Universidade de Brasília
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Biologia Molecular

Juliana Ribeiro Rocha

Análise informacional dos enterramentos atômicos em proteínas globulares

Brasília
2012

Universidade de Brasília
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Biologia Molecular

Juliana Ribeiro Rocha

Análise informacional dos enterramentos atômicos em proteínas globulares

Dissertação de Mestrado apresentada como requisito parcial à obtenção do título de Mestre em Biologia Molecular, pela Universidade de Brasília

Orientador: Antônio Francisco Pereira de Araújo

Brasília
2012

It is our choices [...] that show what we truly are, far more than our abilities.
J. K. Rowling

Agradecimentos

A Deus, por tudo!

Ao meu orientador, Prof. Dr. Antônio Francisco Pereira de Araújo, pela oportunidade, pela orientação e por todos os ensinamentos passados.

Aos colegas do LBTC-UnB, Lindomar, Leandro, Marx, Diogo, Vinícius e Luísa, por proporcionar momentos divertidos e ajudar nos momentos de desespero!

A todos os amigos e familiares, porque cada um, à sua maneira, foram importantes ao longo desses dois anos de mestrado.

Aos meus pais, Rita e Saulo, pela minha formação como pessoa e pelo apoio incondicional.

Ao meu namorado, Pedro, pela ajuda com os computadores e pelo apoio igualmente incondicional.

Ao Programa de Pós-Graduação em Biologia Molecular, à Universidade de Brasília e à CAPES, pela estrutura e pelo apoio financeiro.

Resumo

O estudo do enovelamento de proteínas e a predição de suas estruturas nativas são de grande importância para a ciência. Uma das abordagens possíveis para tal predição tem base nos enterramentos atômicos, entendidos como a distância do átomo até o centro geométrico da proteína normalizada pelo raio de giro da proteína. Esses enterramentos podem ser discretizados em camadas concêntricas e equiprováveis e já foi mostrado que a informação sobre estas camadas é suficiente para levar a proteína ao seu estado nativo (Pereira de Araújo *et al.*, Proteins 70:971-983, 2008; Pereira de Araújo e Onuchic, PNAS 106:19001-19004, 2009), mas não se sabe como esta informação está codificada na sequência. O objetivo do presente trabalho é medir a transinformação, quantidade de informação compartilhada por duas ou mais variáveis aleatórias discretas, entre sequência e enterramentos atômicos a partir de quatro diferentes alfabetos de sequência e de diversos números de níveis de enterramento atômico visando entender as relações entre essas variáveis e, assim, fornecer um máximo teórico para a eficiência dos algoritmos de predição de enterramentos atômicos. Os resultados deste trabalho mostram que a sequência de aminoácidos possui densidade de entropia entre $0,9969 \pm 0,0002$ bit e $4,176 \pm 0,004$ bit, dependendo do número de símbolos do alfabeto considerado (dois, três, cinco ou vinte símbolos), e excesso de entropia sempre próximo a zero. Portanto, infere-se que não existe correlação local entre os elementos da sequência. Por outro lado, os enterramentos atômicos apresentam densidade de entropia variando de $0,617 \pm 0,002$ bit a $2,16 \pm 0,04$ bit para C_α e de $0,735 \pm 0,002$ bit a $2,54 \pm 0,01$ bit para C_β, de acordo com a quantidade de níveis de enterramento considerados (de dois a dez níveis), e excesso de entropia compreendido entre $0,53 \pm 0,01$ bit e $1,4 \pm 0,1$ bit para C_α e entre $0,48 \pm 0,02$ bit e $0,93 \pm 0,04$ bit para C_β. Estes resultados evidenciam que os enterramentos atômicos são correlacionados localmente entre si. Foi encontrada uma relação entre enterramento atômico e estrutura secundária, na qual as α-hélices são distribuídas por todo o espaço ocupado pela proteína, as folhas-β tendem aos níveis mais internos e segmentos sem estrutura secundária regular tendem a ocupar os níveis mais externos. A densidade de entropia da sequência é maior que a dos enterramentos atômicos se considerados alfabetos de mesmo tamanho, ou seja, aquela é potencialmente capaz de armazenar a informação necessária para a determinação deste. Entretanto, a transinformação entre sequência e enterramento (calculada para C_α e para C_β) não é maior que 20% da dúvida do mesmo. Essa porcentagem é aparentemente pequena, mas é possível que mesmo esta quantidade de informação seja suficiente para a predição dos enterramentos nativos, uma vez que a dúvida em relação ao enterramento de um átomo deve diminuir quando o de seu vizinho é conhecido, o que também foi mostrado neste trabalho. A combinação dos resultados deste provê um máximo teórico para os algoritmos de predição e permite saber se eles estão próximos deste máximo.

Palavras-chave: Enovelamento de proteínas, Enterramento atômico, Predição de estrutura terciária de proteínas.

Abstract

Protein folding understanding and protein terciary structure prediction are two main areas in science nowadays. One approach to terciary structure prediction has its basis in atomic burials, understood as the distance from a specific atom to the molecular geometrical center divided by protein's radius of giration. Atomic burials can be discretized in concentric and equiprobable burial layers, and it has been shown that information about these layers cary an amount of information enough to lead protein to its native structure (Pereira de Araújo *et al.*, Proteins 70:971-983, 2008; Pereira de Araújo and Onuchic, PNAS 106:19001-19004, 2009), althougth it is not elucidated how this information is held to the sequence. The intend of this work is to measure mutual information, defined as the amount of information shared between two or more discrete random variables, between sequence and atomic burials considering different combinations of sequence alphabets and number of burial layers, with the aim of understanding the relations between them and, therefore, contribute protein terciary structure prediction algorithms. The obtained results show that aminoacids sequence has entropy density ranging from $0,9969 \pm 0,0002$ bit to $4,176 \pm 0,004$ bit, according to the number of symbols of the considered alphabet (two, three, five or twenty symbols), and excess entropy approximately equal to zero. Taken together, these results demonstrate that sequence elements are not locally correlated. On the other hand, C_α and C_β atomic burials have shown entropy density from $0,617 \pm 0,002$ bit up to $2,16 \pm 0,04$ bit, and from $0,735 \pm 0,002$ bit up to $2,54 \pm 0,01$ bit, respectively, according to the number of burial levels considered (from two to ten levels). Its excess entropy vary from $0,53 \pm 0,01$ bit to $1,4 \pm 0,1$ bit and from $0,48 \pm 0,02$ bit to $0,93 \pm 0,04$ bit, taking C_α and C_β respectively. These results point that atomic burials elements have local correlation. In addition, a relation between atomic burials and secondary structure was evidenced, taken that Loops tend to external levels, β-sheets tend to internal levels and α-helices are homogeneously distributed through the protein radius. Sequence entropy density is major than atomic burials entropy density, i.e. sequence is theoretically capable of holding the amount of information necessary to determine atomic burials. However, mutual information between sequence and burials (calculated considering C_α and C_β) is not greater than 20% of burial uncertainty. This percentage is apparently small, but is it possible that even this few information is enough to a correct prediction of atomic burials once that the entropy with respect to one atom shall decrease once its neigbour's position is known, which has also been shown here. The results obtained here provide a theoretical maximum that can be achieved by prediction algorithms, and therefore permit measure their efficiency.

Key words: Atomic burials, Protein folding, Protein folding prediction.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

$\|\mathcal{A}_X\|$	Número de elementos da distribuição $\mathcal{A}_X$
$\mathcal{A}$	Distribuição $\mathcal{A}$
a	Parâmetro de ajuste
Å	Angstrom
$\mathcal{A}_X$	Alfabeto dos valores possíveis da variável X (Ditribuição $\mathcal{A}_X$)
$\mathcal{A}_Y$	Alfabeto dos valores possíveis da variável Y (Ditribuição $\mathcal{A}_Y$)
b	Parâmetro de ajuste
B	Enterramento atômico
BB	Esqueleto peptídico (*Backbone*)
B_i	Enterramento atômico do átomo da i-ésima posição
B_{i+1}	Enterramento atômico do átomo da $i+1$-ésima posição
c	Parâmetro de ajuste
CASP	*Critical Assessment of Techniques for Protein Structure Prediction*
C_α	Carbono-α
C_β	Carbono-β
E	Energia
E_h	Excesso de entropia
E_{h_B}	Excesso de entropia dos enterramentos atômicos
Estrut. Sec.	Estrutura secundária
G	Energia livre de Gibbs
H	Entalpia (Termodinâmica) ou Entropia (Teoria da informação)
$h(B_{BB})$	Densidade de entropia dos enterramentos dos átomos do esqueleto peptídico
$h(B_{C_\alpha})$	Densidade de entropia dos enterramentos dos átomos de carbono-α
$h(B_{C_\beta})$	Densidade de entropia dos enterramentos dos átomos de carbono-β
$H(B_i\|B_{C_\alpha}, Q)$	Entropia do enterramento do i-ésimo átomo da cadeia lateral condicionada ao enterramento do carbono-α e à identidade do resíduo correspondentes
$H(Q')$	Entropia das sequências de aminoácidos completas
$H(Q'\|T')$	Entropia das sequências de aminoácidos completas condicionada à entropia de suas estruturas terciárias
$H(T'\|Q')$	Entropia das estruturas terciárias completas condicionada à entropia de suas sequências de aminoácidos
$H(X)$	Entropia da variável X
$h(X)$	Densidade de entropia da variável X
$H(X,Y)$	Entropia conjunta das variáveis X e Y
$H(X\|Y)$	Entropia condicional da variável X dada a variável Y
$H(X^N)$	Entropia de um bloco de tamanho N de variáveis X

