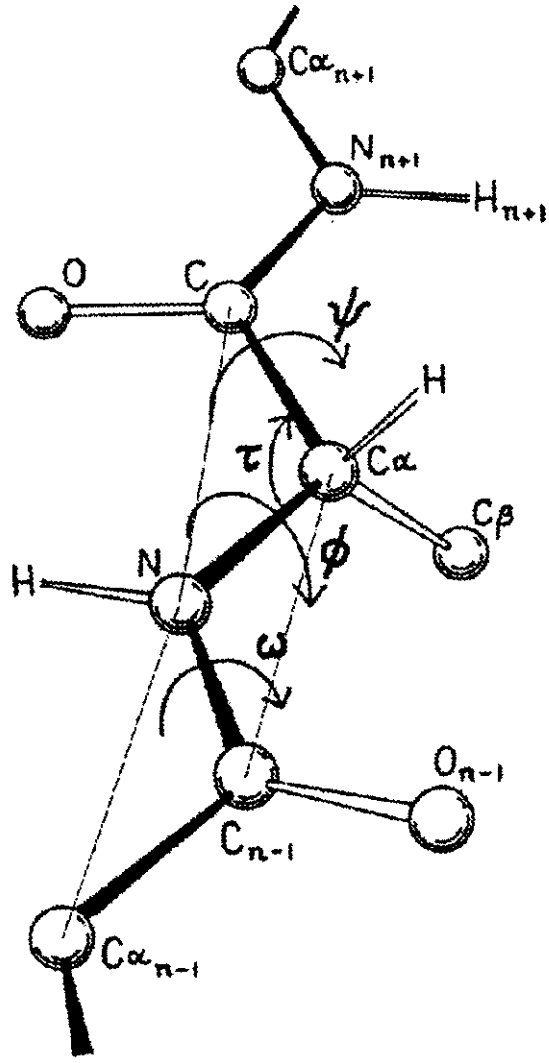


Figura 7 - Ângulos úteis para a compreensão da estrutura protéica. Representamos aqui um segmento da cadeia principal com três aminos-ácidos. A extemidade inferior é N-terminal. Podemos ver alguns átomos dos aminoácidos *n-1*, *n* e *n+1*. Os ângulos representados são do aminoácido central *n*. O nitrogênio que forma ligações peptídicas geralmente faz as três ligações usando orbitais híbridos sp^2, deixando um par de elétrons livres num orbital *p*. Este entra em ressonância com o orbital π da ligação dupla "C=O". A ligação peptídica tem portanto um caráter de dupla ligação, limitando a rotação da cadeia naquele ponto (RICHARDSON & RICHARDSON, 1989). Isto coloca os átomos $C\alpha_{n-1}$, C_{n-1}, N e $C\alpha_n$ num mesmo plano e faz com que o ângulo diédrico ω seja quase sempre próximo a 180 graus. O ângulo plano τ reflete a tensão na cadeia, o quanto C_α se desvia do formato tetraédrico (deformação) e não será tratado neste trabalho. Os ângulos diédricos ϕ e ψ mostram a mudança de direção da cadeia principal na região do C_α. Aquele entre o plano da ligação peptídica com o amino ácido precedente e o plano formado por N, C_α e C é ϕ. Da mesma forma, ψ é o ângulo formado entre o plano da ligação peptídica com o aminoácido seguinte e o plano de N, C_α e C. Adaptado de RICHARDSON & RICHARDSON (1989), p.4.

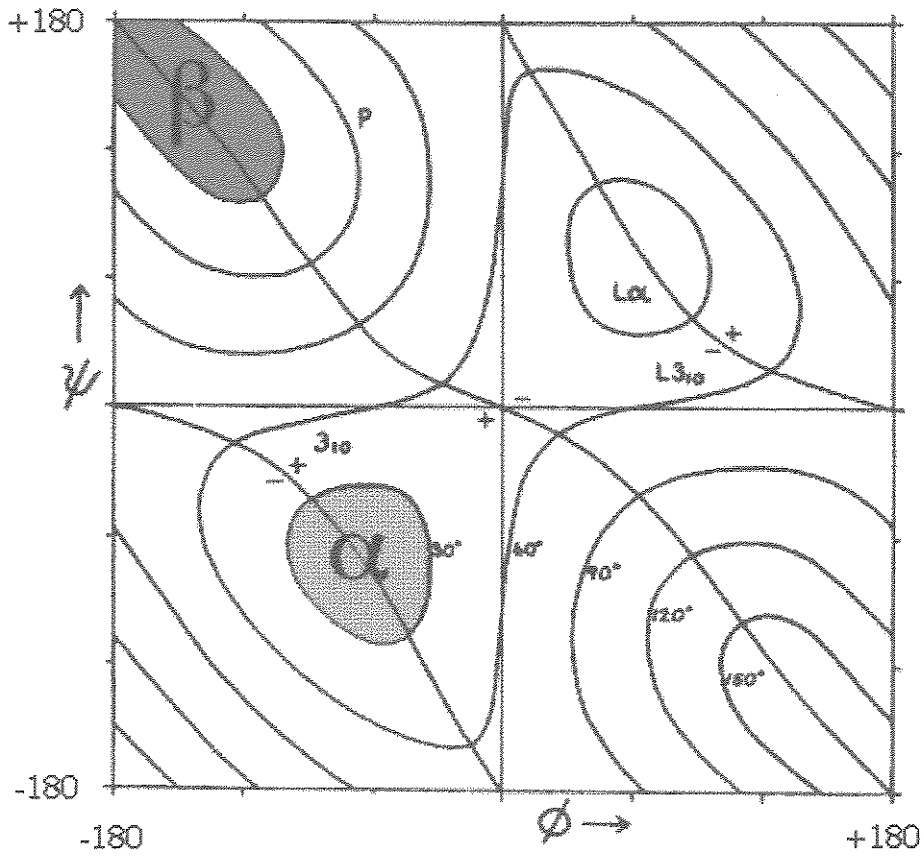


Figura 8 - Gráfico de Ramachandran. Temos os ângulos diédricos ϕ nas abscissas e ψ nas ordenadas (Vide Fig. 7). Foram demarcadas as regiões teoricamente correspondentes aos valores destes ângulos para resíduos nos estados alfa (vermelho) e beta (verde). Outros símbolos encontrados no gráfico são pouco relevantes para o presente trabalho: P = poliprolina, conformação estendida mas não beta, 3_{10} = hélice 1-3, onde as pontes se dão entre os resíduos *i* e *i+3*, L_α = alfa ao espelho (*left handed*) e $L3_{10}$ = hélice 1-3 ao espelho. Os contornos correspondem a ângulos múltiplos de 30º entre duas carbonilas ($C_{\alpha,n}$ e $C_{\alpha,n+1}$). O fato de um resíduo ter ângulos diédricos ocupando os valores demarcados não implica que esteja na estrutura secundária correspondente ou que faça todas as pontes de hidrogênio pressupostas em tais estruturas. Adaptado de RICHARDSON & RICHARDSON (1989), p.8.

regras para a conversão. Este método (KABSCH & SANDER, 1983) tende a subestimar a presença de estruturas regulares (NISHIKAWA & NOGUCHI, 1991). Uma vez que o modelo energético usado para detectar as pontes de hidrogênio se baseia principalmente na distância entre os átomos envolvidos, existe uma sensibilidade muito grande à resolução espacial e ao refinamento energético com que foi determinada a estrutura terciária. Modelos de melhor qualidade terão mais estrutura regular que modelos grosseiros (EISENBERG & HILL, 1989). Da mesma forma, o método deixa de identificar estruturas distorcidas ou com o padrão de pontes de hidrogênio incompleto, mas que ainda são perceptíveis aos olhos do cristalógrafo (NISHIKAWA & NOGUCHI, 1991). Existem ainda outras formas de identificar estrutura secundária, tais como matrizes de distância ou contato e uso de técnicas estatísticas, que fogem ao enfoque do presente trabalho (LEVITT & GREER, 1977; RICHARDS & KUNDROT, 1988; HUNTER & STATES, 1991).

A estrutura terciária, ou conformação tridimensional das proteínas, é geralmente determinada por cristalografia de difração de raios X (BLUNDELL & JOHNSON, 1976; HENDRICKSON, 1986; EISENBERG & HILL, 1989). Mais recentemente também tem sido utilizada a espectroscopia de ressonância magnética nuclear (NMR), combinada com modelagem molecular (*3D-NMR* e *4D-NMR*, CLORE & GRONENBORG, 1989; WÜTRICH 1986). Embora a discussão destas técnicas fuja aos objetivos do presente trabalho, são necessárias algumas noções dos procedimentos envolvidos, uma vez que a estrutura secundária é obtida a partir do resultado desses métodos.

A maioria das estruturas terciárias determinadas até o presente se encontram depositadas em formato eletrônico no *Protein Data Bank* (PDB) mantido pelo *Brookhaven National Laboratory*, E.U.A. (BERSTEIN *et al*, 1977). A cada uma delas corresponde um arquivo digital, com nome codificado. A estrutura interna deste arquivo é baseada no formato dos antigos cartões perfurados, para leitura por programas escritos em FORTRAN. Existem variados detalhes e idiossincrasias na estrutura destes arquivos e da base de dados como um todo. Algumas proteínas são representadas por vários arquivos, chegando mesmo a haver dezenas descrevendo uma única macromolécula. A diferença entre eles pode ser a qualidade, o autor, a presença de um ligante ou a substituição de um aminoácido na sequência. Certos tipos de polipeptídeos, como as globinas por exemplo, correspondem a uma fração significativa do PDB, embora sejam muito semelhantes entre si. Proteínas solucionadas por NMR têm diversas cadeias, com estruturas levemente diferentes, dentro do mesmo arquivo. Além de proteínas, existem muitos polinucleotídeos (DNA e RNA) e carbohidratos. Dentro dos arquivos existem, às vezes, diferenças entre a

sequência apresentada no início e aquela apresentada na porção reservada às coordenadas. Eventualmente, alguns resíduos apresentam múltiplas conformações. Embora complicada, a organização dos arquivos é consistente. Um arquivo retirado do PDB é mostrado no Apêndice 1.

Ambas as técnicas de determinação de estrutura terciária (raio X e NMR) dependem do conhecimento prévio da estrutura primária. No caso da cristalografia, o resultado depende da qualidade da estimativa inicial (ajuste arbitrário da sequência à densidade eletrônica), e do "refinamento energético" do modelo molecular (HENDRICKSON, 1986). Assim, o fato de uma proteína ter sua estrutura depositada no PDB não significa que esta seja isenta de erros. Havendo mais de uma estrutura para uma mesma sequência, há que se escolher a melhor.

1.2. Previsão de Estrutura Protéica

Nos organismos vivos, as proteínas são sintetizadas a partir de moldes polinucleotídicos (RNA, que por sua vez foi moldado do DNA; STRYER, 1981). Técnicas de DNA recombinante permitem a rápida definição da sequência de nucleotídeos de um gene e, consequentemente, da estrutura primária da proteína correspondente (WATSON, 1992). O número de proteínas com estrutura primária conhecida tem crescido exponencialmente (BAIROCH & BOECKMANN, 1992), enquanto o número de estruturas terciárias resolvidas progride em ritmo muito mais lento (BERNSTEIN *et al*, 1977). Acredita-se que a estrutura primária determina a estrutura tridimensional, que por sua vez é responsável por todas as propriedades de um polipeptídeo (ANFINSEN, EPSTEIN, GOLDBERGER, 1963). Existe portanto um grande esforço para prever a conformação de uma cadeia protéica a partir de sua sequência de aminoácidos. Tal problema se apresenta como um dos maiores desafios da biologia molecular (BRANDEN & TOOZE, 1991).

Da mesma forma que a estrutura protéica, os esforços de previsão têm diferentes níveis de complexidade. Enquanto alguns autores se dedicaram aos níveis mais simples, como estrutura secundária (e.g. CHOU & FASMAN, 1974; GIBRAT *et al.*, 1987; ROST & SANDER, 1993), ou simplesmente classe estrutural (LEVITT & CHOTTIA, 1976; LEVITT & GREER, 1977; SHERIDAN *et al*, 1985; KLEIN, 1986; KLEIN & DELISI, 1986; NAKASHIMA, NISHIKAWA & OOI, 1986; DELÉAGE & ROUX, 1987), outros tentaram determinar a estrutura terciária, inclusive a posição de todos os átomos das cadeias laterais, a partir da sequência de aminoácidos (e.g. NEMETHY & SCHERAGA, 1977; BROOKS *et al.*, 1983; BRUCCOLERI &

KARPLUS, 1987; COHEN & KUNTZ, 1989; GUNSTEREN, 1993). Dada a importância da estrutura para o comportamento bioquímico da proteína, atuam neste campo diferentes tipos de especialistas, com diferentes tipos de interesse. Há cristalógrafos que cuidam da determinação precisa da estrutura (e.g. EISENBERG & HILL, 1989). Há químicos teóricos, que estudam o processo de dobramento em si e sua termodinâmica (e.g. BALDWIN, 1986; FREIRE *et al.*, 1992). Há os que atuam como especialistas em química computacional e aplicam um poder de cálculo cada vez maior na minimização de funções de energia (e.g. BROOKS *et al.*, 1983; BRUCCOLERI & KARPLUS, 1987; GUNSTEREN, 1993). Há biologistas moleculares, que desejam uma interpretação estrutural para o comportamento de mutantes que diferem apenas em um ou outro aminoácido da estrutura primária (SHAKHNOVICH & GUTIN, 1991; ZABIN, HORVATH, TERWILLIGER, 1991; LATTMAN & ROSE, 1993). Seja qual for o nível de complexidade em que se deseje fazer a previsão, existem ainda os que se interessam pelos aspectos computacionais, matemáticos ou estatísticos do problema (e.g. KLEIN & DELISI, 1986; QIAN & SEJNOWSKI, 1989). A razão para tamanha atividade neste campo está na abundante disponibilidade de sequências genéticas e protéicas. Um método que permitisse prever a estrutura, mesmo que fosse apenas a secundária, a partir da sequência de aminoácidos, poderia elucidar os mecanismos moleculares da vida: o papel da estrutura protéica nas funções enzimáticas, estruturais e de transdução de sinais, as bases genéticas das doenças, as funções fisiológicas, partindo do nível molecular, chegando até o sistêmico, passando pelo celular. Em resumo, a solução do problema de previsão de estrutura protéica, ou do problema de dobramento de proteínas a ele relacionado, levaria a biologia molecular a uma nova fase de desenvolvimento.

Este trabalho se concentra na previsão de estrutura secundária de proteínas, um problema que tem desafiado a comunidade de inteligência artificial e de aprendizado de máquina (GARNIER & LEVIN, 1991). As primeiras tentativas de previsão foram marcadas pela simplicidade. Entre os pioneiros, CHOU & FASMAN (1974a, 1974b, 1978) desenvolveram um método que logo se popularizou, tanto por dispensar o auxílio do computador, quanto por terem sido apresentadas previsões razoavelmente acuradas pelos autores. Este método se saiu melhor que os "concorrentes" nos "torneios" de previsão de estrutura secundária realizados então (SCHULZ *et al.*, 1974; MATTHEWS, 1975). O pequeno número de estruturas tridimensionais conhecidas na época e a não utilização de técnicas de validação permitiam que houvesse um certo otimismo (ou ingenuidade) quanto à solução do problema. SCHULZ & SCHIRMER (1979) anunciaram, prematura e erroneamente, que os métodos existentes então eram capazes de prever a estrutura secundária de dois terços (66.7%) dos resíduos de uma sequência, e que isto seria suficiente para inferir corretamente a classe estrutural da

proteína (segundo a classificação de LEVITT & CHOTHIA, 1976). Outros métodos pioneiros que merecem citação são os de LIM (1974a e b), MAXFIELD & SCHERAGA (1976) e aquele popularmente denominado GOR (GARNIER, OSGUTHORPE, ROBSON, 1978). Com o crescimento do conjunto de proteínas com estrutura tridimensional conhecida, ficou claro que o poder de previsão dos métodos desenvolvidos na década de 70 não era tão grande assim (KABSCH & SANDER, 1983b). Além disso, o método de CHOU & FASMAN (1974a, 1974b) se mostrou ambíguo e teve de ser modificado para que fosse programado em computador (CHOU & FASMAN, 1978a, 1978b). Quando estes métodos foram testados em novas proteínas, comparados entre si e mesmo combinados, não ultrapassaram a marca dos 56% de acurácia (KABSCH & SANDER, 1983b). Note que isto corresponde, aproximadamente, a mesma percentagem de acerto que obteríamos se supuséssemos que não há estrutura secundária alguma (cerca de 50% dos resíduos são do tipo *coil*, Tab. 1)

A década de 80 foi marcada por uma proliferação de métodos, em geral mais aprimorados, tentando melhorar a qualidade de previsão de estrutura secundária. Diversas técnicas foram aplicadas ao problema: teoria da informação (GARNIER *et al.*, 1978; GIBRAT, GARNIER, ROBSON, 1987; BIOU *et al.*, 1988), redes neurais (HOLLEY & KARPLUS, 1989; QIAN & SEJNOWSKI, 1989; SASAGAWA & TAJIMA, 1993), redes em cascata (KNELLER, COHEN, LANGRIDGE, 1990; NISHIKAWA, 1990; VISVANADHAN, DENCKLA, WEINSTEIN, 1991), sistemas híbridos (ZHANG, MESIROV, WALTZ, 1992), métodos de máxima vizinhança (SALZBERG & COST, 1992), cadeias de Markov (ASAI, HAYAMIZU, HANDA, 1993), aprendizado de máquina (KING & STERNBERG, 1990; STERNBERG *et al.*, 1992), informação mútua (STOLORZ, LAPEDES, XIA, 1992), etc. A maioria destes métodos obteve uma acurácia menor que 65% (quando devidamente validado por cruzamento ou eliminadas semelhanças entre proteínas dos conjuntos de treino e de teste). De fato, até recentemente (ROST & SANDER, 1993), este valor parecia constituir numa barreira intransponível aos métodos preditivos (STERNBERG,1992; RACKOVSKY, 1993). Entre eles, destacou-se o método de GIBRAT *et al.* (1987), também denominado GGR, pelo cuidado com que o conjunto de dados foi montado, pelo modo como foi validado e por sua popularidade e aceitação. Tratava-se de uma extensão do método GOR (GARNIER *et al* , 1978).

1.2.1. O Método GGR

Este método utilizou a teoria da informação (SHANNON & WEAVER 1949; BRILLOUIN, 1956), procurando determinar a quantidade de informação contida numa sequência de símbolos com diferentes probabilidades *a priori*. Sejam **x** e **y** dois eventos, **P(x|y)** a probabilidade condicional de **x** ocorrer dado que **y** ocorreu e **P(x)** a probabilidade de **x**. A informação que **y** contém sobre a ocorrência de **x** é:

$$\mathbf{I(x;y)=\log[P(x|y)/P(x)]} \qquad (1)$$

Isto foi devidamente estendido para um número maior de eventos de forma a calcular a informação que a sequência de aminoácidos $\mathbf{R_{i-8}}$, $\mathbf{R_{i-7}}$, ... , $\mathbf{R_{i+7}}$, $\mathbf{R_{i+8}}$ contém sobre a conformação **X** do resíduo central $\mathbf{R_i}$:

$$\mathbf{I(S_i=X;R_{i-8}, ..., R_{i+8})} \qquad (2)$$

A expressão necessária para o cálculo da equação (2) tem 2 × 8 + 1 = 17 parcelas. O primeiro termo necessita de uma tabela de contingência de 20 células. Como são três estados (alfa, beta e *coil*) este número sobe para 60 células. Os termos seguintes envolvem pares de aminoácidos, de forma que cada um necessitaria de 3 × 20 × 20 = 1200 células. Como a base de dados utilizada continha pouco menos de 12000 resíduos, os autores tiveram que lançar mão de aproximações, e se limitaram a considerar interações entre pares de aminoácidos (2 parcelas no máximo). Duas expressões alternativas foram usadas:

$$I(S_j = X; R_{j-8}, \ldots, R_{j+8}) = \sum_{m=-8}^{m=+8} I(S_j = X; R_{j+m}) \qquad (3)$$

denominada "informação direcional", pois se refere a influência do resíduo na posição **j+m** sobre a conformação da posição **j**, independente da identidade de $\mathbf{R_j}$, e

$$I(S_j = X; R_{j-8}, \ldots, R_{j+8}) = \\ = I(S_j = X; R_j) + \sum_{m=-8}^{m=+8} I(S_j = X; R_{j+m} | R_j) \qquad (4)$$

formada por duas parcelas: $\mathbf{I(S_j=X;R_j)}$, denominada "informação própria", pois reflete a influência da identidade do resíduo sobre sua própria conformação e $\mathbf{I(S_j=X;R_{j+m}|R_j)}$, denominada "informação pareada", pois reflete a interação do par

$\mathbf{R_{j+m}|R_j}$ (influência do resíduo na posição **j+m** sobre a conformação do resíduo na posição **j**, levando em conta a identidade de ambos).

Se as proteínas fossem constituídas por proporções equivalentes de aminoácidos haveriam em média 10 entradas por célula nas tabelas necessárias para estimar suas "informações". Existem, contudo, aminoácidos raros, que levaram a um grande número de células vazias ou pouco habitadas. Os autores contornaram este problema através de dois artifícios. O primeiro foi a seleção de variáveis. Testou-se a independência entre a conformação de $\mathbf{R_j}$ e a identidade do resíduo da posição **j+m** através de χ**2**. Quando o valor de **P** foi superior a um valor arbitrário $\mathbf{P_{max}}$ foi usada a Eq.(3) ao invés da Eq.(4) ("informação direcional" ao invés de "informação pareada").

O segundo artifício foi a inclusão de "observações fictícias" (*dummy observations*), aproximações feitas para pares de aminoácidos pouco frequentes. Nestas casos a Eq.(4) foi aproximada para:

$$I(S_j = X; R_{j-8}, \ldots, R_{j+8}) = \\ = I_o(S_j = X; R_j) + I_D(S_j = X; R_{j+m} | R_j) \quad (5)$$

onde

$$I_D(S_j = X; R_{j+m} | R_j) \approx \\ I(S_j = X; R_j) + I(S_j = X; R_{j+m}) \quad (6)$$

A nova "informação total" inclui a informação observada I_o mais uma parcela I_D calculada para **M** "observações fictícias", contendo a "informação própria" mais a "informação direcional", extrapolando as propriedades observadas quando os resíduos $\mathbf{R_j}$ e $\mathbf{R_{j+m}}$ estavam presentes isoladamente para a situação em que eles interagem, como se as observações fossem independentes.

Além de tudo isto foram adicionadas "constantes de decisão" arbitrárias ao valores estimados. As equações utilizadas resultaram então:

$$\mathbf{I_H(S_j{=}H{:}H, R_{j-8}, \ldots, R_{j+8}) - DC_H} \quad (7)$$

$$\mathbf{I_E(S_j{=}E{:}E, R_{j-8}, \ldots, R_{j+8}) - DC_E} \quad (8)$$

$$\mathbf{I_C(S_j{=}C{:}C, R_{j-8}, \ldots, R_{j+8})} \quad (9)$$

Onde $\mathbf{I_H}$, $\mathbf{I_E}$ e $\mathbf{I_C}$ são respectivamente as "quantidades de informação" que a sequência $\mathbf{R_{i-8}}$, ... , $\mathbf{R_{i+8}}$ tem para que a conformação do resíduo **i** seja alfa, beta ou nenhum dos dois. O maior valor indica o estado mais provável.

Os autores ajustaram por tentativa e erro os valores de **M**, $\mathbf{P_{max}}$, $\mathbf{DC_H}$ e $\mathbf{DC_E}$, escolhendo aqueles que resultavam num maior número de resíduos preditos corretamente. Para estimar o sucesso do método em proteínas que não estavam na base de dados, foi realizada validação por cruzamento: cada cadeia protéica foi retirada do conjunto e as tabelas de contingência, necessárias para o cálculo das "informações", foram construídas com as demais proteínas. O método foi capaz de prever com sucesso 63% dos resíduos (69.7% sem a validação) para $\mathbf{P_{max}}$=1, **M**=255, $\mathbf{DC_H}$ =25 e $\mathbf{DC_E}$=25.

1.2.2. Redes Neurais

Entre os muitos métodos já citados, talvez os mais populares durante a década de 80 e início da década de 90 foram os que utilizavam redes neurais artificiais. Alguns destes métodos clamaram resultados melhores que os 63% de GGR (QIAN & SEJNOWSKI, 1989; SASAGAWA & TAJIMA, 1993; ROST & SANDER, 1993). Com exceção do método de ROST & SANDER (1993), estas cifras não resistiram a uma inspeção mais cuidadosa, nem à utilização em novas estruturas. As principais falácias que reduziram o sucesso de tais métodos aos mesmos aproximados 63% de acurácia de GGR foram: (1) Ausência de validação cruzada. A maioria utilizava a metodologia usual de conjunto de teste *versus* conjunto de treino (QIAN & SEJNOWSKI, 1989; SASAGAWA & TAJIMA, 1993). Os trabalhos que utilizaram de validação cruzada (HOLLEY & KARPLUS, 1989; KNELLER, COHEN, LANGRIDGE, 1990) obtiveram resultados em torno de 63%. (2) Edição do resultado, eliminando segmentos sem sentido, tais como alfa-hélices com menos de 4 resíduos ou interrompidas (KNELLER, COHEN, LANGRIDGE, 1990; SASAGAWA & TAJIMA, 1993; ROST & SANDER, 1993). Este procedimento, quando incluído em outros métodos, também eleva significativamente a acurácia (HOLLEY & KARPLUS, 1991). Para comparar a eficiência de dois ou mais métodos é recomendável a utilização da mesma base de dados, obtidas a partir da mesma versão do PDB, com resultado validado da mesma forma, usando os mesmos critérios de sucesso.

Em setembro de 1993, quando boa parte deste trabalho já havia sido realizada, foi publicado método preditivo que realmente ultrapassou a acurácia de GGR (ROST & SANDER, 1993). Embora tal método utilizasse redes neurais artificiais e padecesse dos mesmos problemas (ausência de validação cruzada, que, neste caso, foi ao menos parcialmente realizada e edição dos resultados) se notabilizou pela maneira como

contornou o maior obstáculo: a escassez de dados. Ao invés de extrapolar resultados, supondo independência de eventos claramente dependentes, como foi feito em GGR, estes autores utilizaram sequências muito similares às das proteínas de estrutura conhecida, para ampliar a bases de dados. A suposição de que sequências muito parecidas devem ter a mesma estrutura secundária é bastante plausível (SCHNEIDER & SANDER, 1991). Este método também contém procedimentos inovadores, que dificultam sua comparação com os resultados da literatura. É provável que tais processos, tais como o embaralhamento das sequências para evitar resultados ruins e a previsão das mesmas sequências várias vezes, usando-se depois uma espécie de média para escolher os resultados ("método do juri", ROST & SANDER, 1992), aumentem a acurácia de qualquer procedimento a que sejam incorporados.

1.3. Hidrofobicidade e Periodicidade

Além dos métodos mais consagrados, citados e descritos acima, que representam o veio principal da pesquisa nesta área, são relevantes outros trabalhos menos conhecidos. Primeiro, devemos ressaltar os esforços de muitos pesquisadores em utilizar escalas de propriedades físicoquímicas dos aminoácidos (revistos em KIDERA *et al.*, 1985). Esta abordagem tem múltiplas origens. Existe uma teoria entre os que trabalham com "dobramento de proteínas" (*protein folding*, processo de enovelamento da cadeia principal, até que atinja a estrutura tridimensional final) de que a principal força que comanda este processo seria o "efeito hidrofóbico", isto é, resíduos poucos solúveis e apolares seriam escondidos no interior da proteína e resíduos polares seriam expostos (RICHARDS, 1977; PRIVALOV & GILL, 1988; SPOLAR, JEUNG-HOI, RECORD, 1989). Elementos de estrutura secundária costumam segregar diferentes tipos de resíduos, dependendo de sua participação na arquitetura do glóbulo protéico. Existem hélices anfipáticas, com resíduos polares de uma lado e apolares de outro (SEGREST *et al.*, 1990), assim como placas beta cobertas em cada lado por um tipo de aminoácido (CHOTHIA & JANIN, 1982). Assim, muitos autores procuraram usar escalas de hidrofobicidade (bem como de outras propriedades, tais como polaridade, tamanho, etc.) em seus métodos preditivos (CHOU & FASMAN, 1974a e b; SHERIDAN *et al.*, 1985; KLEIN & DELISI, 1986; LAMBERT & SCHERAGA, 1989). O uso dessas escalas permite transformar dados discretos (identidade de aminoácidos) em contínuos (valor numérico). Outra possibilidade explorada foi a de "reduzir o alfabeto" de 20 aminoácidos para um número menor de classes (4 a 6 tipicamente), na tentativa de simplificar o problema (e.g. CHARTON & CHARTON, 1982).

Estas escalas foram extensamente revistas por KIDERA *et al.* (1985), que procuraram buscar um conjunto representativo de propriedades linearmente independentes. Este conjunto foi utilizado em método preditivo (LAMBERT & SCHERAGA, 1989) que, como tantos outros, ficou em torno dos 63% de acurácia. Estas escalas não foram geradas somente por pesquisadores interessados em utilizá-las na previsão de estrutura protéica. Duas outras linhas de pesquisa geraram grande parte destas tabelas: (1) Compilação de propriedades químicas dos 20 aminoácidos em laboratório, tais como a solubilidade em determinado solvente (e.g. WOLFENDEN *et al.*, 1981). Na medida em que estruturas tridimensionais foram sendo reveladas, procurou-se uma corroboração desses achados experimentais. A idéia envolvida é de que os contatos entre aminoácidos, ou sua exposição ao solvente aquoso, são diretamente influenciados por estas propriedades (CHOTHIA, 1976). (2) Como a correspondência entre esta dupla "contato/exposição" e as escalas físco-químicas não foi perfeita, outros pesquisadores procuraram utilizar as estruturas tridimensionais conhecidas para obter tabelas empíricas, agora não mais correspondentes a uma propriedade físicoquímica específica, mas baseadas nos contatos entre resíduos e o seu grau de exposição ao solvente (e.g. MANALAVAN & PONNUSWAMY, 1978; NARAYANA & ARGOS, 1984; HERINGA & ARGOS, 1991).

A escala de hidrofobicidade foi empregada ainda de outra forma por pesquisadores interessados em prever não a estrutura secundária, mas a classe estrutural a que pertencem (KLEIN, 1986; KLEIN & DELISI, 1986). Trata-se de problema supostamente mais simples, embora diretamente relacionado. A classe estrutural depende da quantidade e distribuição dos elementos de estrutura secundária ao longo da sequência e sua posição no espaço (LEVITT & CHOTHIA, 1976). Estes autores usaram a análise de Fourier para detectar a periodicidade de propriedades físicoquímicas (essencialmente hidrofobicidade, EISENBERG, WILCOX, MCLACHLAN, 1986; CORNETTE *et al.*, 1987). Estavam principalmente interessados nos períodos correspondentes ao passo das alfa-hélices (3.6) e à alternância de resíduos em cordões beta (2.0). Se resíduos de um mesmo tipo tendem a ocupar um mesmo lado da estrutura, eles estarão separados pelo número correspondente de posições na sequência (3.6 ou 2.0) e a propriedade físicoquímica correspondente deve estar distribuída da mesma forma. As Figuras 9 a 11 mostram o raciocíno envolvido nesta abordagem.

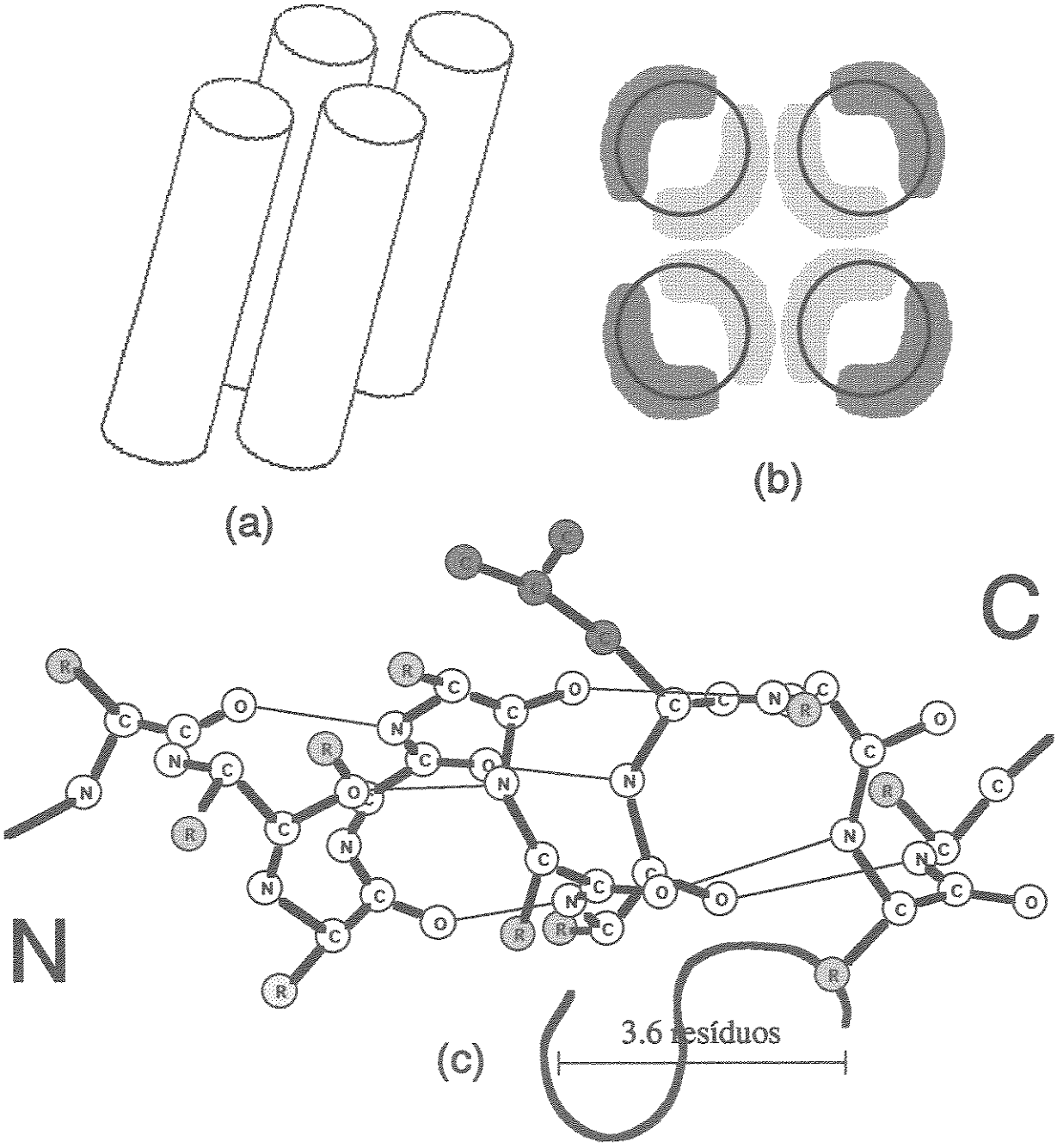


Figura 9 - Anfipaticidade das alfa-hélices (SEGREST *et al.*, 1990). (a) Um feixe de quatro alfa-hélices. Cada cilindro representa uma hélice. (b) Vista superior do mesmo feixe. As faces "escondidas" das hélices interagem entre si e não com o solvente, sendo tipicamente recobertas por cadeias apolares (azul). As faces "expostas" entram em contato com o solvente aquoso e são recobertas por cadeias laterais polares (vermelho). (c) Modelo em bolas e barras de uma alfa-hélice, mostrando que o passo da mesma é de aproximadamente 3.6 resíduos. Isto faz com que as propriedades dos aminoácidos (hidrofobicidade, polaridade) se distribuam com a mesma periodicidade ao longo da sequência. Desenho feito pelo candidato.

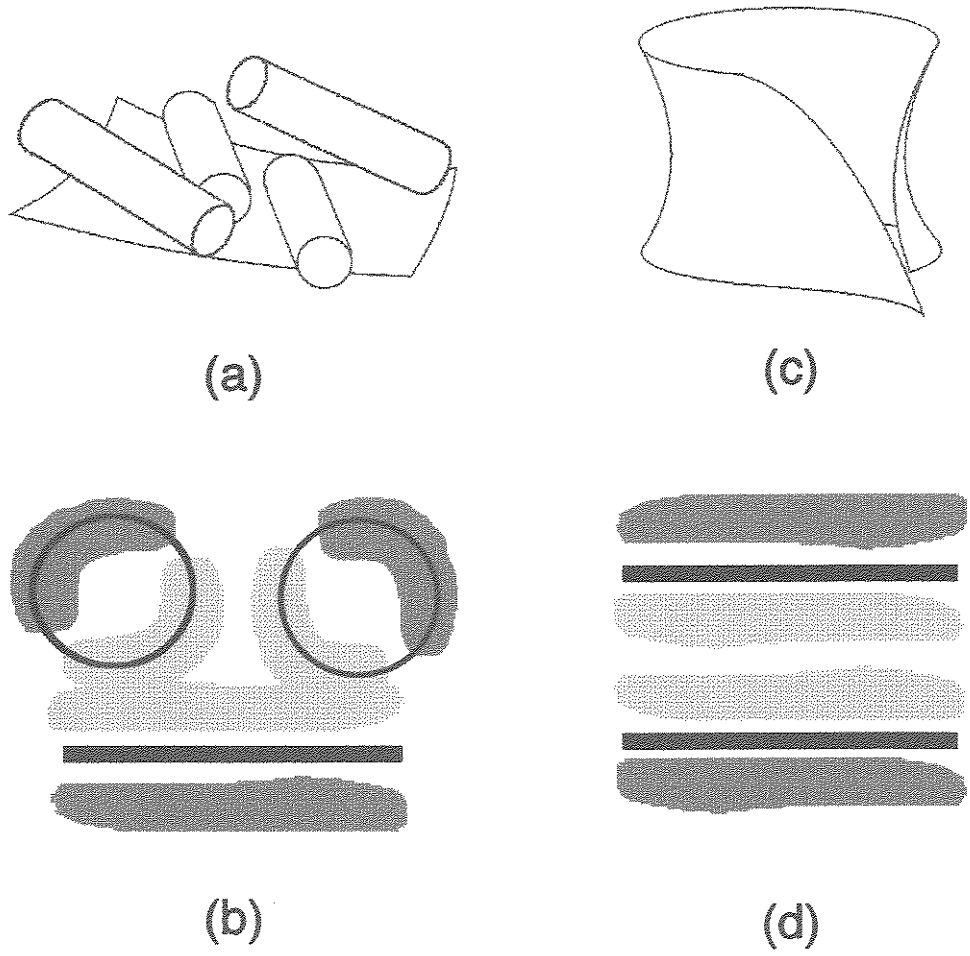


Figura 10 - Anfipaticidade das placas beta (CHOTHIA & JANIN, 1982). (a) Modelo de uma proteína com uma placa beta com uma das faces exposta a solvente e outra recoberta por alfa-hélices. (b) Vista lateral simplificada da mesma proteína. As partes escondidas são hidrofóbicas (apolares, em azul) e as expostas, hidrofílicas (polares, em vermelho) (c) Modelo de uma placa beta que forma um "barril" ou "sanduíche". (d) Vista superior simplificada da mesma estrutura. As faces escondidas são hidrofóbicas (azuis) e as expostas, hidrofílicas (vermelho). A maneira como esta anfipaticidade se reflete na sequência é ilustrada na Figura 11. Desenho feito pelo candidato.

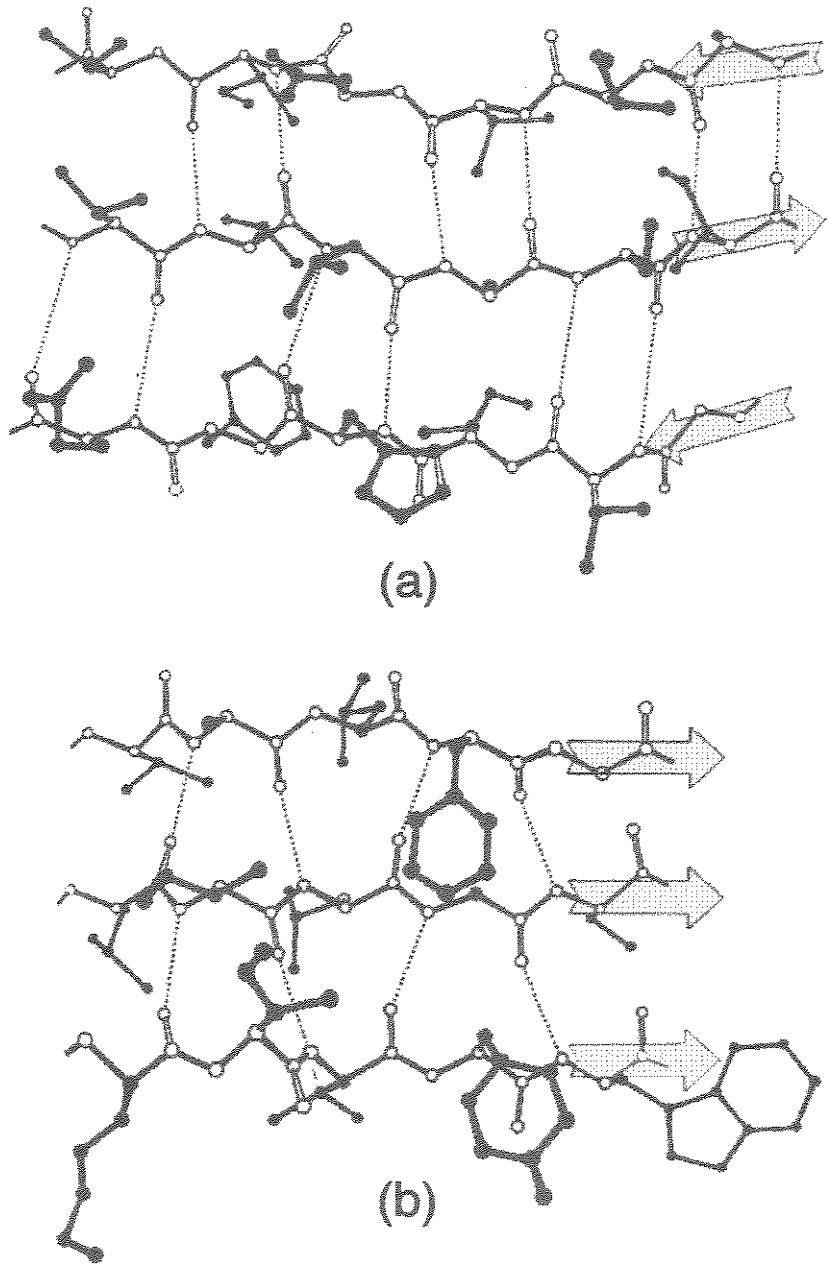


Figura 11 - Anfipaticidade das placas beta (CHOTHIA & JANIN, 1982). (a) Modelo em "bolas e barras" de uma placa beta antiparalela. Note-se como as cadeias laterais se alternam na ocupação das faces: as cadeias coloridas de vermelho ocupam a face superior, enquanto as azuis, a inferior. Se esta placa for anfipática, as cadeias de cores diferentes terão propriedades físicoquímicas diferentes (e.g. hidrofobicidade). O período de alternância das cores ou propriedades ao longo da sequência é de 2.0 resíduos. (b) Modelo em "bolas e barras" de uma placa beta paralela, onde vale o mesmo raciocínio. Adaptado de RICHARDSON & RICHARDSON (1989), p.18-19.

A Figura 9 mostra que, a menos que uma alfa-hélice esteja totalmente exposta a solvente, ou totalmente envolvida por outras porções da proteína, ela terá uma face hidrofílica e outra hidrofóbica. De fato, a maior parte das hélices são anfipáticas. As Figuras 10 e 11 mostram que placas beta que estejam parcialmente expostas a solventes também serão anfipáticas. Os autores que utiliziram este artifício em suas previsões notaram que o fenômeno é muito menos significativo no caso beta, devido ao grande número de placas exclusivamente hidrofóbicas (KLEIN & DELISI, 1986; HUNTER & STATES, 1991).

Enquanto é quase certo que exista uma regra complexa, que permita prever a estrutura secundária a partir da sequência (ANFINSEN *et al,* 1963), métodos estatísticos ou de inteligência artificial parecem estar impedidos de descobrí-la pela limitação dos dados disponíveis (RACKOVSKY, 1993). É provável que o problema só seja resolvido quando uma maior quantidade de conhecimento *a priori* assim como informações a respeito da físicoquímica de proteínas sejam incluídos na máquina preditiva (STERNBERG, 1992).

1.4. Modelos Logísticos e Máxima Verossimilhança

Apesar do grande iinteresse da comunidade de ciências da computação pelo problema, poderosas técnicas de estatística foram ignoradas. Nos pareceu que as soluções empíricas adotadas por GIBRAT *et al.* (1987) mereciam uma abordagem mais moderna e rigorosa. Esse método utilizou as frequências dos aminoácidos como aproximação para as probabilidades (quantidades de informação). Preferimos tratar o problema como uma regressão ou classificação, usando técnicas de otimização do ajuste de um modelo definido, com número de parâmetros e forma da equação de regressão variáveis. Nossa abordagem foi generalizar um procedimento muito conhecido de análise discriminante (AGRESTI, 1990), utilizando uma metodologia híbrida de modelos paramétricos (logísticos lineares e quadráticos) com modelos não paramétricos (técnicas independentes de modelos). Estes sistemas híbridos ou semiparamétricos tem sido extremamente úteis em outras aplicações (GUARDABASSO, MUNSON, RODBARD, 1988). Neles podemos incluir variáveis sabidamente importantes e ainda assim permitir que estruturas gerais, inesperadas, surjam dos dados.

Modelos logísticos são convenientes para a análise de dados que devem ser separados em categorias (BISHOP, FIENBERG, HOLLAND, 1975; AGRESTI 1990), tendo vantagens comprovadas na análise discriminante (SEBER, 1984). Eles surgem

naturalmente quando a distribuição dos resíduos em cada posição de uma janela local é independente, embora tal suposição não seja necessária (STOLORZ *et al*, 1992). Em termos operacionais, modelos logísticos podem incorporar tanto variáveis contínuas quanto categóricas (JOHNSON & WICHERN, 1988).

No método de GIBRAT *et al.* (1987), $\mathbf{P_{max}}$, **M**, $\mathbf{DC_H}$ e $\mathbf{DC_E}$ foram selecionados por tentativa e erro. Ao invés de utilizarmos o número de resíduos preditos corretamente, usamos uma medida quantitativa do ajuste do modelo aos dados: o princípio da máxima verossimilhança (MV). Expressões de verossimilhança são simplesmente o produto das probabilidades marginais. A seleção de um conjunto de parâmetros que maximize tal expressão garante o ajuste ótimo do modelo aos dados (JOHNSON & WICHERN, 1988; EUBANK, 1988). Foi suposto que a estrutura secundária surge de maneira estocástica, com probabilidades fixas, desconhecidas e independentes de se formarem alfa-hélices, cordões beta ou estruturas estendidas (*coil*). O estado mais provável, assim como a confiança da previsão, puderam ser obtidos a partir das estimativas de máxima verossimilhança (EMV) para os parâmetros do modelo. Embora MV seja largamente utilizada pela comunidade estatística, não havia sido empregada no contexto de previsão de estrutura secundária de proteínas (BRYANT & LAWRENCE, 1993). Modelos logísticos tem expressões de verossimilhança muito simples, facilitando o cálculo (AGRESTI, 1990).

O uso de aproximações para extrapolar a informação relacionada aos aminoácidos menos frequentes (GIBRAT *et al.*, 1987), reflete uma discrepância entre a complexidade do modelo (número de parâmetros) e a quantidade de dados disponíveis (AGRESTI, 1990). Lançamos mão de outros meios para redução do número de parâmetros. Um deles foi a penalização da máxima verossimilhança (GOOD & GASKINS, 1971; SIMMONOFF, 1983; TITTERINGTON & BOWMAN, 1985; EUBANK, 1988). Outro foi a seleção gradual de parâmetros (DRAPER & SMITH, 1981).

Nossa abordagem foi portanto um exercício empírico de modelagem, onde o modelo resultou tão complexo quanto necessário para descrever os dados, com uma quantidade de parâmetros adequada ao tamanho da amostra, permitindo a generalização do resultado.

2. Objetivos

Nossos objetivos foram a formulação e o ajuste de um modelo linear logístico e quadrático a dados discretos (sequências), classificados em três categorias, para previsão de estrutura secundária de proteínas. A revisão dos métodos existentes faz suspeitar que o modelo linear seria insuficiente para atingir uma acurácia significativa e que a quantidade de dados seria insuficiente para estimar todos os parâmetros de um modelo quadrático, ainda que muito restrito. Portanto, propusemo-nos a encontrar o número adequado de parâmetros, através de imposição de restrições ao modelo quadrático, penalização da máxima verossimilhança e seleção dos parâmetros mais significativos. Além disso, realizamos validação cruzada para garantir a generalidade.

Mais que resolver um problema que vem desafiando a comunidade de química teórica e inteligência artificial há muito tempo, ou acrescentar mais um aos inúmeros métodos de previsão de estrutura secundária existentes, queríamos abordar o problema através de metodologia estatística rigorosa, coisa que ainda não havia sido feita. Somente o uso de um modelo, onde o número de parâmetros (ou sua equivalência num modelo não paramétrico ou semi-paramétrico), fosse adequado ao tamanho da amostra, que tivesse sido ajustado de modo ótimo, com validação cruzada para estimar o sucesso em novas proteínas, poderia garantir que toda a informação existente fora utilizada, dentro da complexidade possível.

Esperávamos, ainda, que os parâmetros fossem interpretáveis do ponto de vista físicoquímico, confirmando a generalidade do método e revelando novos padrões nos dados.

3. Material e Métodos

3.1. *Hardware*

Foi utilizado um computador Convex C3830 (*Division of Computer Resources and Technology, National Institutes of Health,* Bethesda, Maryland, E.U.A) capaz de processamento paralelo e vetorização. Como terminal de acesso e para executar rotinas mais simples foram utilizados microcomputadores, em geral do tipo Macintosh. De fevereiro de 1993 até o presente o acesso foi feito via Internet, exceto em janeiro e outubro de 1994, quando o candidato esteve no NIH para finalizar o trabalho.

3.2. *Software*

Foram utilizadas rotinas escritas em MATLAB (1993), versão para ConvexOS (UNIX), e FORTRAN 77. Para seleção de parâmetros foi utilizado o pacote JMP (1994), para Macintosh.

3.3. Os Dados

Para podermos comparar nossos resultados com os de GIBRAT *et al.* (1987) foi utilizado o mesmo conjunto de dados, obtido a partir de uma seleção de proteínas do *Protein Data Bank*, mantido pelo *Brookhaven National Laboratory*. (BNL; versão de Outubro de 1993; BERSTEIN *et al* , 1977). O conjunto contém 67 cadeias protéicas, cuja estrutura foi resolvida com alta resolução e qualidade (2.8 angstrons ou menos), sem que houvesse grande similaridade entre as sequências de aminoácidos (50% ou menos). São 11208 resíduos, 29.1% deles no estado alfa-hélice, e 21.4% no estado beta. Haviam 457 hélices e 313 cordões beta. O conjunto é descrito na Tabela 1.

Os arquivos, contendo estruturas tridimensionais, na forma de coordenadas de cada átomo da cadeia, estavam presentes no disco rígido do computador utilizado. Os mesmos arquivos podem ser obtidos através da *Internet* de muitas fontes, entre elas o proprio BNL .

A estrutura secundária foi calculada pelo método de KABSCH & SANDER (1983), utilizando o programa original, DSSP, distribuído pelos autores, que também se encontrava disponível no disco rígido do computador utilizado. O programa foi escrito em PASCAL e teve de sofrer algumas modificações para poder ser compilado no Convex, a fim de ler os arquivos do diretório apropriado, ler a lista de arquivos PDB fornecida a ele e para que fossem desativadas diversas de suas subrotinas, que faziam cálculos demorados e irrelevantes para o presente trabalho.

Já havíamos composto um programa em FORTRAN para integrar os dados obtidos do PDB e do programa DSSP quando implementamos um outro método preditivo. Tal programa, denominado GETPTR.F (Apêndice 2), monta um arquivo com mais dados sobre a estrutura secundária do que o presente trabalho necessita, mas não vimos necessidade de modificá-lo. A Figura 12 mostra uma versão alfanumérica do arquivo montado. Como a linguagem MATLAB necessita de arquivo numérico, a versão realmente utilizada é uma codificação do mesmo conteúdo (Fig. 13).

Tabela 1 - O Conjunto de dados. Cadeias protéicas retiradas do *Protein Data Bank*

N	Código	Cadeia	Nome	Tamanho	% alfa	% beta
1	3APP		*Acid Proteinase Penicillopepsin*	323	10.22	45.51
2	2ACT		*Actinidin*	218	27.52	18.35
3	9WGA	A	*Wheat Germ Agglutinin*	170	11.76	9.41
4	8ADH		*Alcohol Dehydrogenase*	374	23.26	22.72
5	2ALP		*Alpha-Lytic Protease*	198	4.04	52.53
6	6AT1	A	*Aspartate Carbamoyltransferase*	310	35.48	14.84
7	6AT1	B	*Aspartate Carbamoyltransferase*	146	16.44	33.56
8	2AZA	A	*Azurin*	129	11.63	33.33
9	2ABX	A	*Alpha Bungarotoxin*	74	0.00	5.41
10	5CPV		*Calcium-Binding Parvalbumin B*	108	53.70	3.70
11	3ICB		*Calcium-Binding Protein*	75	57.33	0.00
12	2CAB		*Carbonic Anhydrase I Form B*	256	14.45	30.08
13	5CPA		*Carboxypeptidase A*	307	36.16	16.29
14	8CAT	A	*Catalase*	498	29.52	15.46
15	5CHA	A	*Chymotrypsin Alpha*	228	10.96	33.33
16	2CTS		*Citrate Synthase*	437	61.10	1.37
17	1CRN		*Crambin*	46	41.30	8.70
18	1GCR		*Crystallin Gamma II*	174	7.47	44.25
19	1CCR		*Cytochrome C*	111	39.64	0.00
20	2CCY	A	*Cytochrome C'*	127	74.80	0.00
21	2CYP		*Cytochrome C Peroxidase*	293	47.10	5.46
22	3C2C		*Cytochrome C2*	112	42.86	0.00
23	2CDV		*Cytochrome C3*	107	25.23	9.35
24	351C		*Cytochrome C551*	82	46.34	0.00
25	3DFR		*Dihydrofolate Reductase*	162	19.14	31.48
26	2EST	E	*Elastase*	240	7.08	34.17
27	3EBX		*Erabutoxin*	62	0.00	43.55
28	1ECD		*Hemoglobin III*	136	76.47	0.00
29	1FDX		*Ferredoxin*	54	9.26	7.41
30	3FXC		*Ferredoxin*	98	7.14	15.31
31	3FXN		*Flavodoxin*	138	36.23	21.01
32	4FD1		*Ferredoxin*	106	25.47	13.21
33	1GP1	A	*Glutathione Peroxidase*	184	28.80	15.76
34	2HMQ	A	*Hemerythrin*	114	61.40	0.00
35	2HHB	A	*Hemoglobin*	141	76.60	0.00

(Continua)

Tabela 1 - Aqui estão listadas, por ordem alfabética dos nomes originais em inglês, as cadeias protéicas utilizadas para montar o conjunto de dados. Para permitir a comparação dos resultados, o conjunto usado foi aquele selecionado por GIBRAT *et al.* (1987). As proteínas citadas tiveram sua estrutura resolvida com alta resolução e qualidade (2.8 angstrons ou menos), sem que houvesse grande similaridade entre as sequências de aminoácidos (50% ou menos). Na segunda coluna da esquerda para a direita, encontramos os nomes-código dos arquivos PDB que contém as respectivas cadeias. Na coluna seguinte estão as letras-código das cadeias utilizadas, nos casos de proteínas com mais de uma cadeia. Na quarta coluna, temos os

Tabela 1.- O Conjunto de dados. Cadeias protéicas retiradas do PDB Cont.)

N	Código	Cadeia	Nome	Tamanho	% alfa	% beta
36	2HHB	B	*Hemoglobin*	146	77.40	0.00
37	2LHB		*Hemoglobin V*	149	73.15	0.00
38	1HIP		*High Potential Iron Protein*	85	11.76	10.59
39	1MCP	L	*Immunoglobulin Iga Fab Fragment*	220	3.64	45.91
40	1MCP	H	*Immunoglobulin Iga Fab Fragment*	222	0.00	49.10
41	2PKA	A	*Kallikrein A*	80	0.00	45.00
42	2PKA	B	*Kallikrein A*	152	13.82	25.66
43	6LDH		*L-Lactate Dehydrogenase*	329	41.03	16.11
44	1LH1		*Leghemoglobin*	153	77.78	0.00
45	2LZM		*Lysozyme*	164	66.46	8.54
46	1LZ1		*Lysozyme*	130	36.92	7.69
47	2MLT	A	*Melittin*	26	92.31	0.00
48	1MBN		*Myoglobin*	153	77.12	0.00
49	1NXB		*Erabutoxin*	62	0.00	41.94
50	1SN3		*Scorpion Neurotoxin Variant 3*	65	12.31	18.46
51	1OVO	A	*Ovomucoid Third Domain*	56	17.86	21.43
52	1PPD		*Papain*	212	25.00	16.98
53	1BP2		*Phospholipase A2*	123	43.90	6.50
54	1PCY		*Plastocyanin*	99	4.04	35.35
55	2PAB	A	*Prealbumin*	114	7.02	51.75
56	2SGA		*Proteinase A*	181	6.63	54.14
57	3RP2	A	*Rat Mast Cell Protease II*	224	5.36	37.05
58	3RN3		*Ribonuclease A*	124	20.97	33.06
59	5RXN		*Rubredoxin*	54	0.00	14.81
60	2SNS		*Staphylococcyl Nuclease*	141	18.44	19.86
61	1SBT		*Subtilisin Bpn'*	275	30.18	17.82
62	2SOD	O	*Cu Zn Superoxide Dismutase*	151	0.00	38.41
63	3TLN		*Thermolysin*	316	39.56	16.46
64	1TPO		*Trypsin Beta*	223	9.42	32.29
65	4PTI		*Trypsin Inhibitor*	58	20.69	24.14
66	2STV		*Satellite Tobacco Necrosis Virus*	184	9.78	44.57
67	4SBV	A	*Southern Bean Mosaic Virus Protein*	199	12.06	35.18
TOTAL				11208	29.09	21.35

Tabela 1 (cont.) - nomes das proteínas do conjunto de dados, em inglês. Na quinta coluna se encontra o tamanho dos segmentos utilizados. Finalmente, nas duas últimas colunas, temos a proporção de resíduos nos estados alfa e beta de estrutura secundária. Note que, apesar do conjunto de dados como um todo apresentar uma proporção de aproximadamente 30% de resíduos em alfa-hélices, 20% em placas beta e 50% sem estrutura regular, estas proporções variam muito nas cadeias individuais. Existem proteínas sem nenhuma placa pregueada (e.g. *Mellitin*, n.º 47, 92% alfa), bem como sem alfa-hélices (e.g. *Immunoglobulin Iga Fab Fragment*, n.º 40, 49% beta). Da mesma forma, existem proteínas com quantidades equivalentes dos dois tipos de estrutura (e.g. *Alcohol Dehydrogenase*, n.º 4, 23% alfa, 23% beta). Além da proporção entre alfa e beta, a quantidade de estrutura regular como um todo também é bastante variável (e.g. *Acid Proteinase Penicillopepsin*, n.º 1, 10% alfa, 46% beta, 44% *coil*; versus *Ferredoxin*, n.º 30, 7% alfa, 15% beta, 78% *coil*).

3.4. O Modelo Logístico Linear

Os resíduos em uma proteína globular podem estar em um dentre três estados de estrutura secundária: alfa-hélice, cordão beta e *coil* (KABSCH & SANDER, 1983). Foi definido um modelo logístico que associa a cada resíduo da sequência uma probabilidade de estar num destes estados. As probabilidades dependem de uma "janela" local de até 17 aminoácidos.

Para indicar a identidade de um resíduo em uma posição particular usamos um conjunto de variáveis binárias, x_i, i=**1,...,20**, onde a posição do "**1**" indica qual dos 20 aminoácidos que ocorrem naturalmente está presente. Para representar a janela de comprimento 17 necessitamos então de 340 variáveis binárias, x_{ij}, i=**1,...,20**, j=**1,...,17**.

Um valor omisso na posição j é codificado por $x_{ij} = 0$, i=**1,...,20**. Para um modelo logístico que prevê o estado do resíduo central ($j = 9$), primeiro definem-se as variáveis intermediárias:

$$z_a = a_0 + \sum_{i=1}^{20} \sum_{j=1}^{17} a_{ij} x_{ij} \qquad (10)$$

$$z_b = b_0 + \sum_{i=1}^{20} \sum_{j=1}^{17} b_{ij} x_{ij} \qquad (11)$$

Então as probabilidades para os três estados: a (alfa-hélice), b (cordão beta) e c (*random coil*) são dados pelas funções logísticas:

$$p_a = e^{z_a} / (1 + e^{z_a} + e^{z_b}) \qquad (12)$$

$$p_b = e^{z_b} / (1 + e^{z_a} + e^{z_b}) \qquad (13)$$

$$p_c = 1 / (1 + e^{z_a} + e^{z_b}) \qquad (14)$$

```
3app     1    1  999   129 -179 * * UUUUUUUUAASGVATNT    C
3app     2    2  -74   153  175 E t UUUUUUUAASGVATNTP  B C
3app     3    3 -160   159  179 E t UUUUUUAASGVATNTPT    C
3app     4    4 -172   159  179 E t UUUUUAASGVATNTPTA  E E
3app     5    5 -121   127  176 E t UUUUAASGVATNTPTAN  E E
3app     6    6 -112   141  180 E t UUUAASGVATNTPTAND  E E
3app     7    7  -94   140  176 E t UUAASGVATNTPTANDE  E E
3app     8    8 -113   154  178 E t UAASGVATNTPTANDEE  E E
3app     9    9 -135   130  176 E t AASGVATNTPTANDEEY  E E
3app    10   10  -73   146  173 E t ASGVATNTPTANDEEYI  E E
3app    11   11  -84  -175  177 E t SGVATNTPTANDEEYIT    C
3app    12   12  -54   133 -178 E t GVATNTPTANDEEYITP  G C
3app    13   13   62    15 -177 a t VATNTPTANDEEYITPV  G C
3app    14   14   63    36 -178 a t ATNTPTANDEEYITPVT  G C
3app    15   15  -64   -38  180 A t TNTPTANDEEYITPVTI    C
3app    16   16 -153   167 -177 E t NTPTANDEEYITPVTIG    C
3app    17   17 -124   122 -178 E t TPTANDEEYITPVTIGG  E E
3app    18   18 -110   154  180 E t PTANDEEYITPVTIGGT  E E
3app    19   19 -140   141  178 E t TANDEEYITPVTIGGTT  E E
3app    20   20  -78   140 -179 E t ANDEEYITPVTIGGTTL  E E
3app    21   21 -134   131  178 E t NDEEYITPVTIGGTTLN  E E
3app    22   22 -102   123  178 E t DEEYITPVTIGGTTLNL  E E
3app    23   23 -121   117  178 E t EEYITPVTIGGTTLNLN  E E
3app    24   24   52    43  177 a t EYITPVTIGGTTLNLNF  T C
3app    25   25   84   -12  179 a t YITPVTIGGTTLNLNFD  T C
3app    26   26 -101   133  178 E t ITPVTIGGTTLNLNFDT  E E
3app    27   27  -95   130  179 E t TPVTIGGTTLNLNFDTG  E E
3app    28   28 -120   153  180 E t PVTIGGTTLNLNFDTGS  E E
3app    29   29 -102   107 -177 E t VTIGGTTLNLNFDTGSA  E E
3app    30   30 -113   158  179 E t TIGGTTLNLNFDTGSAD  E E
3app    31   31  -93   123  179 E t IGGTTLNLNFDTGSADL  E E
3app    32   32  -83   130 -174 E t GGTTLNLNFDTGSADLW  E E
3app    33   33 -131   109  178 E t GTTLNLNFDTGSADLWV  E E
3app    34   34  -77     1  177 A t TTLNLNFDTGSADLWVF  T C
3app    35   35  -98     1  179 A t TLNLNFDTGSADLWVFS  T C
3app    36   36 -143   162 -178 E t LNLNFDTGSADLWVFST    C
3app    37   37 -122    24  177 A t NLNFDTGSADLWVFSTE    C
3app    38   38 -109   146  174 E t LNFDTGSADLWVFSTEL    C
3app    39   39 -114    94  179 E t NFDTGSADLWVFSTELP  E E
3app    40   40 -118   149 -178 E t FDTGSADLWVFSTELPA  E E
3app    41   41 -140   156  175 E t DTGSADLWVFSTELPAS  E E
3app    42   42  -65   143 -176 E t TGSADLWVFSTELPASQ    C
3app    43   43 -141   175  178 E t GSADLWVFSTELPASQQ  B C
3app    44   44  -74    -6  179 A t SADLWVFSTELPASQQS  T C
3app    45   45  -81   -14  177 A t ADLWVFSTELPASQQSG  T C
3app    46   46  -76   160  174 E t DLWVFSTELPASQQSGH  S C
3app    47   47  -57   137  180 E t LWVFSTELPASQQSGHS    C
3app    48   48  -47   -33 -178 A t WVFSTELPASQQSGHSV  H H
3app    49   49  -71   -28 -178 A t VFSTELPASQQSGHSVY  H H
3app    50   50  -88   -17  178 A t FSTELPASQQSGHSVYN  H H
```

Figura 12 - Exemplo de arquivo de dados gerado pelo programa GENPTR.F, versão alfanumérica. A primeira coluna traz o código PDB da proteína, neste caso 3APP, pepsilopepsina (n.o 1, Tab. 1). Em seguida pode haver a identificação da cadeia protéica utilizada (em branco neste caso). A próxima coluna traz a numeração dos resíduos neste arquivo, seguida da numeração encontrada em PDB (podem ser diferentes). Os três valores seguintes são os ângulos diédricos ϕ, ψ e ω (Fig. 7). A coluna seguinte traz a estrutura secundária de acordo com os ângulos diédricos (LAMBERT & SCHERAGA, 1989), seguida por uma codificação discreta do ângulo ω (*cis* ou *trans*). Nenhuma das colunas referentes a ângulos diédricos é utilizada (veja o texto). Seguem os 17 códigos dos aminoácidos que compõe a janela (Fig. 2, U=valor omisso). A penúltima coluna é a estrutura secundária do resíduo central da janela de acordo com KABSCH & SANDER (1983, oito classes). A última coluna traz a estrutura secundária modificada de acordo com GARNIER *et al.* (1978, três classes, E=beta, H=alfa, C=*coil*).

```
51, 97,112,112, 32,   1,   1, 999, 129,-179, 0, 0,19,19,19,19,19,19,19,19, 1, 1,17, 7,20, 1,18,13,18, 0, 0
51, 97,112,112, 32,   2,   2, -74, 153, 175, 2, 0,19,19,19,19,19,19,19, 1, 1,17, 7,20, 1,18,13,18,14, 3, 0
51, 97,112,112, 32,   3,   3,-160, 159, 179, 2, 0,19,19,19,19,19,19, 1, 1,17, 7,20, 1,18,13,18,14,18, 0, 0
51, 97,112,112, 32,   4,   4,-172, 159, 179, 2, 0,19,19,19,19,19, 1, 1,17, 7,20, 1,18,13,18,14,18, 1, 2, 2
51, 97,112,112, 32,   5,   5,-121, 127, 176, 2, 0,19,19,19,19, 1, 1,17, 7,20, 1,18,13,18,14,18, 1,13, 2, 2
51, 97,112,112, 32,   6,   6,-112, 141, 180, 2, 0,19,19,19, 1, 1,17, 7,20, 1,18,13,18,14,18, 1,13, 4, 2, 2
51, 97,112,112, 32,   7,   7, -94, 140, 176, 2, 0,19,19, 1, 1,17, 7,20, 1,18,13,18,14,18, 1,13, 4, 5, 2, 2
51, 97,112,112, 32,   8,   8,-113, 154, 178, 2, 0,19, 1, 1,17, 7,20, 1,18,13,18,14,18, 1,13, 4, 5, 5, 2, 2
51, 97,112,112, 32,   9,   9,-135, 130, 176, 2, 0, 1, 1,17, 7,20, 1,18,13,18,14,18, 1,13, 4, 5, 5,22, 2, 2
51, 97,112,112, 32,  10,  10, -73, 146, 173, 2, 0, 1,17, 7,20, 1,18,13,18,14,18, 1,13, 4, 5, 5,22, 9, 2, 2
51, 97,112,112, 32,  11,  11, -84,-175, 177, 2, 0,17, 7,20, 1,18,13,18,14,18, 1,13, 4, 5, 5,22, 9,18, 0, 0
```


Figura 13 - Exemplo do arquivo de dados produzido pelo programa GENPTR.F, versão numérica efetivamente utilizada pela rotina em MATLAB. O conteúdo é o mesmo do arquivo mostrado na Fig. 12, codificado. As quatro primeiras colunas correspondem aos códigos ASCII dos caracteres "3", "a", "p", "p", seguidos do código de "espaço em branco", correspondente à ausência de designação de cadeia neste caso. As colunas numéricas são mantidas no mesmo formato. Os aminoácidos e estados de estrutura secundária são numerados de maneira simples, em geral pela ordem alfabética de seus códigos alfanuméricos.

A função exponencial assegura que todas as estimativas p são maiores ou iguais a zero, enquanto que o denominador normaliza as probabilidades, garantindo a soma igual à unidade. Pode ser útil a utilização das razões de probabilidade logarítmica:

$$z_a = \ln(p_a / p_c) \tag{15}$$

$$z_b = \ln(p_b / p_c) \tag{16}$$

onde os zs são também denominados de probabilidades *logito* (AGRESTI, 1990). Assim, o modelo pode também ser chamado de "logito linear". O vetor de parâmetros $\boldsymbol{a} = (a_0, a_{1,1}, a_{1,2}, \ldots a_{20,17})^T$ descreve a diferença entre os estados alfa e *coil*, enquanto o vetor $\boldsymbol{b}$ reflete a diferença entre beta e *coil*. Para evitar uma situação de indeterminação, também foi imposta a restrição

$$\sum_{i=1}^{20} a_{ij} = 0 \tag{17}$$

de forma que resultam 19*17+1=324 parâmetros independentes, e não 20*17+1=341, nos vetores $\boldsymbol{a}$ e $\boldsymbol{b}$.

3.5. O Modelo Logístico Quadrático

O modelo logístico linear foi estendido, incluindo-se termos de ordem superior, a fim de capturar padrões mais complexos presentes nos dados. Descrevemos uma formulação do caso logístico quadrático geral, citando, depois, alguns casos especiais importantes. Embora não pudéssemos calibrar o modelo quadrático completo, esta formulação serve para unificar abordagens já tentadas e sugerir novas.

Primeiro vamos considerar o modelo logístico linear em notação vetorial. Seja $\mathbf{x}$ o vetor coluna das variáveis binárias que descrevem a sequência local. Então o modelo quadrático completo pode ser expresso pelo par de equações vetoriais:

$$\mathbf{z}(\mathbf{x}) = \begin{bmatrix} a_0 + \boldsymbol{a}^{\mathrm{T}}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} \\ b_0 + \boldsymbol{b}^{\mathrm{T}}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x} \\ 0 \end{bmatrix} \qquad (18)$$

$$p(\mathbf{x}) = e^{\mathbf{z}} / \mathbf{1}^{\mathrm{T}} e^{\mathbf{z}} \qquad (19)$$

Aqui, **A** e **B** são matrizes das formas quadráticas, simétricas e contendo um valor para cada uma das possíveis interações entre dois resíduos da janela local. É obvio que **A** e **B** podem ser muito grandes (mais que 100.000 elementos para janelas de 17 aminoácidos), resultando em muito mais parâmetros do que poderiam ser estimados a partir dos dados. De fato, contém mais elementos do que há resíduos em nosso conjunto. Foi portanto necessário fazer restrições a fim de reduzir o número de parâmetros. A primeira delas supõe que aminoácidos separados pela mesma distância na cadeia polipeptídica contribuem da mesma forma para as probabilidades do resíduo central. Isto significa que um par de aminoácidos nas posições 1 e 5 terão o mesmo efeito que nas posições 2 e 6, e assim por diante. Esta restrição equivale a reduzir o número de parâmetros para 400, cada um correspondendo a interação entre um par de aminoácidos, independente de sua posição. Estamos, na realidade, impondo uma nova estrutura interna às matrizes **A** e **B** (veja Fig.14).

Para que esta redução de parâmetros não comprometa a plausibilidade do modelo, uma segunda restrição é necessária, relacionada com a posição dos pares de resíduos na janela. Podemos usar uma matriz quadrada **W**, de dimensão igual à da janela, contendo coeficientes para ponderar o efeito da presença de um par de aminoácidos na estrutura secundária do resíduo central. A Fig.14 ilusta uma das suposições mais simples que podemos fazer: o efeito da interação de um par de aminoácidos varia com a distância que os separa. Isto provoca uma estrutura bandeada nas diagonais de **W** e consequentemente de **A** e **B** (cada elemento de **W** vai multiplicar um bloco de 20 x 20 elementos dentro de **A** e **B**). Muitas outras restrições são possíveis, tais como levar em conta a influência da distância do par ao resíduo central. Uma delas merece descrição em separado, uma vez que utiliza o conceito de anfipaticidade, ou de distribuição periódica de características fisicoquímicas dos aminoácidos ao longo da sequência, e sua relação com o surgimento de estruturas secundárias regulares.

A motivação para tal abordagem é a mesma que levou outros autores a utilizarem índices de anfipaticidade (KLEIN & DELISI, 1986; EISENBERG, WILCOX, MCLACHLAN, 1986; CORNETTE *et al.*, 1987). No caso das alfa-hélices (Fig. 9) as

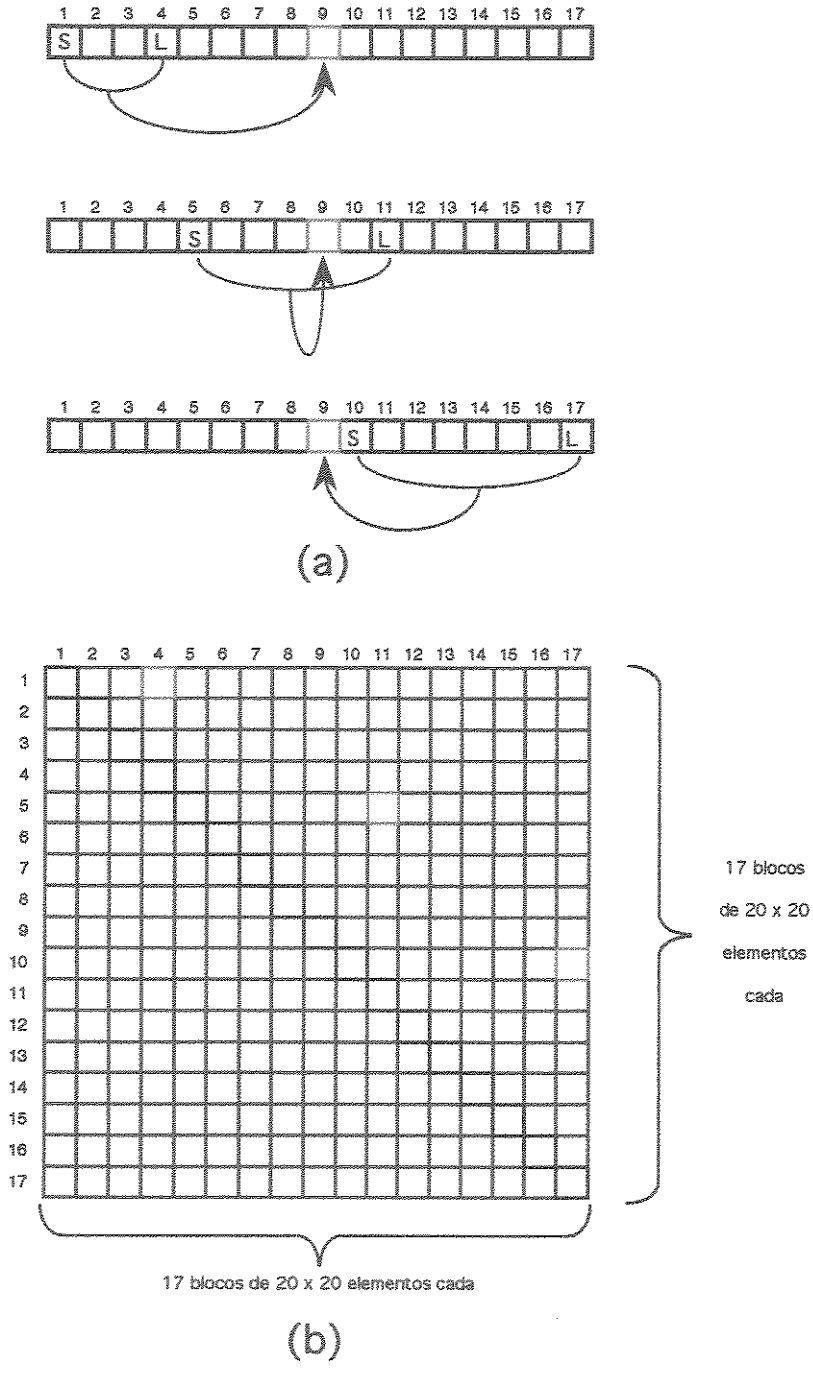


Figura 14 - Restrições às matrizes **A** e **B** das formas quadráticas. (a) Três representações da janela de 17 aminoácidos, ilustrando os efeitos da interação do par serina/leucina (SxL) sobre a estrutura do resíduo central (em verde), independente de sua identidade. Em cada um dos três exemplos, a distância entre S e L varia, assim como sua posição na janela. (b) Representação gráfica de uma das matrizes (e.g. **A**) dividida em 340 blocos de 400 elementos cada. No caso em que todas as interações entre um determinado par de aminoácidos (e.g. S e L, nesta ordem) são consideradas equivalentes, o conteúdo de todos os blocos será o mesmo e o número de parâmetros se reduz aos 400 elementos de cada bloco, cada um representando a influência de um dos possíveis pares sobre o estado do resíduo central. Os blocos da diagonal principal de **A** (coloridos em azul) estão "vazios", pois representam interação entre resíduos na mesma posição da janela. Os blocos marcados em verde correspondem às posições dos blocos contendo os parâmetros correspondentes às situações ilustradas na Fig. 14a.

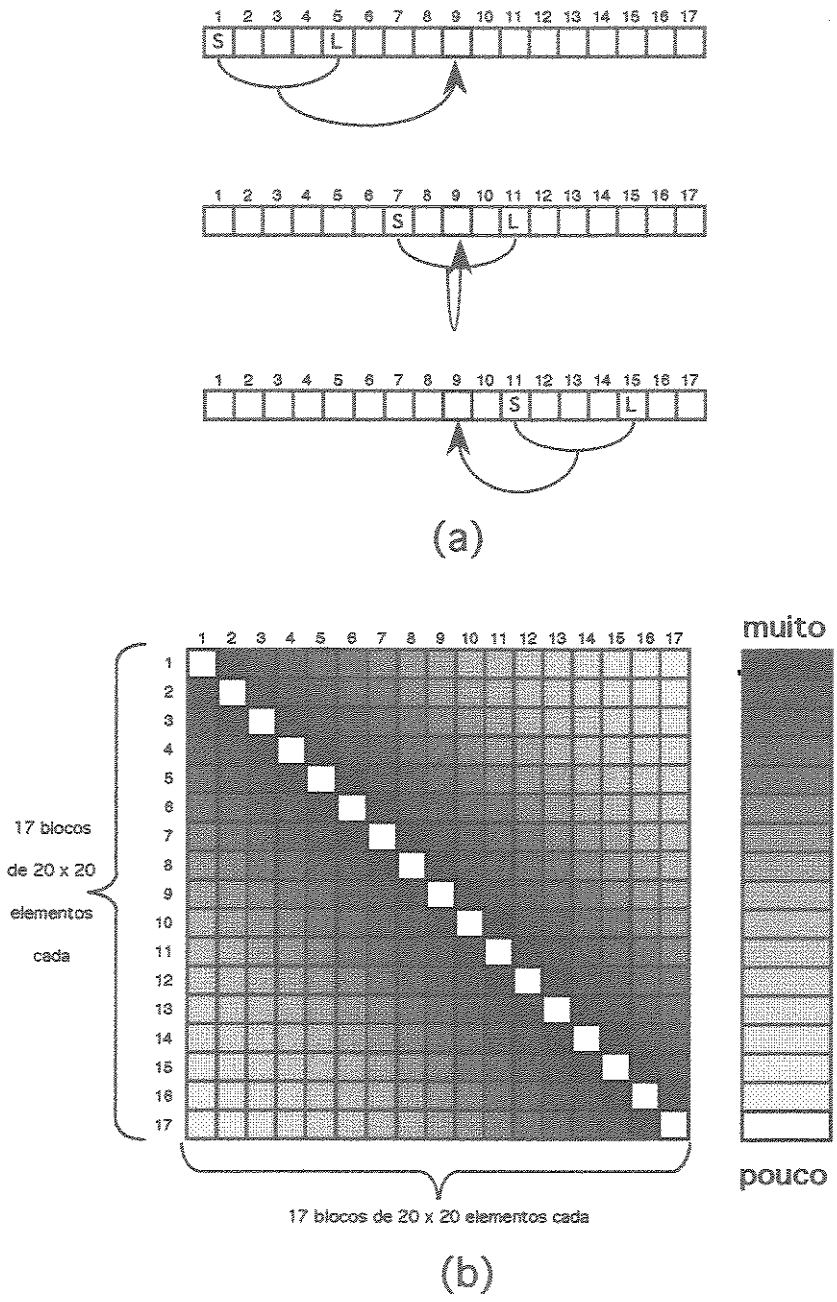


Figura 15 - Restrições às matrizes **A** e **B** das formas quadráticas. (a) Três representações da janela de 17 aminoácidos, ilustrando os efeitos da interação do par serina/leucina (SxL) sobre a estrutura do resíduo central (em verde), independente de sua identidade. Em cada um dos três exemplos a distância entre S e L foi mantida, variando-se apenas sua posição na janela. (b) Representação gráfica de uma das matrizes (e.g. **A**) dividida em 340 blocos de 400 elementos cada. Aqui, as interações entre um determinado par de aminoácidos, separados por uma distância fixa (4 resíduos neste caso), foram consideradas equivalentes. Isto faz com que cada "banda" de blocos paralelos à diagonal principal tenha o mesmo conteúdo. Se associarmos a cada banda um fator multiplicativo, novamente podemos reduzir o número de parâmetros aos 400 elementos de cada bloco, devidamente multiplicados pelo fator apropriado (representados por níveis de cinza, à direita). No presente exemplo, foram usados fatores inversamente proporcionais à distância entre os resíduos do par. Os blocos marcados em preto são os que contém os parâmetros correspondentes aos três exemplos da Fig. 15a. Note-se como elem pertencem à mesma banda, tendo igual conteúdo de acordo com a restrição descrita.

cadeias laterais dos aminoácidos se distribuem de tal forma que o resíduo *i* será vizinho dos resíduos *i-4* e *i+4*, aproxidamente. O passo da hélice é na realidade de 3.6 resíduos. Devido a maneira com as proteínas se dobram em meio aquoso, existe uma tendência a "esconder" cadeias laterais hidrofóbicas no "interior" do glóbulo protéico e a expor as cadeias hidrofílicas na superfície. A menos que a hélice se encontre totalmente envolvida por outras porções da cadeia (no caso de uma proteína muito grande) ou por um meio altamente hidrofóbico (membrana celular) haverá uma tendência de agrupamento de resíduos de características opostas em lados também opostos da estrutura. Isto se refletirá numa periodicidade característica da distribuição dos aminoácidos ao longo da sequência (3.6).

No caso das placas pregueada beta podemos fazer raciocínio semelhante. Na Fig. 11a vemos o que acontece com as cadeia laterais dos resíduos em uma placa antiparalela. As cadeias principais formam um plano e as laterais se alternam, ocupando uma ou outra face deste. A menos que o plano esteja envolvido por outras porções da cadeia (o que não é raro), haverá uma tendência de resíduos de características opostas ocuparem faces também opostas do plano da placa. Se seguirmos uma das cadeias, notaremos que os radicais laterais se alternam de um lado e de outro, de forma que a periodicidade da hidrofobicidade numa placa anfipática (um lado polar, outro apolar) é 2.0. A Fig. 11b mostra que no caso das placas paralelas vale o mesmo tipo de raciocínio, pois as cadeias laterais também se alternam.

Substituindo as identidades dos aminoácidos por seus índices de hidrofobicidade, e calculando a função "potência de Fourier" nessas sequências, com periodicidades 3.6 e 2.0 (períodos das alfa-hélices e dos cordões beta respectivamente), obteríamos as medidas de anfipaticidade de EISENBERG *et al.* (1986) e DeLisi. (CORNETTE *et al.*, 1987). Dado um vetor de índices de hidrofobicidade $\boldsymbol{h_i}$ para cada resíduo da janela local, $i=1,...,17$, a expressão da potência de Fourier com periodicidade de 3.6 resíduos por volta é:

$$\mathrm{Pot}(3.6) = \left(\sum_{i=1}^{17} h_i \cos(2\pi i / 3.6)\right)^2 + \left(\sum_{i=1}^{17} h_i \,\mathrm{sen}(2\pi i / 3.6)\right)^2 \quad (20)$$

Na notação vetorial fazendo $\mathbf{r}=[\cos(2\pi i/3.6)]^2$ $i=1,...,17$ e $\mathbf{s}=[\mathrm{sen}(2\pi i/3.6)]^2$ $i=1...,17$, a expressão se reduz a:

$$\mathrm{Pot}(3.6) = (\mathbf{r}^{\mathrm{T}}\mathbf{h})^2 + (\mathbf{s}^{\mathrm{T}}\mathbf{h})^2 = \mathbf{h}^{\mathrm{T}}\mathbf{W}^{3.6}\mathbf{h} \quad (21)$$

onde $\mathbf{W}^{3.6}$ é a matriz simétrica 17x17 da forma quadrática, com elementos:

$$
\begin{aligned}
w^{3.6}_{|i-j|} &= r_i r_j + s_i s_j \\
&= \cos(2\pi i/3.6)\cos(2\pi j/3.6)+\operatorname{sen}(2\pi i/3.6)\operatorname{sen}(2\pi j/3.6) \\
&= \cos(2\pi |i-j|/3.6). \qquad (22)
\end{aligned}
$$

Assim $\mathbf{W}^{3.6}$ tem um formato bandeado e simétrico, ou seja:

$$
\mathbf{W}^{3.6} = \begin{bmatrix}
w_o & w_1 & w_2 & \cdots & w_{16} \\
w_1 & w_o & w_1 & \ddots & \vdots \\
w_2 & w_1 & w_o & \ddots & w_2 \\
\vdots & \ddots & \ddots & \ddots & w_1 \\
w_{16} & \cdots & w_2 & w_1 & w_o
\end{bmatrix} \qquad (23)
$$

Podemos então usar $\mathbf{W}^{3.6}$ para modificar a estrutura interna de $\mathbf{A}$, substituindo cada elemento escalar por um bloco de 20 por 20, permitindo, assim, a inclusão dos 400 parâmetros correspondentes às "preferências" dos pares de resíduos. Definimos $\mathbf{A}$ como sendo constituída de um arranjo de 17 por 17 blocos, $\mathbf{A}_{ij}=w^{3.6}_{|i-j|}*\mathbf{A}$, onde $\mathbf{A}$ é a matriz 20 por 20 com as preferências. Definindo os multiplicadores $w^{3.6}_{|i-j|}$ como acima, a forma quadrática $\mathbf{x}^T\mathbf{A}\mathbf{x}$ calcula a potência de Fourier na periodicidade 3.6, onde a escala de hidrofobicidade **h** foi substituída pela identidade dos aminoácidos, representada pela codificação da sequência em **x**. Note-se que não vamos supor ou incluir nenhuma informação sobre as propriedades físicoquímicas dos resíduos. Nos métodos já citados (KLEIN & DELISI, 1986; EISENBERG, WILCOX, MCLACHLAN, 1986; CORNETTE *et al.*, 1987) foi utilizado um vetor arbitrário **h** e um valor de **Pot(3.6)** elevado indicava maior probabilidade da ocorrência de alfa-hélices na região. Em nosso caso, tentaremos estimar um parâmetro para cada par de aminoácidos, ponderado pelos valores de $\mathbf{W}^{3.6}$, de forma que o par contribuirá ou não para a ocorrência de uma alfa-hélice de acordo com sua posição. Note-se que há elementos negativos.

Uma matriz similar $\mathbf{W}^{2.0}$ foi usada para construir a matriz $\mathbf{B}$, com periodicidade 2.0, para cordões beta, usando uma matriz **B** de preferências (diferente de **A**). Note que por construção a estrutura das matrizes $\mathbf{A}$ e $\mathbf{B}$ é "bandeada por blocos" (blocos idênticos **A**ij e **B**ij estão dispostos em bandas ao longo das diagonais de $\mathbf{A}$ e $\mathbf{B}$, veja Fig. 15). Esta estrutura supõe de que pares de resíduos separados por uma distância fixa contribuem com o mesmo efeito, independente da posição em que estejam na janela.

Além disso, fizemos a restrição adicional de que as somas das linhas e as somas das colunas de **A** e **B** (blocos que formam **Π** e **B**) sejam zero, o que faz cada bloco contribuir com 19*19=361 parâmetros quadráticos.

A construção do modelo sugere muitas variações possíveis. Pode-se fazer com que algumas das diagonais de **Ш** sejam zero, considerando apenas interações próximas. Pode-se convoluir **Ш** com um envelope gaussiano, ou supor que as matrizes de parâmetros **A** e **B** também devam ser simétricas, isto é, a interação entre dois resíduos deveria contribuir da mesma forma independente de qual deles ocorra antes na sequência (do lado N-terminal).

3.6. Estimativas de Máxima Verossimilhança (EMV)

Para determinar os melhores valores dos parâmetros, primeiro define-se um modelo probabilístico que se aproxime dos valores observados. Supomos que cada resíduo escolha seu estado de acordo com o vetor de probabilidades, isto é, os estados são determinados por uma distribuição de probabilidade trinomial. Uma vez feito isto realiza-se o procedimento de obtenção das EMV. De nada adiantaria reduzir o número de parâmetros da porção quadrática do modelo se tivéssemos que efetuar a multiplicação matricial da forma quadrática $\mathbf{x}^{\mathrm{T}}\mathbf{\Pi}\mathbf{x}$. Para facilitar o cálculo, define-se um novo vetor de parâmetros θ_a combinando-se o vetor $\boldsymbol{a}$ com a matriz **A**. Da mesma forma obtém-se um vetor θ_b. Modificando-se o vetor **x** de modo correspondente, podemos escrever as probabilidades logito **z** como uma função linear dos parâmetros.

$$\mathbf{z} = \begin{bmatrix} \theta_a^{\mathrm{T}}\tilde{\mathbf{x}} \\ \theta_b^{\mathrm{T}}\tilde{\mathbf{x}} \\ 0 \end{bmatrix}. \qquad (24)$$

ou seja

$$\theta_a^{\mathrm{T}}\tilde{\mathbf{x}} = a_0 + \boldsymbol{a}^{\mathrm{T}}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{\Pi}\mathbf{x} \qquad (25)$$

$$\theta_b^{\mathrm{T}}\tilde{\mathbf{x}} = b_0 + \boldsymbol{b}^{\mathrm{T}}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x} \qquad (26)$$

O vetor θ_a foi construído de forma a conter não somente a porção linear, com os 340 parâmetros correspondentes à posição do resíduo na janela, mas também as 400 parâmetros relativos às possíveis interações entre pares de aminoácidos. O vetor **x** por sua vez foi estendido de forma a conter não somente a porção binária, que codifica a

identidade dos aminoácidos pelas posições dos **1**s e **0**s, mas também elementos indicando as interações entre cada par de aminoácidos da janela, já devidamente ponderados pelo valor de **W**, de acordo com sua posição. Além de facilitar os cálculos de estimativa dos parâmetros, a forma linear também facilita o cálculo da expressão de verossimilhança (Eq. 35).

A verossimilhança de um modelo trinomial com N observações pode então ser escrita:

$$\mathcal{L} = \prod_{i=1}^{N} \binom{1}{f_{a,i} \quad f_{b,i} \quad (1-f_{a,i}-f_{b,i})} p_a^{f_{a,i}} p_b^{f_{b,i}} p_c^{(1-f_{a,i}-f_{b,i})} \quad (27)$$

onde a expressão entre grandes parênteses, logo após o sinal de produto, **não** é uma matriz, mas sim um coeficiente trinomial (uma extensão do conceito de coeficiente binomial), neste caso sempre igual a 1. O par ordenado $(\mathbf{f_{a,i}f_{b,i}})$ vale **(1,0)** em caso de alfa-hélice, **(0,1)** para beta e **(0,0)** para *coil* . A contribuição para a verossimilhança feita pela janela i, cujo elemento central está numa alfa-hélice é:

$$v_i = p_{a,i}^1 p_{b,i}^0 p_{c,i}^0 = p_{a,i} = \frac{e^{\theta_a^T \tilde{\mathbf{x}}_i}}{1 + e^{\theta_a^T \tilde{\mathbf{x}}_i} + e^{\theta_b^T \tilde{\mathbf{x}}_i}} \quad (28)$$

e da janela j, com elemento central beta:

$$v_j = p_{a,j}^0 p_{b,j}^1 p_{c,j}^0 = p_{b,j} = \frac{e^{\theta_b^T \tilde{\mathbf{x}}_j}}{1 + e^{\theta_a^T \tilde{\mathbf{x}}_j} + e^{\theta_b^T \tilde{\mathbf{x}}_j}} \quad (29)$$

e da janela k, com elemento central sem estrutura secundária regular:

$$v_k = p_{a,k}^0 p_{b,k}^0 p_{c,k}^1 = (1 - p_{a,k} - p_{b,k}) = \frac{1}{1 + e^{\theta_a^T \tilde{\mathbf{x}}_k} + e^{\theta_b^T \tilde{\mathbf{x}}_k}} \quad (30)$$

As versões logarítmicas das Eqs. 28 a 30 (contribuições de resíduos individuais para a verossimilhança logarítmica) ficam então:

$$\ln v_i = \theta_a^{\mathrm{T}} \tilde{\mathbf{x}}_i - \ln(1 + e^{\theta_a^{\mathrm{T}} \tilde{\mathbf{x}}_i} + e^{\theta_b^{\mathrm{T}} \tilde{\mathbf{x}}_i}) \qquad (31)$$

$$\ln v_j = \theta_b^{\mathrm{T}} \tilde{\mathbf{x}}_j - \ln(1 + e^{\theta_a^{\mathrm{T}} \tilde{\mathbf{x}}_j} + e^{\theta_b^{\mathrm{T}} \tilde{\mathbf{x}}_j}) \qquad (32)$$

$$\ln v_k = -\ln(1 + e^{\theta_a^{\mathrm{T}} \tilde{\mathbf{x}}_k} + e^{\theta_b^{\mathrm{T}} \tilde{\mathbf{x}}_k}) \qquad (33)$$

e a expressão para a verossimilhança logarímitica do modelo se escreve:

$$\ln \mathcal{L} = \sum_{i=1}^{N} f_{a,i} \theta_a^{\mathrm{T}} \tilde{\mathbf{x}}_i + f_{b,i} \theta_b^{\mathrm{T}} \tilde{\mathbf{x}}_j - \ln(1 + e^{\theta_a^{\mathrm{T}} \tilde{\mathbf{x}}_i} + e^{\theta_b^{\mathrm{T}} \tilde{\mathbf{x}}_i}) \qquad (34)$$

A busca de valores ótimos para θ_a e θ_b. é feita pela maximização de $\ln \mathcal{L}$, o que equivale a resolver as equações normais:

$$\partial \ln \mathcal{L} / \partial \theta_a = \sum_{i=1}^{N} (f_{a,i} - p_{a,i}) \tilde{\mathbf{x}}_i = \mathbf{0} \qquad (35)$$

$$\partial \ln \mathcal{L} / \partial \theta_b = \sum_{i=1}^{N} (f_{b,i} - p_{b,i}) \tilde{\mathbf{x}}_i = \mathbf{0} \qquad (36)$$

A técnica de máxima verossimilhança para modelos logísticos multinomiais (assim como modelos loglineares) já foi bem estabelecida, e algoritmos de maximização eficientes são amplamente disponíveis (JOHNSON & WICHERN, 1988). Foi utilizado o método numérico (*Fisher's scoring*, AGRESTI, 1990) que convergiu em menos de 10 iterações mesmo para modelos com mais de 1000 parâmetros. Os cálculos de estimativa foram realizados por um programa escrito em MATLAB, com algumas subrotinas em FORTRAN executados em computadores Convex e Macintosh. O código em MATLAB se encontra nos Apêndices.

3.7. Máxima Verossimilhança Penalizada

Mesmo com as restrições ao modelo quadrático geral, o modelo chega a ter 1370 parâmetros, a serem estimados de uma base de dados com pouco mais de 11000 resíduos. Pode-se esperar instabilidade em alguns dos parâmetros, em especial aqueles

relacionados a pares de aminoácidos raros. Precisávamos de um mecanismo correspondente ao uso de "observações fictícias" feito por GIBRAT *et al* (1987), mas que gozasse de rigor estatístico e não necessitasse suposições inverossímeis a respeito da independência dos dados. Escolhemos um procedimento denominado verossimilhança penalizada (GOOD & GASKINS, 1971; SIMMONOFF, 1983; TITTERINGTON & BOWMAN, 1985; EUBANK, 1988) para reduzir a variância dos parâmetros menos frequentes, baseados na expectativa de que a maioria dos parâmetros de interação contribui pouco para as preferências de estado. Esta abordagem é também denominada "estimativa por achatamento", já que os parâmetros penalizados tendem a ser reduzidos a zero com a aplicação da penalidade, ou "estimativa bayesiana" se pudermos quantificar nossa expectativa de pequena magnitude para os parâmetros (EUBANK 1988).

A penalização é efetuada adicionando-se uma penalidade positiva à logverossimilhança negativa, e minimizando-se a expressão resultante:

$$-\ln \mathcal{L}(\theta)+\lambda\theta^{T}\Lambda\theta \qquad (37)$$

Onde Λ é uma matriz simétrica positiva e definida e $\lambda \geq 0$. Esta parcela com penalidade estabiliza a convergência do algoritmo de minimização, mesmo quando o número de parâmetros do vetor θ é igual ou maior que o número de observações N no conjunto de dados. O parâmetro λ controla a magnitude da penalidade, variando de zero, quando não há penalização alguma, até infinito, quando a penalização domina a expressão de verossimilhança, reduzindo todos os parâmetros para zero. Neste estudo apenas os parâmetros quadráticos do modelo foram penalizados, fazendo com que $\Lambda_{i,i}=1$ onde i corresponde a um parâmetro quadrático e $\Lambda_{j,k}=0$ em todos os outros casos. Isto deixa a porção linear do modelo intacta, mas reduz gradualmente os efeitos de interação entre os resíduos da janela com o aumento de λ.

Diferenciando-se a expressão acima com relação a θ_a e θ_b. e fazendo-a igual a zero obtemos as equações normais da verossimilhança penalizada

$$\sum_{i=1}^{N} (f_{a,i} - p_{a,i})\tilde{\mathbf{x}}_i + \lambda\theta_a = \mathbf{0} \qquad (38)$$

$$\sum_{i=1}^{N} (f_{b,i} - p_{b,i})\tilde{\mathbf{x}}_i + \lambda\theta_b = \mathbf{0} \qquad (39)$$

que são resolvidas da mesma forma que as equações normais não penalizadas.

3.8. Seleção de Modelos

Da mesma forma que GIBRAT *et al.* (1988) escolheram seu parâmetro **M**, referente ao número de "observações fictícias", deveremos escolher o melhor valor para λ. Contudo, utilizaremos como medida de sucesso o valor da verossimilhança, devidamente maximizado por técnicas numéricas, não a acurácia. O problema é análogo à escolha do número de variáveis a serem utilizadas em um problema de modelagem paramétrica, ou ainda, à seleção do grau de alisamento mais adequado para o ajuste de uma função a pontos experimentais. Deve-se dar preferência a fomulações simples do modelo para evitar que abordagens muito complexas permitam a "memorização" dos dados (AGRESTI, 1990). Quando este fenômeno ocorre, o modelo prevê os dados com grande acurácia, mesmo com janelas de 5 ou 6 resíduos. Todavia, esta capacidade preditiva não se estende para sequências fora do conjunto de treino.

O número de parâmetros reflete a complexidade do modelo: quanto menor o número de parâmetros, mais simples e mais elegante ele é. Esta grandeza também reflete, até certo ponto, o equilíbrio ou relação entre o viés e a variância das previsões (EUBANK, 1988). Um índice frequentemente usado na seleção do melhor modelo paramétrico é a raiz do erro quadrático médio (*root mean square error*, RMS), que compara explicitamente o ajuste da previsões (soma dos resíduos quadráticos) com o núimero de graus de liberdade (quantidade de dados, ***N***, menos número de parâmetros, ***npar***; JOHNSON & WICHERN, 1988). Especificamente:

$$\sqrt{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \Big/ (N - npar)} \qquad (40)$$

Evidentemente é necessário mais de uma observação experimental por parâmetro para obter uma previsão adequada. Aliás, aceita-se que 10 observações por parâmetro seja o mínimo necessário para evitar instabilidade na regressão ou "memorização" dos dados (GIBRAT *et al.*, 1987).

A seleção do melhor modelo envolverá portanto a escolha da quantidade adequada de parâmetros, lineares (tamanho da janela) e quadráticos (tamanho e forma das matrizes **A**, **B**, **W**, Λ e do valor λ para a penalidade).