$H(X^N, Y^N)$	Entropia conjunta de blocos de tamanho N de variáveis X e de variáveis Y
$H(X^N\|Y^N)$	Entropia de um bloco de tamanho N de variáveis X condicionada a de um bloco de tamanho N de variáveis Y
$H(Y)$	Entropia da variável Y
$H(Y\|X)$	Entropia condicional da variável Y dada a variável X
$H(X^Y)$	Entropia de um bloco de tamanho N de variáveis Y
$H(Y^N\|X^N)$	Entropia de um bloco de tamanho N de variáveis Y condicionada a de um bloco de tamanho N de variáveis X
$h'(X)$	Densidade de entropia da variável X
HP	Alfabeto constituído de dois símbolos, com os resíduos classificados em hidrofóbicos ou polares
HPN	Alfabeto constituído de três símbolos, com os resíduos classificados em hidrofóbicos, polares ou neutros
HPNhp	Alfabeto constituído de cinco símbolos, com os resíduos classificados em muito hidrofóbicos, pouco hidrofóbicos, muito polares, pouco polares ou neutros
$I(B_i; B_{i+1})$	Transinformação entre os átomos de posição i e de posição $i+1$
$I(B_i; B_{i+1}\|Q_i)$	Transinformação entre os átomos de posição i e de posição $i+1$ condicionada à identidade do i-ésimo resíduo
$I(B_i; B_{i+1}\|Q_i, Q_{I+1})$	Transinformação entre os átomos de posição i e de posição $i+1$ condicionada às identidades dos resíduos de posição i e $i+1$
$i(Q : B)$	Densidade de transinformação entre sequência e enterramentos atômicos
$I(Q_0; B^N)$	Transinformação entre a identidade do resíduo central e um bloco de tamanho N de enterramentos atômicos
$I(Q^1; B^1)$	Transinformação entre a identidade do resíduo e o enterramento de um átomo do mesmo resíduo
$I(Q_1; B_1)$	Transinformação entre a identidade do resíduo e o enterramento de um átomo do mesmo resíduo
$I(Q_i, Q_{I+1}; S_i, S_{i+1})$	Transinformação entre identidades e estruturas secundárias dos resíduos de posição i e $i+1$
$I(Q_i; Q_{i+1})$	Transinformação entre as identidades de dois resíduos vizinhos
$I(Q_i; Q_{i+1}\|B_i)$	Transinformação entre as identidades de dois resíduos vizinhos condicionada ao enterramento de um átomo do primeiro deles
$I(Q_i; Q_{i+1}\|B_1, B_{i+1})$	Transinformação entre as identidades de dois resíduos vizinhos condicionada ao enterramento de um átomo de cada um deles
$I(Q_i; S_i, S_{i+1})$	Transinformação entre a identidade de um resíduo e a estrutura secundária dele e do resíduo seguinte

$I(Q_{i+1}; S_i, S_{i+1})$	Transinformação entre a identidade de um resíduo e a estrutura secundária dele e do resíduo anterior
$I(Q^N; B_0)$	Transinformação entre um bloco de tamanho N de identidades de resíduos e o enterramento de um dos átomos de seu resíduo central
$I(Q^N; B^N)$	Transinformação entre blocos de tamanho N de identidades de resíduos e enterramentos atômicos de um de seus átomos
$I(S_i; S_{i+1})$	Transinformação entre estruturas secundárias de resíduos vizinhos
$I(X; Y)$	Transinformação entre as variáveis X e Y
$I(X^N; Y^N)$	Transinformação entre blocos de tamanho N da variável X e da variável Y
K_B	Constante de Boltzman
LBTC-UnB	Laboratório de Biologia Teórica e Computacional da Universidade de Brasília
m	Tamanho do conjunto amostral $\mathcal{A}$
$m(X^N, Y^N)$	Frequência do par de blocos X^N, Y^N
P	Pressão
$p(x)$	Probabilidade da variável X assumir determinado valor x
$p(x, y)$	Probabilidade conjunta das variáveis X e Y assumirem determinados valores x e y respectivamente
$p(x\|y)$	Probabilidade da variável X assumir determinado valor x dado que a variável Y assumiu determinado valor y
$p(x^N)$	Probabilidade de um bloco de tamanho N da variável X assumir uma sequência de valores x
PDB	*Protein Data-Bank*
Q	Estrutura primária de proteínas
Q_i	Identidade do i-ésimo resíduo de uma proteína
Q_{i+1}	Identidade do $i+1$-ésimo resíduo de uma proteína
r_0	Raio do nível zero de enterramentos atômicos
r_1	Raio do nível um de enterramentos atômicos
r_2	Raio do nível dois de enterramentos atômicos
S	Entropia (Termodinâmica) ou Estrutura secundária de proteínas
T	Temperatura
V	Volume
X	Variável aleatória discreta X
x	Valor assumido pela variável X
x^N	Bloco de tamanho N de valores x assumidos pela variável X
X_n	n-ésima variável X de um bloco
X^N	Bloco de tamanho N formado pela variável X
Y	Variável aleatória discreta Y

y	Valor assumido pela variável Y
α	Parâmetro de ajuste
β	Parâmetro de ajuste
Δ	Variação
Ω	Número de microestados possíveis de serem assumidos por um sistema em determinadas condições

Sumário

1 Introdução

As proteínas estão entre as macromoléculas mais abundantes do planeta porque são presentes em praticamente todos os constituintes celulares e realizam as mais diversas funções nos seres vivos. Por participarem de inúmeros processos celulares, elas apresentam uma variabilidade muito grande de estruturas e funções, o que lhes confere grande interesse nas áreas médica, biotecnológica, industrial, de biocombustíveis, entre outras (<http://www.rcsb.org/pdb/home/home.do>), daí a importância de estudá-las.

Os constituintes básicos das proteínas são os aminoácidos, moléculas orgânicas que contém um grupo amina ($—NH_2$), um grupo carboxila ($—COOH$) e um grupo radical ($—R$), também chamado de cadeia lateral, que varia de acordo com o tipo de aminoácido conferindo-lhes propriedades químicas distintas que podem ser agrupadas em cinco grupos, de acordo com Nelson e Cox (2008). Para essa classificação são usadas características tais como tamanho, estrutura, polaridade e carga elétrica da cadeia lateral. Os grupos amina, carboxila e radical, além de um átomo de hidrogênio, são covalentemente ligados a um átomo de carbono quiral, chamado de carbono-α (C_α), que juntamente com o átomo de nitrogênio da amina e com o átomo de carbono do grupo carboxila formam o esqueleto peptídico ou cadeia principal (*Backbone* em inglês, que será denotada aqui por BB).

Em uma proteína, os aminoácidos estão ligados entre si formando uma cadeia ordenada, daí o nome de cadeia principal. A construção da mesma se dá pela formação de ligações peptídicas, que são ligações simples entre o carbono da carboxila de um aminoácido e o nitrogênio da amina do aminoácido seguinte, com a liberação de uma molécula de água, dando origem a um polímero de resíduos de aminoácidos (Nelson e Cox, 2008). Do conjunto dos heteropolímeros destes resíduos, aqueles que têm a capacidade de se enovelar para um estado de menor energia livre, chamado de estado nativo, e exercem função biológica são chamados de proteínas.