A imposição da penalidade λ à verossimilhança equivale a diminuir o número de parâmetros, na medida em que aqueles mais raros têm sua influência diminuída. A fórmula para o número efetivo de parâmetros para modelos não-paramétricos surge quando os valores previstos $\boldsymbol{p}$ podem ser aproximados por uma função linear dos dados $\boldsymbol{p} = \mathbf{L}\,\mathbf{f}$, que é o caso do modelo logístico penalizado. Assim ***npar*** = **2*traço(L) − traço($\mathbf{L}^T\mathbf{L}$) ≈ traço(L)** (Eq. 41; EUBANK, 1988). Uma certa quantidade de simplificação algébrica está envolvida no cálculo uma vez que **L** excederia a capacidade de armazenamento do computador.

Também aplicamos a seleção de variáveis do modelo quadrático, numa variação do método usualmente utilizado em regressão linear multivariada (DRAPER & SMITH, 1981) . As estratégias mais comuns incluem inclusão gradual, exclusão gradual, e regressão em todos os subconjuntos. Este procedimento pode ser útil quando se espera que algumas das variáveis independentes não contenham informação significativa, contribuindo apenas para aumentar o ruído estatístico.

A seleção de variáveis foi realizada calculando-se o valor de t de Student para os parâmetros contidos em θ_a e θ_b:

$$t_{a,i} = \frac{\theta'_{a,i}}{s_{\theta'_a}}\sqrt{N-1} \quad ; \quad t_{b,i} = \frac{\theta'_{b,i}}{s_{\theta'_b}}\sqrt{N-1} \qquad (42;43)$$

onde θ'_a é o vetor contendo somente os parâmetros quadráticos de θ_a e $s_{\theta'}$ é o desvio padrão dos elementos de θ'. O valor de t reflete a significância com que o parâmetro difere de zero. Toma-se os $n_{qua}/4$ parâmetros quadráticos com os maiores (positivos) e os $n_{qua}/4$ com os menores (negativos) valores de t para montar um novo vetor θ_a, o mesmo acontecendo com θ_b. Serão considerados modelos com $\mathbf{n_{qua}}$ variando de 60 até 400 dos 2*19*19=722 possíveis parâmetros de interação.

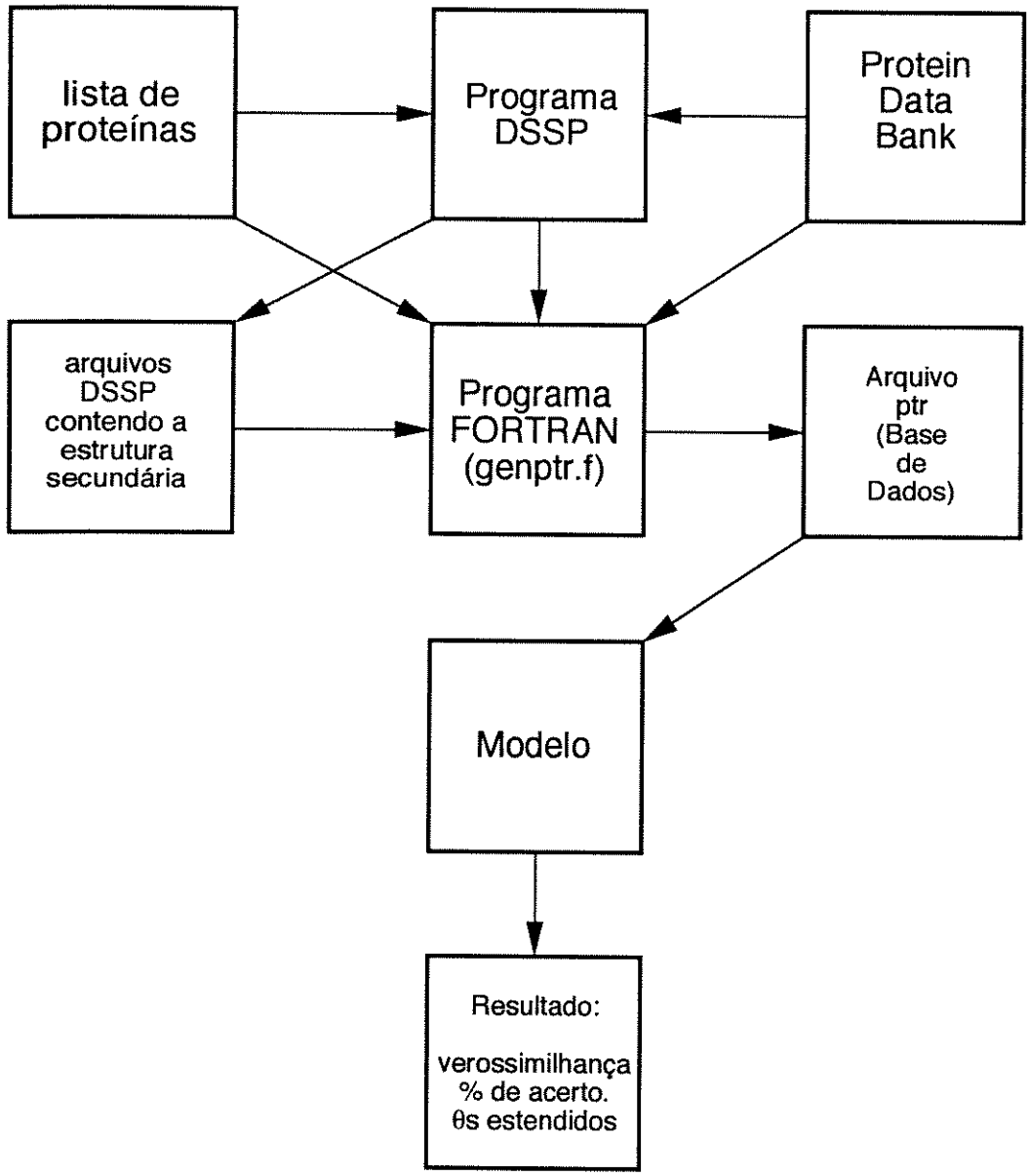


Figura 16 - Fluxograma do processo de ajuste de parâmetros do modelo. A lista contendo as cadeias e segmentos proteicos é fornecida ao programa DSSP (KABSCH & SANDER, 1983a), que gera arquivos contendo a estrutura secundária a partir dos dados em PDB (*Protein Data Bank*). Um segundo programa, GENPTR.F, gera a base de dados a partir dos arquivos DSSP, PDB e da lista de proteínas. A base é usada então por um programa em MATLAB (aqui representado pelo ítem "MODELO") que calcula os parâmetros de uma determinada implementação do modelo (definida pelo tamanho da janela, tamanho e forma das matrizes $\mathbf{A}$, $\mathbf{B}$, $\mathbf{W}$, Λ e do valor λ para a penalidade). O resultado consiste nos parâmetros otimizados, contidos nos vetores θ estendidos. A partir deles e do conjunto de dados podemos obter a acurácia (geral e individual para cada cadeia) e a verossimilhança.

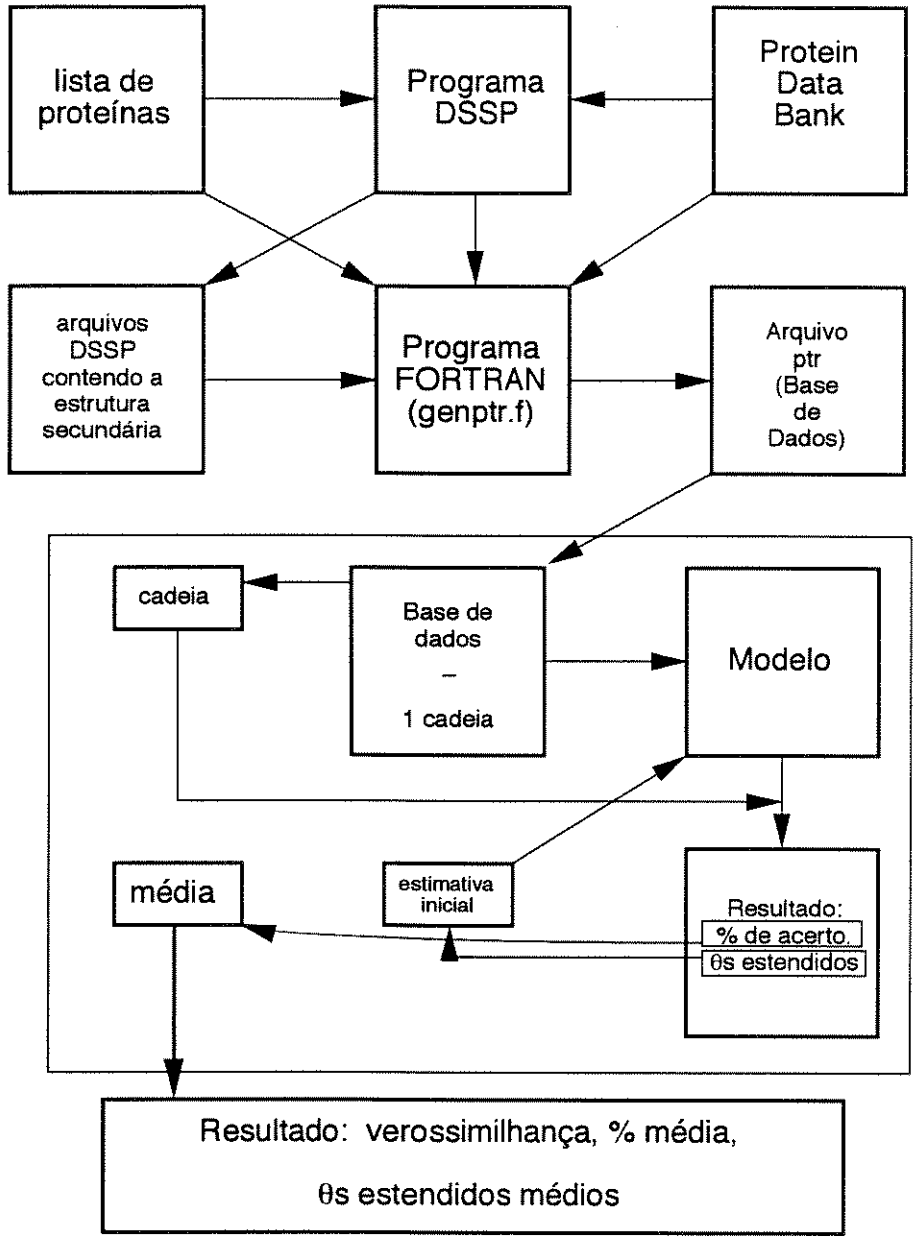


Figura 17 - Fluxograma do processo de validação cruzada. A criação da base de dados é feita como na Fig 16. A base é então usada por um programa em MATLAB (aqui representado pelo retângulo que envolve a parte inferior do fluxograma). Cada cadeia do conjunto é retirada dos dados e a base resultante é utilizada para calcular os parâmetros de uma determinada implementação do modelo (definida pelo tamanho da janela, tamanho e forma das matrizes $\mathbf{A}$, $\mathbf{\Pi}$, $\mathbf{U}$, Λ e do valor λ para a penalidade). Os parâmetros otimizados, contidos nos vetores θ estendidos são usados para prever a estrutura da cadeia retirada. A acurácia obtida, bem como os parâmetros e a verossimilhança, são armazenados para o cômputo das respectivas médias. A cadeia é recolocada nos dados e uma outra é retirada, repetindo-se o processo até que todas as cadeias tenham participado. Para economizar tempo, os parâmetros obtidos num ciclo são utilizados como estimativa inicial no próximo. O resultado consiste nos parâmetros médios, na acurácia média e na LVN. O cálculo é ponderado pelo comprimento das cadeias.

A seleção do melhor modelo também dependerá da acurácia da previsão para novas proteínas, fora do conjunto de dados. Isso foi realizado por meio de procedimento de validação por cruzamento. O conjunto de dados foi repetidamente construído deixando-se uma das sequências protéicas de fora. Calculou-se então a previsão da estrutura secundária desta proteína excluída. A média da acurácia destas previsões (assim como a LNV), ponderada pelo tamanho da sequência prevista, deve refletir a capacidade do modelo em prever novas proteínas. Este método é computacionalmente muito intensivo, requerendo k vezes mais tempo de processamento para cada ajuste do modelo, sendo k o número de cadeias protéicas na bases de dados (67). Os modelos foram testados inicialmente sem a validação cruzada, como exemplifica a Fig. 16. Para podermos realizar a validação por cruzamento, porções do programa tiveram que ser convertidas para FORTRAN. Como o MATLAB só é capaz de usar a vetorização para matrizes de duas dimensões, o procedimento mostrado na Fig. 17 de repetir o ajuste k vezes torna-se muito lento. Rotinas em FORTRAN podem ser mais eficientemente vetorizadas no Convex. Tais sub-rotinas podem ser encontradas nos Apêndices.

Note que a validação cruzada dos modelos com seleção de parâmetros foi feita somente sobre o ajuste dos parâmetros selecionados. Estes modelos devem ser considerados apenas *parcialmente* validados por cruzamento, uma vez que o processo de seleção de parâmetros não foi feito repetidas vezes, retirando-se uma cadeia de cada vez.

Pensou-se em utilizar o índice generalizado de validação cruzada (GCV, EUBANK, 1988) que se baseia no número de parâmetros efetivos. Contudo, já que toda uma sequência era retirada de cada vez (e não simplesmente uma janela local), e que os estados dos resíduos são correlacionados ao longo de uma mesma cadeia, o número de graus de liberdade para uso no cálculo do GCV ficava indeterminado.

3.9. As Constantes de Decisão

GIBRAT *et al.* (1987) também lançaram mão do uso de constantes de decisão, $\mathbf{DC_H}$ e $\mathbf{DC_E}$ (Eqs. 7 e 8). Estes valores foram por eles arbitrariamente escolhidos e refletem claramente a quantidade relativa de cada tipo de estrutura secundária no conjunto de dados.

Em nosso esforço de usar técnicas tradicionais para extrair o máximo de informação sobre a estrutura secundária contida na sequência local, não queríamos incluir um

procedimento desta natureza. De fato, os resultados que são apresentados como ótimos não o incluem. Não obstante, cada modelo também teve acurácias calculadas variando-se sistematicamente o valor de constantes de decisão equivalentes as de GIBRAT *et al.* (1987), ainda que definidas de modo distinto:

Sejam **C1** e **C2**, reais, tais que $0 \leq \mathbf{C1} \leq 1$ e $0 \leq \mathbf{C2} \leq 1$. O resíduo central de cada janela móvel tem sua classe estrutural escolhida da seguinte forma:

alfa: Se $(C2.p_a > C1.p_b)$ e $((1-C1-C2).p_a > C1.(1-p_a-p_b))$ (44)

beta: Se $(C2.p_a \leq C1.p_b)$ e $((1-C1-C2).p_a > C1.(1-p_a-p_b))$ (45)

coil: Nos demais casos.

4. Resultados

4.1. A Seleção do Modelo

Investigamos sistematicamente o equilíbrio entre qualidade de ajuste do modelo e sua complexidade. Um grande número de modelos foi testado, variando-se o tamanho da janela local, incluindo-se ou não o componente quadrático, restrito de diferentes formas (diferentes estruturas internas das matrizes **A** e **B**) e variando-se o grau de penalização causado pelo parâmetro λ. Para comparar estes modelos foi computada a logverossimilhança negativa (LVN), como medida da qualidade de ajuste e o número efetivo de parâmetros (Eq. 41), como medida da complexidade do modelo.

A Tabela 2 descreve os diversos tipos de modelos que foram testados. A primeira classe de modelos (i) são os puramente lineares. Verificamos o efeito do tamanho da janela local, variando seu tamanho de zero (uso das frequências médias de alfa, beta e *coil*) até 17 resíduos. Embora tenhamos utilizado as mesmas cadeias que GIBRAT *et al.* (1988), os arquivos PDB correspondentes não foram os mesmos, devido a atualizações e correções que seus autores publicaram. As versões mais recentes são de qualidade superior e tem estrutura definida com maior resolução. Isto implica numa maior quantidade de estrutura regular (mais alfa e beta, menos *random coil*). Para melhor comparar os resultados, reproduzimos a parte linear do método GGR (GIBRAT *et al.*, 1988, GARNIER *et al.*, 1978) usando a base de dados atualizada, e calculamos a LVN destes resultados. Não realizamos validação cruzada, mas o uso de

estruturas mais refinadas elevou o resultado não validado de 60.9% para 64.1%. A maior LVN, obtida pelo nosso método, mostra que o uso de um modelo otimizado foi capaz de obter um melhor ajuste aos dados. A melhora na acurácia, contudo, foi mínima.

A segunda classe (ii) representa o efeito da penalização (λ) da verossimilhança. Note-se que o modelo não penalizado obtém acurácia não validada e LVN extremamente elavados. Estes resultados não resistem ao processo de validação cruzada. O melhor resultado (maior LVN) foi obtido para λ=1000, o que corresponde a um modelo com cerca de 800 parâmetros efetivos. A acurácia validada foi de 62.5%.

A terceira classe (iii) representa variações no tamanho e estrutura interna das matrizes de parâmetros quadráticos **A** e **B**. Os resultados apresentados representam o efeito da restrição da distância entre pares de aminoácidos (distâncias acima de um teto foram desconsideradas). Nestes casos foi mantida uma janela de 17 aminoácidos para as partes linear e quadrática. O efeito sobre as estruturas de **A** e **B** foi ilustrado na Figura 18. Note-se que o número total de parâmetros permanece 1370. O número efetivo cai devido a menor contagem de contatos entre pares de aminoácidos e consequente maior penalização dos parâmetros correspondentes. Este procedimento não melhorou a acurácia, de forma que o melhor modelo leva em conta todos os contatos entre pares de resíduos na janela de 17 amino ácidos.

A quarta classe (iv) mostra o resultado da seleção de parâmetros. Na segunda coluna podemos observar o número efetivo de parâmetros. Os valores entre parênteses são os parâmetros quadráticos selecionados (n_{qua}). Quando n_{qua}=400, restrições se tornam ativas (somas das linhas e somas das colunas de **A** e de **B** devem ser zero), reduzindo o número de parâmetros independentes. O menor número de parâmetros quadráticos permite o uso de uma penalidade (λ) menor. Note que o melhor modelo tem, como nas seções (ii) e (iii) da mesma tabela, cerca de 800 parâmetros efetivos, atingindo uma acurácia de 65.9%.

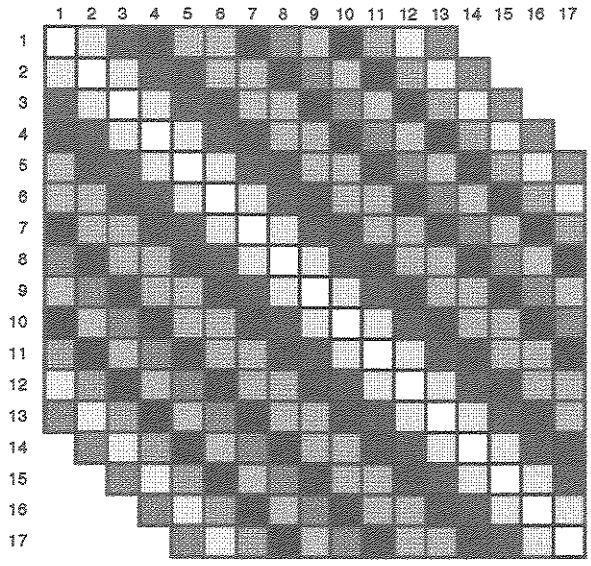
Tabela 2 - Seleção dos melhores modelos preditivos

N	param.	Garnier atualiz.	λ	-Veross.	%	-Veross. cruzada	% cruzada
			(i) Seleção da janela principal (linear)				
0	2	49.5	-	11617	-	-	-
1	40	50.9	-	11056	50.9	11372	46.8
3	116	57.1	-	10186	57.5	10550	54.9
5	192	59.7	-	9706	59.6	9985	58.2
7	268	61.0	-	9444	60.9	9810	58.7
9	344	61.7	-	9267	61.9	9716	59.6
11	420	62.4	-	9110	62.4	9643	58.8
13	496	63.1	-	8955	63.4	9577	60.2
15	572	63.6	-	8825	64.3	9542	60.7
17	648	**64.1**	∞	8714	**64.7**	9525	**61.0**
			(ii) Seleção do fator de penalização				
17	1370		0	6543	75.4	12759	58.1
17	1308		10	6575	75.3	11764	59.0
17	1099		100	6926	74.2	10079	61.4
17	804		1000	7867	69.9	9373	**62.5**
17	672		10000	8558	65.6	9476	61.3
			(iii) Seleção do tamanho da janela (modelo quadrático)				
17	804		1000	7867	69.9	9373	**62.5**
13	754		1000	8140	68.1		
11	729		1000	8307	67.0	9346	62.1
9	705		1000	8453	66.2		
7	683		1000	8577	65.4		
5	665		1000	8661	64.9		
17	1099		100	6926	74.2	10079	61.4
13	1032		100	7171	72.9		
11	988		100	7424	71.6	9910	61.9
9	930		100	7708	70.4		
7	858		100	8039	68.4		
5	773		100	8382	66.5		
			(iv) Seleção de Variáveis				
-	648 (0)		∞	8714	64.8	9525	61.0
17	708 (60)		1	7884	69.9	8973	65.0
17	728 (80)		1	7754	70.2	8952	65.1
17	768 (120)		1	7509	71.1	**8941**	65.5
17	808 (160)		1	7229	72.0	8956	**65.9**
17	1045 (397*)		1	6796	74.4	9911	63.6
17	1370 (722*)		0	6543	75.4	12759	58.1

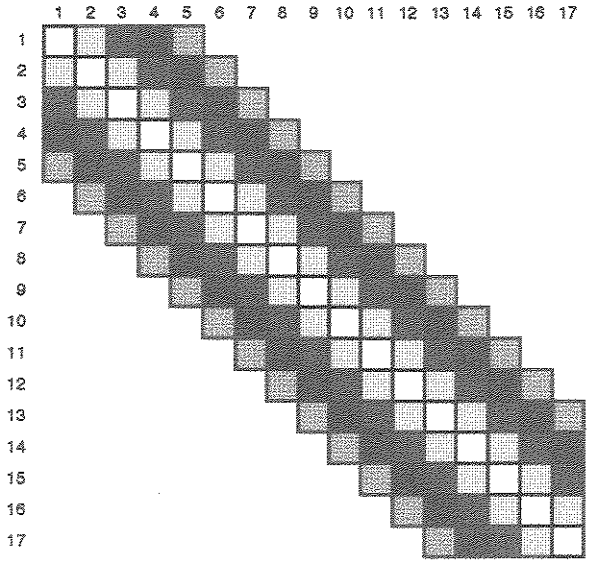
* restrições se tornaram ativas.

Resultado da variação do tamanho da janela local (i), da intensidade da penalização da verossimilhança (ii), da estrutura interna das matrizes **A** e **B** (iii) e da seleção de parâmetros (iv). A primeira coluna (*N*) mostra o tamanho da janela local. No caso dos modelos quadráticos (iii), este tamanho corresponde a uma restrição sobre a máxima distância entre pares de aminoácidos, como mostrado na Fig. 18. A segunda coluna (*param.*) mostra o número de parâmetros de cada modelo. No caso dos modelos penalizados (ii, iii e iv) foi mostrado o número efetivo de parâmetros, calculado de

Tabela 2 (cont.) - acordo com a Eq. 41, (EUBANK, 1990). No caso de seleção de variáveis (iv) o número entre parênteses é a quantidade de parâmetros quadráticos selecionados. A terceira coluna (*Garnier atualiz.*), presente apenas no caso linear (i), é a acurácia obtida pelo método de GARNIER *et al.*, (1978), reproduzido com dados cristalográficos atuais, sem validação cruzada. Este método equivale ao de GIBRAT *et al.* (1987), sem sua porção quadrática. Este resultado deve ser comparado com o obtido pelo nosso método, que se encontra na sexta coluna. A quarta coluna (λ) traz o valor da penalidade λ, aplicável aos modelos que tiveram a verossimilhança penalizada (ii, iii, iv). O modelo puramente linear equivale ao modelo penalizado com $\lambda=\infty$. Somente o modelo linear com N=17 foi assinalado com $\lambda=\infty$ porque todos os modelos penalizados apresentados nesta tabela usaram uma janela local de 17 aminoácidos na sua porção linear, inclusive os do item *iii* (veja Fig. 18). Os valores de λ dos modelos com seleção de variáveis são menores (e.g. $\lambda=1$) porque houve menor necessidade de penalização, devido ao menor número de parâmetros utilizados e a sua melhor sua qualidade (significância dada pelo *t* de Student, Eqs. 42 e 43). A quinta coluna (*LVN*) traz o valor da logverossimilhança negativa (oposto da Eq. 34), que reflete a qualidade do ajuste aos dados (quanto menor, melhor), sem validação cruzada. A sexta coluna (%) traz a acurácia obtida no conjunto de dados, sem validação cruzada. A sétima coluna (*LVN cruzada*) traz o logaritmo da soma das verossimilhanças individuais, com o sinal trocado, obtido após o processo de validação cruzada, refletindo a generalidade do modelo, ou seja, sua capacidade de prever a estrutura secundária de proteínas novas, diferentes daquelas contidas no conjunto de dados. Finalmente, a oitava coluna (*% cruzada*) traz a acurácia média obtida após a validação. Não foi realizada validação cruzada nos modelos do item *iii* que apresentaram LNV elevada. O contraste entre os valores da LVN, com e sem validação, pode ser melhor apreciado na Fig. 19. Note que embora a melhor acurácia (65.9%) tenha sido obtida pelo modelo de 808 variáveis (item iv), a menor LVN corresponde ao modelo de 768 variáveis, que apesar de uma acurácia ligeiramente menor (65.5%), tem maior chance de sucesso em proteínas diferentes.



(a)



(b)

Figura 18 - Efeito da restrição da distância entre pares de aminoácidos sobre a estrutura interna de **A** e **B**. A estrutura bandeada segue o conteúdo de $\mathbf{W}^{3.6}$ (Eq. 24). Embora a janela continue tendo 17 aminoácidos, as interações entre pares de resíduos só são consideradas até uma certa distância *d*. Os exemplos aqui mostrados são para $d = 13$ (a) e $d = 5$ (b). Note-se que isto só reduz o número de parâmetros se houver penalização à verossimilhança. Continuam havendo 400 pares de amino ácidos. Contudo, são contados menos contatos, aumentando a instabilidade (e a consequente penalização) dos pares raros, de forma que um mesmo valor de λ leva um número menor de parâmetros efetivos. Estes modelos correspondem ao item *iii* da Tabela 2.

A Figura 19 mostra uma versão gráfica do conteúdo da Tabela 2, e nos foi especialmente útil na busca do melhor modelo (menor LVN) que tivesse um número adequado de parâmetros (empiricamente 800). Ali também estão representados outros modelos testados mas não mostrados na Tab. 2. Como esperado, o ajuste melhora com o aumento da complexidade do modelo (até 1370 parâmetros), atingindo uma LVN de 6543. A acurácia seguiu uma tendência semelhante (de sentido inverso) aumentando de 51% para 75% (Tab. 2, não mostrado na Fig. 19). O objetivo da seleção de modelos é encontrar aqueles que resultam no melhor ajuste, com um grau de complexidade adequado. Assim, foram adicionados parâmetros aos modelos até que a qualidade do ajuste não melhorasse significativamente. De modo geral, a melhora do ajuste é cada vez menor na medida que aumenta a complexidade. A adição do componente quadrático (648 a 1000, nas abscissas) apresentou um ganho muito maior do que a adição dos últimos parâmetros lineares (400 a 640, nas abscissas). É possível testar a significância estatística da redução da LVN com a complexidade do modelo usando-se o teste χ^2. Este teste revelou que a queda de 2171 pontos na LVN entre o melhor modelo linear (648 parâmetros) e o modelo quadrático completo (1370 parâmetros) foi significativa ($P<.001$).

Uma maneira mais apropriada de selecionar o melhor modelo preditivo é através da validação cruzada da LVN (valor obtido retirando-se a proteína a ser prevista do conjunto de dados e calculando-se os parâmetros). Pela intensidade computacional do método, um número menor de modelos foi testado deste modo. Os resultados estão na Figura 19 (curva superior, **x**). Como esperado, obteve-se uma curva em forma de U, com um mínimo achatado no meio, indicando que a complexidade ideal para o modelo, dado o tamanho do conjunto de dados, é em torno de 800 parâmetros. O braço direito do U apresenta inclinação acentuada, mostrando o efeito do número excessivo de parâmetros. Os modelos mais complexos refletiram aspectos específicos deste conjunto de dados e não seriam capazes de "generalizar" ou obter bons resultados fora dele. Um dos modelos com melhor resultado corresponde ao uso de λ=1000 e resultou em uma acurácia de 62.5% validada por cruzamento (Tab. 2).

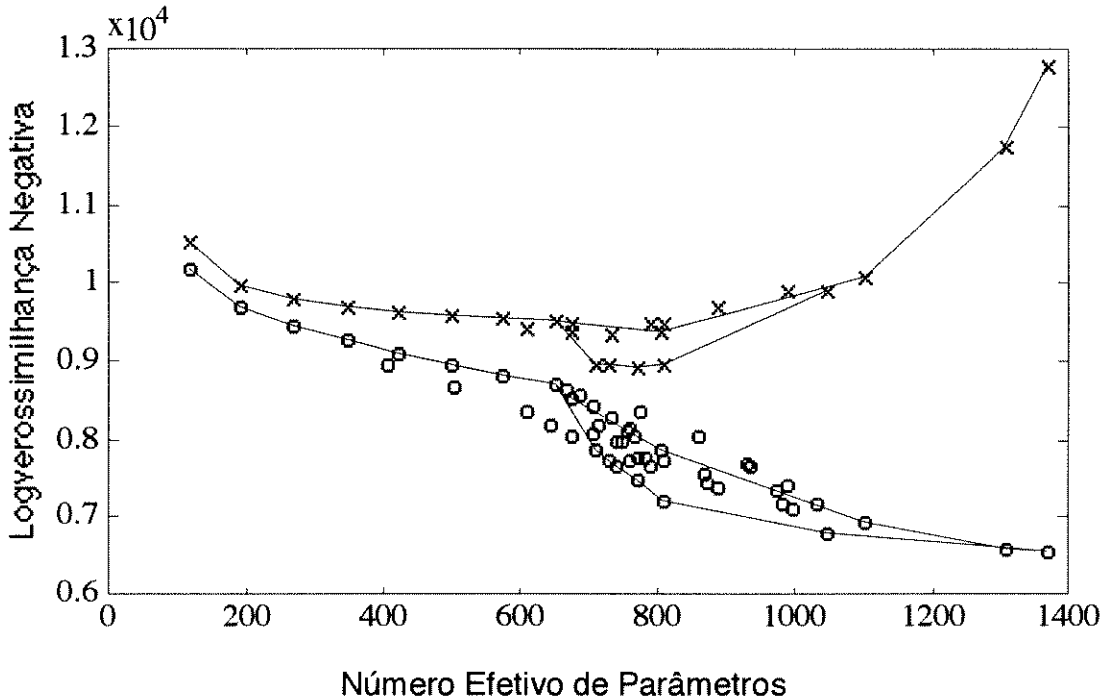


Figura 19 - LVN para modelos com diferentes quantidades de parâmetros. Os resultados não validados (**o**) têm um melhor ajuste (menor LVN). As linhas ligam séries de modelos com complexidade crescente, inicialmente pelo aumento da janela, depois pela adição de parâmetros quadráticos com penalidade decrescente. Os resultados validados por cruzamento (**x**) tem maior LNV. As curvas inferiores em cada série correspondem aos modelos com seleção de parâmetros.

Um resultado ainda melhor foi obtido usando-se a seleção de parâmetros (Fig. 19, porção inferior de ambas as curvas). Aqui, os parâmetros quadráticos foram adicionados um de cada vez na sua ordem de importância, num procedimento análogo à regressão linear escalonada (DRAPER & SMITH, 1981). A melhor seleção de parâmetros teve 768 deles (648 lineares e 120 quadráticos) e atingiu uma acurácia validada de 65.5%. O modelo com 808 parâmetros atingiu 65.9%, a maior acurácia já publicada para este conjunto de dados, embora sua LVN validada fosse ligeiramente maior (pior). Devemos frizar que não foi realizada validação cruzada do processo de seleção de parâmetros, somente no processo de estimativa. Assim o valor validado adequadamente pode ser menor.

Outros modelos foram testados sem que apresentassem resultados satisfatórios. As principais idéias utilizadas foram: (1) Fazer com que a porção linear levasse em conta a distância do resíduo central a cada posição da janela local, inclusive com periodicidade. (2) Variar o tamanho da janela local para ambas as partes do modelo. (3) Diminuir ainda mais o número de parâmetros quadráticos fazendo com que as matrizes **A** e **B** fossem simétricas (efeito da interação entre aminoácidos independeria de sua ordem na sequência). Nenhuma destas suposições apresentou melhor resultado que os modelos apresentados na Tab. 2.

4.2. Os Parâmetros Lineares

Os parâmetros contidos no vetor **a** da Eq. 19 representam a influência do resíduo de posição $i+m$ na janela local sobre a estrutura do resíduo central i. Isto corresponde à "informação direcional" de GIBRAT *et al.* (1988, Eq. 3). Reordenando-se os elementos de **a** pode ser dada uma interpretação físicoquímica aos achados do modelo. A Figura 20 ilustra tal interpretação e a Figura 21 traz o conteúdo de **a** reordenado em forma de matriz e codificado por cores. Note-se que a escala de cores utilizada não é linear, o que intensifica o padrão encontrado, que será discutido adiante. A Figura 22 contém a codificação correspondente para os elementos de **b** (Eq. 19). A escala de cores também é não linear e difere da utilizada para **a**. Os valores numéricos de $a_{i,j}$ e $b_{i,j}$ se encontram nas Tabelas 3 e 4 respectivamente. Os parâmetros apresentados correspondem ao modelo exclusivamente linear, com acurácia validada de 61%, apresentado na décima linha da Tab. 2.

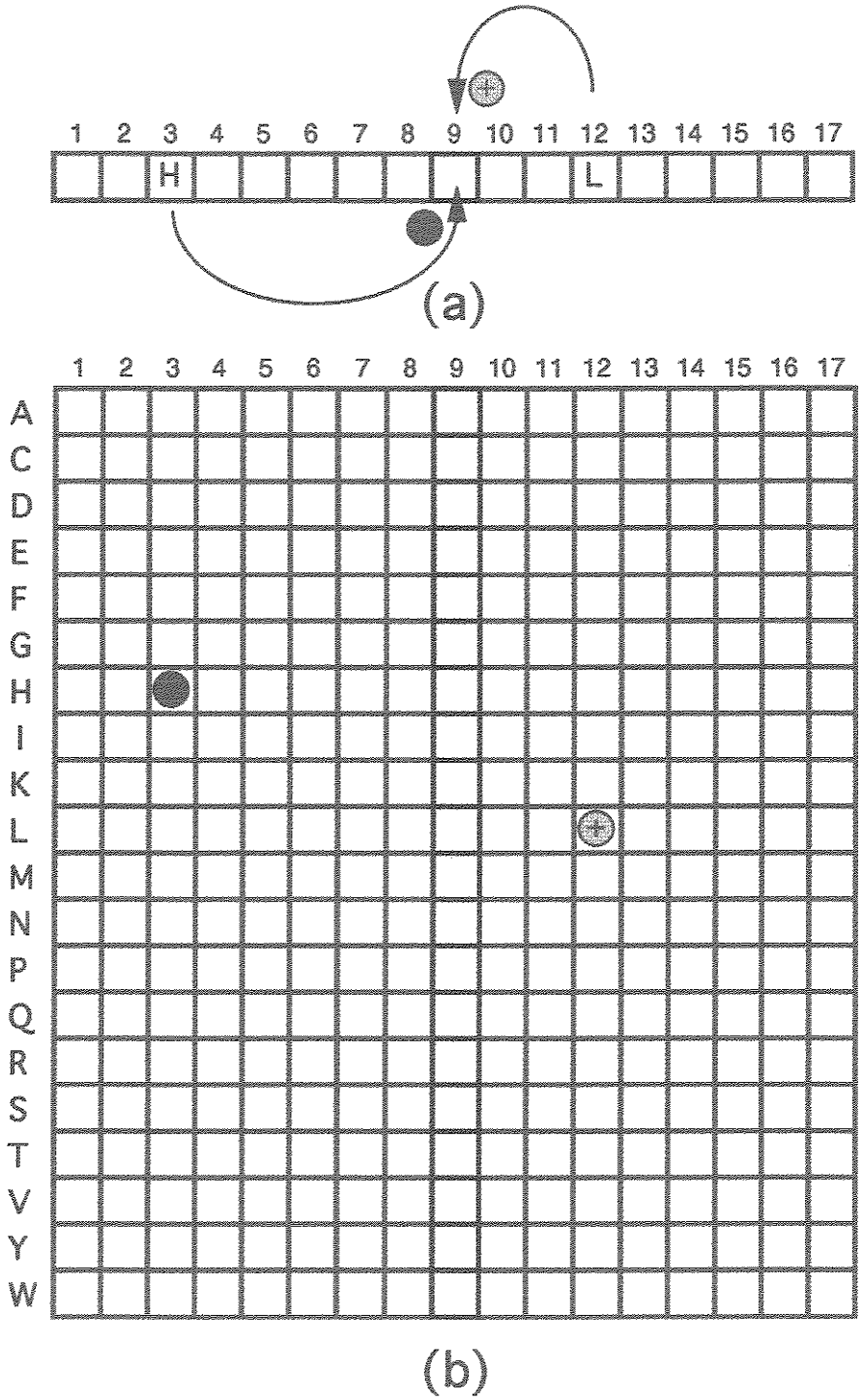


Figura 20 - Interpretação funcional dos parâmeros lineares. (a) Janela de 17 resíduos. Representamos aqui a situação hipotética de um resíduo (histidina, H) que desfavorece a ocorrência de uma determinada estrutura secundária (e.g. alfa) na posição central (9). Aqui também se encontra um outro aminoácido (leucina, L) que, em outra posição (12), favorece o surgimento do mesmo estado. (b) Parâmetros lineares contidos no vetor **a** (Eq. 19), rearranjados em forma matricial para facilitar a interpretação. As linhas correspondem a cada um dos aminoácidos, aqui mostrados por seus códigos. As colunas correspondem às posições da janela local. Cada elemento representa a influência de um aminoácido (rótulo da linha) sobre a estrutura do resíduo central, quando ocupa uma determinada posição da janela local (rótulo da coluna). As situações descritas na Fig. 20a resultam aqui em parâmetro negativo para a histidina na posição 3 ($a_{7,3}<0$) e parâmetro positivo para a leucina na posição 12 ($a_{10,12}>0$).

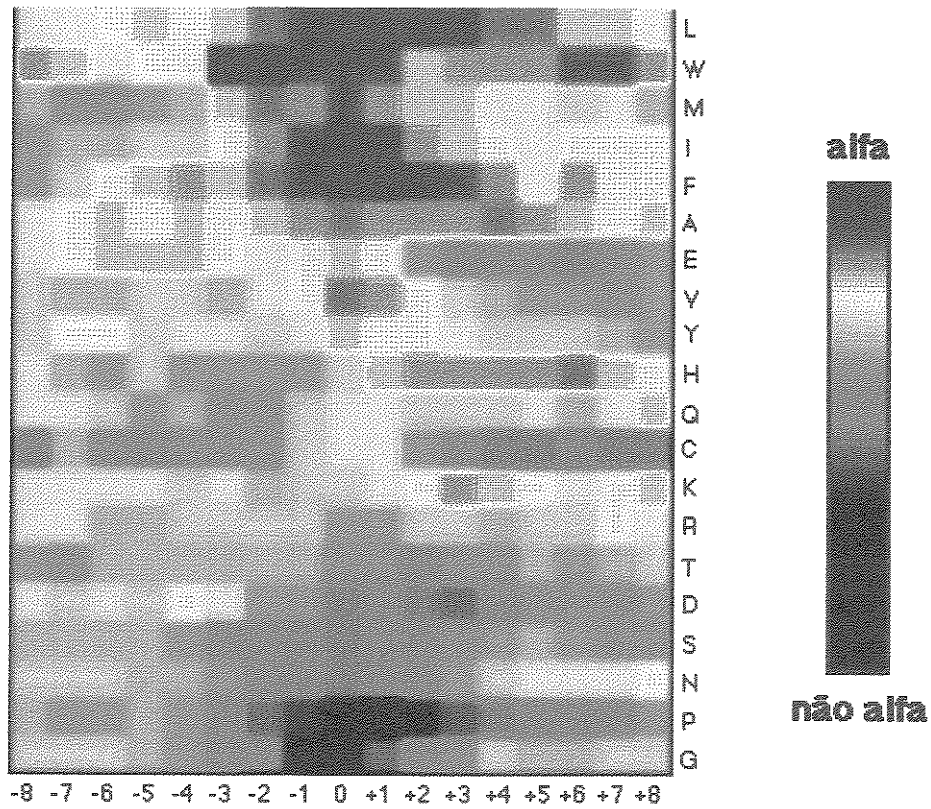


Figura 21 - Representação gráfica dos parâmetros lineares "alfa". Foi usada uma escala colorida, não-linear, mostrada à direita, para codificar os valores dos elementos do vetor **a** (Eq. 19), aqui rearranjados de maneira semelhante à exposta na Fig. 20, porém com linhas ordenadas de acordo com seu valor médio. O envelope da escala tem "forma de sino", assimétrica, que evidencia arbitrariamente os padrões citados no texto. O mesmo pode ser dito do filtro de difusão que foi aplicado à imagem.

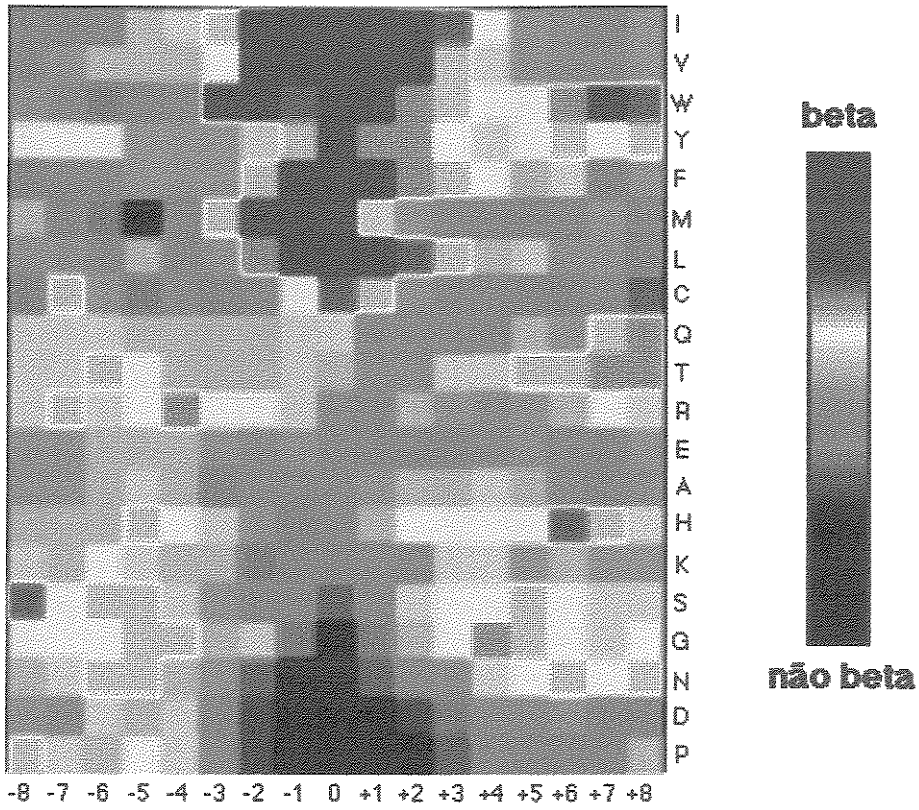


Figura 22 - Representação gráfica dos parâmetros lineares "beta". Foi usada uma escala colorida, não-linear, diferente da usada na Fig. 21, mostrada à direita, para codificar os valores dos elementos do vetor **b** (Eq. 19), aqui rearranjados de maneira semelhante à exposta na Fig. 20, porém com linhas ordenadas de acordo com seu valor médio. O envelope da escala tem "forma de sino", assimétrica, que evidencia arbitrariamente os padrões citados no texto. O mesmo pode ser dito do filtro de difusão que foi aplicado à imagem.

Tabela 3 - Conteúdo do vetor **a** reordenado para interpretação

	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8
L	0.15	0.13	0.20	0.34	0.16	0.25	0.49	0.63	0.83	0.90	0.82	0.73	0.49	0.46	0.33	0.34	0.11
W	0.43	0.30	0.10	0.21	0.24	0.76	0.83	0.55	0.57	0.60	0.28	0.43	0.38	0.41	0.64	0.60	0.41
M	-0.14	-0.29	-0.27	-0.22	-0.09	0.29	0.52	0.40	0.58	0.35	0.35	0.25	0.10	0.13	0.03	0.12	-0.14
I	-0.16	-0.12	-0.15	-0.04	0.00	0.16	0.40	0.55	0.71	0.73	0.38	0.31	0.21	0.09	0.22	0.20	0.25
F	-0.24	-0.04	0.15	0.31	0.36	0.26	0.48	0.57	0.71	0.85	0.69	0.59	0.39	0.12	0.36	0.17	0.16
A	0.14	0.20	0.27	0.24	0.28	0.24	0.33	0.35	0.47	0.44	0.40	0.38	0.52	0.42	0.29	0.19	0.25
E	0.06	0.19	0.31	0.33	0.28	0.17	0.15	0.24	0.28	0.17	-0.15	-0.43	-0.34	-0.31	-0.41	-0.33	-0.22
V	-0.04	-0.06	-0.08	-0.01	-0.03	-0.08	0.08	0.24	0.48	0.35	0.20	-0.01	-0.06	-0.16	-0.21	-0.26	-0.23
Y	0.04	0.17	0.19	-0.03	0.07	0.09	0.01	0.06	0.31	0.16	0.23	0.06	-0.03	0.03	-0.03	-0.13	-0.21
H	0.06	-0.13	-0.17	-0.04	-0.21	-0.16	-0.19	-0.16	-0.01	0.32	0.41	0.41	0.41	0.37	0.46	0.27	0.24
Q	0.14	0.08	0.01	-0.22	-0.14	-0.28	-0.32	0.00	0.14	0.10	0.03	0.02	0.00	0.08	-0.06	0.12	0.26
C	-0.46	-0.12	-0.29	-0.41	-0.31	-0.29	-0.28	-0.02	0.14	0.10	-0.22	-0.43	-0.73	-0.68	-0.63	-0.70	-0.65
K	0.09	0.06	0.10	0.07	0.04	0.06	-0.06	0.03	0.02	0.06	0.18	0.39	0.32	0.21	0.14	0.22	0.28
R	0.12	0.15	-0.09	-0.12	0.03	-0.04	0.05	-0.03	-0.20	-0.24	-0.03	-0.01	-0.09	-0.04	0.00	0.24	0.10
T	-0.19	-0.26	-0.10	-0.06	-0.08	-0.13	-0.08	-0.11	-0.29	-0.31	-0.47	-0.28	-0.27	-0.15	-0.21	-0.11	-0.02
D	0.09	0.02	0.09	0.02	0.20	0.12	-0.20	-0.25	-0.69	-0.56	-0.69	-0.75	-0.46	-0.28	-0.27	-0.27	-0.23
S	-0.06	-0.12	-0.13	-0.12	-0.21	-0.38	-0.46	-0.48	-0.68	-0.48	-0.38	-0.31	-0.18	-0.15	-0.20	-0.26	-0.32
N	0.02	0.02	0.00	-0.08	-0.13	-0.17	-0.35	-0.51	-0.73	-0.55	-0.44	-0.17	0.05	0.11	0.12	0.07	0.19
P	-0.08	-0.15	-0.16	-0.08	-0.20	-0.43	-0.78	-0.96	-1.24	-2.16	-1.21	-0.90	-0.66	-0.52	-0.48	-0.43	-0.24
G	0.01	-0.03	-0.02	-0.09	-0.24	-0.41	-0.60	-1.11	-1.41	-0.83	-0.35	-0.27	-0.05	-0.14	-0.09	-0.04	0.03

Os parâmetros "alfa" (vetor **a**, Eq. 19) do modelo puramente linear (Eqs. 10 e 11) foram listados aqui, no mesmo formato em que se encontram na Fig. 21. Sua influência sobre o resíduo central (coluna *0*, no centro da tabela) pode ser interpretada de acordo com o processo descrito na Fig. 20. Aqui pode-se apreciar a magnitude dos valores numéricos, sem o efeito da tabela de cores ou do filtro de difusão, usados na Fig. 21.

Tabela 4 - Conteúdo do vetor **b** reordenado para interpretação

	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8
I	-0.20	-0.30	-0.30	0.00	0.10	0.30	0.80	1.10	1.20	1.20	0.70	0.50	0.20	-0.10	-0.20	-0.30	-0.10
V	-0.20	-0.10	0.00	0.00	0.00	0.20	0.70	0.90	1.20	1.10	0.70	0.30	0.10	-0.10	-0.20	-0.20	-0.10
W	-0.10	-0.20	-0.50	-0.30	-0.20	0.60	0.60	0.50	0.60	0.70	0.40	0.30	0.20	0.20	0.40	0.60	0.50
Y	0.20	0.20	0.20	-0.10	-0.10	-0.10	0.10	0.30	0.60	0.40	0.40	0.20	0.10	0.20	0.30	0.20	0.30
F	-0.30	-0.20	-0.20	-0.10	-0.10	-0.10	0.30	0.60	0.80	0.80	0.40	0.30	0.20	0.00	0.10	-0.30	-0.10
M	0.00	-0.30	-0.50	-0.90	-0.40	0.30	0.60	0.70	0.80	0.10	-0.10	-0.30	-0.30	-0.30	-0.40	-0.10	-0.30
L	-0.10	-0.10	-0.20	0.00	-0.20	-0.10	0.40	0.60	0.80	0.80	0.50	0.30	0.00	0.10	-0.10	-0.10	-0.30
C	-0.30	0.30	-0.10	-0.50	-0.20	-0.30	-0.20	0.20	0.50	0.30	-0.10	-0.30	-0.40	-0.40	-0.30	-0.20	-0.60
Q	0.10	0.10	0.10	0.00	0.00	0.00	0.00	0.10	0.10	-0.40	-0.40	-0.40	-0.30	0.00	-0.10	0.30	0.40
T	0.10	0.10	0.30	0.20	0.00	0.00	0.00	0.10	0.00	-0.20	-0.20	0.10	0.10	0.30	0.30	0.40	0.40
R	0.10	0.30	0.10	0.20	0.40	0.20	0.20	0.10	-0.20	-0.30	0.00	-0.10	-0.20	-0.10	0.00	0.20	0.10
E	-0.10	-0.10	0.00	0.00	0.00	-0.10	-0.20	-0.10	-0.20	-0.30	-0.30	-0.30	-0.20	-0.10	-0.40	-0.30	-0.10
A	-0.10	-0.10	0.00	0.10	0.00	-0.30	-0.20	-0.20	-0.20	-0.10	0.00	0.00	0.10	0.00	-0.20	-0.30	-0.20
H	0.00	0.00	0.00	0.30	0.20	0.10	-0.10	-0.30	-0.20	0.00	0.20	0.20	0.20	0.20	0.50	0.30	0.10
K	0.10	0.00	0.20	0.10	0.00	0.00	-0.50	-0.30	-0.30	-0.40	-0.30	0.10	0.10	-0.10	0.00	-0.10	-0.10
S	0.50	0.20	0.30	0.30	0.10	-0.10	-0.30	-0.40	-0.60	-0.30	0.10	0.20	0.20	0.30	0.20	0.10	0.10
G	0.20	0.20	0.20	0.30	0.30	0.00	0.10	-0.50	-0.90	-0.50	0.00	0.20	0.40	0.30	0.20	0.10	0.20
N	0.00	0.10	0.30	0.30	0.00	-0.10	-0.60	-1.10	-1.20	-0.60	-0.40	-0.20	0.10	0.20	0.30	0.20	0.30
D	-0.10	-0.20	0.10	0.00	0.10	-0.30	-0.70	-1.00	-1.50	-1.00	-0.70	-0.50	-0.20	-0.20	-0.10	-0.30	-0.30
P	0.30	0.10	0.00	0.20	0.10	-0.20	-0.70	-1.20	-1.40	-1.40	-1.00	-0.60	-0.40	-0.30	-0.30	-0.30	0.00

Os parâmetros "beta" (vetor **b**, Eq. 19) do modelo puramente linear (Eqs. 10 e 11) foram listados aqui, no mesmo formato em que se encontram na Fig. 22. Sua influência sobre o resíduo central (coluna *0*, no centro da tabela) pode ser interpretada de acordo com o processo descrito na Fig. 20. Aqui pode-se apreciar a magnitude dos valores numéricos, sem o efeito da tabela de cores ou do filtro de difusão, usados na Fig. 22.