A estrutura de uma proteína pode ser dividida em quatro níveis de complexidade: estrutura primária, estrutura secundária, estrutura terciária e estrutura quarternária. A estrutura primária é a sequência de resíduos de aminoácidos ligados covalentemente na cadeia peptídica, que é determinada pela sequência de códons presentes no RNA mensageiro e por eventuais modificações pós-traducionais (Mathews *et al.*, 1999).

A estrutura secundária é formada por um arranjo local e particularmente estável da cadeia principal da proteína, sendo esta estabilidade conferida por pontes de hidrogênio entre os átomos pertencentes ao esqueleto peptídico. Ela pode estar presente em grande número ao longo da sequência e apresenta alguns padrões, sendo que os principais são o de α-hélice (estrutura helicoidal com periodicidade de 3,6 resíduos) e o de folha-β (estrutura em formato de vai-e-vem que pode ser paralela ou anti-paralela) (Alberts *et al.*, 2002; Kamtekar *et al.*, 1993; Mathews *et al.*, 1999). As regiões com ausência de estrutura secundária são chamadas de *loops*. A formação deste tipo de estrutura é possibilitada por

uma importante propriedade da ligação peptídica: ser uma ligação simples com caráter de ligação dupla. Essa característica é resultado de um dos elétrons da ligação dupla entre o átomo de carbono e o átomo de oxigênio da carboxila ser atraído pelo átomo de nitrogênio da ligação peptídica, ficando deslocalizado ao longo desta ligação, e tem como consequência o fato de a ligação peptídica ser plana. A planaridade desta ligação restringe os graus de liberdade da cadeia permitindo que o esqueleto peptídico gire somente em torno das ligações do C_α dando origem aos chamados ângulos diedrais, Φ e Ψ (Alberts *et al.*, 2002).

Diferentemente da estrutura sedundária, que é local, a estrutura terciária de uma proteína é a conformação tridimensional dos elementos da estrutura secundária e das cadeias laterais dos resíduos de aminoácidos, resultando no enovelamento da cadeia. Também pode ser chamada de estrutura nativa, ou de estado nativo. A estrutura terciária de uma proteína é determinada por sua estrutura primária, ou seja, uma sequência de aminoácidos adota apenas um padrão de enovelamento, embora sequências bastante distintas possam assumir conformações parecidas. Por fim, o grau mais complexo de organização de uma proteína é a estrutura quaternária, que está presente somente em proteínas formadas por mais de uma cadeia polipeptídica e é entendida como a organização espacial dessas várias cadeias (Mathews *et al.*, 1999; Nelson e Cox, 2008).

Um fato interessante sobre o estado nativo de uma proteína é que algumas cadeias parecem ter um estado nativo quando em monômeros e outro quando em aglomerados. Isso é esperado para proteínas oligoméricas e deve ser o responsável pela oligomerização em alguns casos, entretanto esse fato também pode ser observado em proteínas monoméricas, como a proteína do príon, cujo estado agregado das proteínas em fibras gera a patologia encefalopatia espongiforme bovina, conhecida como doença da vaca louca. Acredita-se que proteínas com este comportamento tenham dois mínimos locais de energia com uma barreira energética razoavelmente pequena entre eles e que a transição entre estes dois estados dependa do ambiente (Chiti e Dobson, 2006).

1.1 Enovelamento de proteínas

Na década de 1960, já era conhecido o fato de uma sequência de aminoácidos adotar um estado nativo, o que não se sabia era como a proteína encontra esse estado. Poderia-se pensar que ela percorre aleatoriamente todas as suas conformações possíveis até encontar a estrutura nativa, mas no ano de 1969 Cyrus Levinthal mostrou que o processo do enovelamento proteico não deve ocorrer de forma aleatória na natureza (Levinthal, 1969). Ele chegou a esta conclusão estimando o número de graus de liberdade de uma proteína formada por 150 resíduos, o que o levou a perceber que o tempo que ela levaria para percorrer todas as conformações até encontrar seu estado nativo seria muito maior do que o tempo que é de fato observado para tal, a isso dá-se o nome de Paradoxo de Levinthal.

No mesmo artigo, o autor sugere que o enovelmento proteico deve ser guiado por interações locais que determinam o caminho de dobramento. Hoje entende-se este processo como regido pelo chamado Colapso Hidrofóbico, que será discutido adiante.

Um experimento que ficou famoso por provar que a sequência de aminoácidos de uma proteína se enovela espontaneamente para uma estrutura nativa foi realizado por Christian Anfinsen, no ano de 1973 (Anfinsen, 1973; Lodish *et al.*, 2000). O autor colocou ribonuclease-A prancreática bovina na presença de ureia e de mercaptoetanol, que são agentes caotrópico e redutor respectivamente, e observou a perda de sua atividade catalítica por conta da desnaturação. Uma vez que esses agentes foram retirados, a proteína recuperou sua atividade, o que evidencia que ela recuperou também sua estrutura nativa. Esse experimento deu origem à Hipótese Termodinâmica do enovelamento protéico. Essa hipótese propõe que a estrutura nativa de uma proteína é aquela capaz de minimizar a energia livre de Gibbs do sistema (Anfinsen, 1973) e é a mais aceita atualmente.

No passado acreditava-se que o enovelamento de proteínas fosse regido por um grande número de pequenas forças, como pontes de hidrogênio, interações iônicas, interações de van der Waals, mas atualmente é aceito que o componente que domina o processo de enovelamento protéico são as interações hidrofóbicas entre as cadeias laterais dos resíduos de aminoácidos e as moléculas de água do meio e entre as cadeias laterais entre si (Dill *et al.*, 2007; Dill *et al.*, 2008), o que é chamado de efeito hidrofóbico.

A explicação termodinâmica para tal efeito surge em termos da energia livre de Gibbs do sistema, conforme proposto por Anfinsen (1973). A entropia S , na termodinâmica, pode ser definida com o grau de desordem de um sistema e é dada por

$$S = K_B \ln \Omega,$$

onde K_B é a constante de Boltzman e Ω é o número de microestados do sistema acessíveis na condição considerada. Desta forma a entropia de um sistema cresce à medida que aumenta o número de microestados acessíveis, intuitivamente diz-se que quando isso ocorre há uma elevação no grau de desorganização do sistema. De acordo com a segunda lei da termodinâmica, em um sistema isolado, um processo espontâneo ocorre com o aumento da entropia.

Por um lado, a entropia é relacionada com a desorganização do sistema. Por outro lado, a entalpia H diz respeito à satisfação de necessidades energéticas do sistema, de modo que quando essas necessidades são satisfeitas a entalpia diminui. Ela é dada por

$$H = E + PV,$$

onde E, P e V são a energia, a pressão e o volume do sistema, respectivamente.

O aumento da entropia só pode ser considerado critério de espontaneidade de processos que ocorrem em sistemas isolados, e como a maioria dos sistemas biológicos não acontece

sob essas condições, a espontaneidade de um processo biológico é medida pela variação na energia livre de Gibbs do sistema, que é dada por

$$\Delta G = \Delta H - T\Delta S$$

e é o critério de espontaneidade em sistemas fechados com temperatura e pressão constantes. Quando essa variação é negativa, o processo é espontâneo; entretanto, quando ΔG é positiva o processo não deve ocorrer de forma espontânea sob as condições consideradas.

De acordo com a equação acima, uma variação positiva na entropia e uma variação negativa na entalpia contribuem para a diminuição da energia livre de Gibbs, sendo resultado de um processo espontâneo. Todavia, é possível que em determinados processos esses termos tenham contribuições opostas para o valor de ΔG, de modo que neste caso o que determina qual termo será dominante é a temperatura do sistema (Eisenberg e Crothers, 1979). Caso ela seja alta, a espontaneidade do processo será regida pela entropia, caso ela seja baixa, a entalpia deve assumir esse papel.

É facilmente observável que a grande maioria das proteínas se enovela corretamente em condições fisiológicas, o que leva a crer que a variação da energia livre de Gibbs desse processo no ambiente celular é menor do que zero. Contudo, a energia livre de Gibbs do estado nativo é apenas de 5 kcal/mol a 10 kcal/mol menor que a dos estados desnaturados (Dill *et al.*, 2008; Onuchic *et al.*, 1997), o que é aparentemente pouco, mas é suficiente para que o enovelamento seja algo inerente às proteínas. O fato de a variação na energia livre de Gibbs ser menor do que zero é resultado do aumento na entropia do sistema durante o enovelamento proteico, que ocorre por conta do efeito hidrofóbico, ou seja, por conta da necessidade que os resíduos hidrofóbicos tem de interagir uns com os outros e de evitar interação com a água (Mathews *et al.*, 1999).

O efeito hidrofóbico é resultante da tendência do sistema de maximizar sua entropia. Quando a proteína está desenovelada, as cadeias laterais hidrofóbicas dos resíduos de aminoácidos estão expostas ao solvente, então as móleculas de água se organizam em torno destas formando clatratos (Mathews *et al.*, 1999), o que reduz a entropia do solvente. Ainda que a entropia da cadeia peptídica desenovelada seja maior que a do estado nativo, a redução na entropia das moléculas de água resulta na redução da entropia do sistema, o que eleva a energia livre de Gibbs. Por outro lado, o enovelamento da cadeia reduz a necessidade da formação de clatrato pois os resíduos hidrofóbicos se agrupam no interior da proteína. A diminuição da entropia da cadeia peptídica é compensada pelo aumento de entropia das moléculas de água, o que eleva a entropia do sistema contribuindo negativamente para a variação da energia livre de Gibbs.

Uma proposta que visa facilitar o entendimento da energia envolvida no processo de enovelamento e da busca da proteína pelos estados de menor energia é o diagrama do funil de enovelamento, proposto por Dill e Chan (Dill e Chan, 1997; Dill *et al.*, 2008). Neste

diagrama, a distância conformacional dos estados possíveis da proteína é representada no eixo horizontal, que não possui orientação positiva, de forma que duas conformações distantes entre si ao longo do eixo horizontal são bastante diferentes umas das outras (sua distância conformacional é grande) enquanto conformações próximas entre si em relação ao eixo horizontal são muito parecidas. No eixo vertical é representada a energia livre, e a curva tem a forma de um funil irregular. A partir dessa imagem percebe-se que alta energia livre corresponde a alta distância conformacional e consequente alta entropia, permitindo à proteína assumir várias conformações; por outro lado, baixa energia livre corresponde a baixa distância conformacional (baixa entropia), o que a leva para o seu estado nativo (correspondente ao mínimo global do funil).

O diagrama do funil de enovelamento permite fazer uma distinção entre o enovelamento proteico e as reações químicas simples. Uma reação química parte de seu reagente R e chega a seu produto P através de uma sequência de estruturas únicas. O mesmo não acontece durante o enovelamento de uma proteína porque seu reagente, o estado desnaturado, é formado por várias estruturas microscópicas e não por uma só. Desta forma, o enovelamento é a transição de um estado desordenado para um estado ordenado, e não de uma estrutura para outra (Dill *et al.*, 2008).

Um ponto fundamental do diagrama do funil de enovelamento é ressaltar que não deve existir um único caminho que leva a proteína do seu estado desenovelado até a estrutura nativa, mas sim múltiplas rotas que visam minimizar a energia do sistema, ainda que uma delas seja mais comum de ocorrer (Dill *et al.*, 2008). Devido a isso, a proteína pode ficar presa em configurações com mínimos locais de energia e atingir estados com enovelamento incorreto, chamados de *misfolded* (Chiti e Dobson, 2006).

A pergunta que segue é o que acontece com a estrutura de uma proteína enquanto ela se dobra até assumir sua conformação nativa e há três principais hipóteses para explicar este processo. A primeira entende o enovelamento como resultado de uma série de eventos ordenados e organizados e é chamada de Hipótese do Glóbulo Fundido (*Molten-Globule* em inglês) (Dolgikhm 1981; Ogushi e Wada, 1983). Segundo ela, nos estágios iniciais do enovelamento ocorre a formação de segmentos locais de estrutura secundária (α-hélice e folha-β), que se agrupam e se organizam até atingir um estágio no qual a cadeia principal está em sua conformação final e as cadeias laterais estão livres. Essa conformação é o glóbulo fundido. Há, então, a organização rígida das cadeias laterais e a formação das pontes de hidrogênio internas, de modo que toda a água seja expelida do interior hidrofóbico (Mathews *et al.*, 1999).

Outra hipótese considera que deve haver a formação de um núcleo de enovelamento, o que propicia a aproximação física de resíduos que estão distantes ao longo da sequência e permite interações não covalentes específicas que levam à estrutura terciária nativa (Abkevich *et al.*, 1994). Esse núcleo é condição suficiente e necessária para o enovelamento pois sua formação faz com que o sistema supere a barreira energética do enovelamento

proteico. Devido a essa última característica, de acordo com esta hipótese, o mecanismo do enovelamento pode ser dividido em duas etapas. A primeira delas é a de formação do núcelo de enovelamento, o que ocorre de maneira estocástica à medida que a cadeia peptídica explora várias conformações. Uma vez que um contato correto é formado ele não se desfaz, de modo que os contatos não precisam se formar todos ao mesmo tempo. Essa primeira etapa é o fator limitante em relação ao tempo de enovelamento pois, como já dito, representa a fase em que o sistema precisa superar a barreira energética da reação. A segunda fase é determinística e rápida, consistindo na formação do restante dos contatos nativos (Abkevich *et al.*, 1994). Treptow e colaboradores (2002) reforçaram a necessidade da formação no núcleo de enovelamento através de um estudo computacional com modelos minimalistas de análise de valores de Φ.

As duas primeiras hipóteses discutidas presumem que o enovelamento é um processo organizado no qual a formação de certas estruturas depende da formação de outras. Por outro lado, a hipótese a do Colapso Hidrofóbico encara o enovelamento proteico como um evento desorganizado resultante das forças participantes do efeito hidrofóbico. De acordo com essa hipótese, dependendo do meio, dois regimes de enovelamento podem ser encontrados. O primeiro deles possui duas etapas: a primeira etapa se trata de um colapso rápido para uma estrutura compacta, seguida de uma etapa de busca lenta pela conformação nativa dentre as estruturas colapsadas. O segundo regime admite que o colapso e a busca pelo estado nativo acontecem simultaneamente (Gutin *et al.*, 1995).

Vale ressaltar que essas hipóteses para as rotas de enovelamento não são mutuamente exclusivas e que esse processo pode acontecer de várias formas, a depender do meio. As interações entre o solvente e as cadeias laterais dos aminoácidos levam à formação da estrutura nativa da proteína, que tem como uma de suas características a segregação espacial parcial nesta estrutura, de modo que os resíduos hidrofóbicos habitem o cerne hidrofóbico da mesma e os resíduos hidrofílicos fiquem na superfície (Kamtekar *et al.*, 1993). Este processo é resultante do balanço entre a entropia e a entalpia do sistema e a necessidade de se reduzir a quantidade de clatrato formada.

1.2 Predição da estrutura terciária de proteínas

"Can we predict how proteins will fold? Out of a near infinitude of possible ways to fold, a protein picks one in just tens of microseconds. The same task takes 30 years of computer time."(Science, 2005)

O experimento de Anfinsen (1973) demonstrou que toda a informação necessária para a determinação da estrutura terciária de uma proteína está em sua sequência de aminoácidos (Fasman, 1989) e, por isso, é possível se predizer aquela a partir desta. Desde então a biofísica molecular tenta entender como essa informação está armazenada ao longo da cadeia peptídica e como ela pode ser interpretada pelos algoritmos de predição de estru-

tura terciária para que seja possível prever o estado nativo de uma proteína conhecendo somente sua sequência de aminoácidos. A solução desse problema pode acelerar a anotação funcional de proteínas em genomas recém sequenciados, o descobrimento de novas drogas (Dill *et al.*, 2008) e o desenho de drogas que produzam menos efeitos colaterais (Mathews *et al.*, 1999).

Os primeiros estudos nessa área foram feitos com o uso de modelos minimalistas. Esses modelos simplificam consideravelmente o problema considerando apenas certas variáveis do sistema que sejam suficientes para se aproximar o modelo da realidade tanto quanto seja necessário ou pretendido. Por exemplo, consideram o ambiente tridimensional a ser explorado pela cadeia peptídica como uma rede quadrada ou cúbica e os resíduos de aminoácidos como esferas com características próprias, como a tendência específica de se aproximarem umas das outras (Karanicolas e Brooks, 2004). A partir do uso desses modelos foi possível entender várias propriedades gerais do enovelamento, tais como compactação da cadeia, a segregação dos resíduos em relação à exposição ao solvente, entre outras.