4.3. Os Parâmetros Quadráticos

Na medida em que existem 20 aminoácidos, existem 20x20=400 possíveis interações, que foram consideradas separadamente para estruturas alfa e beta, elevando seu número para 800. Após as devidas restrições, os modelos mais complexos tinham 722 parâmetros quadráticos. Foi impossível obter estimativas confiáveis para todos eles, dada a limitação no tamanho da base de dados. De fato, sem o uso da penalidade, não foi detectado nenhum padrão coerente no conjunto de parâmetros. Já com a incorporação desta técnica (λ=1000) padrões surgiram na estrutura da matriz de estimativas, relacionada à formação de alfa hélices

Os elementos da submatriz **A** representam a influência da interação entre dois resíduos da janela local sobre a estrutura do resíduo central. A ordem dos resíduos na sequência foi considerada. A influência da distância e posição do par foi dada por **W** durante o processo de ajuste. Note que esta interpretação dos parâmetros quadráticos não encontra correspondência no modelo de GIBRAT *et al.* (1987). Na realidade enquanto estes autores aproximaram a Eq. 2 (equivalente a um modelo logístico quadrático completo, Eq. 19 e 20) através do uso das informações "próprias", "direcionais", "pareadas" e de observações fictícias, nossa estratégia foi a imposição de restrições às matrizes **A** e **B**, usando periodicidade, mais uso de penalização da verossimilhança.

A matriz **A** foi então rearranjada, de forma a agrupar aminoácidos parecidos e facilitar a interpretação. A Fig. 23 mostra o seu conteúdo, codificado por cores. Novamente a escala de cor não é linear. O mesmo foi feito para **B** na Fig. 24, com escala diferente. Esses parâmetros foram obtidos do melhor modelo da seção *ii* da Fig. 2 (que obteve sucesso em 62.5% dos resíduos). Os padrões encontrados em ambas as matrizes são discutidos adiante.

A significância de cada parâmetro quadrático foi calculada conforme descrito (Eqs. 42 e 43). Os principais valores, correpondentes a interação de pares de aminoácidos, são apresentados nas Tabelas 5 a 8 de acordo com o modo como influenciam a ocorrência de cada estrutura secundária. Foram mostrados apenas os parâmetros com valores de $t > 2.0$ (aproximadamente correspondendo à uma significância $P<0.05$). Ali também foram incluídas informações sobre a plausibilidade do efeito da interação sobre a estrutura, assim como comparação com os achados de GIBRAT *et al.* (1987). Foram encontrados duas vezes mais parâmetros signicativos em **A** do que em **B**.

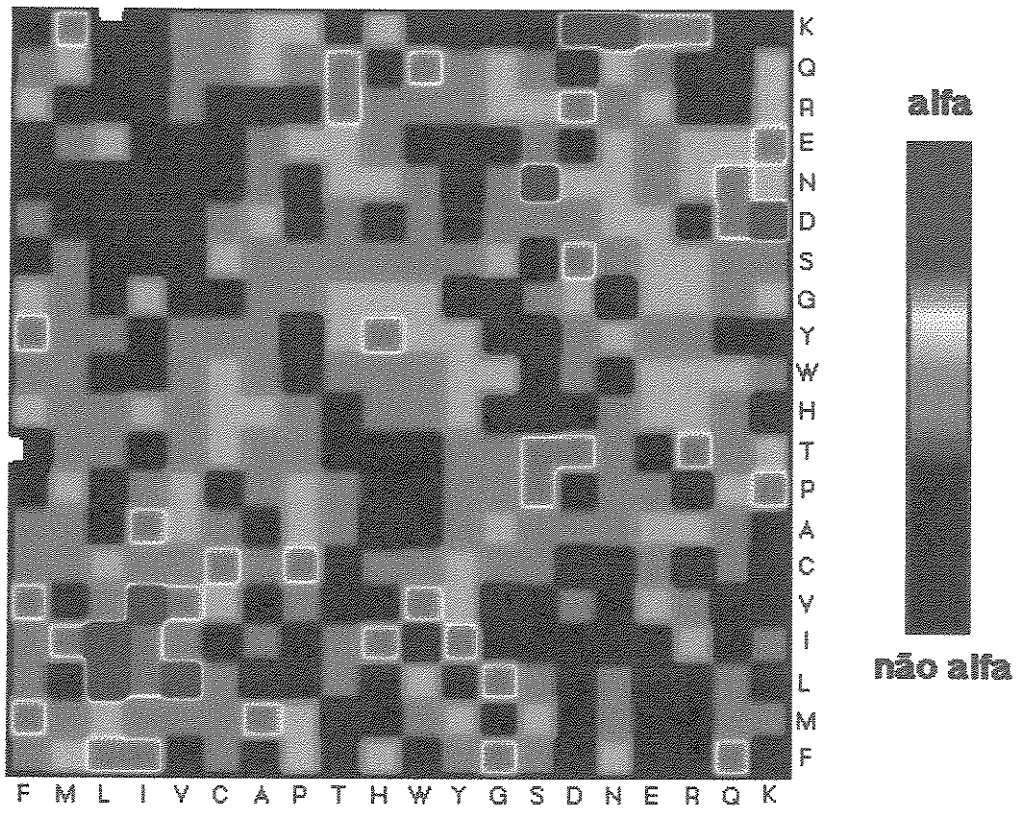


Figura 23 - Representação dos parâmetros quadráticos "alfa". Foi utilizada uma escala de cor, não linear, mostrada à direita, para codificar os elementos da submatriz **A**. Como descrito no texto, esta matriz corresponde a cada um dos 340 blocos que formam a matriz **R**. As linhas e colunas foram rearranjadas de forma a agrupar aminoácidos semelhantes (Fig. 2). O envelope da escala tem "forma de sino", assimétrico, e foi escolhido para evidenciar arbitrariamente os padrões discutidos no texto. O mesmo pode ser dito do filtro de difusão usado na imagem.

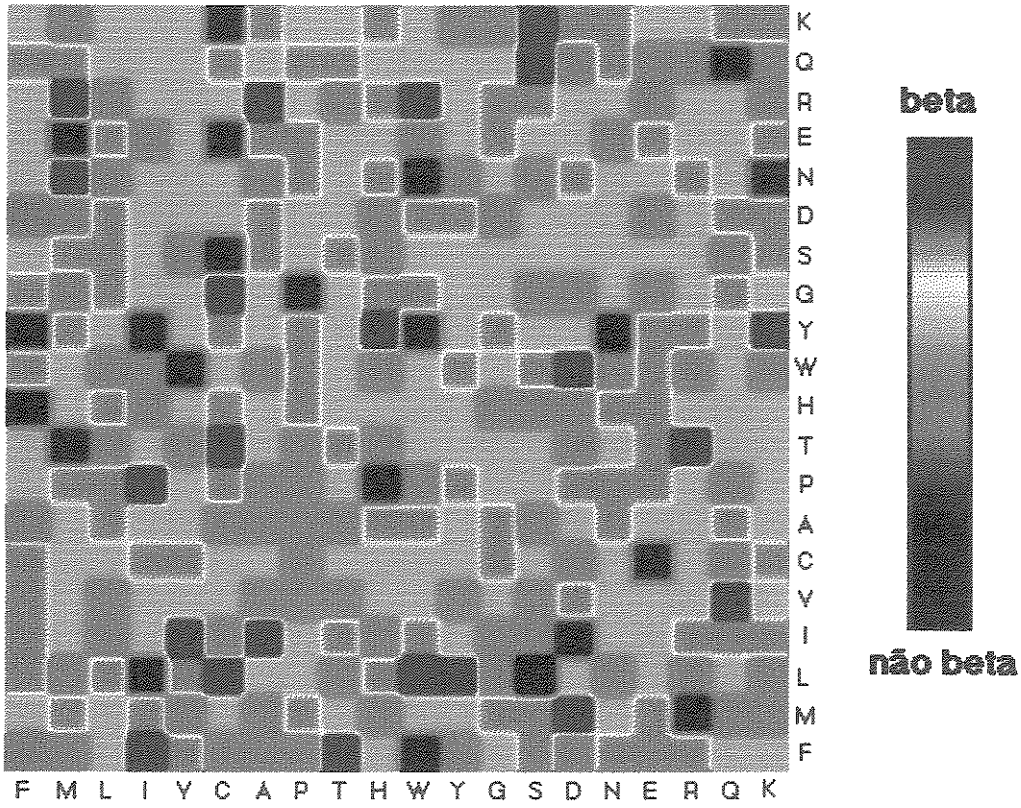


Figura 24 - Representação dos parâmetros quadráticos "beta". Foi utilizada uma escala de cor, não linear, mostrada à direita, diferente da usada na Fig. 23, para codificar os elementos da submatriz **B**. Como descrito no texto, esta matriz corresponde a cada um dos 340 blocos que formam a matriz **B**. As linhas e colunas foram rearranjadas de forma a agrupar aminoácidos semelhantes (Fig. 2). O envelope da escala tem "forma de sino", assimétrico, e foi escolhido arbitrariamente, assim como o filtro de difusão usado na imagem.

Tabela 5 - Parâmetros Quadráticos mais Significativos

Elementos Positivos da Sub-matriz **A**

Res. i	Res. j	t	Aij	Classe i	Classe j	Plausível S	Plausível N	GGR
L	L	4.42	0.22	np	np	*		+
V	L	4.05	0.2	np	np	*		+
L	I	4	0.32	np	np	*		+
I	V	3.94	0.42	np	np	*		
K	D	3.65	0.33	+	-	*		+
K	E	3.19	0.19	+	-	*		
D	K	3.16	0.39	-	+	*		+
V	V	2.99	0.18	np	np	*		
L	F	2.92	0.27	np	np	*		+
E	K	2.89	0.27	-	+	*		+
S	T	2.8	0.32	sp	sp	*		
R	T	2.59	0.35	+	sp	*		
G	A	2.53	0.16	§	†			+/-
Y	I	2.53	0.34	np	np	*		
N	K	2.37	0.3	p	+	*		
S	N	2.35	0.31	sp	p	*		
D	S	2.15	0.25	-	sp	*		+
D	R	2.1	0.44	-	+	*		
E	A	2.04	0.14	-	†			
P	K	2.03	0.4	£	+			+
I	I	2.02	0.14	np	np	*		
D	T	1.93	0.27	-	sp	*		
T	Q	1.79	0.14	sp	p	*		
F	M	1.79	0.92	np	np	*		
F	Y	1.75	0.51	np	np	*		

† - A alanina embora seja pequena e apolar não apresenta grandes aversões.
§ - A glicina embora pequena e polar não tem cadeia lateral para interagir.
£ - A prolina embora pequena e apolar não tem grande aversão por resíduos polares.

Aqui estão listados, por ordem decrescente de valor de *t* de Student (Eqs. 42 e 43), os elementos positivos da matriz **A**, parâmetros associados a um determinado par de resíduos, cuja interação, ponderada de acordo com $\mathbf{W}^{3.6}$, seria favorável ao estado alfa. Nas duas primeiras colunas (*Res. i* e *Res. j*) temos a identidade dos resíduos, sendo que a ordem é a mesma em que aparecem na sequência. Na terceira coluna (*t*) estão os valores de *t* de Student (Eqs. 42 e 43). A quarta coluna (A_{ij}) traz os valor numérico do elemento da matriz **A**. As colunas seguintes (*Classe i* e *Classe j*) informam as características físicoquímicas do par *ij* em questão (p=polar, sp=pequeno e polar, np=apolar,+=carregado positivamente, -=carregado negativamente). Estas informações auxiliam a entender a coluna seguinte (*Plausível S/N*), que traz a plausibilidade da interpretação do parâmetro (S=sim, plausível; N=não). Espera-se que resíduos com a mesma hidrofobicidade e/ou de cargas opostas estejam associados. A última coluna (*GGR*) informa se o elemento A_{ij} está em concordância com os resultados de GIBRAT *et al.* (1987). A linha transversal marca o último elemento significativo ($t>2$ ou $P<0.05$).

Tabela 6 - Parâmetros Quadráticos mais Significativos

Elementos Negativos da Sub-matriz **A**

Res. i	Res. j	t	Aij	Classe i	Classe j	Plausível S	Plausível N	GGR
L	K	-3.53	-0.27	np	+	*		+
L	A	-3.08	-0.17	np	sp	*		
N	G	-2.89	-0.43	p	§			
D	L	-2.71	-0.1	-	np	*		+
A	L	-2.66	-0.15	sp	np	*		+
F	T	-2.6	-0.31	np	sp	*		
V	S	-2.51	-0.24	np	sp	*		+
K	L	-2.51	-0.24	+	np	*		+
K	H	-2.46	-0.69	+	¶	¶		
K	Y	-2.33	-0.53	+	np	*		
E	I	-2.33	-0.41	-	np	*		
L	S	-2.24	-0.18	np	sp	*		
S	H	-2.19	-0.47	sp	¶		¶	+
T	T	-2.15	-0.29	sp	sp		X	
I	S	-2.1	-0.3	np	sp	*		
E	L	-2.09	-0.16	-	np	*		+
D	F	-2.06	-0.21	-	np	*		+
S	I	-2.05	-0.16	sp	np	*		+
K	F	-2.03	-0.32	+	np	*		
K	K	-2.02	-0.3	+	+	*		
L	G	-1.97	-0.2	np	§			
V	N	-1.87	-0.14	np	p	*		
P	I	-1.86	-0.37	£	np			
K	A	-1.83	-0.23	+	†			
D	I	-1.81	-0.11	-	np	*		

† - A alanina embora seja pequena e apolar não apresenta grandes aversões.
§ - A glicina embora pequena e polar não tem cadeia lateral para interagir.
£ - A prolina embora pequena e apolar não tem grande aversão por resíduos polares.
¶ - A histidina tem polaridade intermediária, se comportando geralmente como polar e positivo.

Aqui estão listados, por ordem crescente de valor de *t* de Student (Eqs. 42 e 43), os elementos negativos da matriz **A**, parâmetros associados a um determinado par de resíduos, cuja interação, ponderada de acordo com $\mathbf{W}^{3.6}$, seria desfavorável ao estado alfa. Nas duas primeiras colunas (*Res. i* e *Res j*) temos a identidade dos resíduos, sendo que a ordem é a mesma em que aparecem na sequência. Na terceira coluna (*t*) estão os valores de *t* de Student (Eqs. 42 e 43). A quarta coluna (A_{ij}) traz os valor numérico do elemento da matriz **A**. As colunas seguintes (*Classe i* e *Classe j*) informam as características físicoquímicas do par *ij* em questão (p=polar, sp=pequeno e polar, np=apolar,+=carregado positivamente, -=carregado negativamente). Estas informações auxiliam a entender a coluna seguinte (*Plausível S/N*), que traz a plausibilidade da interpretação do parâmetro (S=sim, plausível; N=não). Espera-se que resíduos com diferentes tipos hidroafinidade ou com cargas do mesmo tipo estejam associados. A última coluna (*GGR*) informa se o elemento A_{ij} está em concordância com os resultados de GIBRAT *et al.* (1987). A linha transversal marca o último elemento significativo (*t* <-2 ou P<0.05).

Tabela 7 - Parâmetros Quadráticos mais Significativos

Elementos Positivos da Sub-matriz **B**

Res. i	Res. j	t	Bij	Classe i	Classe j	Plausível S	Plausível N	GGR
V	I	3.44	0.14	np	np	*		
T	S	3.01	0.1	sp	sp	*		
Y	L	3	0.26	np	np	*		
L	E	2.64	0.2	np	-		X	
C	G	2.61	0.21	snp	§			
I	F	2.49	0.18	np	np	*		
S	Q	2.46	0.27	sp	p	*		+
R	T	2.25	0.23	+	sp	*		
V	C	2.2	0.21	np	snp	*		
P	Q	1.98	0.23	£	p			
C	L	1.85	0.35	snp	np	*		
K	Y	1.82	0.14	+	np		X	
L	L	1.77	0.05	np	np	*		N
M	R	1.76	0.58	np	+		X	
Q	G	1.75	0.17	p	§			
G	G	1.72	0.13	§	§		§	N
Q	I	1.67	0.24	p	np		X	
C	T	1.66	0.22	snp	sp		X	
Q	N	1.62	0.16	p	p	*		
T	Q	1.61	0.15	sp	p	*		
L	G	1.59	0.14	np	§			
R	N	1.58	0.12	+	p	*		
G	N	1.57	0.12	§	p			
F	G	1.55	0.11	np	§			
I	P	1.51	0.16	np	£			
G	P	1.51	0.11	§	£			

§ - A glicina embora pequena e polar não tem cadeia lateral para interagir.
£ - A prolina embora pequena e apolar não tem grande aversão por resíduos polares.

Aqui estão listados, por ordem decrescente de valor de *t* de Student (Eqs. 42 e 43), os elementos positivos da matriz **B**, parâmetros associados a um determinado par de resíduos, cuja interação, ponderada de acordo com $\mathbf{W}^{2.0}$, seria favorável ao estado beta. Nas duas primeiras colunas (*Res. i* e *Res j*) temos a identidade dos resíduos, sendo que a ordem é a mesma em que aparecem na sequência. Na terceira coluna (*t*) estão os valores de *t* de Student (Eqs. 42 e 43). A quarta coluna (B_{ij}) traz os valor numérico do elemento da matriz **B**. As colunas seguintes (*Classe i* e *Classe j*) informam as características físicoquímicas do par *ij* em questão (p=polar, sp=pequeno e polar, np=apolar,+=carregado positivamente, -=carregado negativamente). Estas informações auxiliam a entender a coluna seguinte (*Plausível S/N*), que traz a plausibilidade da interpretação do parâmetro (S=sim, plausível; N=não). Espera-se que resíduos com a mesma hidrofobicidade e/ou de cargas opostas estejam associados. A última coluna (*GGR*) informa se o elemento B_{ij} está em concordância com os resultados de GIBRAT *et al.* (1987). A linha transversal marca o último elemento significativo (*t* >2 ou P<0.05).

Tabela 8 - Parâmetros Quadráticos mais Significativos

Elementos Negativos da Sub-matriz **B**

Res. i	Res. j	t	Bij	Classe i	Classe j	Plausível S	Plausível N	GGR
V	S	-3.56	-0.21	np	sp	*		*
K	N	-3.14	-0.25	+	p		X	*
N	Y	-2.78	-0.32	p	np	*		
I	Y	-2.68	-0.33	np	np		X	
S	L	-2.65	-0.13	sp	np	*		
I	L	-2.48	-0.25	np	np		X	
R	G	-2.37	-0.01	+	§			
T	A	-2.26	-0.15	sp	†			
C	S	-2.25	-0.09	snp	sp	*		
P	G	-2.17	-0.13	£	§			
T	L	-2.16	-0.1	sp	np	*		
D	G	-2.05	-0.11	-	§			
P	P	-2.05	-0.28	£	£			
G	K	-2	-0.13	§	+			+
R	L	-1.95	-0.15	+	np	*		
V	W	-1.87	-0.39	np	&			
L	R	-1.86	-0.19	np	+	*		
Q	P	-1.84	-0.29	p	£			
T	V	-1.8	-0.02	sp	np	*		
E	D	-1.77	-0.2	-	-	*		
G	I	-1.74	-0.14	§	np			
K	T	-1.73	-0.12	+	sp		X	
I	E	-1.64	-0.16	np	-	*		
S	I	-1.59	-0.11	sp	np	*		
E	C	-1.58	-0.19	-	snp	*		+

† - A alanina embora seja pequena e apolar não apresenta grandes aversões.
§ - A glicina embora pequena e polar não tem cadeia lateral para interagir. (veja texto)
£ - A prolina embora pequena e apolar não tem grande aversão por resíduos polares.

Aqui estão listados, por ordem crescente de valor de *t* de Student (Eqs. 42 e 43), os elementos negativos da matriz **B**, parâmetros associados a um determinado par de resíduos, cuja interação, ponderada de acordo com $\mathbb{W}^{2.0}$, seria desfavorável ao estado alfa. Nas duas primeiras colunas (*Res. i* e *Res j*) temos a identidade dosresíduos, sendo que a ordem é a mesma em que aparecem na sequência. Na terceira coluna (*t*) estão os valores de *t* de Student (Eqs. 42 e 43). A quarta coluna (B_{ij}) traz os valor numérico do elemento da matriz **B**. As colunas seguintes (*Classe i* e *Classe j*) informam as características físicoquímicas do par *ij* em questão (p=polar, sp=pequeno e polar, np=apolar,+=carregado positivamente, -=carregado negativamente). Estas informações auxiliam a entender a coluna seguinte (*Plausível S/N*), que traz a plausibilidade da interpretação do parâmetro (S=sim, plausível; N=não). Espera-se que resíduos com diferentes tipos hidroafinidade ou com cargas do mesmo tipo estejam associados. A última coluna (*GGR*) informa se o elemento B_{ij} está em concordância com os resultados de GIBRAT *et al.* (1987). A linha transversal marca o último elemento significativo (*t* <-2 ou P<0.05).

4.4. Constantes de Decisão e Quantidade Relativa de Estrutura Secundária

Ao contrário de GIBRAT *et al.* (1988), decidimos não usar constantes de decisão, por acreditarmos que tal procedimento é válido somente para o conjunto de dados usado para calcular os parâmetros. Qualquer melhora da acurácia devido a seu uso depende do conhecimento prévio da proporção de estrutura secundária existente na proteína testada. As constantes $\mathbf{DC_H}$ e $\mathbf{DC_E}$ de GIBRAT *et al.* (1987) são um reflexo da proporção de alfa, beta e *coil* neste particular conjunto de proteínas (Tab. 1).

Para ilustrar este fato fizemos uma pesquisa sistemática, em cada cadeia, do efeito da ponderação das probabilidades com constantes de decisão (Eqs. 44 e 45). O conhecimento prévio da quantidade de estrutura regular alfa e beta, pode elevar a acurácia até 72%. Os valores de C1 e C2 encontrados, contudo, apresentaram uma variabilidade acentuada (Tabs. 9 a 12). Além disso, os métodos publicados para previsão da proporção alfa/beta (SHERIDAN *et al*, 1985; KLEIN, 1986; KLEIN & DELISI, 1986; NAKASHIMA, NISHIKAWA & OOI, 1986; DELÉAGE & ROUX, 1987), necessária para fazer uso destas constantes em proteínas de estrutura desconhecida, não são suficientemente acurados.

As Tabela 9 a 12 contém os resultados das previsões validadas por cruzamento usando as melhores constantes de decisão para cada cadeia. Proteínas de diferentes classes estruturais (LEVITT & CHOTHIA, 1975) foram listadas separadamente para evidenciar os padrões nos valores encontrados para C1 e C2 (Eqs. 44 e 45). O uso de constantes de decisão diferentes para cada proteína permite elevar a acurácia para 72.2%. Se adotarmos um procedimento semelhante ao de GIBRAT *et al.* (1987), isto é, uso das mesmas constantes (C1= 0.38 e C2= 0.51) para todas as cadeias, a melhora da acurácia é mais modesta, 67.8%.

A Tab. 9 traz os resultados obtidos para proteínas onde predominam as alfa-hélices (e.g. Fig. 25). Nota-se o melhor desempenho do método neste tipo de proteína (70.2%), bem como um aumento significativo da acurácia com o uso constantes de decisão (mais de 8 pontos percentuais). A relação entre os valores médios das constantes (C1<C2) é consistente com as Eqs. 44 e 45, onde C2 é o peso dado à probabilidade alfa (p_a). Uma inspeção dos valores individuais de C1 e C2 obtidos para as diferentes cadeias, porém, mostra intensa variação, inclusive com inversão do padrão (1LZM, n.º 45).

Na Tab. 10 estão as proteínas onde predominam as placas pregueadas (e.g. Fig. 26). A performance do método neste grupo é comparável ao resultado geral (65.5%), e a melhora com o uso de constantes de decisão (7 pontos), comparável ao caso anterior. O valor de médio C1, que multiplica a probabilidade beta (p_b) nas Eq.s 44 e 45, foi maior que o de C2, como esperado. Contudo, este padrão não é representativo, uma vez que existem 15 casos onde C1<C2, mais que o dobro dos casos que acompanham a média.

A Tab. 11 A traz os resultados obtidos para proteínas onde alfa-hélices e cordões beta se alternam ao longo da sequência (e.g. Fig. 27). O desempenho neste grupo (63.3%) é ligeiramente menor que a média e o aumento da acurácia devido ao uso de constantes (menos de 3 pontos percentuais), bem menor que nos dois casos anteriores. O tamanho deste tipo de cadeia (196) é, em geral, maior que nos casos alfa e beta (142 e 146 respectivamente). Existe uma ligeira predominância de resíduos no estado alfa, característico destas proteínas (LEVITT & CHOTHIA, 1975). Os valores médios de C1 e C2, porém, são equivalentes. Como no caso das proteínas beta, este padrão não é representativo dos resultados individuais, onde predominou um ou outro valor (exceto, talvez, 3FXC, n.º 30, com C1=0.9 e C2=0.95, valores duas vezes maiores que as respectivas médias).

Na Tab. 12 estão as proteínas que combinam elementos das três classes anteriores, de diferentes maneiras (e.g. Fig. 28). Como no caso anterior (Tab. 11) a performance foi pior que a média (63.6%) e o aumento da acurácia com o uso de constantes (4 pontos), mais modesto que nos casos alfa e beta (Tabs. 9 e 10). O tamanho médio dessas cadeia (235) é significativamente maior que nos casos α/β, α e β (196, 142 e 146, respectivamente). As quantidades de resíduos no estado alfa e beta foram equivalentes (23%). Apesar disso o valor médio de C2 resultou ligeiramente maior que o de C1. Este foi o padrão predominante nos dados, ainda que com grande variabilidade nos valores absolutos (exceto 3RN3, n.º 58, onde C1>C2).

O padrão observado nos valores médios obtidos para C1 e C2 não se manteve nos casos individuais. As Figuras 25 a 28 ilustram um possível motivo, além de exemplificar as classes estruturais de LEVITT & CHOTHIA (1975). Nelas podemos ver gráficos mostrando o efeito da variação sistemática das constantes de decisão sobre a acurácia. A Fig. 25 traz o resultado para o Citocromo C-551 (351C, n.º 24, Tabela 9), uma proteína contendo apenas resíduos alfa e *coil*. A escolha de C1=0.3 e C2=0.55 eleva a acurácia (valores calculados usando dados das outras 66 cadeias) para 86.6%. Podemos obter o mesmo efeito para a cadeia pesada do fragmento Fab da Imunoglobulina (1MCP, cadeia H, n.º 40, Tabela 10), que tem somente resíduos beta

Tabela 9 - Resultado listado por cadeia. Inclui efeito das constantes de decisão.

Proteínas α

N	Código	C	aa	% alfa	% beta	acurácia	ac.máx.†	C1	C2
10	5CPV		108	53.7%	3.7%	65.7%	66.7%	0.15	0.55
11	3ICB		75	57.3%	0.0%	82.7%	94.7%	0.6	0.75
16	2CTS		437	61.1%	1.4%	74.6%	85.8%	0.85	0.95
17	1CRN		46	41.3%	8.7%	58.7%	76.1%	0.25	0.65
19	1CCR		111	39.6%	0.0%	64.9%	73.9%	0.7	0.85
20	2CCY	A	127	74.8%	0.0%	80.3%	89.8%	0.1	0.5
21	2CYP		293	47.1%	5.5%	69.6%	72.0%	0.85	0.95
22	3C2C		112	42.9%	0.0%	58.0%	68.8%	0.4	0.55
23	2CDV		107	25.2%	9.4%	73.8%	79.4%	0.25	0.55
24	351C		82	46.3%	0.0%	73.2%	86.6%	0.3	0.55
28	1ECD		136	76.5%	0.0%	69.9%	78.7%	0.2	0.6
34	2HMQ	A	114	61.4%	0.0%	78.1%	78.8%	0.25	0.55
35	2HHB	A	141	76.6%	0.0%	68.1%	78.0%	0.45	0.75
36	2HHB	B	146	77.4%	0.0%	58.9%	81.5%	0.65	0.85
37	2LHB		149	73.2%	0.0%	69.8%	77.2%	0.75	0.9
44	1LH1		153	77.8%	0.0%	72.6%	85.0%	0.25	0.65
45	2LZM		164	66.5%	8.5%	62.2%	68.9%	0.85	0.35
46	1LZ1		130	36.9%	7.7%	63.1%	72.3%	0.8	0.9
47	2MLT	A	26	92.3%	0.0%	92.0%	92.0%	- £	- £
48	1MBN		153	77.1%	0.0%	80.4%	83.7%	0.45	0.7
53	1BP2		123	43.9%	6.5%	65.0%	70.7%	0.45	0.75
MÉDIAS			140	59.8%	2.5%	70.2%	78.8%	0.48	0.69

† - A acurácia máxima foi obtida escolhendo-se as constantes de decisão (C1 e C2, mostradas aqui e definidas na Eqs. 44 e 45) para cada cadeia prevista, refletindo a proporção de estruturas alfa e beta neste conjunto de dados.

£ - Casos em que o uso de constantes de decisão não melhorou a acurácia. Quaisquer valores de C1 e C2 resultam na mesma taxa de acerto ou pior.

Resultados da previsão de estrutura secundária listados por cadeia, separados por classe estrutural (LEVITT & CHOTHIA, 1975) e acompanhados do efeito do uso de constantes de decisão (Eqs. 44 e 45). Nesta tabela temos as proteínas α, onde predominam as hélices. Na primeira coluna (*N*) foram mantidos os números originais da Tab. 1. Na segunda, os códigos dos arquivos PDB. Na terceira coluna (*C*) temos a identificação da cadeia utilizada, quando pertinente. Na quarta coluna (*aa*) está o número de aminoácidos. Nas duas colunas subsequentes temos as proporções de resíduos em alfa-hélices e placas beta. Na sétima coluna temos a acurácia obtida com o modelo ajustado às outras 66 cadeias (melhor modelo com seleção 160 parâmetros quadráticos, seção iv da Tab. 2). Na oitava coluna (*ac. máx.*) está o valor máximo que pode ser obtido para a acurácia, variando-se sistematicamente os valores das constantes de decisão. Na duas últimas colunas estão os valores de C1 e C2 que resultaram na acurácia máxima. Note-se que esta procura dos melhores valores para C1 e C2 só é possível quando se conhece a estrutura secundária da proteína, sendo impraticável em proteínas cuja estrutura é desconhecida. Ela foi feita aqui a fim de buscar uma relação entre C1 e C2, as quantidades relativas de estruturas alfa e beta e a classe estrutural da proteína. Na última linha temos as médias, ponderadas quando pertinente, do número de aminoácidos, quantidade de estrutura secundária, acurácia com e sem uso de constantes de decisão, C1 e C2.

Tabela 10 - Resultado listado por cadeia. Inclui efeito das constantes de decisão.

Proteínas β

N	Código	C	aa	% alfa	% beta	acurácia	ac.máx.†	C1	C2
5	2ALP		198	4.0%	52.5%	59.1%	67.2%	0.9	0.2
9	2ABX	A	74	0.0%	5.4%	77.0%	95.9%	0.5	0.55
12	2CAB		256	14.5%	30.1%	65.6%	71.5%	0.05	0.25
18	1GCR		174	7.5%	44.3%	58.1%	65.5%	0.8	0.1
25	3DFR		162	19.1%	31.5%	53.1%	65.4%	0.1	0.25
26	2EST	E	240	7.1%	34.2%	68.3%	72.5%	0.15	0.2
27	3EBX		62	0.0%	43.6%	85.5%	90.3%	0.85	0.2
39	1MCP	L	220	3.6%	45.9%	60.9%	70.5%	0	0.15
40	1MCP	H	222	0.0%	49.1%	71.2%	79.7%	0.05	0.15
41	2PKA	A	80	0.0%	45.0%	62.5%	76.3%	0	0.15
42	2PKA	B	152	13.8%	25.7%	72.9%	73.5%	0.1	0.15
49	1NXB		62	0.0%	41.9%	85.5%	88.7%	0.8	0.25
50	1SN3		65	12.3%	18.5%	75.4%	76.9%	0.35	0.55
51	1OVO	A	56	17.9%	21.4%	57.1%	62.5%	0.35	0.4
54	1PCY		99	4.0%	35.4%	74.8%	82.8%	0.05	0.2
55	2PAB	A	114	7.0%	51.8%	53.5%	64.9%	0	0.15
56	2SGA		181	6.6%	54.1%	55.3%	69.6%	0.8	0.15
57	3RP2	A	224	5.4%	37.1%	74.6%	75.4%	0.05	0.15
59	5RXN		54	0.0%	14.8%	74.1%	85.2%	- §	- §
62	2SOD	O	151	0.0%	38.4%	73.5%	76.2%	0.1	0.2
64	1TPO		223	9.4%	32.3%	72.2%	72.7%	0.3	0.5
65	4PTI		58	20.7%	24.1%	81.0%	84.1%	0.8	0.2
66	2STV		184	9.8%	44.6%	57.1%	67.4%	0.9	0.2
67	4SBV	A	199	12.1%	35.2%	50.3%	60.3%	0.05	0.15
MÉDIAS			146	7.5%	38.1%	65.5%	72.6%	0.35	0.24

† - A acurácia máxima foi obtida escolhendo-se as constantes de decisão (C1 e C2, mostradas aqui e definidas na Eqs. 44 e 45) para cada cadeia prevista, refletindo a proporção de estruturas alfa e beta.
§ - Casos em que grande número de combinações de valores de C1 e C2 resultaram na mesma acurácia máxima.

Resultados da previsão de estrutura secundária listados por cadeia, para proteínas β (LEVITT & CHOTHIA, 1975) e acompanhados do efeito do uso de constantes de decisão (Eqs. 44 e 45). Nesta tabela as cadeias onde predominam as placas pregueadas. As colunas tem o mesmo conteúdo descrito para a Tab. 9. Nota-se que as proteínas β apresentam um tamanho médio (146 aminoácidos) comparável às proteínas α (142). A predominância de resíduos na classe estrutural mais frequente (β) não chega a ser tão predominante (38% β, 7.5% α) quanto no caso anterior (60% α, 2.5% β,Tab. 9). A contrário das proteínas α, C1>C2. (As médias das constantes de decisão não foram ponderadas).

Tabela 11.- Resultado listado por cadeia. Inclui efeito das constantes de decisão.

Proteínas α/β

N	Código	C	aa	% alfa	% beta	acurácia	ac.máx.†	C1	C2
6	6AT1	A	310	35.5%	14.8%	59.0%	60.7%	0.1	0.3
7	6AT1	B	146	16.4%	33.6%	66.4%	66.8%	0.2	0.4
13	5CPA		307	36.2%	16.3%	63.5%	65.1%	0.2	0.5
29	1FDX		54	9.3%	7.4%	74.1%	83.3%	- §	- §
30	3FXC		98	7.1%	15.3%	70.4%	81.6%	0.9	0.95
31	3FXN		138	36.2%	21.0%	65.9%	66.5%	0.85	0.3
32	4FD1		106	25.5%	13.2%	65.1%	71.7%	0.85	0.15
33	1GP1	A	184	28.8%	15.8%	56.5%	63.6%	0.3	0.45
43	6LDH		329	41.0%	16.1%	60.8%	61.5%	0.45	0.7
61	1SBT		275	30.2%	17.8%	66.9%	69.1%	0.3	0.6
MÉDIAS			195	31.1%	17.4%	63.3%	66.1%	0.46	0.48

† - A acurácia máxima foi obtida escolhendo-se as constantes de decisão (C1 e C2, mostradas aqui e definidas na Eqs. 44 e 45) para cada cadeia prevista, refletindo a proporção de estruturas alfa e beta.
§ - Casos em que grande número de combinações de valores de C1 e C2 resultaram na mesma acurácia máxima.

Resultados da previsão de estrutura secundária listados por cadeia, para proteínas α/β (LEVITT & CHOTHIA, 1975) e acompanhados do efeito do uso de constantes de decisão (Eqs. 44 e 45). Nesta tabela as cadeias onde alfa-hélices e placas beta se alternam ao longo da sequência. As colunas tem o mesmo conteúdo descrito para a Tab. 9. Nota-se que as proteínas α/β apresentam um tamanho médio (195 aminoácidos) maior que proteínas α e β (142 e 146 respectivamente). Apesar da predominância de resíduos na classe estrutural α, as constantes de decisão (C1 e C2) tiveram valores equivalentes. (As médias das constantes de decisão não foram ponderadas).

Tabela 12 - Resultado listado por cadeia. Inclui efeito das constantes de decisão.

Proteínas α+β

N	Código	C	aa	% alfa	% beta	acurácia	ac.máx.†	C1	C2
1	3APP		323	10.2%	45.5%	52.0%	59.1%	0.05	0.15
2	2ACT		218	27.5%	18.4%	63.6%	70.5%	0.45	0.65
3	9WGA	A	170	11.8%	9.4%	78.8%	78.8%	- £	- £
4	8ADH		374	23.3%	22.7%	61.0%	62.3%	0.2	0.3
8	2AZA	A	129	11.6%	33.3%	55.0%	59.7%	0.1	0.2
14	8CAT	A	498	29.5%	15.5%	69.7%	69.9%	0.4	0.65
15	5CHA	A	228	11.0%	33.3%	68.0%	69.3%	0.1	0.2
38	1HIP		85	11.8%	10.6%	65.9%	84.7%	0.15	0.3
52	1PPD		212	25.0%	17.0%	69.8%	74.1%	0.5	0.75
58	3RN3		124	21.0%	33.1%	60.5%	67.7%	0.7	0.2
60	2SNS		141	18.4%	19.9%	57.5%	67.4%	0.15	0.3
63	3TLN		316	39.6%	16.5%	59.8%	64.6%	0.6	0.8
MÉDIAS			235	22.3%	23.1%	63.6%	67.7%	0.31	0.41

† - A acurácia máxima foi obtida escolhendo-se as constantes de decisão (C1 e C2, mostradas aqui e definidas na Eqs. 44 e 45) para cada cadeia prevista, refletindo a proporção de estruturas alfa e beta.
£ - Casos em que o uso de constantes de decisão não melhora a acurácia, quaisquer valores de C1 e C2 resultam na mesma taxa de acerto ou pior.

Resultados da previsão de estrutura secundária listados por cadeia, para proteínas α+β (LEVITT & CHOTHIA, 1975) e acompanhados do efeito do uso de constantes de decisão (Eqs. 44 e 45). Nesta tabela as cadeias onde alfa-hélices, placas beta e elementos α/β se misturam de diversas maneiras ao longo da sequência (Fig. 28). As colunas tem o mesmo conteúdo descrito para a Tab. 9. Nota-se que as proteínas α+β apresentam um tamanho médio (235 aminoácidos) maior que proteínas α, β ε α/β (142, 146 e 196 respectivamente). Apesar da quantidade equivalente de resíduos em cada classe estrutural, C1<C2, como nas proteínas α. (As médias das constantes de decisão não foram ponderadas).

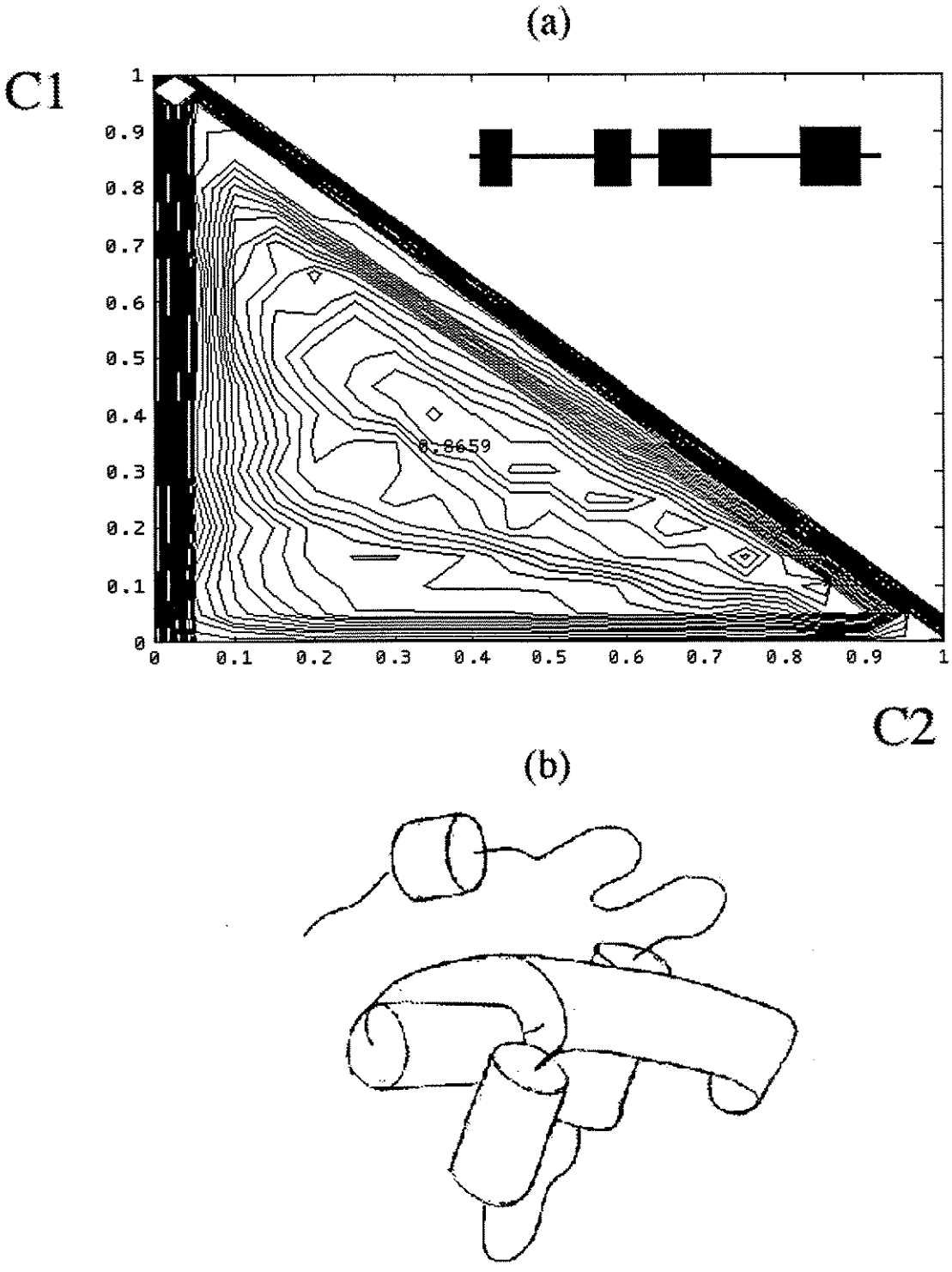
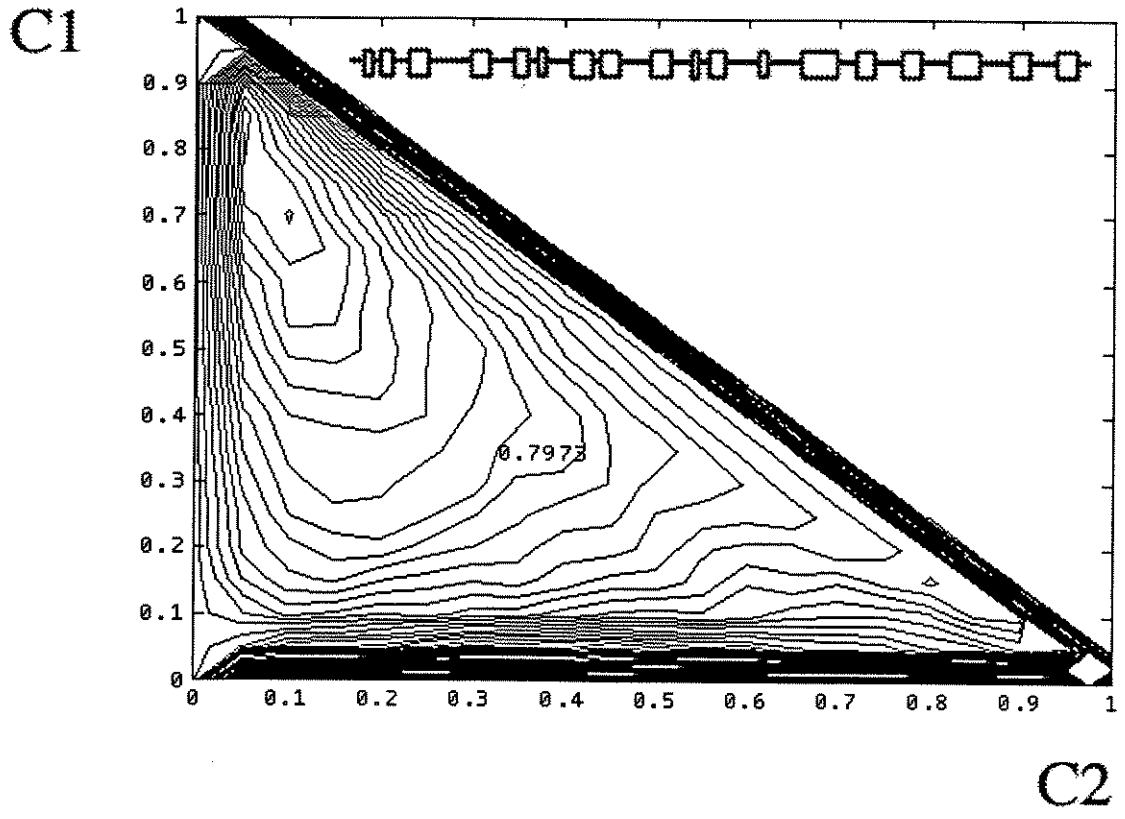


Figura 25 - Efeito do uso de constantes de decisão sobre a acurácia de previsão de estrutura secundária do Citocromo C-551 (351C, n.o 24, Tab. 9), uma proteína α (LEVITT & CHOTHIA, 1975). (a) Gráfico mostrando o efeito de diversos valores de C1 e C2 (Eqs. 44 e 45) sobre a acurácia, aqui representada em curvas de nível. Existem 5 picos com acurácia 0.8659. No canto superior direito da Fig. 25a se encontra representada a estrutura secundária da proteína, composta de 4 hélices. (b) Representação da estrutura terciária da mesma proteína. Note-se que a placa beta aqui representada não satisfaz as condições de KABSCH & SANDER (1983). O gráfico (a) foi produzido por rotina em MATLAB num terminal Tektronics 1145, capturado na tela de um Macintosh e editado para incluir a estrutura secundária. A Fig. 25b foi feita à mão pelo candidato e capturada por um *scanner*.

(a)



(b)

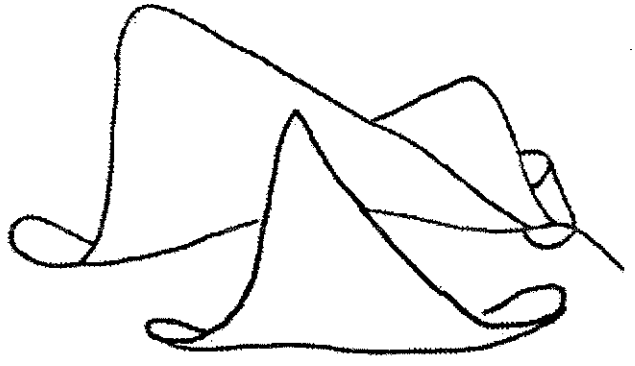


Figura 26 - Efeito do uso de constantes de decisão sobre a acurácia de previsão de estrutura secundária da cadeia pesada do Fragmento Fab da Imunoglobulina (1MCP, n.o 40, Tab. 9), uma proteína β (LEVITT & CHOTHIA, 1975). (a) Gráfico mostrando o efeito de diversos valores de C1 e C2 (Eqs. 44 e 45) sobre a acurácia, aqui representada em curvas de nível. Existem 3 picos com acurácia 0.7973, no canto superior esquerdo. No canto superior direito da Fig. 26a se encontra representada a estrutura secundária da proteína, composta diversos cordões beta. (b) Representação da estrutura terciária da mesma proteína. Note que a alfa-hélice aqui representada não satisfaz as condições de KABSCH & SANDER (1983). O gráfico (a) foi produzido por rotina em MATLAB num terminal Tektronics 1145, capturado na tela de um Macintosh e editado para incluir a estrutura secundária. A Fig. 26b foi feita à mão pelo candidato e capturada por um *scanner*.