Existem duas categorias de algoritmos de predição de estrutura terciária de proteínas. A primeira dessas categorias é a de predição por homologia (Hardin *et al.*. 2002). Quando a proteína a ser predita tem sequência homóloga a outra proteína de estrutura já conhecida, o trabalho de predição é facilitado pois o algoritmo pode fazer um alinhamento entre as duas e estimar a conformação da proteína alvo, ou seja, proteínas já conhecidas podem servir de molde para as novas estruturas (Mathews *et al.*, 1999). Os grupos que trabalham com essa metodologia obtêm resultados bastante bons, mas são alvo de uma crítica constante: para que a predição seja feita é necessário que o banco de dados de estruturas tenha sequências conhecidas semelhantes à que está sendo buscada, além disso, a qualidade da estrutura cristalizada influencia nos resultados da predição, reduzindo sua confiabilidade. Um outro problema igualmente importante é que a predição da estrutura terciária de proteínas feita por homologia não é capaz de prever novas famílias estruturais devido ao fato de estar sempre atrelada ao que já é conhecido e está presente no banco de dados.

A segunda categoria de algoritmos compõe a chamada predição *ab initio*, que utiliza para fazer a predição apenas informações que podem ser obtidas apenas a partir da estrutura primária da proteína, (Hardin *et al.*. 2002; Pereira de Araújo *et al.*, 2008), ou seja, não se baseia em uma estrutura terciária já conhecida. O desenvolvimento de algoritmos que utilizem a metodologia *ab initio* depende do desenvolvimento de um potencial adequado e de um protocolo eficiente de busca pelo estado de menor energia no perfil energético resultante (Bonneau e Baker, 2001). A forma mais precisa de se fazer uma predição *ab initio* é a partir de resultados de cálculos que levam em conta as propriedades físico-químicas dos orbitais atômicos dos resíduos de aminoácidos e de suas interações. Esta abordagem deveria trazer os melhores resultados pois todas as variáveis do sistema

são consideradas, entretanto, ela é computacionalmente dispendiosa, requerendo computadores com grande capacidade de processamento e, por isso, é inviável de ser feita com os computadores atuais. Devido a este fato, os algoritmos que trabalham com esse tipo de predição usam campos de força clássicos, da mecânica molecular, nos quais o sistema é visto como um arranjo newtoniano de massas pontuais unidas por molas, o que é muito mais simples de ser modelado do que orbitais moleculares (Leach, 2001).

Para estimular este ramo da ciência, em 1994, John Moult criou o CASP (*Critical Assessment of Techniques for Protein Structure Prediction* — Avaliação Crítica de Técnicas para Predição de Estrutura de Proteínas em inglês), uma competição bi-anual de predição de estruturas de proteínas. São escolhidas sequências cuja estrutura foi determinada experimentalmente e não foi divulgada e os grupos de pesquisa devem tentar prevê-la com o máximo de acurácia possível. Ao longo dos anos houve melhora significativa nas estruturas com grau de dificuldade considerado intermediário, mas a qualidade da predição de estruturas fáceis e difíceis não foi alterada (Dill e Chan, 1997; Dill *et al.*, 2008).

O desafio, então, continua sendo otimizar os algoritmos de predição de estrutura terciária de proteínas ou pensar em outra forma de se extrair informação a partir da sequência. É neste ponto que os enterramentos atômicos se apresentam como alternativa viável.

Em 2008, Pereira de Araújo e colaboradores propuseram os enterramentos atômicos, entendidos como a distância de cada átomo ao centro geométrico da proteína, como uma forma de se chegar à estrutura terciária das proteínas e mostraram que isso é possível. Eles assumiram que estas distâncias teriam relação com a hidrofobicidade das cadeias laterais dos resíduos e que a informação necessária para determiná-las estaria na sequência. Em 2009, Pereira de Araújo, com outra colaboração, concluiu que, apesar de a sequência não ser capaz de abrigar informação suficiente para indicar as coordenadas atômicas de cada um dos átomos, ela pode carregar quantidade suficiente de informação que indique os níveis de enterramentos atômicos com acurácia o bastante para levar ao correto dobramento da proteína. Usando elementos da Teoria da Informação que serão elucidados adiante, Pereira de Araújo e Onuchic (2009) mostraram que a sequência pode carregar informações para a determinação, não da a estrutura terciária, mas dos enterramentos atômicos e que é possível a partir destes chegar ao estado nativo (Pereira de Araújo, 2008).

O objetivo deste trabalho é entender, à luz da Teoria da Informação, como a informação sobre os enterramentos atômicos está alojada na estrutura primária para, então, fornecer subsídios para o desenvolvimento de algoritmos que sejam capazes de prever os enterramentos atômicos e, com base nesses, a estrutura nativa das proteínas.

2 Teoria da Informação

A Teoria da Informação foi proposta pelo engenheiro americano Claude Shannon em 1948, com a publicação do artigo *A mathematical theory of communication*. Esta teoria foi desenvolvida, a princípio, para o estudo de problemas relativos à engenharia de comunicação, tais como a transmissão de dados entre uma fonte e um receptor ou a criptografia, e usa certas grandezas estatísticas como base, por exemplo probabilidade e probabilidade condicional.

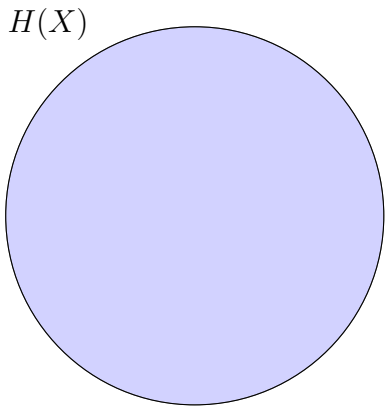
O conceito mais fundamental da teoria da informação é o de entropia. Segundo Shannon (1948), ela mede a informação, assim como os graus de escolha e de incerteza de um sistema. Por exemplo, ao se lançar uma moeda não viciada há duas possibilidades equiprováveis de resultado: cara ou coroa. Como as possibilidades de resultado desse lançamento são equiprováveis, a dúvida em relação ao seu resultado é grande, assim como a informação que pode ser retirada dele. Essa dúvida pode ser medida por

$$H(X) = -\sum_{x \in \mathcal{A}_X} p(x) \log_2 p(x), , \tag{1}$$

onde $H(X)$ é a dúvida em relação ao resultado e $p(x)$ é a probabilidade de cada resultado. Como $p(x)$, nesse caso, é igual a $0,5$ para todas as possilidades, a medida da dúvida é igual a 1 bit.

A mesma idéia pode ser aplicada ao lançamento de um dado não viciado, no qual a probabilidade de cada uma das faces cair voltada pra cima é igual a $\frac{1}{6}$. Nesse caso, a entropia é aproximadamente $2,6$ bit.

Sistematizando o conceito de entropia, ela é definida para variáveis aleatórias discretas e pode ser escrita matematicamente pela equação 1, onde $H(X)$ representa a incerteza acerca da variável X, $\mathcal{A}_X$ representa o alfabeto de valores possíveis de X, onde cada valor pode ser simbolizado por x, e $p(x)$ é a probabilidade associada a cada um desses valores. Pode-se representar graficamente cada uma das medidas da teoria da informação pela área de um Diagrama de Venn (Yeung, 1991). A entropia, então, deve ser representada como a seguir.



A entropia máxima é atingida quando a incerteza do sistema é máxima; ou seja, quando os valores possíveis de uma variável são equiprováveis. Neste caso, a entropia é igual a $H(X) = log_2 |\mathcal{A}_X|$, onde $|\mathcal{A}_X|$ é o tamanho do alfabeto. Ao contrário, a entropia é igual a zero quando, entre todos os possíveis, um determinado valor tem 100% de chance de ocorrer, ou seja, quando não há incerteza com relação a essa variável. Assim,

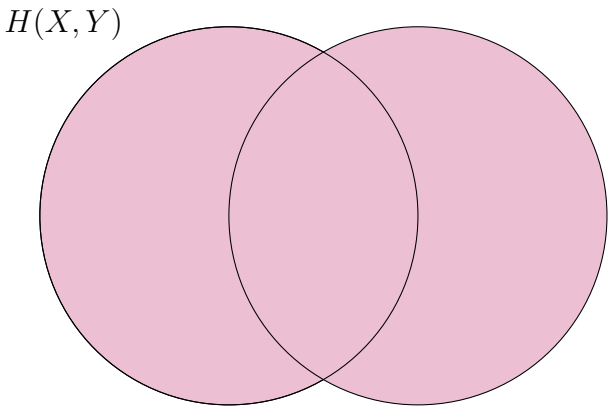
$$0 \leq H(X) \leq \log_2 |\mathcal{A}_X| \tag{2}$$

Assim como é possível falar de probabilidade, probalidade conjunta e probabilidade condicional para duas ou mais variáveis, também pode-se falar em entropia, entropia conjunta e entropia condicional para variáveis aleaórias discretas usando-se a notação correlata. Dessa forma, a entropia conjunta de duas variáveis aleatórias X e Y pode ser definida como

$$H(X,Y) = -\sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} p(x,y) \log_2 p(x,y), \tag{3}$$

e entendida como, por exemplo, o resultado do lançamento de uma moeda e de um dado. Existe uma entropia associada ao lançamento da moeda e uma entropia associada ao lançamento do dado, assim como uma entropia associada ao lançamento dos dois objetos juntos, pode ou não ser a soma das entropias individuais. Caso a entropia conjunta for igual à soma das entropias inviduais, as variáveis em questão não s ao correlacionadas entre si. Por outro lado, se a entropia conjunta for menor que a soma das entropias inviduais, as variáveis possuem correlação.

Graficamente, a entropia conjunta pode ser representada pela área colorida a seguir.

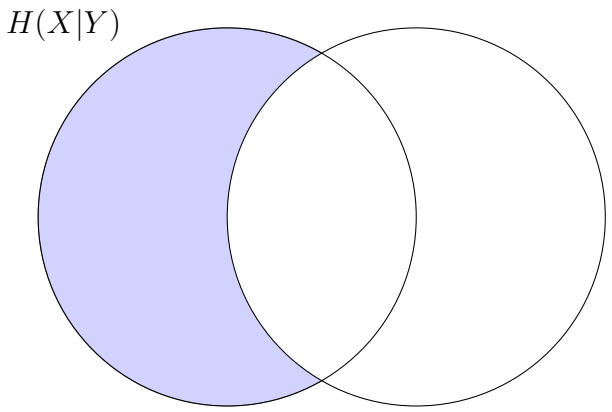


A entropia condicional é entendida como a dúvida que resta sobre uma variável, uma vez que outra variável é conhecida. No exemplo do lançamento da moeda e do dado, se o resultado *cara* da moeda condicionasse o resultado do dado somente às faces pares, por exemplo, e o resultado *coroa* às faces ímpares do dado, uma vez conhecido o resultado do lançamento da moeda, a dúvida em relação ao resultado do dado diminuiria. Assim, a entropia condicional é sempre menor ou igual á entropia de certa variável, não sofrendo redução quando as variáveis em questão forem independentes. Ela é definida por

$$H(X \,|\, Y) = H(X,Y) - H(Y). \tag{4}$$

$$= -\sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} p(x,y) \, \log_2 p(x|y), \tag{5}$$

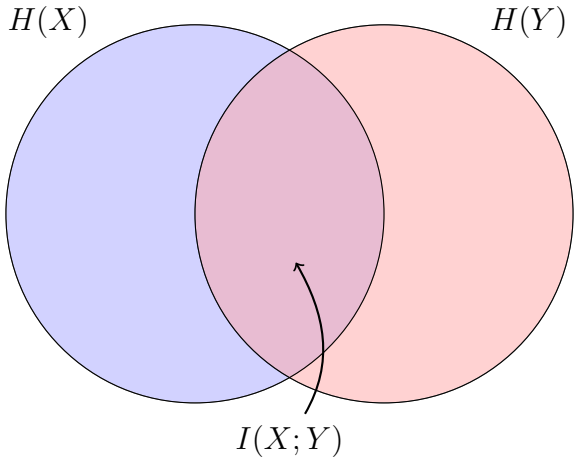
onde $H(X \,|\, Y)$ é a entropia condicional de X dado Y, e pode ser visualizada como a área de cor azul abaixo:



Uma vez que a entropia condicional pode ser menor que a entropia associada a uma variável, essa redução é também medida por uma grandeza da Teoria da Informação chamada de transinformação, ou informação mútua. Ela mede quanta informação uma variável guarda a respeito da outra e é definida matematicamente por

$$I(X;Y) = \sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} p(x,y) \, \log_2 \, \frac{p(x,y)}{p(x)\, p(y)}., \tag{6}$$

e pode ser representada pela intersecção entre os conjuntos abaixo.



Por exprimir uma relação entre as variáveis, a transinformação é igual a zero quando as variáveis são completamente independentes e é máxima quando as variáveis são perfeitamente correlacionadas, sendo igual à medida da entropia da variável menos incerta entre elas (graficamente pode ser representada por um conjunto completamente contido no outro).

A transinformação também pode ser expressa de outras formas, que a relacionam com as entropias simples, conjunta e condicional das variáveis.

$$\begin{aligned} I(X;Y) &= H(X) - H(X \,|\, Y) && (7)\\ &= H(Y) - H(Y \,|\, X) && (8)\\ &= H(X) + H(Y) - H(X,Y). && (9) \end{aligned}$$

Quando de se tratam de blocos de variáveis, de sequências de símbolos obtidos a partir de uma variável aleatória, a entropia dos blocos de tamanho N é definida como

$$H(X^N) = - \sum_{x^N \in (\mathcal{A}_X)^N} p(x^N) \, \log_2 \, p(x^N)\,, \qquad \text{onde } N > 0 \tag{10}$$

e assume valores

$$0 \leq H(X^N) \leq N. \log_2 \, |\mathcal{A}_X| \tag{11}$$

sendo que a igualdade inferior acontece quando somente uma possibilidade tem 100% de chance de ocorrer e a igualdade superior acontece em distribuições uniformes (Crutchfield e Feldman, 2003).

O crescimento da entropia à medida que cresce uma sequência de variáveis aleatórias é dado pela densidade de entropia, que pode ser definida matematicamente de duas formas. A primeira delas é medida com base nas entropias condicionais do bloco, sendo formulada como se segue:

$$h'(X) = \lim_{N\to\infty} H(X_N \,|\, X_{N-1}, X_{N-2} \ldots X_1), \tag{12}$$

quando o limite existe. Ela mede a entropia da última variável do bloco dadas todas as anteriores (Cover e Thomas, 2006).

A segunda forma de se calcular a densidade de entropia é a partir do valor médio da entropia do bloco de variáveis, que é calculada da seguinte forma

$$h(X) = \lim_{N\to\infty} \frac{1}{N} H(X_1, X_2, X_3 \ldots X_N) \tag{13}$$

$$= \lim_{N\to\infty} \frac{H(X^N)}{N} \tag{14}$$

quando o limite existe (Cover e Thomas, 2006). Para valores pequenos de N, essa entropia é melhor representada por

$$h(X) = \frac{H(X^N) - E_h}{N} \tag{15}$$

onde E_h representa o excesso de entropia, que mede quantidade de informação extraída das relações dentro de uma sequência (Crooks e Brenner, 2004). No limite de $N \to \infty$, $h(X)$ e $h'(X)$ tem o mesmo valor (Cover e Thomas, 2006).

Neste contexto, a transinformação também pode ser tratada em termos de densidade de transinformação. Assim,

$$I(X^N; Y^N) = H(X^N) - H(X^N \,|\, Y^N) \tag{16}$$

$$= H(Y^N) - H(Y^N \,|\, X^N) \tag{17}$$

$$= H(X^N) + H(Y^N) - H(X^N, Y^N) \tag{18}$$

e para um bloco X^N onde cada um dos elementos é independente dos demais

$$I(X^N;Y^N) = N.I(X;Y^N). \tag{19}$$

Então,

$$i(X;Y) = \lim_{N\to\infty} I(X;Y^N), \tag{20}$$

se $H(X^N) = N.H(X)$. A densidade de transinformação mede o limite para o qual a transinformação tende quando se aumenta a sequência de variáveis consideradas.

Vale ressaltar que $I(X;Y^N) \neq I(X^N;Y)$ e que o limite, na equação acima, deve ser calculado na variável correlacionada e não na variável não correlacionada. Enquanto na primeira situação, quando $N \to \infty$, obtém-se a densidade de transinformação, na segunda o resultado é a medida da transinformação entre um bloco de X e um elemento de Y desconsiderando-se para tal cálculo as correlações existentes em Y. Matematicamente pode ser calculado da seguinte forma:

$$\begin{aligned} I(X^N;Y) &= H(X^N) - H(X^N \,|\, Y) && (21)\\ &\geq NH(X) - [H(X_1 \,|\, Y) + \cdots + H(X_n \,|\, Y)] && (22)\\ &\geq \sum_{n=1}^{N} [H(X) - H(X_n \,|\, Y)] && (23)\\ &\geq \sum_{n=1}^{N} I(X_n;Y) && (24) \end{aligned}$$

A desigualdade é observada quando a variável X não é correlacionada entre si, mas tem correlação quando condicionada à variável Y.

3 Objetivos

3.1 Objetivo geral

- Contribuir para a predição de estruturas terciárias de proteínas a partir da sequência de aminoácidos, usando como ferramenta os enterramentos atômicos.