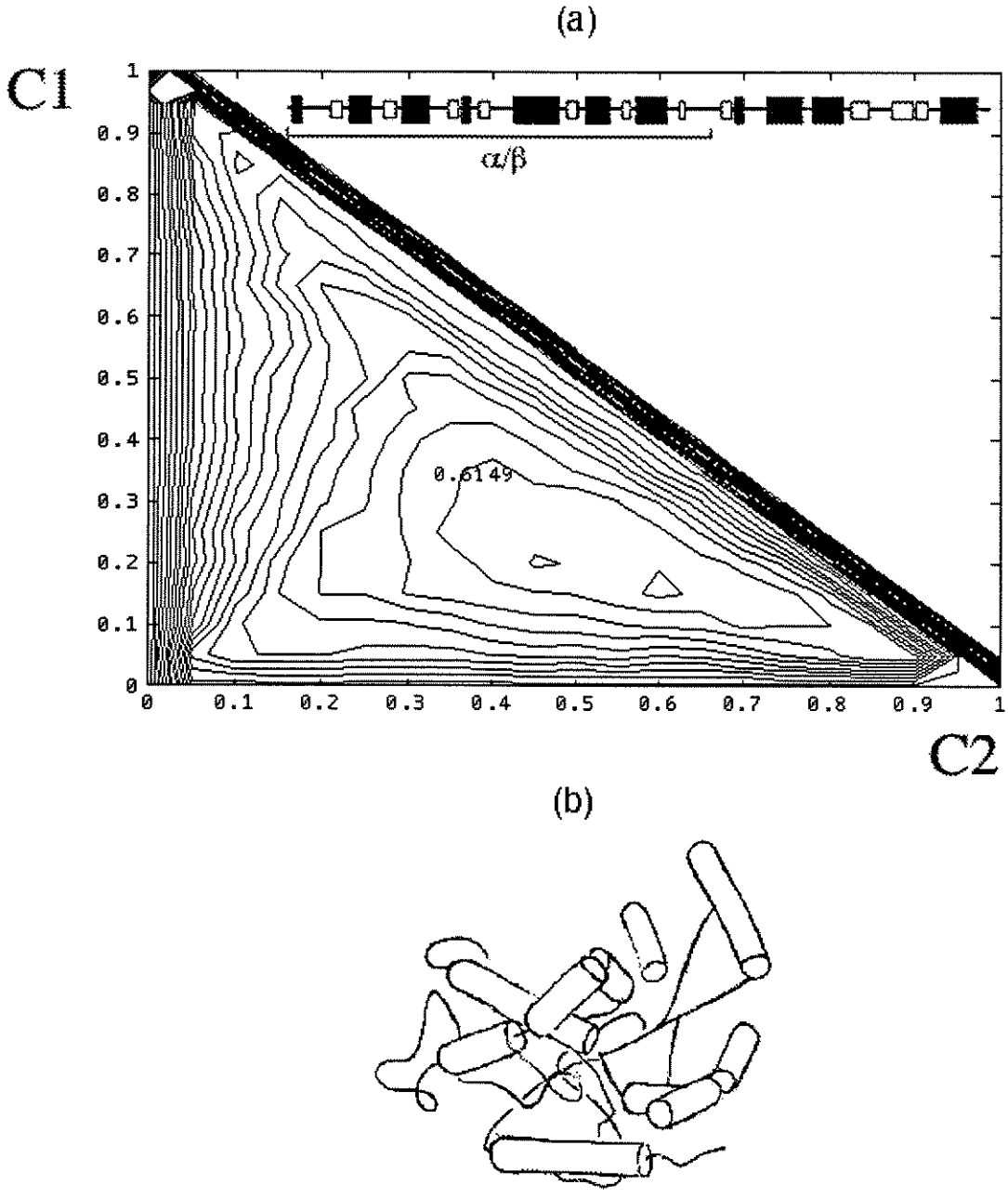


Figura 27 - Efeito do uso de constantes de decisão sobre a acurácia de previsão de estrutura secundária da Desidrogenase Lática (6LDH, n.o 43, Tab. 9), uma proteína α/β (LEVITT & CHOTHIA, 1975). (a) Gráfico mostrando o efeito de diversos valores de C1 e C2 (Eqs. 44 e 45) sobre a acurácia, aqui representada em curvas de nível. Existem quatro picos com acurácia 0.6149. No canto superior direito da Fig. 27a se encontra representada a estrutura secundária da proteína. Existe uma porção α/β típica, onde elementos α e β se alternam, e uma porção C-terminal, mais heterogênea. (b) Representação da estrutura terciária da mesma proteína. O gráfico (a) foi produzido por rotina em MATLAB num terminal Tektronics 1145, capturado na tela de um Macintosh e editado para incluir a estrutura secundária. A Fig. 27b foi feita à mão pelo candidato e capturada por um *scanner*.

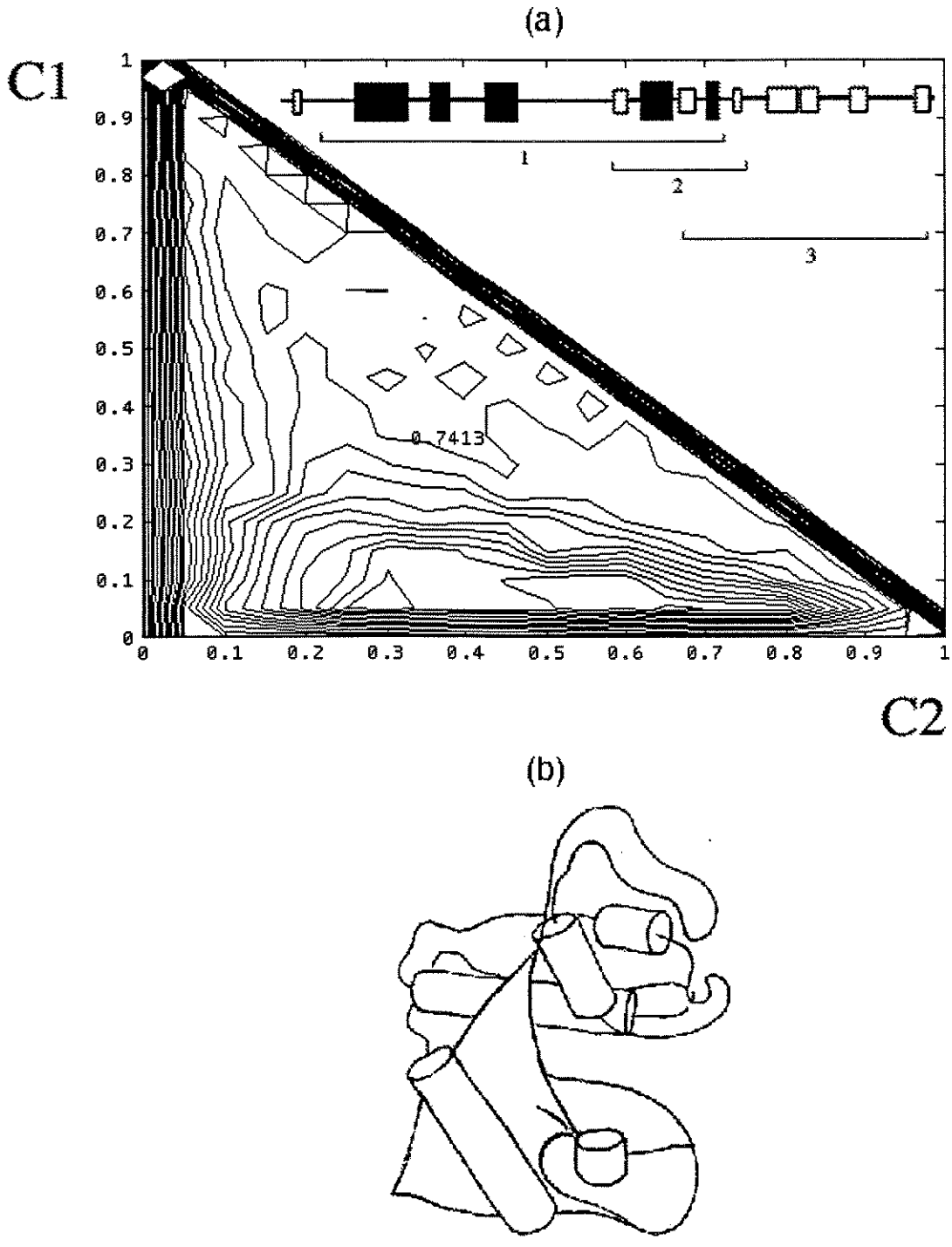


Figura 28 - Efeito do uso de constantes de decisão sobre a acurácia na previsão da estrutura secundária da Papaína (1PPD, n.o 52, Tab. 9), uma proteína $\alpha+\beta$ (LEVITT & CHOTHIA, 1975). (a) Gráfico mostrando o efeito de diversos valores de C1 e C2 (Eqs. 44 e 45) sobre a acurácia, aqui representada em curvas de nível. Existem vários picos com acurácia 0.7413, na porção inferior do gráfico, ao longo do eixo C2. No canto superior direito da Fig. 28a se encontra representada a estrutura secundária da proteína. Existem três porções distintas com características α, β e α/β (1, 2 e 3 respectivamente). (b) Representação da estrutura terciária da mesma proteína. O gráfico (a) foi produzido por rotina em MATLAB num terminal Tektronics 1145, capturado na tela de um Macintosh e editado para incluir a estrutura secundária. A Fig. 28b foi feita à mão pelo candidato e capturada por um *scanner*.

e *coil* (Fig. 26). A seleção de constantes apropriadas resulta na previsão correta de 79.7% dos resíduos. A Fig. 27 traz o resultado obtido para Desidrogenase Lática (6LDH, n.º 43, Tabela 11), que tem estruturas alfa e beta se alternando na sequencia (canto superior direito da mesma figura). A seleção das constantes afeta pouco o valor da acurácia. Na Fig. 28 mostramos o resultado da seleção de constantes de decisão para a Papaína (1PPD, n.º 52 na Tabela 9), que é do tipo $\alpha+\beta$. Esta proteína tem elementos alfa e beta, mas, ao contrário da Desidrogenase Lática (Fig. 27), estes não se distribuem de maneira alternada ao longo da sequência. Da mesma forma, a seleção das constantes elevou pouco a acurácia, de 69.8% para 74.1%.

A relação entre as constantes de decisão que resultam na maior acurácia e a proporção de estruturas alfa e beta é complexa. Ela parece depender não só da proporção α e β, mas também da sua distribuição espacial e da quantidade de estrutura regular. Isto dificulta seu uso para aumentar a acurácia preditiva em proteínas novas.

Ainda na tentativa de estudar o efeito da quantidade de cada um dos tipos de estrutura secundária no modelo preditivo, foi produzido um gráfico de acurácia *versus* porcentagem de estrutura tipo alfa para todas as 67 cadeias (Fig. 29). Como já mostrado, proteínas das classes α e β foram previstas com maior acurácia que as do tipo α/β e $\alpha+\beta$. Contudo, proteínas α e β são em geral pequenas, e desta forma o maior sucesso com que são preditas contribui menos que o resultado obtido para proteínas heterogêneas, em geral maiores. A Figura 30 mostra o número cumulativo de resíduos com estrutura corretamente prevista, onde as cadeias foram ordenadas em ordem decrescente de proporção de conteúdo alfa. As inclinações da curva são maiores nas extremidades, mostrando a maior acurácia com que é prevista a estrutura secundária de resíduos em proteínas α (lado esquerdo da Fig. 30) e β (lado direito).

O tamanho da sequência tem um efeito deletério sobre a acurácia. Poucas proteínas "grandes" chegam a atingir 75% de acurácia, ao contrário do que acontece com as menores. Isto se deve, possívelmente, à estrutura mais complexa (tipo α/β e $\alpha+\beta$), com múltiplos domínios, das estruturas maiores. Outra possibilidade é a maior proporção de resíduos sem estrutura (coil) nestas proteínas.

4.5. Análise das Probabilidades do Modelo

Sempre que se testa um modelo é recomendável a inspeção dos "valores residuais" de seu ajuste aos dados, a fim de buscar pistas que auxiliem a refiná-lo. Embora este

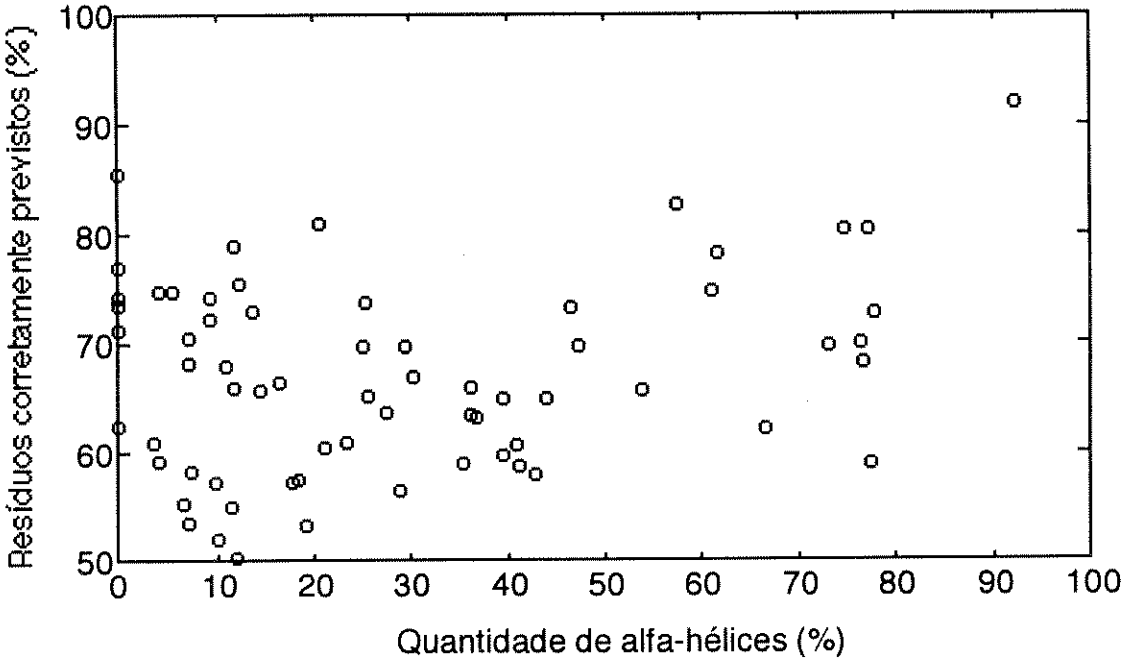


Figura 29 - Influência da proporção de resíduos no estado alfa presentes em uma proteína sobre a acurácia. Cada ponto representa uma das 67 cadeias. Nas ordenadas temos a acurácia com que a estrutura secundária da cadeia foi prevista (validada por cruzamento) e nas abscissas, a quantidade de resíduos no estado alfa que continha.

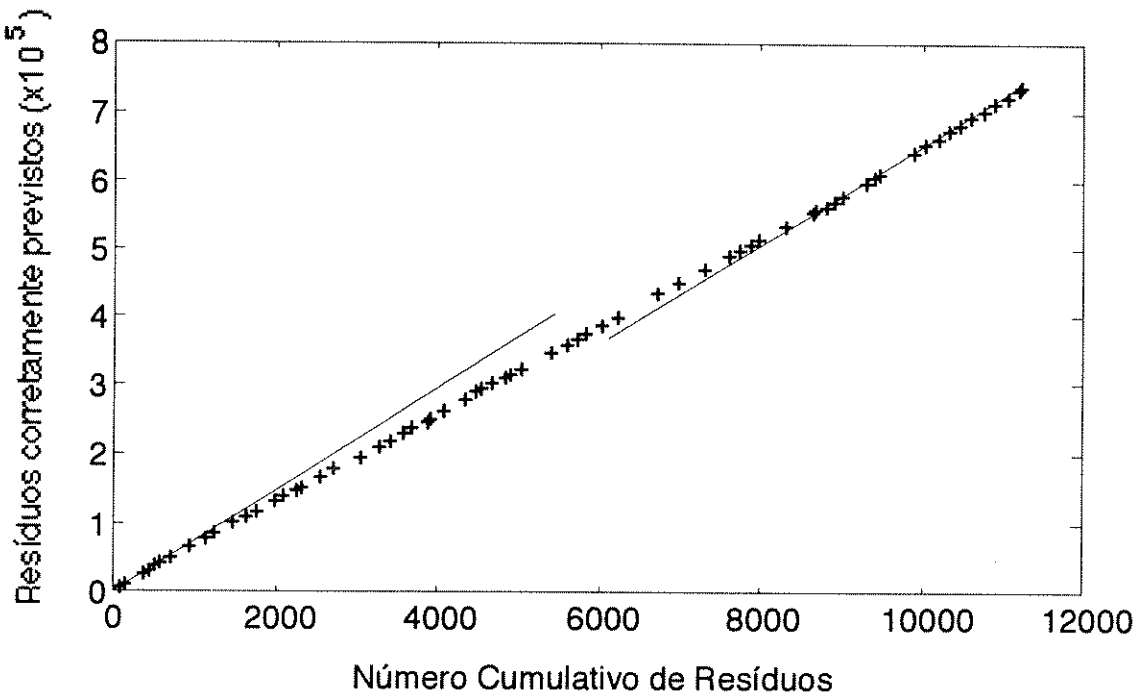


Figura 30 - Influência do conteúdo de estrutura alfa sobre a acurácia. As cadeias foram ordenadas de acordo com a proporção de resíduos no estado alfa (conteúdo decrescente). O gráfico mostra, nas ordenadas, o número de resíduos com estrutura secundária prevista corretamente. Nas abscissas temos o número cumulativo de resíduos. Com a ordenação das cadeias, os resíduos das proteínas com grande conteúdo alfa se encontram à esquerda, enquanto os resíduos das proteínas com pouco ou nenhum conteúdo alfa estão representados à direita. Foram feitas regressões lineares com os primeiros e últimos 1000 resíduos, cujas inclinações mostram a maior acurácia em proteínas α (à esquerda) e β (à direita).

procedimento seja mais adequado a modelos com variáveis contínuas, resolvemos seguí-lo, a fim de testar as suposições de nosso modelo categórico. Foi suposta uma distribuição trinomial para a probabilidade de cada resíduo estar em uma das três possíveis categorias de estrutura secundária. Para cada aminoácido um vetor (p_a, p_b, p_c) determina a probabilidade de estruturas alfa, beta e *coil*, respectivamente. Se agruparmos todos os aminoácidos com probabilidade prevista $\boldsymbol{p}$, de valor similar, cada um destes grupos deveria mostrar uma distribuição trinomial com as proporções devidas. Esta suposição foi testada dividindo os dados em 20 grupos com valores similares de probabilidade. Podemos ver na Figura 31 que o modelo descreve os dados adequadamente. Ali, 5% dos resíduos tem uma alta (aproximadamente 85%) probabilidade de estarem em alfa-hélices. Dentro deste grupo, a porcentagem de resíduos realmente em estado alfa é tão alta ou maior que o previsto. Além disso, a proporção de resíduos em cada uma das conformações de estrutura secundária praticamente confirma a previsão teórica (dados não mostrados para beta e *coil*). Assim, quando o modelo prevê 50% de probabilidade alfa, 50% desses resíduos têm esta estrutura. Desvios pequenos, mas sistemáticos, foram observados para alfa e beta. A proporção observada é na verdade maior do que a prevista, de modo que o modelo é "conservador" ao fazer tais previsões. Isto se deve a grande quantidade de estados *coil* na base de dados (superrepresentação).

A coincidência entre as probabilidades previstas e as proporções observadas possibilitam um cálculo preciso da confiabilidade das previsões. Usando o modelo de verossimilhança multinomial, podemos não somente obter a estimativa de máxima verossimilhança de um estado:

$$\hat{s} = \arg\max_{a,b,c}\left\{\hat{p}_a, \hat{p}_b, \hat{p}_c\right\}, \tag{46}$$

mas também a confiança:

$$\hat{p} = \max_{a,b,c}\left\{\hat{p}_a, \hat{p}_b, \hat{p}_c\right\} \tag{47}$$

e a variância

$$\hat{p}(1-\hat{p}) \tag{48}$$

Pensamos inicialmente que a presença de certos resíduos preditos com alta confiabilidade pudesse ser utilizada. Estes aminoácidos serviriam como "âncoras" ou pontos de "nucleação" para determinar o restante da estrutura secundária. A presença destas regiões é demonstrada nas Figs. 31 e 32, e podem ser consideradas tanto uma consequência do modelo trinomial, como do sucesso do modelo em prever boa parte

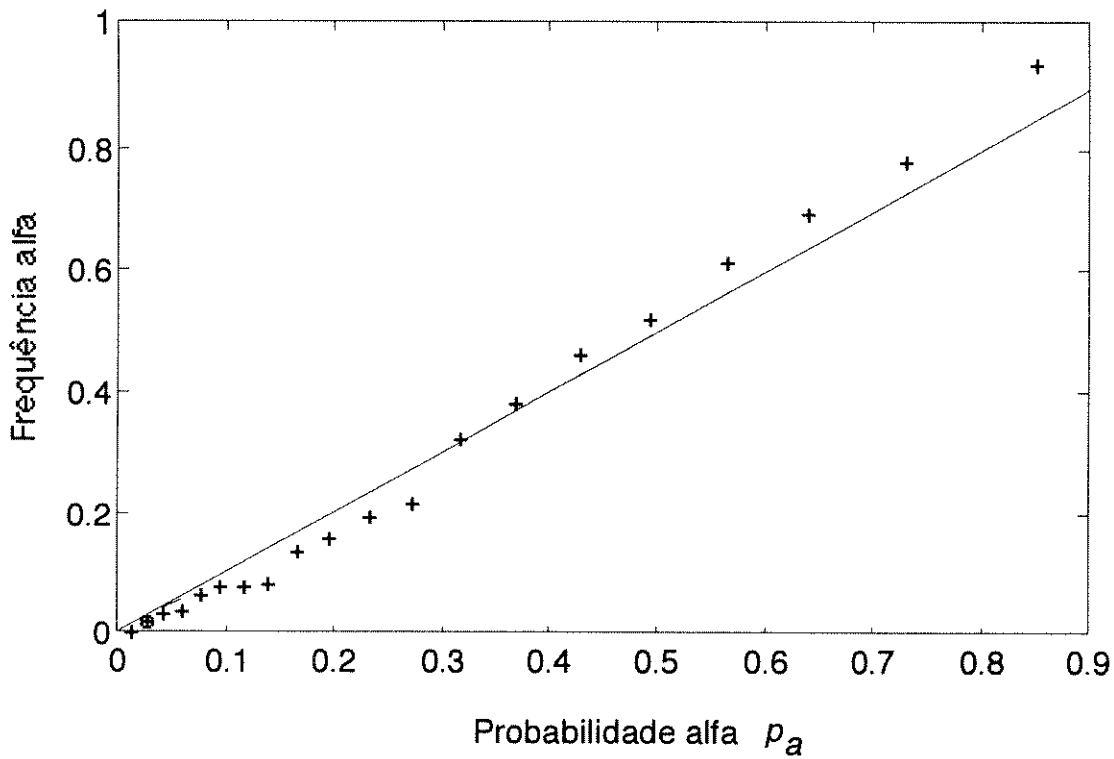


Figura 31 - Verificação da distribuição binomial da probabilidade do estado alfa. Os resíduos foram ordenados de acordo com a probabilidade p_a, do estado alfa, e divididos em vinte grupos, aqui representados nas abscissas. Nas ordenadas temos a frequência de resíduos neste estado em cada um dos grupos. A proximidade dos resultados observados (+) da primeira diagonal (resultados esperados) reflete a adequação do modelo aos dados. Resultados semelhantes foram obtidos para os estados beta e *coil* (não mostrados).

da estrutura secundária. 16% dos resíduos do conjunto de dados tem probabilidades de 85% ou mais. Isto significa, em termos de nível aceitável de acurácia, que podemos prever um em cada seis resíduos. Além disso, podemos dizer quais são os resíduos cujos estados foram preditos com alta probabilidade. Infelizmente, a maioria das previsões com alta probabilidade são para o estado *coil*, de forma que este resultado não é muito útil para o processo de "nucleação". Cerca de 63% da 457 hélices na base de dados contém pelo menos um resíduo com probabilidade maior que 0.5 e apenas uma em cada 20 contém pelo menos uma probabilidade maior que 0.9.

Para ilustrar a qualidade das probabilidades estimadas, e para compará-las com a estrutura secundária real, apresentamos a previsão de estrutura para o Citocromo c' (arquivo 2CCY, cadeia A, n.º 20), que é uma proteína α (LEVITT & CHOTHIA, 1975), ou mais especificamente, um feixe de quatro hélices. Esta proteína contém 75% de seus resíduos no estado alfa e teve sua estrutura prevista com 80% de acurácia validada por cruzamento, de forma que representa um dos casos extremos da Fig. 29. Uma versão gráfica da previsão está representada na Fig. 32. Claramente, a probabilidade de ocorrerem alfa-hélices (terço superior do gráfico) alcança valores máximos em quatro regiões, que correspondem às quatro hélices reais. Uma quinta alfa-hélice, curta, na porcão N-terminal da cadeia, não foi corretamente prevista. A probabilidade beta (traçado intermediário) foi geralmente pequena, o que é correto e esperado, uma vez que se trata de uma proteína alfa. A probabilidade do resíduo não ter qualquer estrutura (*coil*, terço inferior do gráfico) é alta justamente nas regiões onde não existem nem alfa nem beta, incluindo uma porção isolada na posição 103. Nesta proteína, é possível observar como probabilidades elevadas de um tipo de estrutura secundária (alfa-hélices) poderiam servir de pontos de nucleação, que ajudariam a determinar o aspecto geral da estrutura secundária. O ponto de terminação de algumas das hélices pode ser inferido por picos na probabilidade *coil*. Trata-se porém de caso extremo. A qualidade da previsão de outras cadeias do conjunto de dados e, principalmente, de proteínas não contidas ali merece maior estudo.

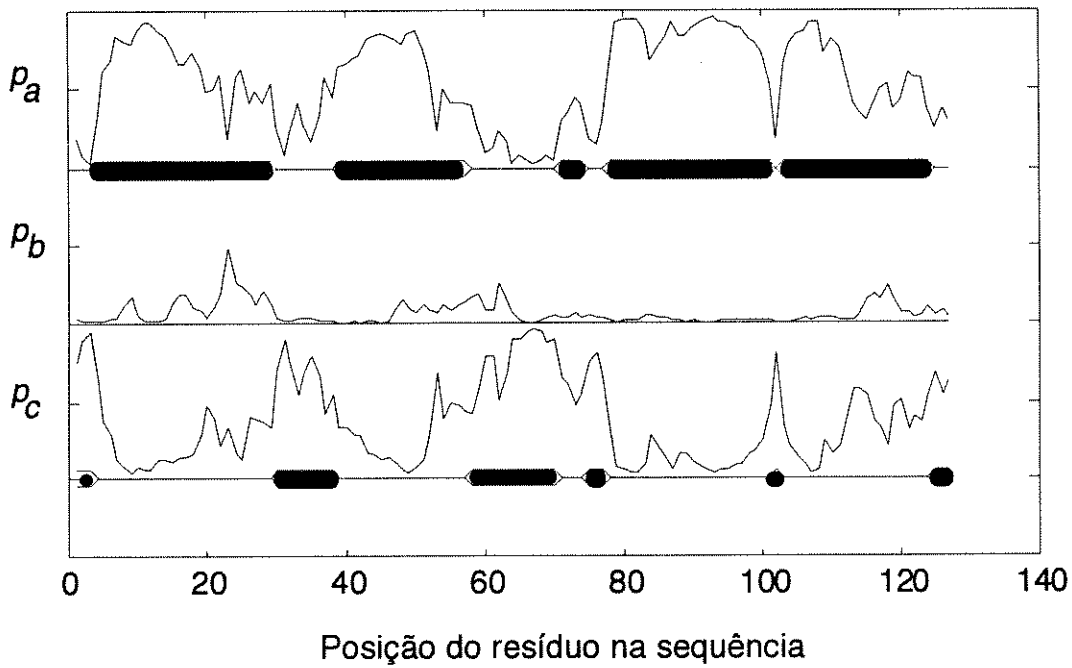


Figura 32 - Estrutura secundária prevista e observada para o Citocromo c' (2CCY, n.º 20, Tabela 9). O traçado superior corresponde à probabilidade alfa (p_a), o traçado intermediário, à probabilidade beta (p_b) e o inferior, à probabilidade *coil* (p_c). As barras sólidas sob os traçados superior e inferior representam as alfa-hélices e regiões sem estrutura (*coil*) observadas na proteína.

5. Discussão

Embora esperássemos conseguir um maior sucesso na previsão de estrutura secundária a partir da sequência, nosso objetivo foi antes metodológico: utilizar técnicas modernas de modelagem estocástica para atacar um problema que, apesar de antigo e alvo da atenção de muitos pesquisadores, ainda não fora abordado desta forma. Depois de fazê-lo, esperamos ter demonstrado sermos capazes de implementar modelos preditivos estatística e computacionalmente sofisticados. Os conceitos e métodos aqui utilizados são comuns na análise de sinais e em modelos preditivos em geral.

Até 1992, modelos logísticos com máxima verossimilhança praticamente não tinham sido explorados na previsão de estrutura secundária, embora se tratasse de procedimento estatístico consagrado (BRYANT & LAWRENCE, 1993). Procuramos então explorar seu poder e conveniência. A acurácia atingida foi comparável, senão superior, àquela obtida por teoria da informação (GIBRAT *et al.*, 1987). A capacidade de medir a confiabilidade das estimativas representou um avanço. Como o modelo é estritamente probabilístico, foi possivel testar a suposições feitas, até certo ponto. A estrutura secundária se comportou aproximadamente como um processo probabilístico trinomial (Fig. 31). Todavia, tal processo não é independente ao longo de resíduos adjacentes. O presente modelo não lida adequadamente com isto e, provavelmente, a incorporação de dependência, como série temporal ou cadeia de Markov (HUNTER

& STATES, 1991), pode melhorar os resultados. Ambas abordagens podem utilizar máxima verossimilhança, somente o modelo probabilístico mudaria.

O modelo logístico puramente linear é virtualmente idêntico a uma rede neural sem camada intermediária (QIAN & SEJNOWSKI, 1989) e à abordagem de GARNIER *et al.* (1978, GOR) usando "teoria da informação". A incorporação do componente quadrático encontra correspondência na extensão que GIBRAT *et al.* (1988) deram ao modelo GOR. Os dados estruturais disponíveis no presente não permitem considerar todas interações entre pares de aminoácidos de uma região da cadeia. Nossa inovação foi o uso de técnicas mais rigorosas na formulação de um modelo probabilístico, permitindo tanto uma aproximação mais geral e menos arbitrária do modelo quadrático completo (Eq. 19), pela imposição de restrições (bandeamento por blocos, periodicidade, penalização), quanto uma estimativa confiável de seus parâmetros. Outros autores (BIOU *et al.*, 1988; GIBRAT *et al.*, 1987; ZHANG *et al.*, 1992) incluíram termos de ordem superior em seus modelos. GIBRAT *et al.* (1988) consideraram inicialmente todos os pares de interações que incluiam o resíduo central de uma janela local (Eq. 2) mas depois se limitaram às informações "próprias", "direcionais" e "pareadas" (Eqs. 3 a 6). ZHANG *et al.* (1992) e BIOU *et al.* (1988) consideraram a contribuição da anfipaticidade. Nossa abordagem foi mais geral, uma vez que apenas a identidade dos resíduos foi utilizada e não alguma escala ou índice de propriedade bioquímica (e.g. hidrofobicidade). As matrizes de parâmetros quadráticos **A** e **B** não só incluem os três tipos de "informação" citados, como podem representar relações mais complexas, inclusive uma generalização do conceito de anfipaticidade (EISENBERG *et al.*, 1986; CORNETTE *et al.*, 1987).

Alta confiabilidade computacional foi obtida mesmo com o grande número de parâmetros estimados. Assim, mesmo com 1370 variáveis, foi possível calcular as estimativas com 6 ou 7 iterações do algoritmo de minimização (com tolerância de 0.0001). Confiabilidade estatística só foi obtida com valores de λ maiores que 1000. Na medida em que mais dados sejam utilizados, o valor ótimo de λ poderá diminuir e a complexidade do modelo poderá aumentar.

A maior vantagem da presente abordagem sobre o uso de redes neurais artificiais é a interpretabilidade dos parâmetros baseada em propriedades físico-químicas dos aminoácidos. Estes padrões emergiram dos dados independente de conhecimento *a priori* incorporado ao modelo. Apenas a identidade e a ordem dos aminoácidos em cada janela local foi utilizada, embora o conhecimento de bioquímica de proteínas tenha sido utilizado para melhor formular o modelo (anfipaticidade).

5.1. Interpretação dos Parâmetros Lineares

Os parâmetros lineares, mostrados nas Figs. 21 e 22, refletem a preferência do resíduo *j* por um estado alfa na posição *i*. Quando *i=j* (coluna central da matriz **a**, Fig.21) eles correspondem à "informação própria" de GIBRAT *et al.* (1988, Eq. 4) e, nas demais colunas, à "informação direcional" (Eq. 3). A Fig. 20 ilustra como interpretar seus valores. A posição vertical das linhas de **a** foi ordenada de acordo com seu valor médio.

Na parte inferior da Fig. 21 encontramos o mais simples dos aminoácidos, a glicina (G), que tem como cadeia lateral apenas um átomo de hidrogênio. Na ausência de um radical lateral maior, a cadeia principal da proteína tem grande liberdade de rotação neste ponto, o que sabidamente desfavorece a ocorrência de estrutura secundária (RICHARDSON & RICHARDSON, 1989). Além disso, este amino ácido é o único a ter seu C_α simétrico, aumentando o número de conformações possíveis. Os outros aminoácidos são sempre levógiros (STRYER, 1981). A Fig.33 mostra os ângulos diédricos assumidos por glicinas em um conjunto de proteínas solucionadas com alta resolução. Ali podemos observar que conformações "invertidas" ($\phi>0$) são comuns. Embora ocorram glicinas em alfa-hélices, este aminoácido representa um ponto fraco na estrutura helicoidal devido à mobilidade da cadeia principal.

A segunda linha de **a** de baixo para cima corresponde à prolina (P). Ao contrário da glicina este é o mais rígido de todos os aminoácidos, não pelo tamanho de sua cadeia lateral, mas por ter seu C_α dentro de uma estrutura cíclica (Fig. 2, DOOLITTLE, 1985). Isto impede que assuma os ângulos diédricos típicos de uma alfa-hélice, embora apresente valores próximos, como mostra o gráfico de Ramachandran da Fig. 34 (RICHARDSON & RICHARDSON, 1989). Além de restringir a movimentação da cadeia principal neste ponto, a presença física do anel e a ausência do nitrogênio doador de elétrons, impedem a realização de ponte de hidrogênio com os resíduos precedentes (do lado N-terminal). Assim a prolina desfavorece o estado alfa, exceto talvez quando ocorre nas primeiras três posições da hélice, que não fazem pontes de hidrogênio com resíduos anteriores, ou nas primeiras posições após a hélice, quando a interrompe, pelo mesmo motivo (RICHARDSON & RICHARDSON, 1988). Isto confere uma assimetria à linha, com valores mais negativos do lado C-terminal (direito) e valores mais neutros do lado N-terminal.

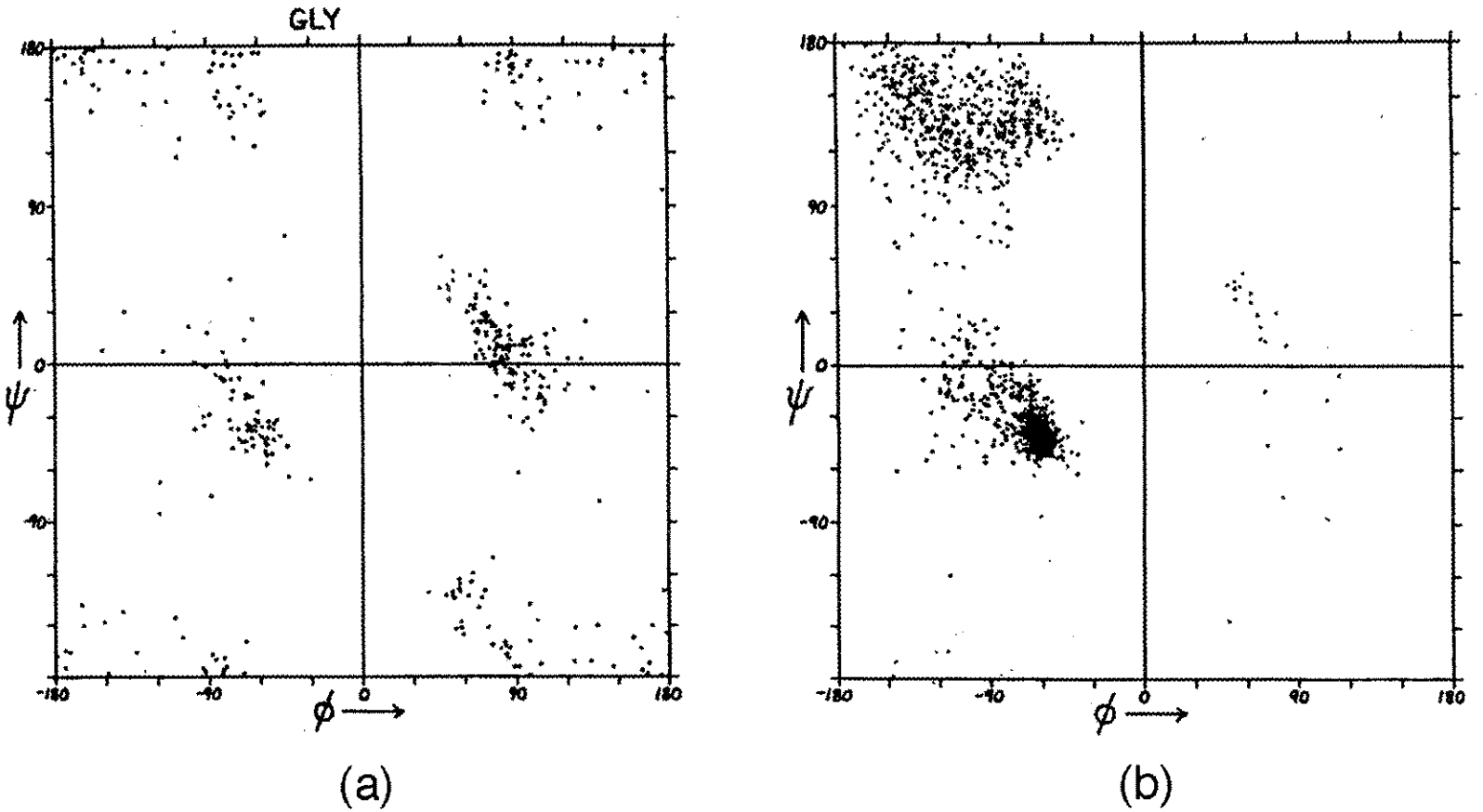


Figura 33 - Mobilidade e simetria da glicina. À esquerda temos o gráfico de Ramachandran mostrando os ângulos diédricos (Fig. 7) da glicina em um conjunto de proteínas de estruturas conhecidas com alta precisão. À direita temos o mesmo tipo de gráfico, obtido no mesmo conjunto de proteínas, com os ângulos diédricos de todos os aminoácidos, com exceção de glicina e prolina. Enquanto a maioria dos aminoácidos têm seus valores nas regiões α e β (explicadas na Fig. 8), a glicina apresenta uma distribuição simétrica, ocupando também os demais quadrantes. Isto se deve a ausência de cadeia lateral neste aminoácido, que aumenta a flexibilidade da cadeia principal. O fato de possuir C_α simétrico também permite a occorência de estruturas "invertidas" (pontos do lado direito do gráfico). Dados retirados de RICHARDSON & RICHARDSON (1988).

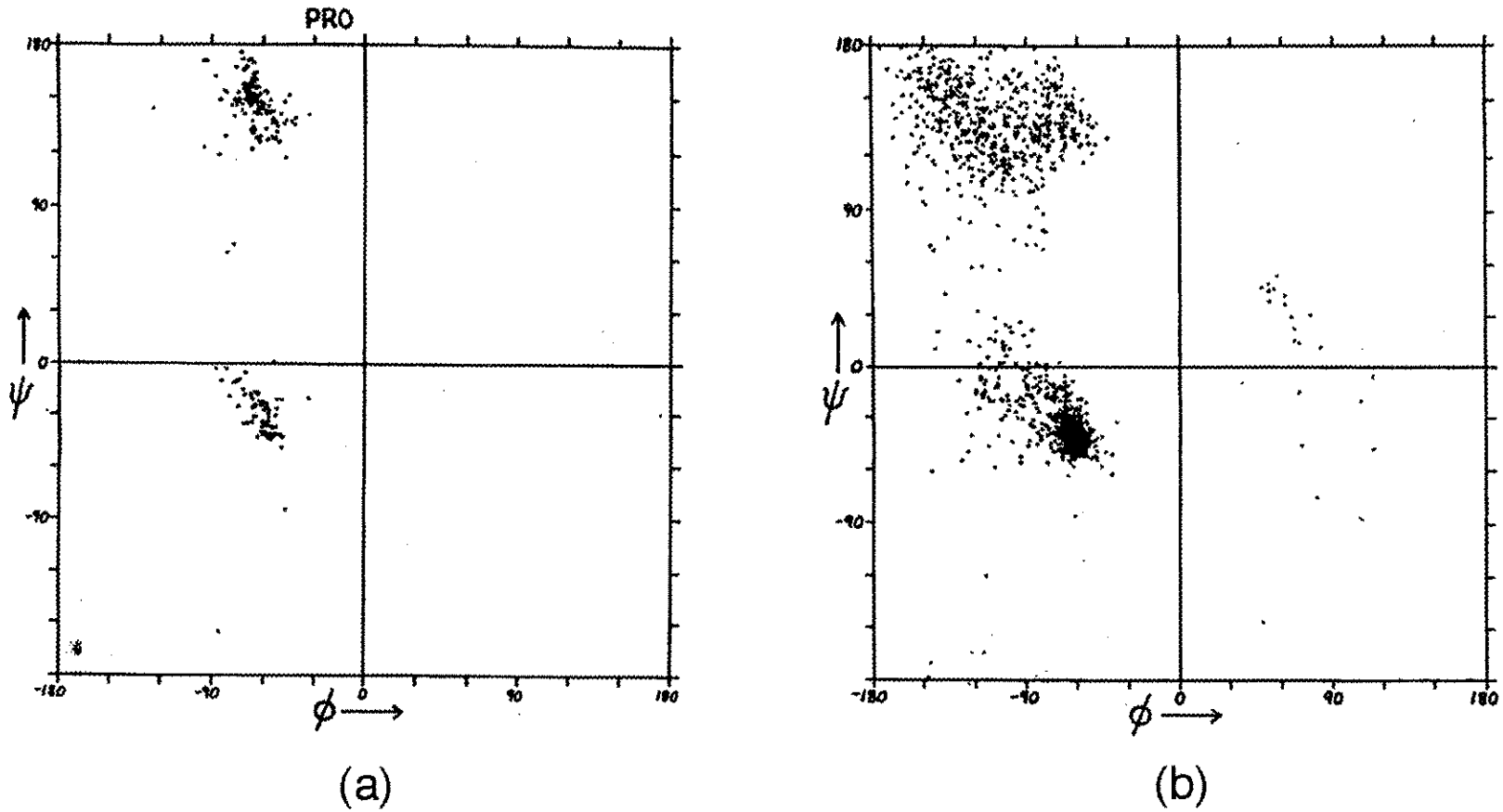


Figura 34 - Rigidez da prolina. À esquerda temos o gráfico de Ramachandran mostrando os ângulos diédricos (Fig. 7) da prolina em um conjunto de proteínas de estruturas conhecidas com alta precisão. À direita temos o mesmo tipo de gráfico, obtido no mesmo conjunto de proteínas, com os ângulos diédricos de todos os aminoácidos, com exceção de glicina e prolina. A estrutura cíclica deste aminoácido impede que assuma os ângulos diédricos típicos de alfa-hélices e placas beta, embora seus valores se situem próximos às regiões correspondentes (explicadas na Fig. 8). Note o menor "espalhamento" dos pontos no gráfico da esquerda, correspondente à menor flexibilidade da cadeia principal neste ponto. Dados retirados de RICHARDSON & RICHARDSON (1988).

A asparagina (N) tem uma cadeia lateral longa e é um aminoácido que tende a desfavorecer alfa hélices devido a interações com a cadeia principal, embora não seja rara sua ocorrência neste estado (RICHARDSON & RICHARDSON, 1988). Sua cadeia lateral é considerada um "pseudopeptídeo", por conter um grupo amina semelhante ao C_α. Isto faz com que ocorra frequentemente na posição imediatamente precedente ao primeiro elemento de alfa hélices, por permitir que se estabeleçam pontes de hidrogênio com padrão de alfa hélice entre sua cadeia lateral e o terceiro resíduo da estrutura (RICHARDSON & RICHARDSON, 1989). A linha correspondente de **a** apresenta valores ligeiramente negativos no centro e valores moderamente positivos nas extremidades. Isto era o esperado, exceto pelos parâmetros do lado N-terminal serem mais favoráveis. Uma provável justificativa para a positividade dos parâmetros do lado N-terminal, é que N frequentemente precede alfa-hélices, sendo extremamente rara nas primeiras três posições desta estrutura, o que atenuaria a intensidade dos parâmetros do lado C-terminal, já que se trata de janela móvel. A linha tem o mesmo padrão desviado para a direita justificado pela orientação das cadeias laterais mostrada na Fig. 9c (veja comentário adiante). O resultado acompanha os achados de GIBRAT *et al.* (1987), baseados exclusivamente na frequência. Talvez exista discrepância no comportamento de N no presente conjunto de dados e naquele utilizado por RICHARDSON & RICHARDSON (1988).

A serina (S) é semelhante à asparagina (N) no que se refere a estruturas alfa. Também é frequente na primeira posição antes da hélice. Em geral desfavorece esta estrutura, embora seja ali muitas vezes encontrada, inclusive nas primeiras posições (RICHARDSON & RICHARDSON, 1988). A linha correspondente em **a** apresenta um padrão coerente. Com valores moderadamente negativos no centro, à direita e ligeiramente positivos à esquerda.

O ácido aspártico (D) difere da asparagina devido à sua carga negativa (STRYER, 1981). Este é um resíduo que favorece regiões expostas, sem estrutura secundária (*turns*, ROSE, GIERASCH, SMITH, 1985; COHEN, PRESNELL, COHEN, 1991). Apesar disto ocorre frequentemente em hélices, em especial na sua porção N-terminal (RICHARDSON & RICHARDSON, 1988). A justificativa por esta preferência é devida a interação de sua carga com o dipolo dessas estruturas, em geral positivas na região N-terminal e negativas na C-terminal, devido à composição dos efeitos dos dipolos individuais dos resíduos (PITITSYN, 1969; HOL, VAN DUIJNEN, BERENDSEN, 1978; PFLUGRATH & QUIOCHO, 1985). Assim a linha correspondente de **a** apresenta valores moderamente positivos à esquerda, valores negativos na porção centro-direita e valores neutros na extrema direita.

A treonina (T), como a serina (S), tem um grupo *OH* na cadeia lateral e é uma versátil participante em pontes de hidrogênio (STRYER, 1981). Assim também é frequente como iniciadora de hélices. Não é tão avessa ao estado alfa como a serina, nem é rara nos primeiros resíduos como a asparagina (RICHARDSON & RICHARDSON, 1989). Sua linha apresenta valores neutros, um pouco positivos do lado N-terminal.

Arginina(R), lisina (K) e histidina (H), sendo positivas, tendem a ocupar o lado C-terminal de hélices, devido ao efeito dipolo citado anteriormente (HOL *et al.*, 1985). A arginina é a mais apolar e a maior entre elas, não favarocendo o estado alfa (DOOLITTLE, 1985). A linha correspondente é a que mostra a menor assimetria. Já as linhas da lisina e principalmente da histidina mostram valores marcadamente mais altos do lado direito.

A cisteína (C) não apresenta forte preferência por qualquer estrutura secundária (RICHARDSON & RICHARDSON, 1989), seu papel marcante é o de estabelecer pontes dissulfídricas entre porções distintas da cadeia (STRYER, 1981). A linha correspondente de **a** é ligeiramente positiva, com valores mais altos no centro e ligeira assimetria, com preferência pelo lado C-terminal.

A glutamina (E), a despeito de sua semelhança estrutural com a arginina, é um aminoácido sem grandes preferências ou aversões, interagindo bem com quase tudo. É frequente em hélices, devido a sua cadeia longa e flexível que se ajusta bem na superfície exposta dessas estruturas, sem competição desestabilizante por pontes de hidrogênio com a cadeia principal (RICHARDSON & RICHARDSON, 1989). A linha correspondente de **a** é marcadamente assimétrica, provavelmente pela orientação das cadeias laterais (Fig. 9c, comentada adiante).

Tirosina (Y), fenilalanina (F) e triptofano (W) são aminoácidos aromáticos apolares (Fig. 2), que geralmente formam o núcleo hidrofóbico das proteínas globulares (PRIVALOV, P.L. & GILL, 1988). Sua aversão por estruturas expostas faz com que sejam frequentes nos lados apolares de hélices e faces apolares de placas beta. Devido ao seu tamanho e apolaridade participam mais de interações com as cadeias laterais que com a cadeia principal (RICHARDSON & RICHARDSON, 1989). A posição mais baixa da linha da tirosina provavelmente se deve ao seu caráter menos hidrofóbico (é capaz de fazer uma ponte de hidrogênio). É a que melhor tolera exposição a solvente entre os três, sendo quase indiferente à estar escondida ou exposta (DOOLITTLE, 1985). A fenilalanina é completamente apolar e tem grande preferência por estruturas regulares. O triptofano também é capaz de uma ponte de hidrogênio, mas isto tem pouco efeito em seu caráter altamente apolar. Triptofano é o

segundo amino ácido mais raro e a tirosina, o terceiro (STRYER, 1981), o que pode comprometer um pouco a significância dos achados.

Valina (V), isoleucina (I), metionina (M) e leucina (L) são aminoácidos alifáticos, apolares (Fig. 2), que apresentam forte preferência por estruturas regulares e são frequentemente encontrados no núcleo apolar de proteína globulares (PRIVALOV, P.L. & GILL, 1988). Seus tamanhos provocam diferentes preferências pelas faces hidrofóbicas de hélices ou de placas. Leucina e metionina sendo maiores, preferem alfa-hélices, embora sejam frequentes em placas. Valina e isoleucina, menores, tem suas linhas em posições inferiores às dos outros dois. Isto se justifica pelo modo distinto de empacotamento de cadeias laterais nas faces hidrofóbicas destes dois tipos de estrutura secundária. Nas hélices é maior a separacão entre as cadeias laterais, e há interdigitação entre as mesmas na face "escondida" (Fig.9b, em azul), favorecendo cadeias grandes que preencham todos os espaços, excluindo o solvente (RICHARDSON & RICHARDSON, 1989). A metionina é o amino ácido mais raro.

O ácido glutâmico (E) tem preferência pelo lado polar das hélices e por outras estruturas expostas devido à sua carga negativa (Fig. 2). A maior afinidade pela extremidade N-terminal da hélice, representada pela assimetria de sua linha em **a**, se justifica pelo efeito de dipolos já citado para outros resíduos polares (D, H, K e R).

Embora a alanina (A) prefira a conformação alfa, sua flexibilidade e falta de reatividade faz com que não apresente grande aversão por nenhum estado, podendo estar em placas beta e regiões expostas. As primeiras preferem resíduos maiores, as últimas, resíduos mais hidrofílicos ou com maior capacidade de participar em pontes. Na ausência de influências ou restrições será encontrada no estado alfa. Polipeptídeos poliprolina formam hélices invariavelmente. Aliás, foi o amino ácido mais frequentemente encontrado nas posições centrais de hélices em estudo feito por RICHARDSON & RICHARDSON (1988).