3.2 Objetivos específicos

- Calcular a entropia e a densidade de entropia para estruturas primárias, secundárias e enterramentos atômicos em proteínas globulares compactas;
- Medir correlações nestes ambientes;
- Calcular a transinformação entre sequência e enterramentos atômicos considerando vários alfabetos diferentes para cada um deles;
- Medir qual o máximo de informação sobre os enterramentos atômicos pode ser extraído da sequência proteica.

4 Metodologia

4.1 Bancos de dados

4.1.1 PDB-select

O maior banco de dados de estruturas de proteínas disponível atualmente é o Protein Data-Bank (PDB) (<http://www.rcsb.org/pdb/home/home.do>), que conta com aproximadamente 75.000 entradas relativas a proteínas e complexos proteicos. Por conta de todo este volume de informação, é natural que haja redundância entre as estruturas ali presentes e é essa questão que o PDB-select (Griep e Hobohm, 2009) visa melhorar.

O PDB-select é um banco de dados de proteínas criado em 1992 e baseado no PDB, mas no qual são aceitas sequências que tenham homologia de sequência igual ou inferior a 25%, de forma a maximizar a cobertura e reduzir a redundância. Para a obtenção do banco de dados usado neste trabalho foi usada a versão de novembro de 2009 do PDB-select (Griep e Hobohm, 2009).

4.1.2 Filtros aplicados ao banco de dados

Das estruturas presentes no banco de dados referido, foram escolhidas as que atendessem aos seguintes critérios: a) ter resolução melhor ou igual a $2,5$Å; b) ter sido resolvida por difração de raios-X; c) não ser relativa a alguma proteína de membrana; e d) atender ao critério de globularidade descrito por Gomes e colaboradores (2007). Assim, o banco de dados usado conta com 1499 proteínas, em um total de aproximadamente 263.000 resíduos.

A preocupação com as características experimentais da resolução da estrutura tridimensional da proteína ao se montar o banco de dados tem o objetivo de representar da maneira mais fiel possível a variedade estrutural encontrada na natureza, por isso a escolha de estruturas com alta resolução. Pelo mesmo motivo deu-se a seleção de estruturas resolvidas por difração de raios-X, uma vez que por ressonância magnética nuclear não é possível resolver a estrutura de proteínas com número de resíduos elevado. O fato de os enterramentos atômicos terem papel fundamental neste estudo inviabiliza o trabalho com proteínas membranares ou não-globulares pois estas apresentam uma relação entre tipo de resíduo, acessibilidade ao solvente e enterramento atômico diferente das globulares não-membranares.

4.2 Alfabetos

Por conta do tamanho limitado do banco de dados, os diferentes tipos de resíduos de aminoácidos foram agrupados de acordo com sua hidrofobicidade em quatro alfabetos de tamanhos distintos, sendo três deles reduzidos, conforme mostra a Tabela 1. O

primeiro alfabeto, chamado de HP, classifica os resíduos em hidrofóbicos (H) ou polares (P). O segundo, HPN, classifica-os em hidrofóbicos (H), polares (P) ou neutros (N). Da mesma frma, o terceiro alfabeto, HPNhp, classifica-os em cinco grupos de acordo com a hidrofobicidade de cada resíduo.

Tabela 1: Divisão dos alfabetos de resíduos de aminoácidos de acordo com sua hidrofobicidade.

Resíduos	Alfabetos			
	HP	HPN	HPNhp	20 letras
ILE	H	H	H	I
LEU	H	H	H	L
PHE	H	H	H	F
TRP	H	H	H	W
VAL	H	H	H	V
CYS	H	H	h	C
MET	H	H	h	M
TYR	H	H	h	Y
ALA	H	N	N	A
GLY	H	N	N	G
HIS	P	N	N	H
SER	P	N	N	S
THR	P	N	N	T
ARG	P	P	p	R
ASN	P	P	p	N
GLN	P	P	p	Q
PRO	P	P	p	P
ASP	P	P	P	D
GLU	P	P	P	E
LYS	P	P	P	K

4.3 Enterramentos atômicos

É chamada de enterramento atômico a razão entre a distância de um átomo em questão até o centro geométrico da proteína e o raio de giro da proteína (Gomes *et al.*, 2007). Uma vez calculados os enterramentos atômicos de todos os átomos de interesse de uma proteína, os valores são discretizados em níveis (ou camadas) de enterramento, que são esferas concêntricas conténdo o mesmo número de átomos, ou seja, a distribuição dos átomos de todo o banco de dados é equiprovável (Gomes *et al.,* 2007). Neste trabalho foram utilizados de dois a dez níveis de enterramentos atômicos e a Fig. 1 é um exemplo da divisão espacial de uma proteína de acordo com esses níveis. Neste exemplo foram considerados em três níveis de enterramentos.

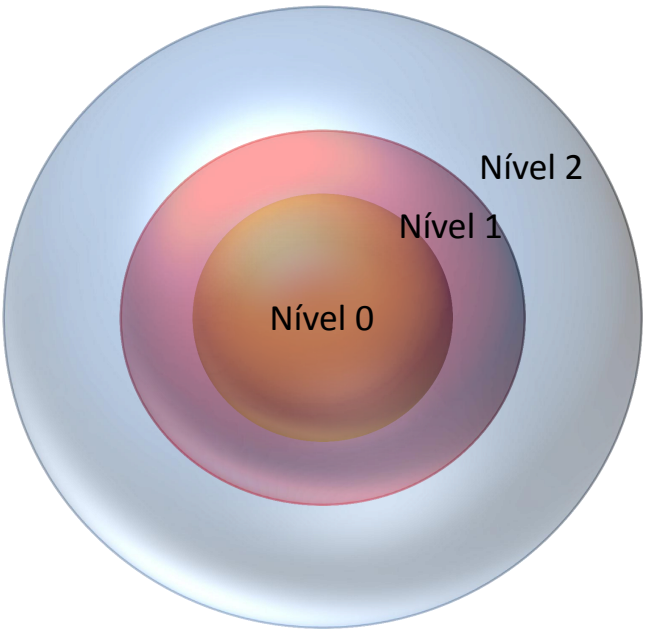


Figura 1: Organização dos níveis de enterramento atômico. Os níveis, ou camadas, consistem em esfera concêntricas e equiprováveis, ou seja, a probabilidade de um átomo estar um determinado nível é independente deste. Isso é obtido construindo esferas cujas medidas dos raios não são múltiplos inteiros da medida do raio da menor esfera, conforme mostrado na figura. Assim, $r_1 \neq 2r_0$ e $r_2 \neq 3r_0$, onde r_0, r_1 e r_2 são os raios das esferas correspondentes aos níveis 0, 1 e 2 respectivamente.

4.4 Estimativa das probabilidades

As probabilidades de ocorrência de cada evento foram estimadas a partir das frequências observadas no banco de dados, assumindo que as probabilidades são iguais às frequências corrigidas por *Bootstrap* e pseudocontagem.

4.4.1 *Bootstrap*

Por se tratar de um banco de dados composto por uma amostra da população de proteínas existentes, é natural que essa amostra apresente um desvio (*bias* em inglês) em relação à população. A estatística propõe uma técnica para a correção deste desvio, chamada de *Bootstrap*.

O conjunto amostral, neste caso o banco de dados considerado, será denotado por $\mathcal{A}$ e seu tamanho por m, de modo que cada elemento de $\mathcal{A}$ tem probabilidade igual a $\frac{1}{m}$. São escolhidos, ao acaso e com reposição, m elementos de $\mathcal{A}$, dando origem à primeira repetição do *Bootstrap*. São feitas quantas repetições forem necessárias com o objetivo de se obter a melhor estimativa do erro do conjunto amostral em relação à população. Em geral, 50 repetições são suficientes para se obter uma boa estimativa (Cover e Thomas, 2006; Crooks e Brenner, 2004).

Assim como o grupo amostral possui um desvio em relação à população, as repetições do *Bootstrap* possuem um desvio médio em relação à amostra, que é usado para estimar

o primeiro. Dessa forma, cada uma das quantidades medidas na amostra pode ter seu desvio corrigido pela média da mesma quantidade medida nas repetições do *Bootstrap*.

4.4.2 Pseudocontagem

A limitação no tamanho do banco de dados traz prejuízos aos cálculos de entropia e transinformação. Isso porque, à medida que o tamanho das janelas consideradas cresce, apenas algumas combinações de janelas estão presentes no banco de dados, e como considera-se que a probabilidade de ocorrência de determinado evento é igual à sua frequência no banco de dados, algumas das janelas possíveis passam a ter probabilidade estimada próxima ou igual a um e enquanto outras tem probabilidade estimada próxima ou igual a zero.