Além das justificativas achadas para os casos individuais, a própria disposição das cadeias laterais ao longo de uma alfa hélice (Fig.9c) sugere um motivo para a assimetria encontrada na maioria das linhas de **a**. Numa alfa hélice as cadeias laterais apontam para o lado N-terminal, de forma que os resíduos situados depois do resíduo central devem ter maior influência sobre seu estado. Foram exceções E, Y, V, T, S, D e G, explicáveis, na sua maioria: E e D com base no efeito dipolo, S e T por iniciarem hélices, G por não apresentar cadeia lateral. Não pudemos explicar Y e V, assim como o fato de N não ser uma exceção como S e T. O efeito dipolo também serve como justificativa alternativa para a assimetria das linhas de R, K e H.

Da mesma forma que **a**, **b** foi rearranjada em forma matricial (Fig. 22) para possibilitar a interpretação dos parâmetros (Fig. 20). A preferência pelo estado alfa não implica na aversão pelo estado beta. Dependendo das características do amino ácido pode haver até uma concordância nesta preferência. O padrão encontrado em **b** mostra um contraste entre preferências por estruturas regulares (alfa e beta) e não regulares (*coil*). Cordões beta são em geral curtos, de 4 a 6 aminoácidos e separados por porções expostas (BRANDEN & TOOZE, 1991). As linhas de **b** mostram um padrão correspondente. À exceção de E e D, aminoácidos grandes e com carga negativa, todas as outras linhas tem o centro com sinal oposto às extremidades. Aqueles que favorecem o estado beta, quando ocorrem nas posições centrais (apolares), desfavorecem-no, quando nas extremidades, provavelmente por impedir que ali ocorra uma região exposta. Aqueles que desfavorecem o estado beta nas regiões centrais (os polares, G e P), favorecem-no nas extremidades, por que ali costumam estar expostos, contribuindo para a estabilização da placa. As linhas em que o padrão foi menos nítido ou ausente foram justamente aquelas em que, ao invés de haver uma oposição entre estrutura regular e não regular, havia uma oposição entre exposição ou não a solvente, por se tratar de aminoácidos grandes e polares (Fig. 2), que preferem estar expostos em hélices ou em *turns*, mas não em placas beta (E, D, K e R; RICHARDSON & RICHARDSON, 1989).

Novamente, prolina (P) e glicina (G) foram bastante desfavoráveis à formação de estrutura regular, a primeira devido à sua rigidez (Fig. 34), a segunda devido a sua flexibilidade (Fig. 33). O padrão citado acima não foi evidente na linha da prolina, do lado direito. Já a glicina favorece fortemete um estado beta para o resíduo central quando está nas extremidades da janela.

A asparagina (N) desfavorece estruturas beta por ser grande e polar (negativa, Fig. 2). Como comentado, deixou de apresentar o padrão alternado das outras linhas, exceto talvez por uns valores positivos na porção N-terminal. O ácido glutâmico (E) também é grande e negativo e foi quase inteiramente neutro. Uma maior aversão ao estado beta era esperada (RICHARDSON & RICHARDSON, 1989).

Asparagina (N), serina (S) e treonina (T) se comportaram como fazedoras de pontes de hidrogênio. Este comportamento desestabiliza estruturas regulares e favorece a exposição dos resíduos. S e T apresentaram o padrão alternado citado acima e as preferências pelo estado beta (T>S>N) acompanharam aquelas achadas para alfa (Fig. 21), de acordo com a polaridade desses resíduos.

Lisina (K) e arginina (R), aminoácidos grandes e positivos (Fig. 2), tiveram uma certa aversão pelo estado beta, bem menor que a o ácido aspártico (D, grande e negativo) e comparável à do ácido glutâmico (E). Ao contrário de D e E, mostraram o padrão alternado citado, até certo ponto.

A histidina (H) teve comportamento intermediário, assim como é o seu caráter (Fig. 2). Trata-se de amino ácido aromático (caráter apolar), heterocíclico (caráter polar). Como no caso de **a** (parâmetros lineares alfa), acompanhou o comportamento dos aminoácidos grandes e positivos, mostrando uma certa aversão pelo estado beta nas posições centrais e favorecendo-o quando nas extremidades, com mais intensidade que K e R.

A alanina teve sua linha em **b** quase que toda neutra, de acordo com sua indiferença (RICHARDSON & RICHARDSON, 1989). Não apresentou padrão alternado, provavelmente porque porções expostas requerem aminoácidos maiores, mais reativos ou mais polares. Outros aminoácidos que não seguiram o padrão alternado foram a glutamina (E) e a cisteína (C), que não apresentam fortes preferências, tendo também ocupado posições intermediárias em **a** e na lista de parâmetros de GIBRAT *et al.* (1987).

Os aminoácidos grandes e apolares L, I, M e V, assim como os aromáticos F, W e Y (Fig. 2), favoreceram o estado beta quando nas posições centrais. Uma certa ruptura do padrão alternado ocorreu em Y, W e até certo ponto M, justamente os aminoácidos mais raros (STRYER. 1981). Como citado anteriormente, isoleucina e valina favorecem as placas beta, dentre as estruturas regulares, devido a serem menores que L e M, invertendo as posições que suas linhas tiveram em **a** (Fig. 22). Nas faces hidrofóbicas de placas beta o empacotamento de cadeias laterais é mais compacto, favorecendo resíduos pequenos, ao contrário do que acontece com alfa (RICHARDSON & RICHARDSON, 1989).

5.2. Interpretação dos Parâmetros Quadráticos

Índices de anfipaticidade se correlacionam com a tendência de um segmento assumir a configuração de alfa hélice ou cordão beta (EISENBERG *et al.*, 1986; CORNETTE *et al.*, 1987). O modelo quadrático explora uma generalização deste conceito. Considerando a periodicidade, não de um índice, mas das interações de qualquer par de aminoácidos ao longo da janela local, o método mede a tendência destas interações ocorrerem em um dos "lados" de uma alfa hélice ou de uma placa beta. O modelo

utiliza portanto uma propriedade genérica, representada pela própria identidade dos aminoácidos. Voltemos então à Fig.23, onde **A** teve suas linhas e colunas reordenadas de forma a agrupar resíduos de tamanho e polaridade semelhantes. Nela encontramos áreas contendo valores com mesmo sinal para a tendência de formar alfa hélices. No canto superior esquerdo, as estimativas relacionadas a interação de resíduos apolares são geralmente positivas. Isto corresponde à tendência dos resíduos apolares de se agruparem de um dos lados de uma hélice anfipática (SEGREST *et al.*, 1990). Da mesma forma, embora menos intensa, interacões entre resíduos polares parecem reforçar a formação de hélices quando na periodicidade adequada (elementos amarelos e vermelhos no canto inferior direito da Fig. 24). Estes parâmetros interativos podem ser interpretados, até certo ponto, como "propensão a formação de contatos", uma vez que refletem a tendência das cadeias laterais se agruparem de um mesmo lado de alfa hélices e placas betas. Propensões a formação de contatos foram calculadas para estruturas protéicas por SINGH & THORNTON (1992) e deveriam ser comparáveis aos nossos parâmetros. A Fig. 35b mostra as propensões citadas acima, mostrando uma estrutura completamente simétrica, lembrando o padrão da Fig. 35a, que nada mais é do que uma versão da Fig. 23 em branco e preto. Escolhemos esta representação monocromática para evitar o uso de escalas de cor, que teriam de ser diferentes para as duas tabelas.

No canto superior esquerdo de ambas figuras podemos observar um bloco representando as interações entre resíduos não polares (L, I, V, e até certo ponto, F) favoráveis a formação de alfa hélices. Os valores para metionina (M) são geralmente próximos a zero (Fig. 35b) devido ao efeito da penalização, mais intenso nos aminoácidos raros. No canto inferior direito duas figuras aparecem nas duas áreas positivas, correspondentes às interações entre resíduos carregados positivamente (K e R) e aqueles carregados negativamente (D e E). Além disso, o valores negativos (K e K, R com R) podem ser atribuidos à repulsão eletrostática. Àreas com valores negativos nos cantos inferior esquerdo e superior direito, correspondem a contatos entre aminoácidos polares e não polares, pouco frequentes na Fig.35a e desfavoráveis a estruturas alfa na Fig. 35b. Para melhor verificar a similiraridade entre estas matrizes, a correlação de Pearson foi calculada, resultando significativa, apesar de modesta (0.366, $P<0.01$).

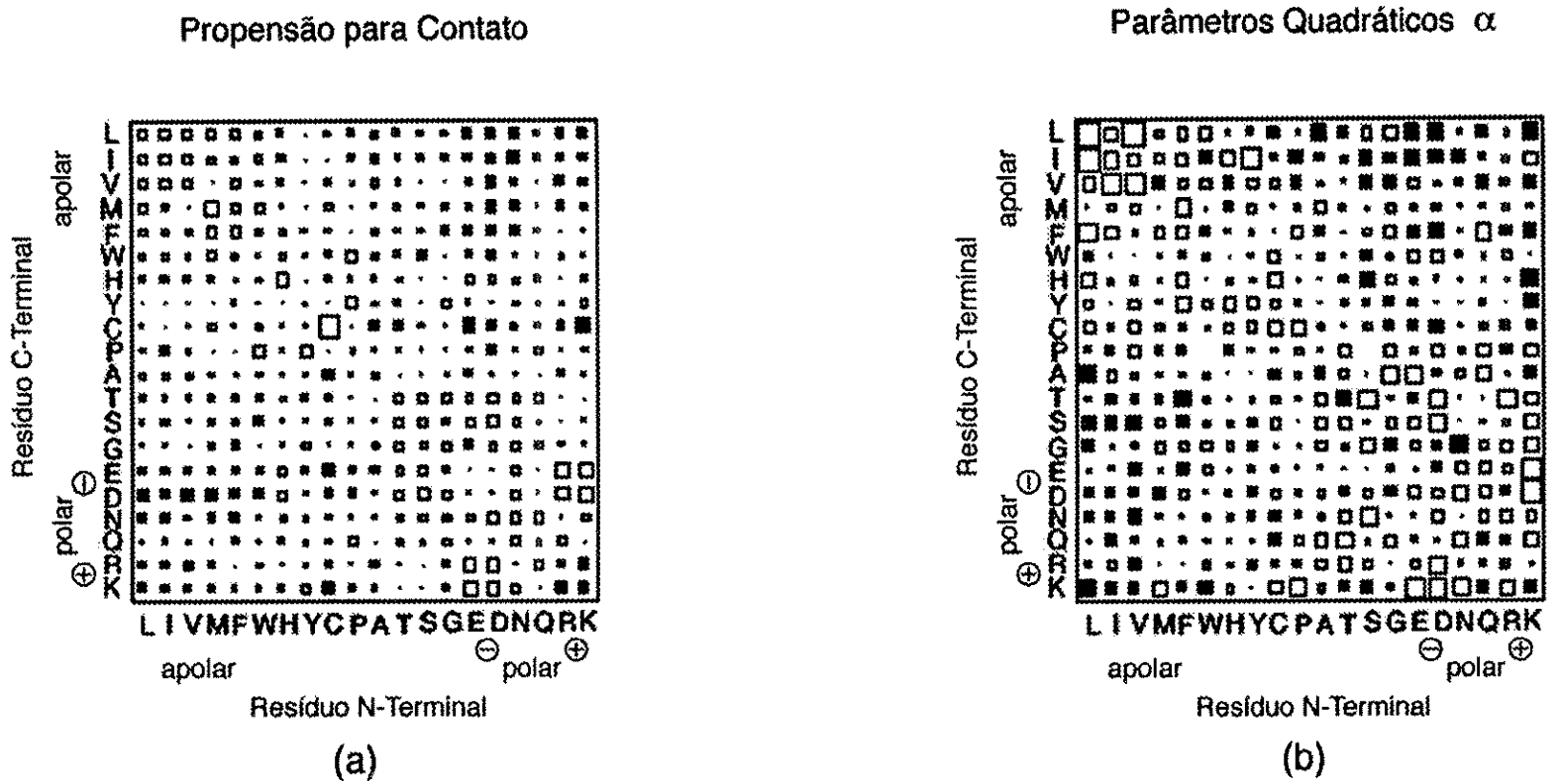


Figura 35 - Semelhança entre (a) Coeficientes calculados por SINGH & THORNTON (1992), baseados na frequência de contatos entre cadeias laterais de aminoácidos no interior de proteínas globulares e (2) Parâmetros quadráticos "alfa" (já mostrados na Fig. 23). Linhas e colunas foram ordenadas agrupando resíduos com propriedades semelhantes (como feito na Fig. 2). Ambas mostram uma predominância de valores positivos (❑), nos cantos superior esquerdo e inferior direito, correspondentes a interações entre aminoácidos com o mesmo tipo de hidropatia (hidrofóbicos e hidrofílicos, respectivamente). Valores negativos (■) predominam nos cantos inferior esquerdo e superior direito, correspondentes a interações entre resíduos de natureza oposta. Alguns valores negativos ocorrem nas interações entre aminoácidos polares com cargas de mesmo sinal, supostamente devido a repulsão eletrostática (e.g. RxK). Os gráficos foram produzidos com auxílio do programa JMP para Macintosh.

Apontamos esta semelhança a fim de argumentar que os parâmetros quadráticos positivos correspondem a pares de aminoácidos que tendem a ser vizinhos de um mesmo lado de alfa-hélices (periodicidade 3.6), fazendo "contato lateral" entre si. É claro que este tipo de contato deve diferir ligeiramente daquele analisado por SINGH & THORNTON (1992), que levava em consideração todos os tipos de interação entre resíduos. Daí algumas diferenças entre as duas matrizes. Note-se, por exemplo, a elevada tendência de contato entre cisteínas (C) na Fig. 35a, devida a frequente ocorrência de pontes dissulfídricas. Como este tipo de associação nunca ocorre dentro de uma alfa-hélice (BRANDEN & TOOZE, 1991), a positividade do parâmetro quadrático correspondente (Fig. 35b) não chama tanto a atenção.

A matriz **B** com os parâmetros beta não mostrou padrão semelhante. Também não há semelhança entre ela e a matriz de "propensão a formação de contatos". Isto não é de todo surpreendente, uma vez que anfipaticidade beta é muito menos frequente do que a anfipaticidade alfa (KLEIN & DELISI, 1986; CORNETTE *et al.*, 1987). Como ilustra a Fig. 36, existem placas beta, paralelas e antiparalelas, recobertas por resíduos apolares em ambos os lados (CHOTHIA & JANIN, 1982). Uma outra razão para a existência em placas beta de segmentos sem a periodicidades esperada é a ocorrência de "protuberâncias" (*beta buldges*, Fig. 37), onde a exclusão de um resíduo quebra o padrão de alternância das cadeias laterais (RICHARDSON, GETZOFF, RICHARDSON, 1978; Fig. 10). Existem 9 protuberâncias beta no conjunto de dados. Portanto, nas placas beta, a periodicidade 2.0 não implica na proximidade espacial das cadeias laterais, o que impede que um padrão semelhante ao da Fig. 35a surja na estrutura da matriz **B**. Isto não significa que os contatos entre os aminoácidos, que ocupam diferentes lados de placas beta, não sigam aproximadamente o padrão descrito por SINGH & THORNTON (1992).

Poderia ser sugerido que a já citada configuração dos cordões beta (trechos curtos de 4 a 6 aminoácidos apolares separados por porções onde a cadeia muda de direção, geralmente com aminoácidos polares e expostos) poderia ser utilizada na forma de outra periodicidade (aproximadamente 10 resíduos) para $\mathbf{W}_b$. Vários motivos nos dissuadiram de tal esforço: (1) Tal periodicidade dependeria do tamanho dos cordões beta na proteína. O padrão surgido no vetor de parâmetros lineares **b** reflete a média encontrada no conjunto de dados, que não é generalizável para proteínas em geral. COHEN *et al.* (1986) já mostraram que depende do tamanho dos domínios estruturais da cadeia protéica. Além disso, a mesma cadeia pode ter placas beta com cordões de tamanho variável (RICHARDSON, 1981). (2) O raciocínio que motivou a utilização da periodicidade baseava-se na manutenção do valor 3.6 ou 2.0 ao longo de cadeia contínua de resíduos (EISENBERG, WILCOX, MCLACHLAN, 1986; CORNETTE

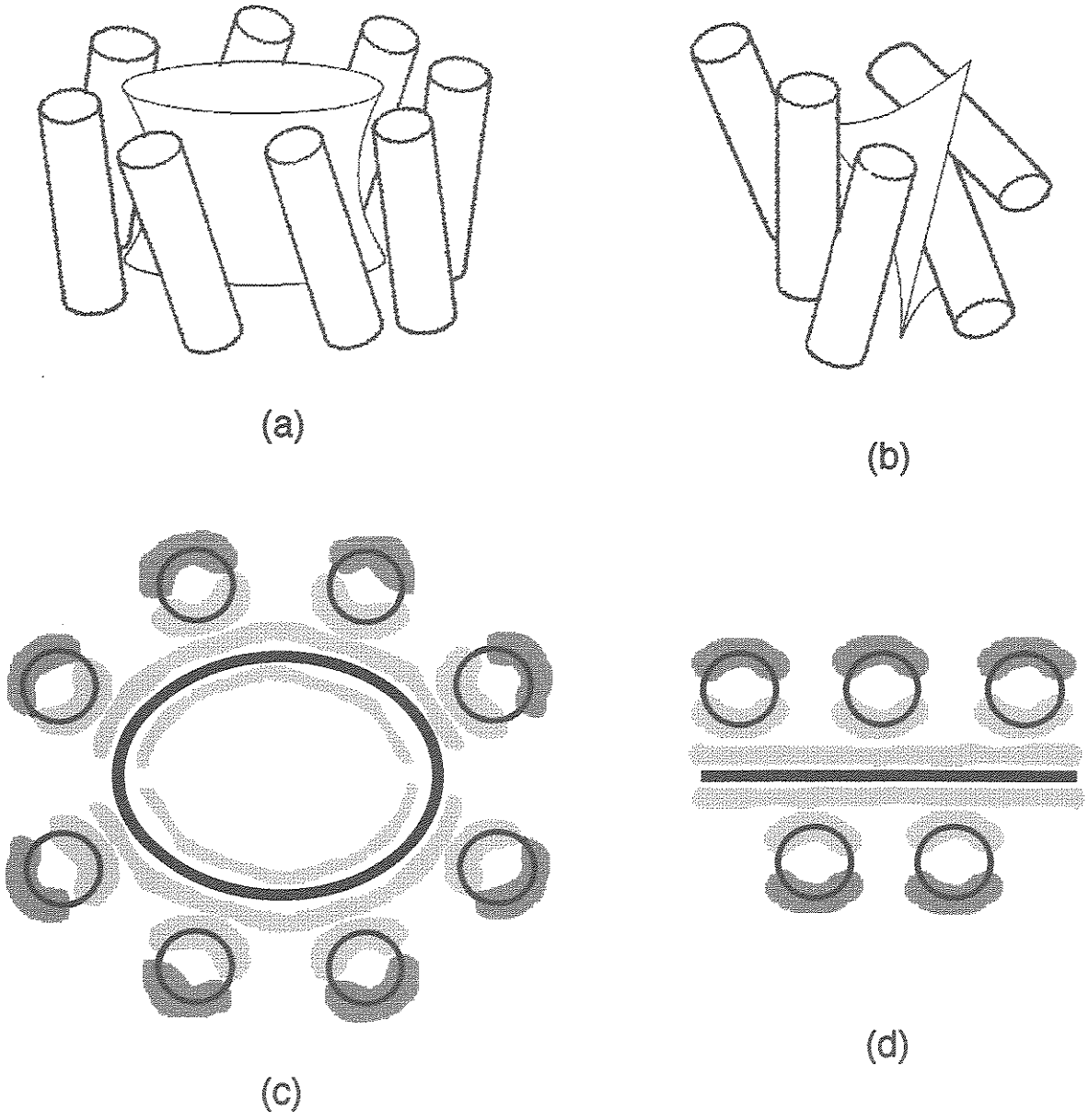


Figura 36 - Placas beta apolares, portanto sem anfipaticidade. (a) Desenho representando "barril" beta cercado de hélices (RICHARDSON, 1981). (b) Placa beta com ambas as faces recobertas por hélices (RICHARDSON, 1981). (c) Vista superior da proteína representada na Fig. 36a, mostrando a placa beta totalmente apolar (azul) e as hélices anfipáticas (face polar em vermelho, face apolar em azul). (d) Vista superior da proteína representada na Fig. 36b, mostrando a placa apolar e as hélices anfipáticas. Existem vários exemplos de placas apolares em PDB e em nosso cojunto de dados. Os segmentos que delas participam não apresentam periodicidade 2.0 na distribuição dos diferentes aminoácidos. Compare-se com a Fig. 11.

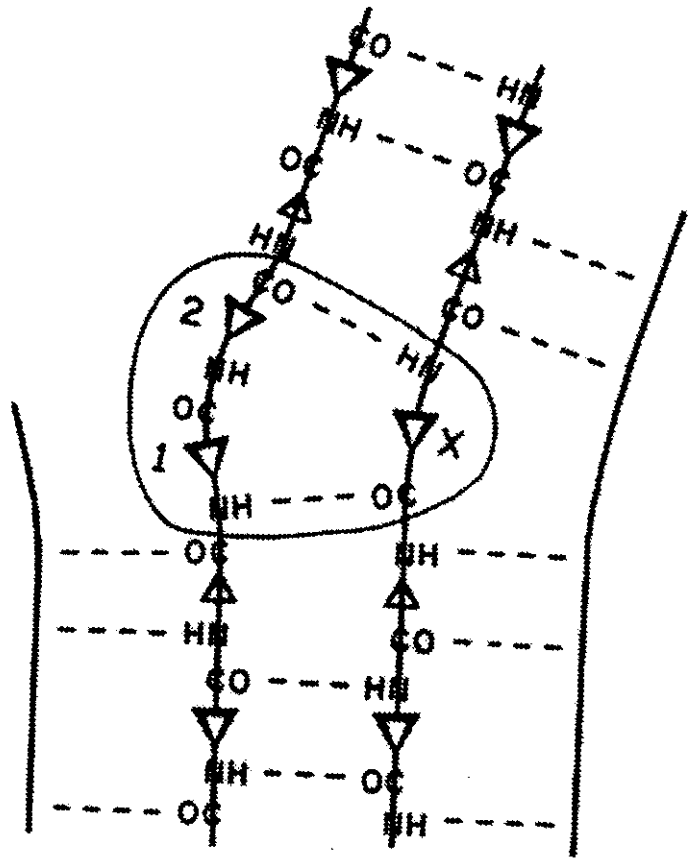


Figura 37 - Protuberância beta. Aqui temos representada uma placa beta com quatro cordões. As cadeias laterais estão representadas por triângulos. Os maiores representam as cadeias acima do plano da página, os menores, aquelas abaixo. Cada aminoácido está representado pelo seu grupo amina (NH), sua cadeia lateral (triângulo) e seu grupo carboxila (CO), nesta ordem (veja o aminoácido X, no terceiro cordão). Trata-se de uma placa antiparalela. No segundo cordão, da esquerda para a direita, o grupo CO do aminoácido 1 e o grupo NH do aminoácido 2 são excluídos da rede de pontes de hidrogênio. Isto faz com que duas cadeias laterais consecutivas (1 e 2) ocupem o mesmo lado da placa, quando o normal é que se alternem de um lado e de outro (Fig. 10). Se a placa for anfipática, isto quebrará o padrão de periodicidade no segundo cordão. Retirado de RICHARDSON & RICHARDSON (1988).

et al., 1987). As placas beta encontradas na natureza raramente são formadas pela simples alternância de cordões e porções expostas. Não só a ordem destes cordões difere, na sequência e na placa, como também, costuman ocorrer em porções distantes uma das outras havendo interposição de outros segmentos, de forma a destruir a periodicidade (BRANDEN & TOOZE, 1991).

A seleção de variáveis foi mais uma maneira de reduzir seletivamente o número de parâmetros do modelo. A acurácia obtida para modelos mais simples tem maior chance de se estender a novas proteínas. Os parâmetros mais significativos ($t > 2$) também foram aqueles mais facilmente interpretáveis com base na biofísica. Quando usamos os 80 parâmetros quadráticos mais significativos em cada uma das matrizes **A** e **B** obtivemos uma acurácia de 65.9%, mais de 3 pontos percentuais acima do melhor modelo com penalização apenas. Curiosamente, o modelo com maior verossimilhança (menor LVN, Tab. 3, seção *iv*, 120 parâmetros quadráticos selecionados), não foi o de maior acurácia (65.5%). Teoricamente, este é o modelo com maior capacidade preditiva em proteínas ausentes do conjunto de dados.

Embora tivéssemos validado o ajuste dos parâmetros por cruzamento, o mesmo não foi feito para o processo de seleção em si. Acreditamos que este pequeno aumento na acurácia também ocorra quando o método for aplicado em novas proteínas. Se houvesse uma grande diferença, entre as diversas proteínas, nas interações em que se baseou a seleção, deveríamos notar uma grande variação nas acurácias individuais antes e depois da seleção, o que não ocorreu.

A ordenação dos parâmetros quadráticos pelo valor de t, feita para orientar a seleção, também foi útil na interpretação da plausibilidade de seus valores. A maioria dos parâmetros significativos pode ser interpretada com base na hidrofobicidade e polaridade dos aminoácidos correspondentes, como mostram as Tabelas 5 a 8. Pares de mesma natureza (hidrofobicidade semelhante) e de polaridade oposta, tenderam a favorecer a estrutura secundária. O contrário ocorreu para pares com hidrofobicidade muito diferente ou cargas do mesmo sinal. Isto mostra que a utilização da periodicidade permitiu ao modelo detectar "propensão de contatos" entre resíduos (Fig. 35). Encontramos duas vezes mais parâmetros significativos em **A** do que em **B**. A maior parte dos valores implausíveis ou indiferentes foram encontrados em **B**. Isto é mais uma indicação que a periodicidade é mais significativa nas alfa hélices, como já discutido. Ainda assim a eliminação completa do modelo quadrático, ou da periodicidade da parte beta, assim como o uso de um único bloco de parâmetros para alfa e beta, não melhoram o desempenho do modelo (dados não mostrados).

Avaliamos a plausibilidade das interações baseados exclusivamente na hidrofobicidade e polaridade, o que pode não corresponder a realidade. Na Tabela 6. podemos ver que somente um par com t >2 foi considerado implausível, por representar a interação de dois aminoácidos polares. De acordo com o raciocínio que motivou o uso da periodicidade, dois resíduos de mesma natureza, do mesmo lado da alfa-hélice, deveriam favorecer a ocorrência deste tipo de estrutura. Tratava-se do par TxT. Além de não favorecer o estado alfa, a treonina prima por ter uma cadeia lateral que tenta fazer pontes de hidrogênio com a cadeia principal. Isto desestabiliza estruturas regulares (RICHARDSON & RICHARDSON, 1989) e poderia ser a razão para o valor negativo deste parâmetro.

Outros aminoácidos tem seu comportamento influenciado por outros fatores que não sua hidrofobicidade. A glicina prima por sua flexibilidade, apesar de pequena e polar não tem cadeia lateral para interagir com outos aminoácidos. A prolina é muito rígida e apesar de pequena e apolar tem praticamente não tem como interagir com outras cadeias. Desfavorece estruturas regulares, no sentido de não poder assumir uma determinada conformação ou não poder formar pontes de hidrogênio. A alanina apesar de pequena e polar não tem grande aversão por resíduos polares (RICHARDSON & RICHARDSON, 1989). No caso destes três aminoácidos nada foi incluído nas Tabelas 5 a 8 sobre a plausibilidade dos parâmetros associados.

Aminoácidos de polaridade intermediária, tais como H, Y e W (Fig. 2), foram tratados da seguinte forma: a histidina foi considerada polar e positiva, embora sua carga dependa do pH e tenha considerável caráter apolar, especialmente quando não ionizada (STRYER, 1981). A tirosina e o triptofano foram considerados não polares.

Na Tabela 5, que mostra os parâmetros quadráticos positivos mais significativos encontrados em **A**, apenas três interações (GxA, ExA e PxK) envolvendo aminoácidos pequenos ficaram sem uma explicação razoável (para t >2, acima da linha divisória). G e A são aminoácidos pequenos, e podem favorecer hélices se ocorrerem do lado hidrofílico da mesma, aumentando a accessibilidade ao solvente. Contudo, ainda que a alanina ocorra muitas vezes na porção exposta de hélices, certamente prefere ambientes mais hidrofóbicos (RICHARDSON & RICHARDSON, 1988).

Entre as implausibilidades encontradas na Tabela 6, das interações que deveriam desfavorecer alfa hélices, encontramos os pares TxT (já discutido) e SxH. A primeira já foi discutida. Serina e histidina deveriam ocorrer no lado hidrofílico das hélices, embora a primeira prefira a porção N-terminal e a segunda a porção C-terminal. NxG não tem motivo aparente para ser desfavorável, exceto talvez o fato desses serem os

dois aminoácidos mais frequentemente encontrados na região L_α do gráfico de Ramachandran (RICHARDSON & RICHARDSON, 1989; Fig. 8) e ambos tem aversão pelo estado alfa. Então, por que não vemos GxN entre os parâmetros significativos ?

Apenas nove elementos positivos de B tiveram t >2. Entre eles, LxE não pode se justificar pela hidrofobicidade. Também não encontramos justificativa para a interação favorável entre CxG, embora ela não seja implausível. Já entre os elementos negativos (Tab. 8) achamos 14 significativos. Apenas 5 deles se justificam. Dos demais 3 foram implausíveis por envolverem resíduos de caráter semelhante: KxN, IxYe IxL. O primeiro par é formado por aminoácidos que desfavorecem a estrutura beta individualmente. Os dois outros têm a isoleucina, pequena e apolar, que individualmente é o amino ácido que mais favorece a estrutura beta, junto com dois aminoácidos grandes e apolares, também muito frequentes nas faces apolares de placas (RICHARDSON & RICHARDSON, 1989). A justificativa para sua interação desfavorável talvez seja o maior empacotamento das cadeias laterais hidrofóbicas, já citado antes, que favorece resíduos apolares pequenos em detrimento dos grandes. Todas as outras 6 interações significativas da mesma Tabela 8 envolveram aminoácidos pequenos A, P e G, o que é curioso uma vez que eles não tem cadeias laterais para interagir. Todos eles desfavorecem estruturas beta, em especial G e P.

5.3. Constantes de Decisão

Para completar nossa tentativa de reproduzir o trabalho de GIBRAT *et al.*, (1987) usando estatística convencional, usamos "constantes de decisão" C1 e C2 (Eqs. 44 e 45). Seu cálculo difere bastante de $\mathbf{DC_H}$ e $\mathbf{DC_E}$ (Eqs. 7 e 8), pois o modelo já contava com os parâmetros aditivos a_0 e b_0 (Eqs. 10 e 11), estimados por MV. Ao refletir a proporção de resíduos nos estados alfa e beta existentes no conjunto de treino, estas constantes ($\mathbf{DC_H}$, $\mathbf{DC_E}$, C1 e C2) permitem um aumento espúrio da acurácia. Seu uso só faria sentido se pudéssemos prever esta proporção. Uma vez feito isso, poderíamos usar constantes de decisão apropriadas para prever a estrutura de novas proteínas. O uso deste artifício também foi capaz de elevar a acurácia de nosso modelo (67.8%). Além de selecionar valores para o conjunto de dados como um todo, selecionamos um par de valores para cada uma das cadeias estudadas, a fim de revelar uma relação entre o valor das constantes e o tipo estrutural da cadeia protéica. Embora isso resultasse numa acurácia de 72.2%, não fomos capazes de estabelecer uma relação consistente entre os valores das constantes de decisão e o tipo estrutural da cadeia. Portanto, esta acurácia não se estende a proteínas fora do conjunto de dados.

Devemos afirmar que, apesar de espúria, é essa (67.8%) a acurácia a ser comparada com o valor de 63% originalmente publicado por GIBRAT *et al.* (1987), que também utilizou constantes de decisão. Todavia, não sabemos qual acurácia validada seria atingida com a atualização dos dados.

Talvez ainda melhor seria comparar resultados sem constantes de decisão. Modelo logístico quadrático restrito por periodicidade, com penalização da verossimilhança, com seleção de variáveis, com validação cruzada: 65.9%. Modelo de GIBRAT *et al.* (1987), usando teoria da informação, variáveis fictícias, com seleção de variáveis e validação cruzada: 61.8% .

5.4. Perspectivas de Continuidade do Trabalho

O uso de modelos logísticos, com otimização por máxima verossimilhança, com número e tipo de parâmetros adequados à quantidade de informação disponível, validados por cruzamento, permitiu a obtenção de resultados ligeiramente mais acurados que aqueles anteriormente publicados para o mesmo conjunto de dados. Também nos permite supor que exploramos adequadamente toda a informação estrutural contida em segmentos de até 17 aminoácidos. Nos resta então propor novas abordagens que possam levar a uma acurácia ainda maior.

A maior limitação encontrada foi a quantidade dados disponível, pequena frente à complexidade dos modelos que desejávamos utilizar. A disponibilidade de novas estruturas protéicas certamente permitirá o uso de maior número de parâmetros, possibilitando um melhor desempenho. Uma forma inteligente de amplificar os dados disponíveis foi implementada por ROST & SANDER (1993), cujo método detém a maior capacidade preditiva até o momento. Tal trabalho foi publicado durante a fase de conclusão deste. O presente modelo foi extendido de forma a usar o mesmo tipo de informação (sequências com semelhança superior a 80%), mantendo nossa abordagem de modelo logístico quadrático penalizado de máxima verossimilhança. Ao contrário do uso de redes neurais, modelos logísticos permitem a interpretação dos parâmetros, assim como a substituição de procedimentos, tais como o "júri" (ROST & SANDER, 1992), por técnicas bem estabelecidas de estatística computacionalmente intensiva. Além disso, um método moderno de previsão necessita validação por cruzamento, que foi apenas parcialmente realizada por esses autores. Os resultados preliminares apontam para um desempenho comparável e não foram apresentados aqui por terem

sido obtidos pela colega Valentina di Francesco, que deve ser considerada a autora principal, apesar de nossa participação.

Uma segunda forma de melhorar a acurácia seria levar em conta informações contidas fora da janela local, especialmente aquelas que consideram interações a longa distância. Podemos fazer isto prevendo o tipo estrutural da proteína. Muitos métodos já surgiram baseados tanto em janelas locais quanto em proporção de aminoácidos (SHERIDAN *et al*, 1985; DELÉAGE & ROUX, 1987). O maior problema destes métodos é a suposicão de que podemos substituir classe estrutural por conteúdo relativo de estrutura secundária (KLEIN, 1986; KLEIN & DELISI, 1986; NAKASHIMA, NISHIKAWA & OOI, 1986). Os achados das Tabelas 9 a 12, também discutidos nas Figs. 25 a 28, apontam para o equívoco de tal abordagem. Enquanto se espera que as proteínas com predominância de alfa hélices ou placas beta se sujeitem a tal simplificação, o mesmo não acontece com proteínas do tipo α/β e $\alpha+\beta$ de LEVITT & CHOTHIA (1975), assim como cadeias com múltiplos domínios (KLEIN, 1986).

Uma explicação para a ausência de padrão consistente nos valores de C1 e C2, seria a existência de diferentes tipos de segmentos dentro de uma mesma cadeia, como ilustrado nas Figuras 27 e 28. A Fig 28a mostra que uma proteína $\alpha+\beta$ (LEVITT & CHOTHIA, 1975) tem elementos α, β e α/β em sua sequência (numerados 1, 2 e 3 no canto superior direito). O mesmo pode acontecer com cadeias de cada uma das outras classes, como demonstra a Fig. 27a, onde uma proteína a/b, tem um segmento C-terminal onde elementos α e β não apresentam a alternância característica. A solução estaria em prever a participação de cada porção de sequência na estrutura global da proteína, usando não somente a informação sequencial local, como também a informação contida na cadeia como um todo (classe estrutural). Já realizamos grande estudo de classificação das proteínas contidas em PDB, com ênfase na identificação dos segmentos de sequência que participam de diferentes tipos de estrutura, que pretendemos utilizar para a continuação deste trabalho.

Uma terceira forma de melhorar o modelo seria através de outros tipos de restrições ao modelo logístico, em especial à sua porção quadrática. Estas poderiam ser baseadas na biofísica, tais como preferências do comprimento de hélices e cordões, ou em dados filogenéticos, tais como indicadores de regiões conservadas. Um outro modelo quadrático para a porção beta seria desejável. A natureza deste tipo de estrutura secundária, contudo, implica na interação entre pares de aminoácidos de porções muitas vezes distantes na sequência, o que exigiria utilização de janelas muito grandes ou múltiplas e ,consequentemente, modelos mais complexos.

Obviamente esses três tipos de abordagem não são exclusivos, devendo ser combinados.

6. Conclusões

O uso de metodologia estatística permitiu uma melhora de pouco mais de 3 pontos percentuais na acurácia de métodos preditivos de estrutura secundária de proteínas.

Este resultado representa um avanço muito modesto no sentido de prever a estrutura protéica a partir da sequência de aminoácidos. Contudo, o uso de modelos probabilísticos, onde as densidades marginais foram estimadas com máxima verossimilhança, com validação cruzada, onde foi realizada uma seleção da melhor formulação, procurando adequar a complexidade à quantidade de dados, nos permite afirmar que utilizamos toda a informação estrutural contida em segmentos de até 17 aminoácidos, na complexidade permitida.

Somente a utilização de informação adicional, contida em novas proteínas que venham a ter sua estrutura resolvida, em sequências homólogas ou ainda em porções da sequência situadas além da vizinhança já explorada, poderá elevar a acurácia além dos 65-70 pontos percentuais.

O uso de modelos logísticos quadráticos é promissor para prosseguir nas direções apontadas acima.

7. Referências Bibliográficas

AGRESTI, A.- **Categorical Data Analysis**. John Wiley & Sons, New York, 1990. 559p.

ANFINSEN, C.B.; EPSTEIN, C.J.; GOLDBERGER, R.F. - The genetic control of tertiary protein structure: studies with model systems. **Cold Spring Harbor Symp. Quant. Biol., 28:** 439-449, 1963.

ASAI, K.; HAYAMIZU, S.; HANDA, K. - Prediction of protein secondary structure by the hidden Markov model. **Compt. Appl. Biosc., 9:** 141-6, 1993.

BALDWIN, R.L. - Temperature dependence of the hydrophobic interaction in protein folding. **Proc. Natl. Acad. Sci. U.S.A., 83:** 8069-8072, 1986.

BAIROCH, A. & BOECKMANN, B.- The SWISS-PROT protein sequence data bank. **Nucl. Acids Res., 20:** 2019-2022, 1992.

BERNSTEIN, F.C.; KOETZLE, T.F.; WILLIAMS, G.H.B; MEYER, E.F.; BRICE, M.D.; RODGERS, J.R.; KENNARD, O.; SHIMANOUCHI, T.; TASUMI, M. - The Protein Data Bank: a computer-based archival file for macromolecular structures. **J. Mol. Biol., 112:** 535-542, 1977.

BIOU, V.; GIBRAT, J. F.; LEVIN, J.M.; ROBSON, B.; GARNIER, J. - Secondary structure prediction: combination of three different methods. **Protein Engineering, 2**:185-191, 1988.

BISHOP, Y.M.M.; FIENBERG, S.E.; HOLLAND, P.W. - **Discrete mulitvariate analysis: theory and practice**. The Mit Press, Cambridge, Mass., 1975, 535p.

BLUNDELL, T.L. & JOHNSON, L.N. - **Protein crystallography**. London, Academic Press, 1976. 602p.

BRANDEN, C. & TOOZE, J. - **Introduction to protein structure.** New York, Garland, 1991. 302p.

BRILLOUIN, L. - **Science and information theory.** Academic Press, New York, 1956. 432p.

BROOKS, B.R.; BRUCCOLERI, R.E.; OLAFSON, B.D. - CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. **J. Comp. Chem., 4:** 187-217, 1983.

BRUCCOLERI, R.E. & KARPLUS, M. - Prediction of the folding af short polypeptide segments by uniform conformational sampling. **Biopolymers, 26:** 137-168, 1987.

BRYANT, S.H.; LAWRENCE, C.E. - An empirical energy function for threading protein sequence through folding motifs. **Proteins, 5:** 29-35, 1993.

CHARTON, M. & CHARTON, B.I. - The structural dependence of amino acid hydrophobicity parameters. **J. Theor. Biol., 99:** 629-644, 1982.

CHOTHIA, C. - The nature of accessible and buried surfaces in proteins. **J. Mol. Biol., 105:** 1-14, 1976.

CHOTHIA, C. & JANIN, J. - Orthogonal packing of β-pleated sheets in proteins. **Biochemistry, 21:** 3957-3965, 1982.

CHOU, P.Y. & FASMAN, G.D. - Prediction of Protein Conformation I. **Biochemistry, 13:** 211-222, 1974a.

CHOU, P.Y. & FASMAN, G.D. - Prediction of Protein Conformation II. **Biochemistry, 13:** 222-245, 1974b.

CHOU, P.Y. & FASMAN, G.D. - Prediction of the secondary structure of proteins from their amino acid sequence. **Adv. Enzymol. and Related Areas Mol. Bio., 47:**45-148, 1978a.

CHOU, P.Y. & FASMAN, G.D. - Empirical Predictions of Protein Conformation. **Rev. Biochem., 47:** 251-273, 1978b.

CLEVELAND, W.S. & DEVLIN, S.J. - Locally weighted regression: an approach to regression analysis by local fitting. **J. of the Amer. Stat'l. Assn., 83:** 596-610, 1988.

CLORE, G.M. & GRONENBORG. A.M. - Determination of threedimensional structure in proteins and nucleic acids in solution by nuclear magnetic resonance spectroscopy. **CRC Crit. Rev. Biochem., 24:** 479-564, 1989.

COHEN, B.I.; PRESNELL, S.R., COHEN, F.E. - Pattern-based approaches to protein structure prediction. **Meth. Enzymol., 202:** 252-268, 1991.

COHEN, F.E.; ABARBANEL, R.M.; KUNTZ, I.D.; FLETTERICK, R.J. - Pattern-matching algorithm for prediction of turns in proteins. **Biochemistry, 25:** 266-275, 1986.

COHEN, F.E. & KUNTZ, I.D. - Terciary structure prediction. In: FASMAN, G.F - **Prediction of protein structure and the principles of protein conformation.** New York, Plenum, 1989. p.647-705.

CORNETTE, J.L.; CEASE, K.B.; MARGALIT, H.; SPOUGE, J.L.; BERZOFSKY, J.A.; DELISI, C. - Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. **J. Mol. Biol., 195:** 659-85, 1988.

DELÉAGE, G. & ROUX, B. - An algorithm for protein secondary structure prediction based on class prediction. **Protein Eng., 1:** 289-294, 1987.

DOOLITTLE, R.F. - Proteins. **Scientific American, 253(4):** 74-83, 1985.

DRAPER, N.L. & SMITH, H. **Applied regression analysis.** 2.ed. John Wiley & Sons, New York, 1981. 407p.

EISENBERG, D.; WILCOX, W.; MCLACHLAN, A.D. - Hydrophobicity and amphiphilicity in protein structure. **J. Cell. Biochem., 31:** 11-17, 1986.

EISENBERG, D. & HILL, C.P. - Protein Crystallography: more surprises ahead. **Trends Biochem. Sci., 14:** 260-264, 1989.

ENGH, H.R.; LOEBERMANN, M.; SCHNEIDER, G.; WIEGAND, R.; HUBER, R. - The S variant of human alpha-1-antitrypsin, structure and implications for function and metabolism. **Protein Eng.,** 2: 407, 1989.

EUBANK, R.L. - **Spline smoothing and nonparametric regression.** Marcel Dekker, Inc. New York, 1988. 457p.

FREIRE, E.; MURPHY, K.P.; SANCHEZ-RUIZ, J.M.; GALISTEO, M.L.; PRIVALOV, P.L. - The molecular basis of cooperativity in protein folding: thermodynamic dissection of interdomain interactions in phosphoglycerate kinase. **Biochemistry, 31:** 250-256, 1992.

GARNIER, J. & LEVIN, J.M. - The protein structure code: what is its present status? **Cabios, 7**: 133-142, 1991.

GARNIER, J.; OSGUTHORPE, D.J.; ROBSON, B. - Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. **J. Mol. Biol., 120:** 97-120, 1978.

GIBRAT, J.-F.; GARNIER, J.; ROBSON, B. - Further developments of protein secondary structure prediction using information theory new parameters and consideration of residue pairs. **J. Mol. Biol., 198:** 425-443, 1987.

GIBRAT, J.-F.; ROBSON, B.; GARNIER, J. - Influence of the local amino acid sequence upon the zones of the torsional angles adopted by residues in proteins. **Biochemistry, 30:** 1578-1586, 1991.

GEER, S. VAN DE - Estimating a regression function. **Ann. of Stat., 18:** 907-924, 1990.

GOOD, I.J. & GASKINS, R.A. - Nonparametric roughness penalties for probability densities. **Biometrika, 58:** 255-277, 1971.

GUARDABASSO, V.; MUNSON, P.J.; RODBARD, D. - A versatile method for simultaneous analysis of families of curves. **FASEB J., 2(3):** 209-215, 1988.

GUNSTEREN, W.F. VON - Molecular dynamics studies of proteins. **Curr. Opin. Str. Biol., 3:** 167-174, 1993.

HARDLE, W.; HALL, P.; MARRON, J.S. - How far are automatically chosen regression smoothing parameters from their optimum. **J. Amer. Stat. Assoc., 83(401):** 86-101, 1988.

HENDRICKSON, W - X-ray diffraction. In: OXENDER, D. & FOX, C.F. Eds. - **Protein Engineering.** New York, Liss, 1986. p341-423.

HERINGA, J. & ARGOS, P. - Side chain clusters in protein structures and their role in protein folding. **J. Mol. Biol., 220:** 151-171, 1991.

HOL, W.; VAN DUIJNEN, P.; BERENDSEN, H. - The α-helix dipole and the properties of proteins. **Nature, 273:** 443-446, 1978.

HOLLEY, L.H. & KARPLUS, M. - Protein secondary structure prediction with a neural network. **Proc. Natl. Acad. Sci. U.S.A., 86:** 152-156, 1989.

HOLLEY, L.H. & KARPLUS, M. - Neural networks for protein secondary structure prediction. **Meth. Enzymol., 202:** 204-224, 1991.

HUNTER, L. & STATES, D.J. - Bayesian classification of protein structural elements. **Proceedings of the 25th Annual Hawaii International Conference on Systems Sciences**, 1991, IEEE. 595-604.

JMP User's Manual. SAS Institute, Inc., Cary, NC, 1994. 354p.

JOHNSON, R.A. & WICHERN, D.W - **Applied multivariate statistical analysis.** 2.ed. Englewood Cliffs, New Jersey, Prentice Hall, 1988. 607p.

KABSCH, W. & SANDER, C. - Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. **Biopolymers, 22:** 2577-2637, 1983a.

KABSCH, W. & SANDER, C. - How good are predictions of protein secondary structure ? **FEBS Lett., 155:** 179-182, 1983b.

KIDERA, A.; KONISHI, Y.; OKA, M.; OOI, T; SCHERAGA, H.A. - An orthogonal set of parameters to describe amino acid properties. **J. Prot. Chem., 3:** 23-95, 1985.

KING, R.D. & STERNBERG, M.J. - Machine learning approach for the prediction of protein secondary structure. **J. Mol. Biol., 216:** 441-57, 1990.

KLEIN, P - Prediction of protein structural class by discriminant analysis. **Bioch. Biophys. Acta, 874:** 205-215, 1986.

KLEIN, P. & DELISI, C. - Prediction of protein structural class from amino acid sequence. **Biopolymers 25:** 1659-1672, 1986.

KNELLER, D.G.; COHEN, F.E.; LANGRIDGE, R. - Improvements in protein secondary structure prediction by an enhanced neural network. **J. Mol. Biol., 214:** 171-182, 1990.

LAMBERT, M.H. & SCHERAGA, H.A. - Pattern recognition in the prediction of protein structure I: tripeptide conformational probabilities calculated from the amino acid sequence. **J. Comput. Chem. 10:** 770-797, 1989.

LATTMAN, E.E. & ROSE, G.D. - Protein folding: what's the question ? **Proc. Natl. Acad. Sci. U.S.A., 90:** 439-441, 1993.

LEVITT, M. & CHOTHIA, C. - Structural patterns in globular proteins. **Nature, 261:** 552-558, 1976.

LEVITT, M. & GREER, J. - Automatic identification of secondary structure in globular protein. **J. Mol. Biol., 114:** 181-293, 1977.

LIM, V.I. - Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. **J. Mol. Biol., 88:** 857-872, 1974a.

LIM, V.I. - Algorithms for prediction of α-helical and β-structural regions in globular proteins. **J. Mol. Biol., 88:** 873-894, 1974b.

MACIMDAD Version 4 manual. Yeda, Calif., Molecular Applications Group, 1993. 231p.

MANAVALAN, P. & PONNUSWAMY, P.K. - A study of the preferred environment of amino acid residues in globular proteins. **Arch. Biochem. Biophys., 184:** 476-487, 1977.

MATLAB User's Manual. South Natick, Mass., The Math Works, Inc., 1989. 323p.

MATTHEWS, B.W. - Comparison of the predicted and observed secondary structure of T4 phage lysozyme. **Biochim. Biophys. Acta, 405:** 442-451, 1975.

MAXFIELD, F.R. & SCHERAGA, H.A. - Status of empirical methods for the prediction of protein backbobe topography. **Biochemistry, 15:** 5138-5153, 1976.

NAKASHIMA, H.; NISHIKAWA, K.; OOI, T. - The folding type of a protein is relevant to the amino acid composition. **J. Biochem. (Tokyo), 99:** 153-162, 1986.

NARAYAMA S.V.L. & ARGOS, P. - Residue contacts in protein structures and implications for protein folding. **Int. J. Peptide Protein Res., 24:** 25-39, 1984.

NEMETHY, G. & SCHERAGA, H.A. - Potein Folding. **Q. Rev. Biophys., 3:** 239-352, 1977.

NIERMANN, T. & KIRSCHNER, K. - Improving the prediction of secondary structure of 'TIM-barrel' enzymes. [Artigo corrigido e republicado. Original em Protein Eng. 4: 137-47, 1990]. **Protein Eng., 4:** 359-370, 1991.

NISHIKAWA, K. - [Prediction of protein secondary structure by a new joint method]. **Seikagaku, 62:** 1490-6, 1990.

NISHIKAWA, K. & NOGUCHI, T. - Predicting protein secondary structure based on amino acid sequence. **Meth. Enzymol., 202:** 31-44, 1991.

PFLUGRATH, J. & QUIOCHO, F. - Sulphate sequestered in the sulphate-binding protein of *Salmonella typhimurium* is bound solely by hydrogen bonds. **Nature, 314:** 257-260, 1985.

PIMENTEL, G.C. & SPRATLEY, R.D. - **Química, um tratamento moderno.** v.2. São Paulo, Edgard Blücher, Ed. da Universidade de São Paulo e INL, 1974. 777p.

PRIVALOV, P.L. & GILL, S.J. - Stability of protein structure and hydrophobic interaction. **Adv. Prot. Chem., 39:** 193-234, 1988.

PTITSYN, O. - Statistical Analysis of the distribution of amino acids among helical and non-helical regions in globular proteins. **J. Mol. Biol., 42:** 501-510, 1969.

QIAN, N. & SEJNOWSKI, T.J. - Predicting the secondary structure of globular proteins using neural network models. **Cabios, 5:** 163-178, 1989.

RACKOVSKY, S. - On the nature of the protein folding code. **Proc. Natl. Acad. Sc. U.S.A., 90:** 664-648, 1993.

RAMACHANDRAN, G.N. & SASSEKHARAN, V - Conformation of polypeptides and proteins. **Adv. Protein Chem., 23:** 283-437, 1968.

RICHARDS, F.M - Areas, volumes, packing and protein structure. **Annu. Rev. Biophys. Bioeng., 6:** 151-176, 1977.

RICHARDS, F.M & KUNDROT, C.E. - Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. **Proteins, 3:** 71-84, 1988.

RICHARDSON, J.S. - The anatomy and taxonomy of protein structure. **Adv. Protein Chem., 34:** 167-339, 1981.

RICHARDSON, J.S.; GETZOFF, E.; RICHARDSON, D.C - Tht β buldge: a common small unit of nonrepetitive protein structure. **Proc. Natl. Acad. Sci. U.S.A., 75:** 2574-2578, 1978.

RICHARDSON, J.S. & RICHARDSON D.C - Amino acid preferences for specific locations at the end of α-helices. **Science, 240:** 1648-1652, 1988.

RICHARDSON, J.S. & RICHARDSON D.C - Principles and patterns of protein conformation. In: FASMAN, G.F - **Prediction of protein structure and the principles of protein conformation.** New York, Plenum, 1989. p. 1-98.

ROSE, G.D.; GIERASCH, L.M.; SMITH, J.A. - Turns in peptides and proteins. **Adv. Protein Chem., 37:** 1-109, 1985.

ROST, B. & SANDER, C. - Jury returns on structure prediction. **Nature, 361:** 540, 1992.

ROST, B. & SANDER, C. - Prediction of protein secondary structure at better than 70% accuracy. **J. Mol. Biol., 232:** 584-599, 1993.

SALZBERG, S. & COST, S. - Predicting protein secondary structure with a nearest-neighbor algorithm. **J. Mol. Biol., 227:** 371-347, 1992.

SASAGAWA, F. & TAJIMA, K. - Prediction of protein secondary structures by a neural network. **Comput. Appl. Biosc., 9:** 147-52, 1993.

SCHNEIDER, R. & SANDER, C. - Database of homology-derived structures and the structurally meaning of sequence alignment. **Proteins, 9:** 56-68, 1991.

SCHULZ, G.E.; BARRY, C.D.; FRIEDMAN. J.; CHOU, P.Y.; FASMAN, G.D.; FINKELSTEIN, A.V.; LIM, V.I.; PTITSYN, O.B.; KABAT, E.A.; WU, T.T.; LEVITT, M.; ROBSON, B.; NAGANO, K. - A critical evaluation of methods for prediction of protein secondary structures. **Nature, 250:** 140-142, 1974.

SEGREST, J.P; DE LOOF, H.; DOHLMAN, J.G.; BROUILLETTE, C.G.; ANANTHARAMAIAH, G.M. - Amphipathic helix motif: classes and properties. **Proteins, 8:** 103-117, 1990.

SHACKHNOVICH, E.I. & GUTIN, A.M. - Influence of point mutations on protein structure: probability of a neutral mutation. **J. Theor. Biol., 149:** 537-546, 1991.

SHANNON, C.E. & WEAVER, W. - **The mathematical theory of communication.** University of Illinois Press, Urbana, 1949. 343p.

SHERIDAN, R.P.; DIXON, J.S.; VENKATARAGHAVAN, R.; KUNTZ, I.D.; SCOTT; K.P. - Amino acid composition and hydrophobicity patterns of protein domains correlate with theis structure. **Biopolymers, 24:** 1995-2023, 1985.

SIMMONOFF, J. - A penalty function approach to smoothing large sparse contingency tables. **J. Am. Stat. Assoc., 68:** 478-482, 1983.

SINGH, J. & THORNTON, J.M. - **Atlas of Protein Side-chain Interactions.** v.1. IRL Press at Oxford University Press, Oxford, 1992. 123p.

SPOLAR, R.S.; JEUNG-HOI, H.; RECORD Jr., J.T. - Hydrophobic effect in protein folding and other noncovalent processes involving proteins. **Proc. Natl. Acad. Sci. U.S.A., 86:** 8382-8385, 1989.

STERNBERG, M.J.- Secondary structure prediction. **Curr. Opin. Str. Biol., 2:** 237-241, 1992.

STERNBERG, M.J.; LEWIS, R.A.; KING, R.D.; MUGGLETON, S. - Modelling the structure and function of enzymes by machine learning. **Faraday Dicuss., 93:** 269-80, 1992.

STOLORZ, P.; LAPEDES, A.; XIA, Y. - Predicting protein secondary structure using neural net and statistical methods. **J. Mol. Biol., 225:** 363-377, 1992.

STRYER, L. - **Biochemistry**. 2.ed. San Francisco, W.H. Freeman and Company, 1981. 949p.

TITTERINGTON, D.M. & BOWMAN, A.W. - A comparative study of smoothing procedures for ordered categorical data. **J. Stat. Comput. Simul., 21:** 291-312, 1985.

VISVANADHAN, V.N.; DENCKLA, B.; WEINSTEIN, J.N. - New joint prediction algorithm (Q7-JASEP) improves the prediction of protein secondary structure. **Biochemistry, 30:** 11164-11172, 1991.

WATSON, J.D. - **Recombinant DNA.** 2.ed. New York, Sci. Am. Books/Freeman, 1992. 557p.

WOLFENDEN, R.; ANDERSON, L.; CULLIS, P.M.; SOUTHGATE, C.C.F. - Affinities of amino acid side chains for solvent water. **Biochemistry, 20:** 849-855, 1981.

WÜTRICH, K. - Protein structure determination in solution by nuclear magnetic resonance spectroscopy. **Science, 243:** 45-50, 1989.

ZABIN, H.B.; HORVATH, M.P.; TERWILLIGER, T.C. - Approaches to predicting effects of single amino acid substitutions on the function of a protein. **Biochemestry, 30:** 6230-6240, 1991.

ZHANG, X.; MESIROV, J.P.; WALTZ, D.L. - Hybrid system for protein secondary structure prediction. **J. Mol. Biol., 225:** 1049-1063, 1992.

Apêndices

Apêndice 1

Arquivo PDB

Listamos aqui o conteúdo do arquivo 1FDX, com dados estruturais sobre a Ferredoxina (n.º 29, Tabela1), obtido do *Protein Data Bank* (BERSTEIN *et al.*, 1977). As linhas tem o formato dos antigos cartões perfurados IBM. As primeiras sete letras representam o rótulo, que define o conteúdo. As cinco últimas, o número da linha.

O significados dos rótulos são os seguintes:

ATOM:	Linha com as coordenadas do átomo.
AUTHOR:	Autores que resolveram a estrutura.
COMPND:	Nome da Proteína.
CONECT:	Ligações covalentes entre diferentes átomos.
CRYST:	Parâmetros cristalográficos.
END:	Final do arquivo.
FORMUL:	Quantidade e número de coordenação do heteroátomo.
HEADER:	Primeira linha do arquivo, com o título.
HET:	Informações sobre heteroátomo.
HETATM	Coordenadas de heteroátomo.
JRNL:	Principal referência comunicando a solução da estrutura.
MASTER:	Parâmetros cristalográficos.
ORIGX:	Parâmetros cristalográficos.
REMARK:	Comentários e adições ao arquivo.
REVDAT:	Informações acrescentadas em revisões.
SCALE:	Parâmetros cristalográficos.
SEQRES:	Sequência de aminoácidos.
SOURCE:	Organismo de onde foi extraída a proteína.
TER	Término das coordenadas.

Nas linhas rotuladas com REMARK, se encontram referências bibliográficas, assim como a resolução com que foi determinada a estrutura e se houve refinamento energético (linhas G1 a G4).

Nas linhas rotuladas com ATOM temos o número do átomo, sua identificação (e.g. CA=Carbono alfa), tipo de aminoácido a que pertence, número do aminoácido na sequência, coordenadas espaciais **x**,**y** e **z** do átomo, e por último, dois, coeficientes cristalográficos.

```
HEADER    ELECTRON TRANSPORT                      01-AUG-76   1FDX      1FDX   3
COMPND    FERREDOXIN                                                    1FDX   4
SOURCE    (PEPTOCOCCUS AEROGENES)                                       1FDXF  1
AUTHOR    E.T.ADMAN,L.C.SIEKER,L.H.JENSEN                               1FDXD  1
REVDAT   9   30-SEP-83 1FDXH   1       REVDAT                           1FDXH  1
REVDAT   8   31-DEC-80 1FDXG   1       REMARK                           1FDXH  2
REVDAT   7   05-MAR-80 1FDXF   1       SOURCE                           1FDXH  3
REVDAT   6   23-MAY-78 1FDXE   3       HET    FORMUL HETATM             1FDXH  4
REVDAT   5   01-NOV-77 1FDXD   1       AUTHOR JRNL   REMARK FORMUL      1FDXH  5
REVDAT   4   09-SEP-77 1FDXC   1       REMARK                           1FDXH  6
REVDAT   3   13-JUN-77 1FDXB   1       HET                              1FDXH  7
REVDAT   2   03-JAN-77 1FDXA   3       ATOM                             1FDXH  8
REVDAT   1   04-AUG-76 1FDX    0                                        1FDXH  9
JRNL        AUTH   E.T.ADMAN,L.C.SIEKER,L.H.JENSEN                      1FDXD  2
JRNL        TITL   STRUCTURE OF PEPTOCOCCUS AEROGENES FERREDOXIN,       1FDXD  3
JRNL        TITL 2 REFINEMENT AT 2 ANGSTROMS RESOLUTION                 1FDXD  4
JRNL        REF    J.BIOL.CHEM.                 V. 251  3801 1976      1FDXD  5
JRNL        REFN   ASTM JBCHA3  US ISSN 0021-9258                  071  1FDXD  6
REMARK   1                                                              1FDXD  7
REMARK   1 REFERENCE 1                                                  1FDXD  8
REMARK   1  AUTH   E.T.ADMAN,L.C.SIEKER,L.H.JENSEN                      1FDXD  9
REMARK   1  TITL   THE STRUCTURE OF A BACTERIAL FERREDOXIN              1FDXD 10
REMARK   1  REF    J.BIOL.CHEM.                 V. 248  3987 1973      1FDXD 11
REMARK   1  REFN   ASTM JBCHA3  US ISSN 0021-9258                  071  1FDXD 12
REMARK   1 REFERENCE 2                                                  1FDXD 13
REMARK   1  EDIT   R.J.FELDMANN                                         1FDXD 14
REMARK   1  REF    ATLAS OF MACROMOLECULAR               148 1976      1FDXD 15
REMARK   1  REF  2 STRUCTURE ON MICROFICHE                              1FDXD 16
REMARK   1  PUBL   TRACOR JITCO INC.,ROCKVILLE,MD.                      1FDXD 17
REMARK   1  REFN                 ISBN 0-917934-01-6             434  1FDXD 18
REMARK   1 REFERENCE 3                                                  1FDXD 19
REMARK   1  EDIT   M.O.DAYHOFF                                          1FDXD 20
REMARK   1  REF    ATLAS OF PROTEIN SEQUENCE     V.   5    62 1976      1FDXD 21
REMARK   1  REF  2 AND STRUCTURE,SUPPLEMENT 2                           1FDXD 22
REMARK   1  PUBL   NATIONAL BIOMEDICAL RESEARCH FOUNDATION              1FDXD 23
REMARK   1  PUBL 2 SILVER SPRING,MD.                                    1FDXD 24
REMARK   1  REFN                 ISBN 0-912466-05-7             435  1FDXD 25
REMARK   2                                                              1FDX  15
REMARK   2 RESOLUTION. 2.0 ANGSTROMS.                                   1FDXG  1
REMARK   3                                                              1FDX  17
REMARK   3 REFINEMENT. BY DIFFERENCE FOURIER METHOD WITH CONSTRAINTS.   1FDXG  2
REMARK   3  THE COORDINATES HERE ARE IDEALIZED AFTER STEP IV.3, FIG. 1  1FDXG  3
REMARK   3  OF JRNL CITATION ABOVE.                                     1FDXG  4
REMARK   4                                                              1FDX  21
REMARK   4 THE IRON AND INORGANIC SULFUR ATOMS OF METAL CLUSTERS I AND  1FDX  22
REMARK   4 II (NOTATION OF REFERENCE 1) ARE GROUPED AS PSEUDO-RESIDUES  1FDX  23
REMARK   4 CL1 AND CL2 RESPECTIVELY.                                    1FDX  24
REMARK   5                                                              1FDX  25
REMARK   5 RESIDUE 23 IS INCLUDED HERE AS ILE ALTHOUGH REPORTED TO BE   1FDX  26
REMARK   5 GLN IN THE CHEMICAL SEQUENCE. SEE JRNL CITATION.             1FDX  27
REMARK   6                                                              1FDXA  1
REMARK   6 CORRECTION. THE SG ATOMS OF CYSTEINES 11, 14, 18, 35, 38,    1FDXA  2
REMARK   6  41, 45 HAD INCORRECT Y AND Z COORDINATES. (WHEN THE VALUE   1FDXA  3
REMARK   6  WAS GREATER THAN 9.999 THE HIGH-ORDER DIGIT WAS LOST.)      1FDXA  4
REMARK   6  03-JAN-77.                                                  1FDXA  5
REMARK   7                                                              1FDXB  1
REMARK   7 CORRECTION. MOVE COMMENT ON HET RECORDS TO PROPER COLUMNS.   1FDXB  2
REMARK   7  13-JUN-77.                                                  1FDXB  3
```

```
REMARK   8                                                              1FDXC  1
REMARK   8 CORRECTION. FIX MASTER RECORD TO SHOW PROPER NUMBER OF       1FDXC  2
REMARK   8  REMARKS. 09-SEP-77.                                        1FDXC  3
REMARK   9                                                              1FDXD 26
REMARK   9 CORRECTION. REFORMAT HEADER INFORMATION TO MEET NEW         1FDXD 27
REMARK   9  SPECIFICATIONS.                                            1FDXD 28
REMARK   9  ADD FORMUL RECORDS.                                        1FDXD 29
REMARK   9  01-NOV-77.                                                 1FDXD 30
REMARK  10                                                             1FDXE  1
REMARK  10 CORRECTION. STANDARDIZE NAMING OF HETATMS. 23-MAY-78.       1FDXE  2
REMARK  11                                                             1FDXF  2
REMARK  11 CORRECTION. STANDARDIZE FORMAT OF SOURCE RECORD.            1FDXF  3
REMARK  11  05-MAR-80.                                                 1FDXF  4
REMARK  12                                                             1FDXG  5
REMARK  12 CORRECTION. STANDARDIZE FORMAT OF REMARKS 2 AND 3.          1FDXG  6
REMARK  12  31-DEC-80.                                                 1FDXG  7
REMARK  13                                                             1FDXH 10
REMARK  13 CORRECTION. INSERT REVDAT RECORDS. 30-SEP-83.               1FDXH 11
SEQRES   1     54  ALA TYR VAL ILE ASN ASP SER CYS ILE ALA CYS GLY ALA  1FDX  28
SEQRES   2     54  CYS LYS PRO GLU CYS PRO VAL ASN ILE ILE GLN GLY SER  1FDX  29
SEQRES   3     54  ILE TYR ALA ILE ASP ALA ASP SER CYS ILE ASP CYS GLY  1FDX  30
SEQRES   4     54  SER CYS ALA SER VAL CYS PRO VAL GLY ALA PRO ASN PRO  1FDX  31
SEQRES   5     54  GLU ASP                                              1FDX  32
HET    FES      1       8     FE/S (INORGANIC) CLUSTER NUMBER 1         1FDXE  3
HET    FES      2       8     FE/S (INORGANIC) CLUSTER NUMBER 2         1FDXE  4
FORMUL   2  FES    2(FE4 S4)                                           1FDXE  5
CRYST1  30.520   37.750   39.370  90.00  90.00  90.00 P 21 21 21    4  1FDX  35
ORIGX1      1.000000  0.000000  0.000000        0.000000               1FDX  36
ORIGX2      0.000000  1.000000  0.000000        0.000000               1FDX  37
ORIGX3      0.000000  0.000000  1.000000        0.000000               1FDX  38
SCALE1       .032765  0.000000  0.000000        0.000000               1FDX  39
SCALE2      0.000000   .026490  0.000000        0.000000               1FDX  40
SCALE3      0.000000  0.000000   .025400        0.000000               1FDX  41
ATOM      1  N   ALA     1      17.186  -1.593  15.748  1.00  0.00      1FDX  42
ATOM      2  CA  ALA     1      18.608  -1.306  15.701  1.00  0.00      1FDX  43
ATOM      3  C   ALA     1      18.813    .030  15.016  1.00  0.00      1FDX  44
ATOM      4  O   ALA     1      17.851    .242  14.264  1.00  0.00      1FDX  45
ATOM      5  CB  ALA     1      19.490  -2.167  14.803  1.00  0.00      1FDX  46
ATOM      6  N   TYR     2      19.954    .623  15.287  1.00  0.00      1FDX  47
ATOM      7  CA  TYR     2      20.387   1.903  14.595  1.00  0.00      1FDX  48
ATOM      8  C   TYR     2      21.181   1.503  13.295  1.00  0.00      1FDX  49
ATOM      9  O   TYR     2      21.544    .295  13.268  1.00  0.00      1FDX  50
ATOM     10  CB  TYR     2      21.392   2.526  15.559  1.00  0.00      1FDX  51
ATOM     11  CG  TYR     2      20.528   3.360  16.213  1.00  0.00      1FDX  52
ATOM     12  CD1 TYR     2      19.560   2.828  17.028  1.00  0.00      1FDX  53
ATOM     13  CD2 TYR     2      20.748   4.730  16.055  1.00  0.00      1FDX  54
ATOM     14  CE1 TYR     2      18.794   3.813  17.638  1.00  0.00      1FDX  55
ATOM     15  CE2 TYR     2      19.914   5.538  16.614  1.00  0.00      1FDX  56
ATOM     16  CZ  TYR     2      18.987   5.119  17.382  1.00  0.00      1FDX  57
ATOM     17  OH  TYR     2      18.220   6.146  17.902  1.00  0.00      1FDX  58
ATOM     18  N   VAL     3      21.300   2.416  12.413  1.00  0.00      1FDX  59
ATOM     19  CA  VAL     3      21.935   2.186  11.095  1.00  0.00      1FDX  60
ATOM     20  C   VAL     3      22.698   3.530  10.799  1.00  0.00      1FDX  61
ATOM     21  O   VAL     3      22.075   4.538  10.945  1.00  0.00      1FDX  62
ATOM     22  CB  VAL     3      20.964   1.789   9.972  1.00  0.00      1FDX  63
ATOM     23  CG1 VAL     3      19.951   2.846   9.496  1.00  0.00      1FDX  64
ATOM     24  CG2 VAL     3      21.721   1.352   8.721  1.00  0.00      1FDX  65
ATOM     25  N   ILE     4      23.958   3.315  10.492  1.00  0.00      1FDX  66
ATOM     26  CA  ILE     4      24.816   4.428  10.287  1.00  0.00      1FDX  67
ATOM     27  C   ILE     4      24.694   4.677   8.870  1.00  0.00      1FDX  68
ATOM     28  O   ILE     4      24.923   3.752   8.043  1.00  0.00      1FDX  69
ATOM     29  CB  ILE     4      26.256   3.979  10.571  1.00  0.00      1FDX  70
ATOM     30  CG1 ILE     4      26.754   4.666  11.850  1.00  0.00      1FDX  71
ATOM     31  CG2 ILE     4      27.325   4.364   9.571  1.00  0.00      1FDX  72
ATOM     32  CD1 ILE     4      27.584   3.813  12.705  1.00  0.00      1FDX  73
ATOM     33  N   ASN     5      24.358   5.866   8.433  1.00  0.00      1FDX  74

                                    ...

ATOM    364  OE2 GLU    53      14.351  -1.638  10.988  1.00  0.00      1FDX 405
ATOM    365  N   ASP    54      17.342   -.313  10.866  1.00  0.00      1FDX 406
ATOM    366  CA  ASP    54      17.964   -.914   9.673  1.00  0.00      1FDX 407
ATOM    367  C   ASP    54      16.868   -.529   8.638  1.00  0.00      1FDX 408
ATOM    368  O   ASP    54      15.733    .034   8.724  1.00  0.00      1FDX 409
ATOM    369  CB  ASP    54      18.318  -2.325   9.736  1.00  0.00      1FDX 410
ATOM    370  CG  ASP    54      18.217  -2.646  11.185  1.00  0.00      1FDX 411
ATOM    371  OD1 ASP    54      19.276  -2.963  11.898  1.00  0.00      1FDX 412
```

```
ATOM    372  OD2 ASP    54      17.134  -2.699  11.929  1.00  0.00      1FDX 413
ATOM    373  OXT ASP    54      16.749   -.630   7.276  1.00  0.00      1FDX 414
TER     374      ASP    54                                               1FDX 415
HETATM  375 FE1  FES     1      30.535   7.886  11.642  1.00  0.00      1FDXE  6
HETATM  376 FE2  FES     1      31.762   9.524  13.595  1.00  0.00      1FDXE  7
HETATM  377 FE3  FES     1      30.328   7.308  14.228  1.00  0.00      1FDXE  8
HETATM  378 FE4  FES     1      29.104   9.577  13.205  1.00  0.00      1FDXE  9
HETATM  379  S1  FES     1      32.144   7.244  13.138  1.00  0.00      1FDXE 10
HETATM  380  S2  FES     1      30.715  10.272  11.819  1.00  0.00      1FDXE 11
HETATM  381  S3  FES     1      28.710   7.561  12.803  1.00  0.00      1FDXE 12
HETATM  382  S4  FES     1      30.270   9.336  15.260  1.00  0.00      1FDXE 13
HETATM  383 FE1  FES     2      25.417    .955  20.244  1.00  0.00      1FDXE 14
HETATM  384 FE2  FES     2      23.030    .627  19.142  1.00  0.00      1FDXE 15
HETATM  385 FE3  FES     2      23.394   2.329  21.268  1.00  0.00      1FDXE 16
HETATM  386 FE4  FES     2      24.129   2.929  18.791  1.00  0.00      1FDXE 17
HETATM  387  S1  FES     2      23.821    .049  20.972  1.00  0.00      1FDXE 18
HETATM  388  S2  FES     2      24.999   1.167  18.020  1.00  0.00      1FDXE 19
HETATM  389  S3  FES     2      25.091   3.307  20.791  1.00  0.00      1FDXE 20
HETATM  390  S4  FES     2      22.002   2.677  19.520  1.00  0.00      1FDXE 21
CONECT   60   59  375                                                    1FDX 432
CONECT   79   78  376                                                    1FDX 433
CONECT   94   93  377                                                    1FDX 434
CONECT  125  124  383                                                    1FDX 435
CONECT  248  247  384                                                    1FDX 436
CONECT  270  269  385                                                    1FDX 437
CONECT  286  285  386                                                    1FDX 438
CONECT  310  309  378                                                    1FDX 439
CONECT  375   60  379  380  381                                          1FDX 440
CONECT  376   79  379  380  382                                          1FDX 441
CONECT  377   94  379  381  382                                          1FDX 442
CONECT  378  310  380  381  382                                          1FDX 443
CONECT  379  375  376  377                                               1FDX 444
CONECT  380  375  376  378                                               1FDX 445
CONECT  381  375  377  378                                               1FDX 446
CONECT  382  376  377  378                                               1FDX 447
CONECT  383  125  387  388  389                                          1FDX 448
CONECT  384  248  387  388  390                                          1FDX 449
CONECT  385  270  387  389  390                                          1FDX 450
CONECT  386  286  388  389  390                                          1FDX 451
CONECT  387  383  384  385                                               1FDX 452
CONECT  388  383  384  386                                               1FDX 453
CONECT  389  383  385  386                                               1FDX 454
CONECT  390  384  385  386                                               1FDX 455
MASTER       58    0    2    0    0    0    0    6  389    1   24    5  1FDXH 12
END                                                                      1FDX 457
```

Apêndice 2

Programa Gerador do Conjunto de Dados

Listamos a seguir GENPTR.F, programa escrito em FORTRAN para criar arquivos como os das Figs. 12 e 13, a partir de arquivos PDB e DSSP.

```
C23456789012345678901234567890123456789012345678901234567890123456789012
C     GENPTR.F
C     Creates a .sif file (.siq + fourier-amphipathicity index)
C     Version 3.1
C     Modified to deal with pseudo-multiple conformers. i.e.
C     residues with duplicated numbers due to numbering scheme
C     based on a different protein
C     Modified to be able to extract chains and residues (NpdbC####-####)
C     Modified to deal with 1gp1 seleninic acid
C     Uses a .fil list of proteins, chains, and segments
C     Raul Neder Porrelli
C     Bethesda, May 22, 1992
      CHARACTER*14 F
      CHARACTER*20 G,H
      INTEGER P
      WRITE (*,*) 'File name:'
      READ (*,'(A20)') G
      OPEN (7, FILE = G)
      P=INDEX(G,'.')
      H=G(:P)//'ptr'
      OPEN (2, FILE = H)
    5 READ (7,'(A14)') F
      WRITE (*,'(A14)') F
      IF (F(:3).NE.'END') THEN
                          CALL GF (F)
                          GOTO 5
      END IF
      CLOSE (7)
      WRITE (2,'(A3)') 'END'
```

```
      CLOSE (Z)

      END


      SUBROUTINE GF(F)

      CHARACTER*4 RN,FI,SI,OM,X,Y,Z,N,LN
      CHARACTER*3 RA,TRI,KRI
      CHARACTER*20 G
      CHARACTER*28 L
      CHARACTER*33 AS
      CHARACTER*36 AU
      CHARACTER*69 AT
      CHARACTER*23 O
      CHARACTER*1 RO,LO,S,K,AA,ST,CT,GAR,GR,CH,CO,SS,CC,KH,LC,LD
      CHARACTER*131 D
      CHARACTER*17 tist
      CHARACTER*17 SEQ,SQ,LS
      CHARACTER*14 F

      INTEGER LI,NR,PHI,PSI,OMG,KC,CB,NC,ND,K4,K5,BG,ED,K6

      DIMENSION X(2000),Y(2000),Z(2000),N(2000),K(2000),AA(2000),GAR(2000)
      DIMENSION SQ(2000),CO(2000),CB(20),SS(2000),CC(2000),PHI(2000)
      DIMENSION PSI(2000),OMG(2000)

      K4=0
      K5=0
      K6=0
      IF (F(5:5).NE.' ') THEN
                              KH=F(5:5)
                              K4=1
      END IF
      IF (F(10:10).EQ.'-') THEN
                                WRITE(X(1),'(A4)') F(6:9)
                                WRITE(Y(1),'(A4)') F(11:14)
                                READ (X(1),'(I4)') BG
                                READ (Y(1),'(I4)') ED
                                K6=1
      END IF
      G='opo/'//F(:4)//'.ppo'
      OPEN (1, FILE = G)

      G='dssp/'//F(:4)//'dssp.dat'
      OPEN (3, FILE = G)

      AS='ALAASXCYSASPGLUPHEGLYHISILELYSLEU'
      AU='METASNPROGLNARGSERTHRVALTRPUNKTYRGLX'
      AT=AS//AU
      O='ABCDEFGHIKLMNPQRSTVWUYZ'
      SEQ='UUUUUUUU'

      DO 10, I=1, 3
             READ (1, '(A28)') L
10    CONTINUE
      tist=' '
      DO 11 WHILE (tist .NE. '  #  RESIDUE AA S')
                                READ (3,'(A131)') D
                                tist = D(:17)
11    CONTINUE
      K1=0
      GG=0
      KC=0
      NC=0
      I=1
      DO 350, W=1, 10000
              READ (1,1000) RA,CH,RN,FI,SI,OM
              IF (I.EQ.1) LC=CH
              IF ((K4.NE.0).AND.(K5.EQ.0).AND.(CH.NE.KH)) GOTO 350
              IF ((RA.EQ.'END').OR.((K5.EQ.1).AND.(CH.NE.KH))) THEN
                                    LI=I-1
                                    GOTO 700
              END IF
              IF ((K4.EQ.1).AND.(CH.EQ.KH)) K5=1
              IF (KC.NE.0) KC=KC+1
              IF (CH.NE.LC) THEN
```

```
                              SEQ='UUUUUUUU'
                              KC=1
                              NC=NC+1
                              CB(NC)=I-1
                              ND=1
            END IF
            LC=CH
            DO 15, J=1, 23
                   IF (RA.EQ.AT((J-1)*3+1:J*3)) THEN
                                                RO=O(J:J)
                                                GOTO 17
                   END IF
15          CONTINUE
            RO='X'
17          IF ((RN .EQ. '   0').OR.
     +          ((RN.EQ.LN).AND.(FI.EQ.' 999'))) THEN
                              GOTO 310
            END IF
            IF (KC.EQ.0) THEN
                   IF (I.LE.9) THEN
                              SEQ(8+I:8+I)=RO
                   ELSE
                              SEQ=SEQ(2:17)//RO
                   END IF
            ELSE
                   IF (KC.LE.9) THEN
                              SEQ(8+KC:8+KC)=RO
                   ELSE
                              SEQ=SEQ(2:17)//RO
                   END IF
            END IF
            IF ((I.GT.8).AND.(KC.EQ.0)) SQ(I-8)=SEQ
            IF ((KC.NE.0).AND.(I.GT.CB(NC)+8)) SQ(I-8)=SEQ
            IF (RO.EQ. 'X') THEN
               IF (I.EQ.1) THEN
                              GOTO 310
                        ELSE
                              S= '*'
                              GOTO 80
               END IF
            END IF
            SK=0
            LD = ' '
70          IF (SK .NE. 1) READ (3, '(A131)') D
            DD=VIXEN(D(7:10))+6
            NN=VIXEN(RN)
            LD=D(:5)
            IF ((D(DD:10).GT.RN(NN:)).AND.(RN(NN:).NE.'1').AND
     +      .((K4.EQ.0).OR.((K4.EQ.1).AND.(D(12:12).EQ.KH))))  THEN
                                    S = '*'
                                    SK=1
                                    GOTO 80
            END IF
            IF ((D(DD:10).NE.RN(NN:)).OR.((K4.EQ.1).AND.
     +      (D(12:12).NE.KH)))   GOTO 70
            S= D(17:17)
            SK=0
80          WRITE (N(I),'(A4)') RN
            AA(I) = RO
            K(I) = S
            CO(I)=CH
            WRITE (X(I),'(A4)') FI
            WRITE (Y(I),'(A4)') SI
            WRITE (Z(I),'(A4)') OM
   IF ((S.NE.'G').AND.(S.NE.'H').AND.(S.NE.'E').AND.(S.NE.'*')) THEN
                   S='C'
                   GG=0
                   K1=0
                   GOTO 300
   END IF
   IF (S.EQ.'G') THEN
                   IF (K1.EQ.1) THEN
                              S='H'
                              GOTO 300
                   END IF
                   IF (GG.EQ.3) GOTO 400
                   IF (GG.LT.3) THEN
```

```
                                GG=GG+1
                                S="C"
                                K1=0
                    END IF
                    GOTO 300
    END IF
    IF (S.EQ.'H') THEN
                    IF (GG.GT.0) GOTO 500
220                 K1=1
                    GG=0
    END IF
300 GAR(I)=S
305 I=I+1
310 LN=RN
    LO=RO
350 CONTINUE
    GOTO 700
400 DO 410, J=I-GG, I-1
       GAR(J)='H'
410 CONTINUE
    S='H'
    K1=1
    GG=0
    GOTO 300
500 DO 510, J=I-GG, I-1
       GAR(J)='H'
510 CONTINUE
    GOTO 220

700 CLOSE (1)
    CLOSE (3)
    K5=0
    DO 800,I=1, LI
       READ (N(I),'(I4)') NR
       RO = AA(I)
       S = K(I)
       CH=CO(I)
       READ (X(I),'(I4)') PHI(I)
       READ (Y(I),'(I4)') PSI(I)
       READ (Z(I),'(I4)') OMG(I)
       GR=GAR(I)
       SEQ=SQ(I)
       IF ((K4.NE.0).AND.(K5.EQ.0).AND.(CH.NE.KH)) GOTO 800
       IF((K5.EQ.1).AND.(CH.NE.KH)) GOTO 810
       IF ((K4.EQ.1).AND.(CH.EQ.KH)) K5=1
       IF ((PHI(I).EQ.999).OR.(PSI(I).EQ.999)) THEN
                                               ST='*'
                                               CT='*'
                                               GOTO 790
       END IF
       IF ((PHI(I).GE.0).AND.(PHI(I).LE.160)) THEN
                IF ((PSI(I).GT.-90).AND.(PSI(I).LT.110)) THEN
                        ST='a'
                ELSE
                        ST='e'
                END IF
       ELSE
                IF ((PSI(I).GT.-120).AND.(PSI(I).LT.40)) THEN
                        ST='A'
                ELSE
                        ST='E'
                END IF
       END IF
       IF ((OMG(I).GT.-90).AND.(OMG(I).LT.90)) THEN
                CT='c'
       ELSE
                CT='t'
       END IF
790    IF (I.GT.LI-8) SEQ=LS(2:17)//'U'
       IF (KC.NE.0) THEN
          IF (ND.LE.NC) THEN
             IF ((I.GT.CB(ND)-9).AND.(I.LT.CB(ND))) SEQ=LS(2:17)//'U'
          END IF
       END IF
       IF (I.EQ.CB(ND)) ND=ND+1
       LS=SEQ
       SQ(I)=SEQ
```

```
      SS(I)=ST
      CC(I)=CT
800 CONTINUE
810 IF (KC.NE.0) ND=1
    K5=0
    DO 900, I=1,LI
      READ (N(I),'(I4)') NR
      RO = AA(I)
      S = K(I)
      CH=CO(I)
      GR=GAR(I)
      SEQ=SQ(I)
      ST=SS(I)
      CT=CC(I)
      IF ((K4.NE.0).AND.(K5.EQ.0).AND.(CH.NE.KH)) GOTO 900
      IF((K5.EQ.1).AND.(CH.NE.KH)) GOTO 910
      IF ((K4.EQ.1).AND.(CH.EQ.KH)) K5=1
      IF ((K6.NE.0).AND.(NR.LT.BG)) GOTO 900
      IF ((K6.EQ.1).AND.(NR.GT.ED)) GOTO 910
      IF (I.EQ.CB(ND)+1) ND=ND+1
897 WRITE (2,1300) F(:4),CH,NR,I,PHI(I),PSI(I),OMG(I),ST,CT,SEQ,S,GR
900 CONTINUE
910 K6=K6
1000 FORMAT (1X,A3,1X,A1,1X,A4,3X,A4,1X,A4,1X,A4)
1100 FORMAT (' ',A4,' ',A1,' ',A1,3(' ',A4))
1300 FORMAT (A4,' ',A1,5(' ',I4),2(' ',A1),' ',A17,2(' ',A1))
    END

    FUNCTION VIXEN(char)
    CHARACTER*4 char
    INTEGER I
    DO 300, I=1, LEN(char)
          IF (ICHAR(char(I:I)) .NE. 32) GOTO 310
300 CONTINUE
    VIXEN = LEN(char)
    RETURN
310 VIXEN = I

    END
```

Apêndice 3

Rotina MATLAB para Ajuste do Modelo com Penalização

Listamos a seguir a rotina **jelicle.m**, escrita em MATLAB. Ela executa o ajuste do modelo quadrático, usando um fator λ=1000 para a penalização da máxima verossimilhança. As porções que foram posteriormente implementadas em FORTRAN (programa BOSS), para agilizar o processo de validação cruzada, estão em negrito.

```
disp('beginning Jelicle')
%
% Raul Porrelli & Peter Munson
% Bethesda, MD, February 12, 1992.

flops(0);

load garnier.ptr
lgarnier=length(garnier)

format compact
alpha=(garnier(:,31)==1);
beta=(garnier(:,31)==2);

% Prepare T
ind3=[1 21 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 21 18 19 20 21]';
clear T

%%%%%%%%%%%%%%%%%%%%
kmin=-8;
kmax=8;
flattails=0
%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%
anyinteractions=1
kincr=0
%%%%%%%%%%%%%%%%%%%%
```

```
% if kincr=0, all pairwise interaction
% with every possible increment, else
% kincr defines increment
% in interaction pairs
% This creates a 17x17 matrix for a quadratic form A
% sensitive to period 3.6 signals
%%%%%%%%%%%%%%%%%%%%
l1=sin((0:16)*2*pi/3.6)
l2=cos((0:16)*2*pi/3.6)
ll=[l1;l2];
A1=ll'*ll;
A1=A1(1:kmax-kmin+1,1:kmax-kmin+1);
l1=sin((0:16)*2*pi/2.0)
l2=cos((0:16)*2*pi/2.0)
ll=[l1;l2];
A2=ll'*ll;
A2=A2(1:kmax-kmin+1,1:kmax-kmin+1);


T=zeros(lgarnier,kmax-kmin+1);
for k=kmin:kmax
T(:,k-kmin+1)=ind3(garnier(:,21+k));
end

dada=(T(:,:)==21);
'dada done'
duda=dada';
'duda done'
deda=sum(duda);
'deda done'
dind=deda';
'dind done'
didi=(dind~=0);

T(didi,:)=[];
clear garnier

f1=alpha;
f1(didi)=[];
f2=beta;
f2(didi)=[];
length(f1)

disp('The number of residues with alpha=1 is ')
disp(sum(f1))
disp('The number of residues with beta=1 is ')
disp(sum(f2))

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%  A program to calculate the MLE for a logistic linear
%  model for any number of independent vars
%  Optimized for Matlab 1-13-92
%  Modified to reduce number of params from 341 to 181 1-29-92
%
%  T is a table with
%  ncases by nvar
%  using consequtive integers 1, 2, 3, ..., nlevs(i)
%  for levels of variable(i)
%  f is a vector of length ncases, giving success (0 or 1)
%  X is the implicit "design" matrix
%  ncases by npar-nvar,
%  nparmain = 1+sum(levels in each main effect variable)
%  nvarmain = number of nominal  main effect variables
%  No provision is made for weights(frequencies) of each case
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

[ncases,nvarmain]=size(T)
nlevs=max(T)
nparmain=1+sum(nlevs)
cnlevs=tril(ones(nvarmain,nvarmain),-1)*nlevs';
Toffset=1+cnlevs';
nvar=nvarmain;
npar=nparmain;
ww1=[];%vector of weights for interactions
ww2=[];

if anyinteractions
```

```
%  kincr is the increment between interaction vars
%  Define k1, k2 to be two variables involved in interaction, if any
%  modify T, Theta, assuming columns k1,k2 are the interaction
%  Set up code table for recoding interactions as a means of
%  incorporating constraints
%  Works only if nlevs==20
   if any(nlevs~=20)
      disp('some variables do not have 20 levels, a problem')
      pause
   end
   codeT=zeros(nlevs(1));
   kk=0;
   for ii=1:20
       for jj=ii:20
           kk=kk+1;
           codeT(ii,jj)=kk;
           codeT(jj,ii)=kk;
       end
   end

%%%begin looping over all increments

   if kincr==0 %this signals use of all increments
      kincrmin=1;
      kincrmax=(kmax-kmin);
   else
      kincrmin=kincr;
      kincrmax=kincr;
   end

   for kincr1=kincrmin:kincrmax
       kincr1
       for k1=1:nvarmain-kincr1
           k2=k1+kincr1;

%          extend the T array

           Toffset=[Toffset Toffset(nvarmain)+nlevs(nvarmain)];
           T(:,nvar+1)=nlevs(k2)*(T(:,k1)-1)+T(:,k2);
           nvar=nvar+1;

           ww1=[ww1 A1(k1,k2)]; % Pick out appropriate weighting from
           ww2=[ww2 A2(k1,k2)]; % Matrix of quadratic form A
       end
   end %%% of kincr1 loop
   if kincr~=0
      ww1=ones(1,nvar-nvarmain);% use weight of 1.0 if only one kincr is used
      ww2=ww1;
   end

   www1=[ones(1,nvarmain) ww1]
   www2=[ones(1,nvarmain) ww2]
   nvar
   nl1=nlevs(k1);
   nl2=nlevs(k2);
   npar=nparmain+nl1*nl2

   Thetafull=[Thetafull; zeros(nl1*nl2,1)];
   Theta=[Theta; zeros((nl1-1)*(nl2-1),1)];
end
npairs=nvar-nvarmain

% initial estimate for Intercept only

pbar1=sum(f1)/ncases;
Mu01=log(pbar1/(1-pbar1))
Theta1=[Mu01; zeros(nparmain-nvarmain-1,1)];
XTheta1=Mu01*ones(ncases,1);
pbar2=sum(f2)/ncases;
Mu02=log(pbar2/(1-pbar2))
Theta2=[Mu02; zeros(nparmain-nvarmain-1,1)];
Theta=[Theta1 ; Theta2];
XTheta2=Mu02*ones(ncases,1);

lnL=-inf;

%%%  first time
```

```
XpwX1=zeros(npar,npar);
Xpfmp1=zeros(npar,1);
XpwX2=zeros(npar,npar);
Xpfmp2=zeros(npar,1);
Q12=zeros(npar,npar);

z1=exp(XTheta1);
z2=exp(XTheta2);
p1=z1./(1+z1+z2);
p2=z2./(1+z1+z2);
w11=p1.*(1-p1);
w22=p2.*(1-p2);
w12=-p1.*p2;

oldlnL=lnL;
lnL=f1'*XTheta1+f2'*XTheta2-sum(log(1+z1+z2))

T2aoffset=nl1*nl2*(0:npairs-1);
T51j=zeros(ncases,npairs);
T52j=zeros(ncases,npairs);
ndistinctpairs=zeros(ncases,1);\

disp('beginning pass thru data')

ttime=clock;

for j=1:ncases
    T2=[1 T(j,:)+Toffset];
    T2a=T(j,nvarmain+1:nvar);
    T3=zeros(npar,1);
    T3(T2)=ones(1+nvar,1);%sets only distinct elements = 1.
    T3a=T3(nparmain+1:npar);
    ndistinctpairs(j)=sum(T3a);
    T41=zeros(npar-nparmain,npairs);
    T42=zeros(npar-nparmain,npairs);
    T41(T2a+T2aoffset)=ww1;
    T42(T2a+T2aoffset)=ww2;
    T51j(j,1:ndistinctpairs(j))=sum(T41(T3a,:)');
    T52j(j,1:ndistinctpairs(j))=sum(T42(T3a,:)');
    T51=[ones(1,1+nvarmain) T51j(j,1:ndistinctpairs(j))];
    T52=[ones(1,1+nvarmain) T52j(j,1:ndistinctpairs(j))];
%   sum only over the nonzero columns of T4'
    XpwX1(T3,T3)=XpwX1(T3,T3)+w11(j)*T51'*T51;
    Xpfmp1(T3)=Xpfmp1(T3)+T51'*(f1(j)-p1(j));
    XpwX2(T3,T3)=XpwX2(T3,T3)+w22(j)*T52'*T52;
    Xpfmp2(T3)=Xpfmp2(T3)+T52'*(f2(j)-p2(j));
    Q12(T3,T3)=Q12(T3,T3)+w12(j)*T51'*T52;
end

disp(etime(clock,ttime))
disp('finished pass thru data')

% Now check for zero cols of XpwX1

emptycol=(sum(XpwX1~=0)==0);
if sum(emptycol)>0
  'there are empty cols in XpwX'
   sum(emptycol)
end

% constraints
% Projection matrix P which enforces constraints on
% sum(Theta)=zero for each main effect variable

P=eye(nparmain);
for i=1:nvarmain
   P(Toffset(i)+nlevs(i),Toffset(i)+(1:nlevs(i)))=-ones(1,nlevs(i));
end
P(:,nlevs+Toffset(1:nvarmain))=[];
ninter=0;

if anyinteractions
%  Define k1, k2 to be two variables involved in interaction, if any
   P1=eye(nl1*nl2);
   offset1=nl2*(0:(nl1-1));
   for i=1:nl1
```

```
        P1(offset1(i)+nl2,offset1(i)+(1:nl2))=-ones(1,nl2);
    end
    for i=1:(nl1-1)
%       set last row to negative of last block
        P1(offset1(nl1)+(1:nl2),offset1(i)+(1:nl2))=...
        -P1(offset1(nl1)+(1:nl2),offset1(nl1)+(1:nl2));
    end
%   delete unneeded columns
    dind=zeros(1,nl1*nl2);
    dlist=[nl2*(1:nl1-1) (nl1-1)*nl2+(1:nl2)];
    dind(dlist)=ones(length(dlist),1);
%   we don't expect any empty cols of X'wX in first nparmain elts
    emptycol(1:nparmain)=[];
    if length(dind)~=length(emptycol)
      'there is an error in emptycol'
       pause
    end
%   Now check if emptycol matches dind
    if any(emptycol+dind>1)
      'you are trying to delete a column of P1 twice!'
       pause
    end
    dind=dind|emptycol;
    P1(:,dind)=[];
    [nrowsP1,ninter]=size(P1)
%   finally, modify P
    P=[[P zeros(nparmain,ninter)]; [zeros(nrowsP1,nparmain-nvarmain) P1]];
end
QP=P;
clear P
nparams=nparmain-nvarmain+ninter

%  New code for penalization

D=diag([zeros(nparmain,1); ones(nrowsP1,1)]);
QPpDQP=QP'*D*QP;
QPpDQP=[QPpDQP zeros(QPpDQP); zeros(QPpDQP) QPpDQP];

Theta=[Theta1(1); zeros(nparams-1,1);Theta2(1); zeros(nparams-1,1)];

lambda=1000    %  smoothing parameter

for i=1:7
%   patch up for constraints
%   QP is projection matrix
    disp('begin linear algebra')
    ttime=clock;
    XpwX1=QP'*XpwX1*QP;
    Xpfmp1=QP'*Xpfmp1;
    XpwX2=QP'*XpwX2*QP;
    Xpfmp2=QP'*Xpfmp2;
    Q12=QP'*Q12*QP;
    ZpwZ=[XpwX1 Q12; Q12' XpwX2];
    Zpfmp=[Xpfmp1; Xpfmp2];

%   Here is the heart of the updating procedure

    Theta=Theta+(ZpwZ+lambda*QPpDQP)\(Zpfmp-lambda*QPpDQP*Theta);

    Thetafull1=QP*Theta(1:length(Theta)/2);
    Thetafull2=QP*Theta(length(Theta)/2+1:length(Theta));
    disp('end of linear algebra')
    disp(etime(ttime,clock))

%   Need to form XpwX, Xpfmp,XTheta
%   First for larger version of XpwX, Xpfmp
%   Then  project to smaller using P
%   which implies constraints on Thetafull
    XpwX1=zeros(npar,npar);
    Xpfmp1=zeros(npar,1);
    XpwX2=zeros(npar,npar);
    Xpfmp2=zeros(npar,1);
    Q12=zeros(npar,npar);
    XTheta1=zeros(ncases,1);
    XTheta1(:)=Thetafull1(1)*ones(ncases,1);
    XTheta2=zeros(ncases,1);
    XTheta2(:)=Thetafull2(1)*ones(ncases,1);
```

```
    for k=1:nvar
%%%     Note presence of www, a weight giving the fixed influence
%%%     of each parameter, esp. in the interactions.
        XTheta1(:)=XTheta1(:)+www1(k)*Thetafull1(T(:,k)+Toffset(k));
        XTheta2(:)=XTheta2(:)+www2(k)*Thetafull2(T(:,k)+Toffset(k));
    end

    z1=exp(XTheta1);
    z2=exp(XTheta2);
    p1=z1./(1+z1+z2);
    p2=z2./(1+z1+z2);

    w11=p1.*(1-p1);
    w22=p2.*(1-p2);
    w12=-p1.*p2;
    oldlnL=lnL;
    lnL=f1'*XTheta1+f2'*XTheta2-sum(log(1+z1+z2))
    if abs(lnL/oldlnL-1)<1e-5
       break                % Break out of loop if converged
    end

    T2aoffset=nl1*nl2*(0:npairs-1);
    disp('beginning pass thru data')
    ttime=clock;
    for j=1:ncases
        T2=[1 T(j,:)+Toffset];
        T2a=T(j,nvarmain+1:nvar);
        T3=zeros(npar,1);
        T3(T2)=ones(1+nvar,1);%  sets only distinct elements = 1.
%       we removed redundant code 4-12-92
        T51=[ones(1,1+nvarmain) T51j(j,1:ndistinctpairs(j))];
        T52=[ones(1,1+nvarmain) T52j(j,1:ndistinctpairs(j))];
%       sum only over the nonzero columns of T4'
        XpwX1(T3,T3)=XpwX1(T3,T3)+w11(j)*T51'*T51;
        Xpfmp1(T3)=Xpfmp1(T3)+T51'*(f1(j)-p1(j));
        XpwX2(T3,T3)=XpwX2(T3,T3)+w22(j)*T52'*T52;
        Xpfmp2(T3)=Xpfmp2(T3)+T52'*(f2(j)-p2(j));
        Q12(T3,T3)=Q12(T3,T3)+w12(j)*T51'*T52;
    end
    disp(etime(clock,ttime))
    disp('finished pass thru data')
end

Mufinal1=Thetafull1(1)
Mufinal2=Thetafull2(1)

%  Here is some graphical output to interpret
%  the quality of the model

percencorr=zeros(21,21);
for i=0:20
    c1=i/20;
    for j=0:20-i
        c2=j/20;
        d12=(c2*p1>c1*p2);
        d13=((1-c1-c2)*p1>c1*(1-p1-p2));
        d23=((1-c1-c2)*p2>c2.*(1-p1-p2));
        class1=d12&d13;
        class2=~d12&d23;
        class3=(1-class1-class2);
        percencorr(i+1,j+1)=sum(f1.*class1+f2.*class2+(1-f1-f2).*class3)/(ncases);
   end
end

contour(flipud(percencorr),0:.02:1,0:.05:1,0:.05:1)
title('Quality of prediction versus cutoff value')
ylabel('beta cutoff')
xlabel('alpha cutoff')

maxpercencor=max(max(percencorr))
text(.33,.33,num2str(maxpercencor))
flops

save Th1sym100000.m Thetafull1 /ascii
save Th2sym100000.m Thetafull2 /ascii
```

Apêndice 4

Rotinas para Validação Cruzada

Listamos a seguir as rotinas **crosslog.m**, **grone.m** e **gruna.m**, escritas em MATLAB, assim como BOSS.F, SERPE.F, SERPESUB.F e POLLOEST.F, escritas em FORTRAN, usadas para o processo de validação cruzada descrito na Fig. 17.

crosslog.m

Controla o processo de validação cruzada, retirando uma cadeia e estimando os parâmetros com os demais dados.

```
disp('beginning  crosslog')

% updated Feb 7, 1992 to handle pooled pairwise interaction terms
% updated Feb 20 to MLEcat3 to handle three states
% updated Aug 27, 1992 to pass parameters to fortran
% updated Aug 31, 1992 to make (cici)lina take one out


load garnier.ptr
lgarnier=length(garnier)
lALLNEW=lgarnier;


load myres1.4real1000
load myres2.4real1000

myres110ascii=myres1;
myres210ascii=myres2;

%%%%%%%%%%%%%%%%
kmin=-8
kmax=8
flattails=0
%%%%%%%%%%%%%%%%
anyinteractions=1
```

```
kincr=3
%%%%%%%%%%%%%%% if kincr=0, all pairwise interaction
%  with every possible increment, else
% kincr defines increment
% in interaction pairs

alpha=(garnier(:,31)==1);
size(alpha)
beta=(garnier(:,31)==2);
size(beta)
aa=alpha;
bb=beta;

% Prepare T
ind3=[1 21 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 21 18 19 20 21]';
clear T

% This creates a 17x17 matrix for a quadratic form A
% sensitive to period 3.6 signals
l1=sin((0:16)*2*pi/3.6);
l2=cos((0:16)*2*pi/3.6);
ll=[l1;l2];
A1=ll'*ll;
A1=A1(1:kmax-kmin+1,1:kmax-kmin+1);
l1=sin((0:16)*2*pi/2.0);
l2=cos((0:16)*2*pi/2.0);
ll=[l1;l2];
A2=ll'*ll;
A2=A2(1:kmax-kmin+1,1:kmax-kmin+1);

T=zeros(lALLNEW,kmax-kmin+1);
for k=kmin:kmax
    T(:,k-kmin+1)=ind3(garnier(:,21+k));
end

TT=T;
size(T)


jo=1;
logall=0;
last=zeros(1,5);
%last=[51 97 112 112 32];
for iot=1:lgarnier
      it=garnier(iot,1:5);
      if any(it~=last)
            last=it
            iot
            jo
            dum=(garnier(:,1:5)==(ones(lgarnier,1)*it));
            didi=(all(dum'==1))';
            alpha=aa;
            beta=bb;
            T=TT;
            gruna
            bigred(jo,:)=[it lgarnier-ncases maxpercencor 0 0];
         tata1(jo,:)=Thetafull1';
         tata2(jo,:)=Thetafull2';
            grone
            pbmax=percencorr(idx)
            bigred(jo,9)=max(pbmax);
            bigred(jo,8)=max(max(percencorr))

            save /scratch/boto/bigred bigred /ascii
            save /scratch/boto/tata1 tata1 /ascii
            save /scratch/boto/tata2 tata2 /ascii
           jo=jo+1;
       end
end
la=bigred(:,6);
res=bigred(:,7);
resla=res.*la;
jad=sum(resla)/sum(la)
```

gruna.m

Ajusta o modelo com penalização. Equivalente a **jelicle.m**, mas utilizando rotinas externas em FORTRAN

```
disp('beginning gruna')


inizio=clock
!rm /scratch/boto/ppall.mat
!rm /scratch/boto/QP.mat
!rm /scratch/boto/XpwX1.mat
!rm /scratch/boto/lop.mat
!rm /scratch/boto/sav.mat
!rm /scratch/boto/param.mat
!rm /scratch/boto/p1p2.mat
!rm /scratch/boto/t51.mat
!rm pulce0
!rm /scratch/boto/myres*.*
if jo==1
!cp blank iter
end


Ta=T;
dada=(didi==0);
T(didi,:)=[];
Ta(dada,:)=[];
alphi=alpha;
beti=beta;
alpha(didi)=[];
alphi(dada)=[];
beta(didi)=[];
beti(dada)=[];
f1=alpha;
f2=beta;
f3=alphi;
f4=beti;
length(f1)
disp('The number of residues with alpha=1 is ')
disp(sum(f1))
disp('The number of residues with beta=1 is ')
disp(sum(f2))
disp('The protein taken out has that much')
disp(sum(dada))

% A program to calculate the MLE for a logistic linear
% model for any number of independent vars
%
% T is a table with
% ncases by nvar
% using consequtive integers 1, 2, 3, ..., nlevs(i)
% for levels of variable(i)
% f is a vector of length ncases, giving success (0 or 1)
% X is the implicit "design" matrix
% ncases by npar-nvar,
% nparmain = 1+sum(levels in each main effect variable)
% nvarmain = number of nominal  main effect variables
% No provision is made for weights(frequencies) of each case

[ncases,nvarmain]=size(T)

nl=21
nlnotmissing=20
nl1=20
nlevs=nl*ones(1,nvarmain)
nparmain=1+sum(nlevs)
cnlevs=tril(ones(nvarmain,nvarmain),-1)*nlevs';
Toffset=1+cnlevs';
nvar=nvarmain;
npar=nparmain;
```

```
nfreeparmain=nparmain-2*nvarmain % This anticipates two constraints on each
                                 % variable (for misssing variable category)
ww1=[];
ww2=[]; % vector of weights for interactions

if anyinteractions
%  kincr is the increment between interaction vars
%  Define k1, k2 to be two variables involved in interaction, if any
%  modify T, Theta, assuming columns k1,k2 are the interaction

%  begin looping over all increments

if kincr==0 %this signals use of all increments
   kincrmin=1;
   kincrmax=(kmax-kmin);
else
   kincrmin=1;
   kincrmax= nvarmain - 2*kincr;
end

for kincr1=kincrmin:kincrmax
      for k1=kincr:nvarmain-kincr1-kincr+1
          k2=k1+kincr1;
              [k1 k2]
%         extend the T array
          Toffset=[Toffset Toffset(nvarmain)+nlevs(nvarmain)];
          T(:,nvar+1)=(T(:,k1)==21|T(:,k2)==21)*401+...
                (T(:,k1)~=21&T(:,k2)~=21).*(nl1.*(T(:,k1)-1)+T(:,k2));
          nvar=nvar+1;
          ww1=[ww1 A1(k1,k2)]; % Pick out appropriate weighting from
          ww2=[ww2 A2(k1,k2)]; % matrix of quadratic form A
      end
end       % of kincr1 loop

www1=[ones(1,nvarmain) ww1];
www2=[ones(1,nvarmain) ww2];
nvar
npar=nparmain+nl1*nl1+1
% add one to npar for missing interaction

Thetafull=[Thetafull;zeros(nl1*nl1,1)];

end     % of anyinteractions if

npairs=nvar-nvarmain

Thetafull1 = myres110ascii ;
Thetafull2 = myres210ascii ;

% initial estimate for Intercept only

XTheta1=zeros(ncases,1);
XTheta1(:)=Thetafull1(1)*ones(ncases,1);
XTheta2=zeros(ncases,1);
XTheta2(:)=Thetafull2(1)*ones(ncases,1);
for k=1:nvar
%   Note presence of www, a weight giving the fixed influence
%%%   of each parameter, esp. in the interactions.

          XTheta1(:)=XTheta1(:)+www1(k)*Thetafull1(T(:,k)+Toffset(k));

          XTheta2(:)=XTheta2(:)+www2(k)*Thetafull2(T(:,k)+Toffset(k));

end

lnL=-inf;
%%% first time
z1=exp(XTheta1);
z2=exp(XTheta2);
p1=z1./(1+z1+z2);
p2=z2./(1+z1+z2);
oldlnL=lnL;
lnL=f1'*XTheta1+f2'*XTheta2-sum(log(1+z1+z2))

parameters = [nvarmain nparmain nvar npar ncases npairs 0] ;

% parameters(7) = 0 means :  fortran program calledf for
```

```
% the first time so that 'pollo' will be called by 'boss'.

tog = [ Thetafull1(1) Thetafull2(1) lnL] ;
disp('beginning pass thru data')
echo on
init=clock
save /scratch/boto/ppall T Toffset f1 f2 ww1 ww2 tog
save /scratch/boto/param parameters
save /scratch/boto/p1p2 p1 p2

aftmatsav = etime(clock, init)
t = clock
!boss > pulce0
loadsave = etime(clock,t)
t1 = clock
load /scratch/boto/XpwX1
aftmatload = etime(clock,t1)
echo off

disp('finished pass thru data')

%  Now check for zero cols of XpwX1
emptycol=(sum(XpwX1~=0)==0);
if sum(emptycol)>0
   'there are empty cols in XpwX'
   sum(emptycol)
end

%  constraints
%  Projection matrix P which enforces constraints on
%  sum(Theta)=zero for each main effect variable

P=eye(nparmain);
for i=1:nvarmain
    P(Toffset(i)+nl1,Toffset(i)+(1:nl))=-ones(1,nl);
%   set row corresponding to last category to -1
end
P(:,[nl+Toffset(1:nvarmain) nl1+Toffset(1:nvarmain)])=[]; % delete the last
                                                          % category column
ninter=0;

if anyinteractions
%  Define k1, k2 to be two variables involved in interaction, if any
   P1=eye(nl1*nl1+1); % 401x401
   offset1=nl1*(0:(nl1-1));
   for i=1:nl1
       P1(offset1(i)+nl1,offset1(i)+(1:nl1))=-ones(1,nl1);
   end
   for i=1:(nl1-1)
%      set next to last row to negative of last block
       P1(offset1(nl1)+(1:nl1),offset1(i)+(1:nl1))=...
      -P1(offset1(i)+(1:nl1),offset1(i)+(1:nl1));
   end
%  delete unneeded columns
   dind=zeros(1,nl1*nl1+1);
   dlist=[nl1*(1:nl1) (nl1-1)*nl1+(1:nl1) nl1*nl1+1];
   dind(dlist)=ones(length(dlist),1);
%  we don't expect any empty cols of X'wX in first nparmain elts
   emptycol(1:nparmain)=[];
   if length(dind)~=length(emptycol)
      'there is an error in emptycol'
   end
%  Now check if emptycol matches dind
   if any(emptycol+dind>1)
      'you are trying to delete a column of P1 twice!'
   end
   dind=dind|emptycol;
   P1(:,dind)=[];
   [nrowsP1,ninter]=size(P1)

%  finally, modify P
   P=[[P zeros(nparmain,ninter)]; [zeros(nrowsP1,nfreeparmain) P1]];
end

QP=P;
'QP is ready'
clear P
```

```
nparams=nfreeparmain+ninter
parameters(7) = nparams;   % the fortran program will be called for the 2 time
lambda = 1000
semacros=1;
parameters = [parameters nrowsP1 lambda semacros] ;
parameters

disp('beginning pass thru data')
echo on
init=clock
save /scratch/boto/QP QP
save /scratch/boto/param parameters

aftmatsav = etime(clock, init)
t = clock
!boss >> iter

loadsave = etime(clock,t)
t1 = clock
load /scratch/boto/lop

aftmatload = etime(clock,t1)
echo off
%disp(etime(clock,ttime))
disp('finished pass thru data')

end

Mufinal1=Thetafull1(1)
Mufinal2=Thetafull2(1)

percencorr=zeros(21,21);
for i=0:20
    c1=i/20;
    for j=0:20-i
        c2=j/20;
        d12=(c2*p1>c1*p2);
        d13=((1-c1-c2)*p1>c1*(1-p1-p2));
        d23=((1-c1-c2)*p2>c2.*(1-p1-p2));
        class1=d12&d13;
        class2=~d12&d23;
        class3=(1-class1-class2);
        percencorr(i+1,j+1)=sum(f1.*class1+f2.*class2+(1-f1-f2).*class3)/(ncases);
    end
end

maxpercencor=max(max(percencorr))
idx=find(percencorr==maxpercencor)
percencorr(idx)

fine=etime(clock,inizio)
ora=clock
```

grone.m

Esta rotina verifica acurácia e verossimilhança da previsão de estrutura secundária da cadeia protéica retirada do conjunto de dados, usando parâmetros ajustados às cadeias remanescentes.

```
% tests the paramenters on a T set

% Sept 3, 1992

disp('cone here')

[ncases,nvarmain]=size(Ta)
nvar=nvarmain;

%  begin looping over all increments
```

```
if kincr==0 %this signals use of all increments
   kincrmin=1;
   kincrmax=(kmax-kmin);
else
   kincrmin=1;
   kincrmax= nvarmain - 2*kincr;
end

for kincr1=kincrmin:kincrmax
      for k1=kincr:nvarmain-kincr1-kincr+1
          k2=k1+kincr1;
%         extend the T array
          Ta(:,nvar+1)=(Ta(:,k1)==21|Ta(:,k2)==21)*401+...
                (Ta(:,k1)~=21&Ta(:,k2)~=21).*(nl1.*(Ta(:,k1)-1)+Ta(:,k2));
         nvar=nvar+1;
      end
end      % of kincr1 loop
nvar

end      % of anyinteractions if

XTheta1=zeros(ncases,1);
XTheta1(:)=Thetafull1(1)*ones(ncases,1);
XTheta2=zeros(ncases,1);
XTheta2(:)=Thetafull2(1)*ones(ncases,1);
for k=1:nvar
%   Note presence of www, a weight giving the fixed influence
%   of each parameter, esp. in the interactions.
    XTheta1(:)=XTheta1(:)+www1(k)*Thetafull1(Ta(:,k)+Toffset(k));
    XTheta2(:)=XTheta2(:)+www2(k)*Thetafull2(Ta(:,k)+Toffset(k));
end

z1=[];
z2=[];
p1=[];
p2=[];
z1=exp(XTheta1);
z2=exp(XTheta2);
p1=z1./(1+z1+z2);
p2=z2./(1+z1+z2);
w11=p1.*(1-p1);
w22=p2.*(1-p2);
w12=-p1.*p2;

lnL=f3'*XTheta1+f4'*XTheta2-sum(log(1+z1+z2))
logall=logall+lnL

percencorr=zeros(21,21);
for i=0:20
    c1=i/20;
    for j=0:20-i
        c2=j/20;
        d12=(c2*p1>c1*p2);
        d13=((1-c1-c2)*p1>c1*(1-p1-p2));
        d23=((1-c1-c2)*p2>c2.*(1-p1-p2));
        class1=d12&d13;
        class2=~d12&d23;
        class3=(1-class1-class2);
        percencorr(i+1,j+1)=sum(f3.*class1+f4.*class2+(1-f3-f4).*class3)/(ncases);
    end
end
```

ROTINAS FORTRAN

Listamos a seguir as rotinas externas escritas em FORTRAN, que permitiram uma execução mais rápida da validação cruzada, por permitir vetorização de operações em mais de duas dimensões. O programa usado para compilar e ligar os objetos está listado a seguir. Depois listamos os programas fonte individuais. O programa **floadsave.c** é uma rotina em C fornecida junto com o aplicativo MATLAB (1993) para intercâmbio de dados.

makefile

```
boss:  boss.o floadsav.o serpe.o serpesub.o polloest.o
        fc  -O2 -db -pa -fi -o boss boss.o floadsav.o serpe.o serpesub.o \
        polloest.o -lveclib
boss.o: boss.f
        fc -O2  -pa -db -fi -c boss.f
serpe.o: serpe.f
        fc  -O2 -pa -db -fi -c serpe.f
serpesub.o: serpesub.f
        fc  -O2 -pa -db -fi -c serpesub.f
floadsav.o: floadsav.c
        cc -db -fi -c floadsav.c
polloest.o: polloest.f
        fc  -O2 -pa -db -fi -c polloest.f
```

BOSS.F

```
      program boss

      parameter (MNMAX = 2000000)
      double precision AR(MNMAX), ww1(137), ww2(137),
     +  Xpfmp1(759), Xpfmp2(759), XpwX1(759,759), XpwX2(759,759),
     +  Q12(759,759), w11(11205), w12(11205), w22(11205),
     +  p1(11205), p2(11205), T51(153), T52(153),
     +  T51bis(153), T52bis(153), T51bbis(153),
     +  T41(419,136), T42(419,136), T51j(11205,136), T52j(11205,136),
     +  ndistinctpairs(11205), Zpfmp(1371), QP(759,685),
     +  QPtr(685,759), QPtrXpwX1(685,759), QPtrXpwX2(685,759),
     +  QPtrQ12(685,759), ZpwZ(1371,1371), work(1371), Theta(1371),
     +  Thetafull1(759), Thetafull2(759), Xtheta1(11205),
     +  Xtheta2(11205), www1(153), www2(153), X1QP(685,685),
     +  X2QP(685,685), Q12QP(685,685), QPpDQP(1371,1371), QPtr0(685,759),
     +  PpD(685,685)
      integer type, m, n, imagf
      integer*4 T(11205,153), Toffset(153), f1(11205), initT41(137),
     +  f2(11205), T3(155)
      real*4 tarray(2), t1, t4, t5, t6, t7
      character*20 name
      character*24 systime1, systime2, savetime1

      call MOPEN('/scratch/boto/param.mat', 'r', ierr)
      if (ierr .ne. 0) stop 'Cannot open /scratch/boto/param.mat'

      call MLOAD(type, name, m, n, imagf, AR)

      nvarmain = AR(1)
      nparmain = AR(2)
      nvar = AR(3)
      npar = AR(4)
      ncases = AR(5)
      npairs = AR(6)
      nparams = AR(7)
```

```
      write(*,100)
 100  format ('sono nel boss! ')
      write (*,*) nvarmain,nparmain,nvar,npar,ncases,npairs,nparams
      if (nparams .eq. 0) then

      call pollo (nvarmain, nparmain, nvar, npar, ncases, npairs, ww1,
     +  ww2, Xpfmp1, Xpfmp2, XpwX1, XpwX2, Q12, T51, T52, T51bis,
     +  T52bis, T51bbis, T41, T42, T51j, T52j, ndistinctpairs, T,
     +  Toffset, f1, f2, initT41, T3, p1, p2 , w11, w12, w22)

      else

      nrowsP1 = AR(8)
      lambda = AR(9)
      write (*,*) nrowsP1, lambda

      call serpe (nvarmain, nparmain, nvar, npar, ncases, npairs,
     +  nparams,  Xpfmp1, Xpfmp2, XpwX1, XpwX2, Q12, T51, T52, T51bis,
     +  T52bis, T51bbis, T51j, T52j, ndistinctpairs, T, Toffset, f1,
     +  f2, T3, w11, w12, w22, p1, p2, Zpfmp, QPtr, QPtrXpwX1, QP,
     +  QPtrXpwX2, QPtrQ12, ZpwZ, work, Theta, Thetafull1, Thetafull2,
     +  Xtheta1, Xtheta2, www1, www2, X1QP, X2QP, Q12QP, nrowsP1, lambda,
     +  QPpDQP, QPtr0, PpD)

      end if

      stop
      end
```

POLLOEST.F

```
c     This is pollo, modified in order to be able to use Thetafull1,
c     Thetafull2, vectors of estimates calculated by previous runs of the
c     matlab version. Pollo is not loading Thetafull1 and Thetafull2,
c     but p1 and p2 .


      subroutine pollo(nvarmain, nparmain, nvar, npar, ncases, npairs,
     +  ww1, ww2, Xpfmp1, Xpfmp2, XpwX1, XpwX2, Q12, T51, T52, T51bis,
     +  T52bis, T51bbis, T41, T42, T51j, T52j, ndistinctpairs, T,
     +  Toffset, f1, f2, initT41, T3, p1, p2, w11, w12, w22)
c
c   Example program for subroutines MOPEN, MCLOSE and MLOAD.
c   See initial comments in FLOADSAV.C for subroutine specification.
c   Also, see FSAVEXX.F for an MSAVE example.
c
c   The actual code that implements these routines is written in C.
c   The source file is called FLOADSAV.C.
c
c   For UNIX machines:
c
c   To test the program, first generate a few matrices in MATLAB and save
c   them with the "save" command.  This creates a file called matlab.mat.
c   Then exit MATLAB and run floadxx, or run floadxx in another window,
c   or run floadxx with a "!" shell escape from within MATLAB.
c
      parameter (MNMAX = 2000000)
      double precision AR(MNMAX), ww1(npairs), ww2(npairs),
     +  Xpfmp1(npar), Xpfmp2(npar), XpwX1(npar,npar),XpwX2(npar,npar),
     +  Q12(npar,npar),  T51(nvar), T52(nvar),
     +  T51bis(nvar), T52bis(nvar), T51bbis(nvar), sum1, sum2,
     +  T41(npar-nparmain,npairs), T42(npar-nparmain,npairs),
     +  T51j(ncases,npairs), T52j(ncases,npairs),
     +  ndistinctpairs(ncases), w11(ncases), w12(ncases), w22(ncases),
     +  p1(ncases), p2(ncases)
      integer type, m, n, imagf
      integer*4 T(ncases,nvar), Toffset(nvar), f1(ncases),
     +  initT41(npairs), f2(ncases), T3(nvar)
      real*4 tarray(2), t1, t4, t5, t6
      character*20 name, tog
      character*24 systime1, systime2, savetime1
c
c     Open data file for reading.
```

```
c
      t1 = dtime (tarray(1),tarray(2))
      call fdate(systime1)
      write (*,*) t1, tarray(1), tarray(2), systime1
      newload = 0
      neof = 0
      call MOPEN('/scratch/boto/ppall.mat', 'r', ierr)
      if (ierr .ne. 0) stop 'Cannot open /scratch/boto/ppall.mat'
c
c     Continue loading and printing matrices until file is empty.
c
   10 call MLOAD(type, name, m, n, imagf, %ref(AR))
      t1 = dtime(tarray(1),tarray(2))
      write (*,95) name, t1, tarray(1), tarray(2)
   95 format(' time of loading ',a,' is ',f10.7,'  ',f10.7,'  ',f10.7)
      if (newload .eq. 1) go to 11
      write(*,20) m,n,name
      if (m*n .gt. MNMAX) write(*,30) m*n, MNMAX
c
      if (name(:1) .eq. 'T'.and. name(2:5).eq. '    ') then
          do 200 jj=1,n
              do 200 ii=1,m
  200             T(ii,jj) = AR((jj-1)*m + ii)
      elseif (name(:7) .eq. 'Toffset') then
              do 210 jj=1,n
  210             Toffset(jj) = AR(jj)
      elseif (name(:2) .eq. 'f1') then
          do 230 ii=1,m
  230             f1(ii) = AR(ii)
      elseif (name(:2) .eq. 'f2') then
          do 240 ii=1,m
  240             f2(ii) = AR(ii)
      elseif (name(:3) .eq. 'ww1') then
              do 250 jj=1,n
  250             ww1(jj) = AR(jj)
      elseif (name(:3) .eq. 'ww2') then
              do 260 jj=1,n
  260             ww2(jj) = AR(jj)
      elseif (name(:3) .eq. 'tog') then
          newload = 1
          call MCLOSE()
          call MOPEN('/scratch/boto/p1p2.mat','r',ierr)
          if(ierr .ne. 0) stop 'Cannot open /scratch/boto/p1p2.mat'
          go to 10
      end if
      t1 = dtime(tarray(1), tarray(2))
      write (*,97) t1, tarray(1), tarray(2)
      go to 10

11    write (*,20) m, n, name
      if (name(:2) .eq. 'p1') then
          do 499 ii=1,m
499           p1(ii) = AR(ii)
      elseif (name(:2) .eq. 'p2') then
          do 500 ii=1,m
500           p2(ii) = AR(ii)
         neof = 1
      end if

      t1 = dtime(tarray(1), tarray(2))
      write (*,97) t1, tarray(1), tarray(2)
      if (neof .eq. 0) go to 10

97    format ('time for assigning values  ', f10.7,f10.7, f10.7 )
   20 format(' Loaded a ',i7,' -by- ',i7,' matrix named ',a,i4)
   30 format(' ',i7,' > ',i7,' Overwrote array storage.')
c
c
      t4 = dtime(tarray(1), tarray(2))
      call fdate(systime2)
      write (*,*) t4, tarray(1), tarray(2), systime2
      t5 = etime(tarray(1), tarray(2))
      write (*,98) t5
   98 format ('total time for the fortran loading program  ',f10.7)

      T3(1) = 1
      do 1050 i=1, nvarmain+1
```

```
          T51(i) = 1
1050      T52(i) = 1

      do 2000 i=1, ncases
          w11(i) = p1(i) * (1 - p1(i))
          w22(i) = p2(i) * (1 - p2(i))
2000      w12(i) = -p1(i) * p2(i)


      do 1000 j=1, ncases

c      if (j .gt. 1) stop
      ndistinctpairs(j) = 0
      kk = 2
      do 1055 ii= 1, nvar - nvarmain
          T41(initT41(ii),ii) = 0.0
1055      T42(initT41(ii),ii) = 0.0
      do 1060 ii=2, nvar+2
1060      T3(ii) = 0

      do 1010 k=1, nvar-nvarmain
          initT41(k) = T(j,k+nvarmain)
          T41(T(j,k+nvarmain),k) = ww1(k)
          T42(T(j,k+nvarmain),k) = ww2(k)
1010  continue

c      do 1005 i=2, nvar+1
c          ndummy = T(j,i-1) + Toffset(i-1)
c          do 1035 kj= 1, kk - 1
c              if (T3(kj) .eq. ndummy) then
c                   go to 1005
c              else if (ndummy .gt. T3(kj) .and. ndummy .lt. T3(kj+1)) then
c                   do 1065 ki=kk, kj+1, -1
c1065                   T3(ki) = T3(ki-1)
c                       T3(kj+1) = ndummy
c                       go to 1006
c               end if
c1035      continue
c          T3(kk) = ndummy
c1006      kk = kk + 1
c1005  continue

      do 1007 i=2,nvar+1
          ndummy = T(j,i-1) + Toffset(i-1)
          do 1037 kj=1, kk - 1
              if (T3(kj) .eq. ndummy) go to 1007
1037      continue
          T3(kk) = ndummy
          kk = kk + 1
1007  continue
      kk = kk - 1

      call isort('A',kk,T3,1)

c      kk = kk - 1
      do 1070 i=1, kk
          sum1 = 0
          sum2 = 0
          if (T3(i) .gt. nparmain) then
              ndistinctpairs(j) = ndistinctpairs(j) + 1
              nrow = T3(i) - nparmain
c             write (*,*) T3(i), nrow
              do 1015 k=1, npairs
                   sum1 = sum1 + T41(nrow,k)
1015               sum2 = sum2 + T42(nrow,k)
              nca = 1 + nvarmain + ndistinctpairs(j)
              T51(nca) = sum1
              T52(nca) = sum2
          end if
1070  continue

      do 1075 i=1, ndistinctpairs(j)
          T51j(j,i) = T51(1+nvarmain+i)
1075      T52j(j,i) = T52(1+nvarmain+i)


      do 1025 i=1, nca
```

```
            T51bis(i) = w11(j)*T51(i)
            T52bis(i) = w22(j)*T52(i)
1025        T51bbis(i) = w12(j)*T51(i)
c$dir no_recurrence
      do 1020 i=1, nca
            Xpfmp1(T3(i)) = Xpfmp1(T3(i)) + T51(i) * (f1(j) - p1(j))
            Xpfmp2(T3(i)) = Xpfmp2(T3(i)) + T52(i) * (f2(j) - p2(j))
c$dir no_recurrence
            do 1020 kj=1, nca
                XpwX1(T3(i),T3(kj))=XpwX1(T3(i),T3(kj))+(T51bis(i)*T51(kj))
               XpwX2(T3(i),T3(kj))=XpwX2(T3(i),T3(kj))+(T52bis(i)*T52(kj))
               Q12(T3(i),T3(kj))=Q12(T3(i),T3(kj)) + (T51bbis(i)*T52(kj))
1020  continue


1000  continue
      t5 = dtime(tarray(1), tarray(2))
      t4 = etime(tarray(1), tarray(2))
      write (*,94) t4, t5
   94 format ('total time  ',f15.7, '   time loop j  ',f15.7)

c
c     Open data file for writing.
c
      call MOPEN('/scratch/boto/sav.mat', 'w', ierr)
      if (ierr .ne. 0) stop 'Cannot open /scratch/boto/sav.mat'
c
c     Use a subroutine with two-dimensional subscripting.
c     The m*n elements must be contiguous in memory so they
c     can be saved with a single write statement.
c     type is 0 for IEEE floating point with Intel byte order.
c     type is 1000 for IEEE floating point with Motorola byte order.
c     type is 2000 for VAX D-floating or 3000 for VAX G-floating
c
      type = 1000
      imagf = 0

      m = npar
      n = npar
      name='XpwX2'
      call MSAVE(type, name, m, n, imagf, XpwX2 )
      write(*,32) m,n,name
      name='Q12'
      call MSAVE(type, name, m, n, imagf, Q12 )
      write(*,32) m,n,name

      n = 1
      name='Xpfmp1'
      call MSAVE(type, name, m, n, imagf, Xpfmp1 )
      write(*,32) m,n,name
      name='Xpfmp2'
      call MSAVE(type, name, m, n, imagf, Xpfmp2 )
      write(*,32) m,n,name
      call MCLOSE()

      call MOPEN('/scratch/boto/XpwX1.mat','w', ierr)
      if (ierr .ne. 0) stop 'Cannot open /scratch/boto/XpwX1.mat'
      m = npar
      n = npar
      name='XpwX1'
      call MSAVE(type, name, m, n, imagf, XpwX1 )
      write(*,32) m,n,name
      call MCLOSE()

      call MOPEN('/scratch/boto/t51.mat', 'w', ierr)
      if (ierr .ne. 0) stop 'Cannot open scratch/boto/t51.mat'

      m = ncases
      n = npairs
      name='T51j'
      call MSAVE(type, name, m, n, imagf, T51j )
      write(*,32) m,n,name
      name='T52j'
      call MSAVE(type, name, m, n, imagf, T52j )
      write(*,32) m,n,name

      n = 1
```

```
      name='ndistinctpairs'
      call MSAVE(type, name, m, n, imagf, ndistinctpairs)
      write(*,32) m,n,name

      call MCLOSE()
c
   32 format('Saved a ',i7,' -by- ',i7,'  matrix named ',a)
      call fdate(savetime1)
      write(*,*) savetime1
      t6 = etime(tarray(1), tarray(2))
      write(*,96) t6-t4, t6
   96 format ('time for saving only is ',f15.7,'  total time  ',f15.7)
      return
      end
```

SERPE.F

```
      subroutine serpe (nvarmain, nparmain, nvar, npar, ncases, npairs,
     +  nparams, Xpfmp1, Xpfmp2, XpwX1, XpwX2, Q12, T51, T52, T51bis,
     +  T52bis, T51bbis, T51j, T52j, ndistinctpairs, T, Toffset, f1,
     +  f2, T3, w11, w12, w22, p1, p2, Zpfmp, QPtr, QPtrXpwX1, QP,
     +  QPtrXpwX2, QPtrQ12, ZpwZ, work, Theta, Thetafull1, Thetafull2,
     +  Xtheta1, Xtheta2, www1, www2, X1QP, X2QP, Q12QP, nrowsP1, lambda,
     +  QPpDQP, QPtr0, PpD)
c
c   Example program for subroutines MOPEN, MCLOSE and MLOAD.
c   See initial comments in FLOADSAV.C for subroutine specification.
c
c   The actual code that implements these routines is written in C.
c   The source file is called FLOADSAV.C.
c
c
      parameter (MNMAX = 2000000)
      double precision AR(MNMAX), Xpfmp1(npar),
     +  Xpfmp2(npar), XpwX1(npar,npar), XpwX2(npar,npar),
     +  Q12(npar,npar), w11(ncases), w12(ncases),
     +  w22(ncases), Zpfmp(nparams*2),
     +  p1(ncases), p2(ncases), ndistinctpairs(ncases),
     +  T51j(ncases,npairs), T52j(ncases,npairs),QP(npar,nparams),
     +  QPtr(nparams,npar), QPtrXpwX1(nparams,npar),
     +  QPtrXpwX2(nparams,npar), QPtrQ12(nparams,npar),
     +  ZpwZ(nparams*2,nparams*2), work(nparams*2), Theta(nparams*2),
     +  Thetafull1(npar), Thetafull2(npar), Xtheta1(ncases),
     +  Xtheta2(ncases), www1(nvar), www2(nvar),
     +  X1QP(nparams,nparams), X2QP(nparams,nparams),
     +  Q12QP(nparams,nparams), QPpDQP(nparams*2,nparams*2),
     +  QPtr0(nparams,npar), PpD(nparams, nparams),
     +  rcond, Mu01, Mu02, oldlnL, lnL, ff1, ff2, sum,dumexp1,
     +  dumexp2, alpha, beta, alphanew, betanew
      integer type, m, n, imagf
      integer*4 T(ncases,nvar), Toffset(nvar),lda,ldb,ldc,
     +  f1(ncases), f2(ncases), ier, nparams2, m1,n1, lambda
      real*4 tarray(2), t1, t4, t5, t6, t7
      character*20 name, tag
      character*24 systime1, systime2, savetime1
c
c     Open data file for reading.
c
      t1 = dtime (tarray(1),tarray(2))
      call fdate(systime1)
      write (*,*) t1, tarray(1), tarray(2), systime1
      newload = 0
      neof = 0
      call MOPEN('/scratch/boto/ppall.mat', 'r', ierr)
      if (ierr .ne. 0) stop 'Cannot open ppall.mat'
c
c     Continue loading and printing matrices until file is empty.
c
   10 call MLOAD(type, name, m, n, imagf, %ref(AR) )
      t1 = dtime(tarray(1),tarray(2))
      write (*,95) name, t1, tarray(1), tarray(2)
   95 format(' time of loading ',a,' is ',f10.7,'  ',f10.7,'  ',f10.7)
      if (newload .eq. 1) then
```

```
          go to 11
      elseif (newload .eq. 2) then
           go to 12
      elseif (newload .eq. 3) then
           go to 13
      end if
      write(*,20) m,n,name
      if (m*n .gt. MNMAX) write(*,30) m*n, MNMAX
c
      if (name(:1) .eq. 'T'.and. name(2:5).eq. '    ') then
          do 200 jj=1,n
              do 200 ii=1,m
  200             T(ii,jj) = AR((jj-1)*m + ii)
      elseif (name(:7) .eq. 'Toffset') then
               do 210 jj=1,n
  210              Toffset(jj) = AR(jj)
      elseif (name(:2) .eq. 'f1') then
           do 230 ii=1,m
  230              f1(ii) = AR(ii)
      elseif (name(:2) .eq. 'f2') then
           do 240 ii=1,m
  240              f2(ii) = AR(ii)
      elseif (name(:3) .eq. 'ww1') then
           do 245 jj=nvarmain+1, nvar
245             www1(jj) = AR(jj-nvarmain)
      elseif (name(:3) .eq. 'ww2') then
           do 246 jj=nvarmain+1, nvar
246             www2(jj) = AR(jj-nvarmain)
      elseif (name(:3) .eq. 'tog') then
           Mu01 = AR(1)
           Mu02 = AR(2)
           lnL = AR(3)
           do 247 jj=1, nvarmain
              www1(jj) = 1
247           www2(jj) = 1
              newload = 1
              call MCLOSE()
              call MOPEN('/scratch/boto/sav.mat', 'r', ierr)
        if (ierr .ne. 0) stop 'Cannot open /scratch/boto/sav.mat'
              go to 10
      end if
      t1 = dtime(tarray(1), tarray(2))
      write (*,97) t1, tarray(1), tarray(2)
      go to 10


11    write(*,20) m,n,name
      if (name(:5) .eq. 'XpwX2') then
           do 270 jj=1,n
               do 270 ii=1,m
270                XpwX2(ii,jj) = AR((jj-1)*m + ii)
      elseif (name(:3) .eq. 'Q12') then
           do 280 jj=1,n
               do 280 ii=1,m
280                Q12(ii,jj) = AR((jj-1)*m + ii)
      elseif (name(:6) .eq. 'Xpfmp1') then
           do 290 ii=1,m
290           Xpfmp1(ii) = AR(ii)
      elseif (name(:6) .eq. 'Xpfmp2') then
           do 300 ii=1,m
300           Xpfmp2(ii) = AR(ii)
              newload = 3
              call MCLOSE()
              call MOPEN('/scratch/boto/QP.mat', 'r', ierr)
              if (ierr .ne. 0) stop 'cannot open QP.mat'
              go to 10
      end if
      t1 = dtime(tarray(1), tarray(2))
      write (*,97) t1, tarray(1), tarray(2)
      go to 10


13    write(*,20) m,n,name
      if (name(:2) .eq. 'QP') then
          do 250 jj=1,n
              do 250 ii=1,m
250               QP(ii,jj) = AR((jj-1)*m + ii)
```

```
          call MCLOSE()
          call MOPEN ('/scratch/boto/XpwX1.mat', 'r', ierr)
          if (ierr .ne. 0) stop 'cannot open XpwX1.mat'
          go to 10
      else
          do 251 jj=1,n
              do 251 ii=1,m
251               XpwX1(ii,jj) = AR((jj-1)*m + ii)
      end if
      newload = 2
      t1 = dtime(tarray(1), tarray(2))
      write (*,97) t1, tarray(1), tarray(2)
      call MCLOSE()
      call MOPEN('/scratch/boto/t51.mat', 'r', ierr)
      if (ierr .ne. 0) stop 'cannot open /scratch/boto/t51.mat'
      go to 10


12    write(*,20) m,n,name
      if (name(:4) .eq. 'T51j') then
          do 201 jj=1,n
              do 201 ii=1,m
 201              T51j(ii,jj) = AR((jj-1)*m + ii)
      elseif (name(:4) .eq. 'T52j') then
          do 202 jj=1,n
              do 202 ii=1,m
 202              T52j(ii,jj) = AR((jj-1)*m + ii)
      elseif (name(:14) .eq. 'ndistinctpairs') then
          do 203 jj=1,m
203               ndistinctpairs(jj) = AR(jj)
      neof = 1
      endif
      t1 = dtime(tarray(1), tarray(2))
      write (*,97) t1, tarray(1), tarray(2)
      if (neof .eq. 0 ) go to 10
c
97    format ('time for assigning values  ', f10.7,f10.7, f10.7 )
   20 format(' Loaded a ',i7,' -by- ',i7,' matrix named ',a,i4)
   30 format(' ',i7,' > ',i7,' Overwrote array storage.')
c
c
      t4 = dtime(tarray(1), tarray(2))
      call fdate(systime2)
      write (*,*) t4, tarray(1), tarray(2), systime2
      t5 = etime(tarray(1), tarray(2))
      write (*,98) t5
   98 format ('total time for the fortran loading program  ',f10.7)


      nparams2 = nparams * 2
      m1=nparams
      n1=npar
      lda=npar
      ldb=nparams
      ldc = nparams
      alpha = 1.0
      beta = 0.0
      alphanew = -lambda
      betanew = 1.0
      Theta(1) = Mu01
      Theta(nparams+1) = Mu02
      do 310 i=1,nparams
          do 310 j=1,npar
310           QPtr(i,j) = QP(j,i)

c      if (lambda .ne. 0) then
c          do 255 i=1, nparams
c              do 260 j=i, nparams
c                  do 265 jk = nrowsP1, npar
c265                   QPpDQp(i,j) = QPpDQP(i,j) + QP(jk,i)*QP(jk,j)
c                  QPpDQP(j,i) = QPpDQP(i,j)
c                  QPpDQP(i+nparams,j+nparams) = QPpDQP(i,j)
c                  QPpDQP(j+nparams,i+nparams) = QPpDQP(i,j)
c260          continue
c255      continue
c      end if
```

```
      if (lambda .ne. 0) then
           do 255 i=1,nparams
               do 255 j=nparmain+1, npar
255                QPtr0(i,j) = QPtr(i,j)
           call dgemm('n','n',m1,m1,n1,alpha,QPtr0,ldb,QP,lda,
     +                beta,PpD,ldc)
           do 260 i=1, nparams
               do 260 j=i, nparams
                   QPpDQP(i,j) = PpD(i,j)
                   QPpDQP(j,i) = PpD(i,j)
                   QPpDQP(i+nparams,j+nparams) = PpD(i,j)
260                QPpDQP(j+nparams,i+nparams) = PpD(i,j)
      end if


      do 1000 iter=1,7

      print 1001,iter
1001  format('this is iteration n.',iter)
      do 320 j=1, nparams
               Zpfmp(j) = 0.0
               Zpfmp(j+nparams) = 0.0
           do 320 i=1,npar
               Zpfmp(j) = Zpfmp(j) + QP(i,j)*Xpfmp1(i)
320            Zpfmp(j+nparams) = Zpfmp(j+nparams) + QP(i,j)*Xpfmp2(i)

c     QPtrXpwX1 = QPtr * XpwX1, where XpwX1 is simmetric
c     QPtrXpwX2 = QPtr * XpwX2, where XpwX2 is simmetric

      call dsymm('right','lower',m1,n1,alpha,XpwX1,lda,
     +   QPtr,ldb,beta,QPtrXpwX1,ldc)
      call dsymm('right','lower',m1,n1,alpha,XpwX2,lda,
     +   QPtr,ldb,beta,QPtrXpwX2,ldc)

c     QPtrQ12 = QPtr * Q12, where Q12 is not symmetric
      call dgemm('n','n',m1,n1,n1,alpha,QPtr,ldb,Q12,lda,
     +   beta,QPtrQ12,ldc)


      call dgemm('n','n',m1,m1,n1,alpha,QPtrXpwX1,ldb,QP,lda,beta,
     +   X1QP,ldc)
      call dgemm('n','n',m1,m1,n1,alpha,QPtrXpwX2,ldb,QP,lda,beta,
     +   X2QP,ldc)
      call dgemm('n','n',m1,m1,n1,alpha,QPtrQ12,ldb,QP,lda,beta,
     +   Q12QP,ldc)

      do 331 i=1, nparams
           do 331 j=1, nparams
               ZpwZ(i,j) = X1QP(i,j)
               ZpwZ(i+nparams,j+nparams) = X2QP(i,j)
               ZpwZ(i,j+nparams) = Q12QP(i,j)
331            ZpwZ(i+nparams,j) = Q12QP(j,i)

      if (lambda .ne. 0) then
           do 335 i=1, nparams
               do 335 j=i, nparams
                   ZpwZ(i,j) = ZpwZ(i,j) + lambda * QPpDQP(i,j)
335                ZpwZ(i+nparams,j+nparams) = ZpwZ(i+nparams,j+nparams) +
     +                    lambda * QPpDQP(i,j)

           do 330 i=1, nparams
c$dir no_recurrence
               do 334 j=i, nparams
                   ZpwZ(j,i) = ZpwZ(i,j)
334                ZpwZ(j+nparams,i+nparams) = ZpwZ(i+nparams,j+nparams)
330        continue

           call dsymv ('lower',nparams2,alphanew,QPpDQP,nparams2,Theta,
     +                 1,betanew,Zpfmp,1)
      end if

c     ****** Cholesky factorization of ZpwZ, positive definite matrix,
c           with the evaluation of the condition number for the
c            first iteration only.
      if (iter .eq. 1) then
           call dpoco (ZpwZ,nparams2,nparams2,rcond,work,ier)
           if (1.0+rcond .eq. 1.0) stop 'ZpwZ is singular'
```

```
      else
           call dpofa (ZpwZ,nparams2,nparams2,ier)
      end if

      if (ier .ne. 0) stop 'ZpwZ is not positive definite'

c     ****** The following routine solves the linear system
c            ZpwZ*x=Zpfmp, where ZpwZ is a positive definite matrix.
c            The solution vector x overwrites Zpfmp.
      call dposl (ZpwZ,nparams2,nparams2,Zpfmp)

      do 370 i=1, nparams2
370       Theta(i) = Theta(i) + Zpfmp(i)

      do 380 i=1, npar
           Thetafull1(i) = 0.0
           Thetafull2(i) = 0.0
           do 380 j=1, nparams
              Thetafull1(i) = Thetafull1(i) + QP(i,j)*Theta(j)
380           Thetafull2(i) = Thetafull2(i) + QP(i,j)*Theta(j+nparams)

      sum = 0.0
      ff1 = 0.0
      ff2 = 0.0
      do 390 j=1, ncases
           Xtheta1(j) = Thetafull1(1)
           Xtheta2(j) = Thetafull2(1)
           do 400 k=1, nvar
               Xtheta1(j) = Xtheta1(j) +
     +                      www1(k)*Thetafull1(T(j,k)+Toffset(k))
400            Xtheta2(j) = Xtheta2(j) +
     +                      www2(k)*Thetafull2(T(j,k)+Toffset(k))
          dumexp1 = dexp(Xtheta1(j))
          dumexp2 = dexp(Xtheta2(j))
          p1(j) = dumexp1 / (1 + dumexp1 + dumexp2)
          p2(j) = dumexp2 / (1 + dumexp1 + dumexp2)
          sum = sum + dlog(1 + dumexp1 + dumexp2)
          ff1 = ff1 + f1(j) * Xtheta1(j)
390       ff2 = ff2 + f2(j) * Xtheta2(j)

      do 410 j=1, ncases
           w11(j) = p1(j) * (1 - p1(j))
           w22(j) = p2(j) * (1 - p2(j))
410        w12(j) = -p1(j) * p2(j)

      oldlnL = lnL
      lnL = ff1 + ff2 - sum
      write (*,500) lnL
500   format ('lnL is ',f12.6)
      if (dabs((lnL/oldlnL)-1) .lt. 1.0d-5) go to 420
      call serpesub (nvarmain, nparmain, nvar, npar, ncases,
     +  npairs, T, Toffset, f1, f2, p1, p2, w11, w12, w22, T51j, T52j,
     +  ndistinctpairs, Xpfmp1, Xpfmp2, XpwX1, XpwX2, Q12, T3, T51,
     +  T52, T51bis, T52bis, T51bbis)

1000  continue

c
c     Open data file for writing.
c
420   call fdate(savetime1)
      write(*,*) savetime1
      call MOPEN('/scratch/boto/lop.mat', 'w', ierr)
      if (ierr .ne. 0) stop 'Cannot open /scratch/boto/lop.mat'
c
c     Use a subroutine with two-dimensional subscripting.
c     The m*n elements must be contiguous in memory so they
c     can be saved with a single write statement.
c     type is 0 for IEEE floating point with Intel byte order.
c     type is 1000 for IEEE floating point with Motorola byte order.
c     type is 2000 for VAX D-floating or 3000 for VAX G-floating
c
      type = 1000
      imagf = 0

      m = npar
      n = 1
```

```
      name='Thetafull1'
      call MSAVE(type, name, m, n, imagf, Thetafull1 )
      t7 = dtime(tarray(1), tarray(2))
      write(*,32) m,n,name, t7
      name='Thetafull2'
      call MSAVE(type, name, m, n, imagf, Thetafull2 )
      t7 = dtime(tarray(1), tarray(2))
      write(*,32) m,n,name, t7
      m = ncases
      name = 'p1'
      call MSAVE(type, name, m, n, imagf, p1 )
      t7 = dtime(tarray(1), tarray(2))
      write(*,32) m,n,name, t7
      name='p2'
      call MSAVE(type, name, m, n, imagf, p2)
      t7 = dtime(tarray(1), tarray(2))
      write(*,32) m,n,name, t7

      call MCLOSE()

c
   32 format('Time to save a ',i7,' -by- ',i7,'  matrix named ',a,
     +    '  is  ',f10.7)
      call fdate(savetime1)
      write(*,*) savetime1
      t6 = etime(tarray(1), tarray(2))
      write(*,96) t6-t4, t6
   96 format ('time for saving only is ',f15.7,'  total time  ',f15.7)
      return
      end
```

SERPESUB.F

```
      subroutine serpesub (nvarmain, nparmain, nvar, npar, ncases,
     +  npairs, T, Toffset, f1, f2, p1, p2, w11, w12, w22, T51j, T52j,
     +  ndistinctpairs, Xpfmp1, Xpfmp2, XpwX1, XpwX2, Q12, T3, T51, T52,
     +  T51bis, T52bis, T51bbis)
c
c   Example program for subroutines MOPEN, MCLOSE and MLOAD.
c   See initial comments in FLOADSAV.C for subroutine specification.
c   Also, see FSAVEXX.F for an MSAVE example.
c
c   The actual code that implements these routines is written in C.
c   The source file is called FLOADSAV.C.
c
c   For UNIX machines:
c
c   To test the program, first generate a few matrices in MATLAB and save
c   them with the "save" command.  This creates a file called matlab.mat.
c   Then exit MATLAB and run floadxx, or run floadxx in another window,
c   or run floadxx with a "!" shell escape from within MATLAB.
c
      double precision T51j(ncases,npairs), T52j(ncases,npairs),
     +  Xpfmp1(npar), Xpfmp2(npar), XpwX1(npar,npar), XpwX2(npar,npar),
     +  Q12(npar,npar), w11(ncases), w12(ncases), T51(nvar),
     +  T52(nvar), w22(ncases), p1(ncases), p2(ncases),
     +  ndistinctpairs(ncases), T51bis(nvar), T52bis(nvar),
     +  T51bbis(nvar)
      integer*4 T(ncases,nvar), Toffset(nvar), f1(ncases),
     +  f2(ncases), T3(nvar)
      real*4 tarray(2), t1, t4, t5
      character*24 systime1, savetime1
c
c     Open data file for reading.
c
      t1 = dtime (tarray(1),tarray(2))
      call fdate(systime1)
      write (*,*) t1, tarray(1), tarray(2), systime1


      do 1100 i=1, npar
           Xpfmp1(i) = 0.0
           Xpfmp2(i) = 0.0
```

```
          do 1100 j=1, npar
              XpwX1(i,j) = 0.0
              XpwX2(i,j) = 0.0
1100          Q12(i,j) = 0.0

      T3(1) = 1
      do 1050 i=1, nvarmain+1
           T51(i) = 1
1050       T52(i) = 1

      do 1000 j=1, ncases

      kk = 2
      do 1060 ii=2, nvar+2
1060       T3(ii) = 0

      do 1005 i=2, nvar+1
           ndummy = T(j,i-1) + Toffset(i-1)
           do 1035 kj= 1, kk - 1
               if (T3(kj) .eq. ndummy) then
                    go to 1005
               else if (ndummy .gt. T3(kj) .and. ndummy .lt. T3(kj+1)) then
                    do 1065 ki=kk, kj+1, -1
1065                     T3(ki) = T3(ki-1)
                        T3(kj+1) = ndummy
                         go to 1006
               end if
1035       continue
           T3(kk) = ndummy
1006       kk = kk + 1
1005  continue

      kk = kk - 1
      nca = 1+ nvarmain + ndistinctpairs(j)

      do 1075 i=1, ndistinctpairs(j)
           T51(1+nvarmain+i) = T51j(j,i)
1075       T52(1+nvarmain+i) = T52j(j,i)


      do 1025 i=1, nca
           T51bis(i) = w11(j)*T51(i)
          T52bis(i) = w22(j)*T52(i)
1025       T51bbis(i) = w12(j)*T51(i)
c$dir no_recurrence
      do 1020 i=1, kk
           Xpfmp1(T3(i)) = Xpfmp1(T3(i)) + T51(i) * (f1(j) - p1(j))
           Xpfmp2(T3(i)) = Xpfmp2(T3(i)) + T52(i) * (f2(j) - p2(j))
c$dir no_recurrence
           do 1020 kj=1, kk
               XpwX1(T3(i),T3(kj))=XpwX1(T3(i),T3(kj))+(T51bis(i)*T51(kj))
              XpwX2(T3(i),T3(kj))=XpwX2(T3(i),T3(kj))+(T52bis(i)*T52(kj))
              Q12(T3(i),T3(kj))=Q12(T3(i),T3(kj)) + (T51bbis(i)*T52(kj))
1020  continue

1000  continue
      t5 = dtime(tarray(1), tarray(2))
      t4 = etime(tarray(1), tarray(2))
      write (*,94) t4, t5
   94 format ('total time  ',f15.7, '   time loop j  ',f15.7)


      call fdate(savetime1)
      write(*,*) savetime1
      return
      end
```