Para minimizar esse efeito, as probabilidades foram estimadas com o uso de pseudocontagem. A pseudocontagem atribui um peso para a frequência do elemento central da janela, de modo que quando o número de janelas contadas é grande (tamanhos de janela pequenos) esse peso tem efeito praticamente nulo, e quando o número de janelas é pequeno (tamanhos de janelas grandes) a pseudocontagem mimetiza uma contagem maior que a real, distanciando todas as probalidades dos extremos iguais a zero ou um.

A pseudocontagem foi feita atribuindo-se um peso igual a 20 para a frequência do elemento central da janela, de acordo com a equação a seguir.

$$p(X^N|Y^M) = \frac{m(X^N, Y^M) + 20p(X^N|Y_0)}{Y^M + 20}.$$

4.5 Equação usada para a estimativa da densidade de entropia a partir dos pontos obtidos

As densidades e excessos de entropia foram calculadas a partir de ajustes feitos ao gráfico correspondente, de acordo com a eq. 15, que pode ser reescrita como

$$H(X^N) = Nh(X) + E_h.. \tag{25}$$

4.6 Equações usadas para a estimativa da densidade de transinformação a partir dos pontos obtidos

As densidades de transinformação foram calculadas de quatro formas distintas, sempre com ajustes exponenciais ou sigmóides. A primeira forma é a mesma usada por Crooks e Brenner (2004) e pode ser escrita como

$$I(Q;B^N) = a - b \times e^{(-N/c)} \tag{26}$$

ou

$$I(Q;B^N) = i(X;Y) - b \times e^{(-N/c)}, \tag{27}$$

onde c é característico do tamanho de janela no qual o enterramento de um resíduo ainda pode ser sentido por outro, b tem relação com a importância dessa influência, e a é a densidade de transinformação entre a identidade do resíduo central e os enterramentos de uma janela com tamanho tendendo ao infinito. A forma exponencial dos ajustes foi obtida experimentalmente por Crooks e Brenner (2004).

A segunda forma também é uma exponencial, mas esta contém apenas dois parâmetros, o que permite um melhor ajuste quando se tem poucos pontos (Alon, 2006). Ela é escrita por

$$I(Q;B^N) = \frac{\beta}{\alpha} + (I(Q^1;B^1) - \frac{\beta}{\alpha}) \times e^{(-\alpha(N-1))} \tag{28}$$

ou

$$I(Q;B^N) = i(Q;B^N) + (I(Q^1;B^1) - i(Q;B^N) \times e^{(-\alpha(N-1))}, \tag{29}$$

pois agora a densidade de transinformação é a razão $\frac{\beta}{\alpha}$.

A terceira forma, assim como a segunda, possui apenas dois parâmetros. A diferença consiste em que nesta o primeiro ponto ($I(Q^1, B^1)$) não aparece multiplicando o termo exponencial, assim não influencia a forma da curva, enquanto na anterior ocorre o contrário. A terceira equação é dada por

$$I(Q;B^N) = I(Q^1;B^1) + \frac{\beta}{\alpha}(1 - e^{(-\alpha(N-1))}) \tag{30}$$

ou

$$I(Q;B^N) = I(Q^1;B^1) + i(Q;B^N)(1 - e^{(-\alpha(N-1))}). \tag{31}$$

Já a quarta equação é uma sigmóide dada por

$$I(Q;B^N) = a + \frac{b}{1+e^{-\alpha(N-\beta)}}, \tag{32}$$

onde a densidade de transinformação é igual a $a+b$, e foi usada porque se ajusta melhor em alguns conjuntos de pontos do que a eq. 26.

A escolha da equação utilizada foi feita com base no número de pontos disponíveis para o ajuste e na disposição desses pontos. Para os pontos obtidos sem a utilização de *Bootstrap* e pseudocontagem, a densidade de entropia foi calculada a partir das equações 26, 28 e 30 e foi considerada a estimativa produzida pela equação 30, pois é a que melhor se adequa a um ajuste feitos com poucos pontos. Já os pontos obtidos com a correção do *Bootstrap* e da pseudocontagem foram ajustados usando-se as equações 26 ou 32, de acordo com a disposição dos pontos.

4.7 *Scripts*

Alguns dos *scripts* utilizados aqui já haviam feito parte de outros trabalhos do Laboratório de Biologia Teórica e Computacional da Universidade de Brasília (LBTC-UnB), de modo que foram escritos pelos ex-alunos Me. Antonio Luiz Cruz Gomes e Paulo Henrique Azevêdo e pelo orientador deste trabalho Prof. Dr. Antônio Francisco Pereira de Araújo. Outros *scripts* foram escritos ao longo deste trabalho pela aluna que defende esta dissertação e pelo aluno de Iniciação Científica, Diogo César Ferreira.

Os *scripts* usados foram feitos em duas linguagens: *shell* e *awk*. Eles tratam o banco de dados para que sejam consideradas apenas as proteínas de interesse (globulares não-membranares) e para que eventuais falhas ao longo das cadeias peptídicas presentes nos arquivos do PDB-select (Griep e Hobohm, 2009) considerados não sejam computadas. Esses *scripts* montam os diferentes alfabetos de aminoácidos a partir do código de três letras de cada um, discretizam os enterramentos atômicos no número de níveis requerido, montam os blocos de tamanho crescente, fazem o *Bootstrap* e computam as frequências usando ou não pseudo-contadores.

5 Resultados

Primeiramente, alguns dos resultados obtidos por Crooks e Brenner (2004) foram reproduzidos para servirem de controle positivo da metodologia empregada. A Tabela 2 compara os valores de densidade de entropia obtidos no artigo citado e neste trabalho. Os valores da densidade de entropia da estrutura primária e do excesso de entropia da estrutura secundária estão de acordo com a literatura. No entanto, a densidade de entropia da estrutura secundária apresenta-se diferente, o que será justificado adiante.

Tabela 2: Comparação entre as densidades de entropia calculadas por Crooks e Brenner (2004) e as calculadas aqui.

Grandeza	Crooks e Brenner	Presente trabalho
Estrutura Primária		
Densidade de Entropia (bit)	$4,173 \pm 0,003$	$4,176 \pm 0,004$
Estrutura Secundária		
Densidade de Entropia (bit)	$0,598 \pm 0,001$	$0,6812 \pm 0,0007$
Excesso de Entropia (bit)	$0,997 \pm 0,006$	$0,943 \pm 0,005$

Para facilitar o entendimento, a partir deste ponto a estrutura primária será denotada por Q, a estrutura secundária por S e os enterramentos atômicos por B.

5.1 Entropia da estrutura primária

As densidades de entropia para cada um dos alfabetos de sequência, mencionados na sub-seção 4.2, foram calculadas a partir da eq. 15. Para tal foram construídas janelas crescentes a partir da estrutura primária das proteínas do banco de dados e a entropia de cada uma dessas janelas foi medida. A inclinação da reta que contém os pontos corresponde à densidade de entropia. A Tabela 3, a Figura 2 e a Figura 3 mostram os resultados obtidos. Assim como em todas as figuras que mostram pontos relativos às entropias, a suavização observada nas janelas de tamanho maior é resultado da saturação do banco de dados por conta de seu tamanho limitado.

Tabela 3: Densidade e excesso de entropia medidos para cada um dos alfabetos de sequência. É mostrada também a razão entre a densidade de entropia medida e a máxima possível (dada por $log_2|\mathcal{A}|$), que evidencia quando a densidade observada está próxima de seu máximo teórico.

Alfabeto	$H(Q_1)$	$h(Q)$ (bit/resíduo)	E_{h_Q} (bit)	$\frac{h(Q)}{log_2\|\mathcal{A}\|}$
HP	$1,00000 \pm 0,00009$	$0,9969 \pm 0,0002$	$0,0096 \pm 0,0007$	$0,9969 \pm 0,0002$
HPN	$1,5806 \pm 0,0004$	$1,5734 \pm 0,0007$	$0,018 \pm 0,003$	$0,9927 \pm 0,0007$
HPNhp	$2,187 \pm 0,002$	$2,1817 \pm 0,0004$	$0,011 \pm 0,001$	$0,9396 \pm 0,0004$
20 letras	$4,185 \pm 0,002$	$4,176 \pm 0,004$	$0,010 \pm 0,005$	$0,966 \pm 0,004$

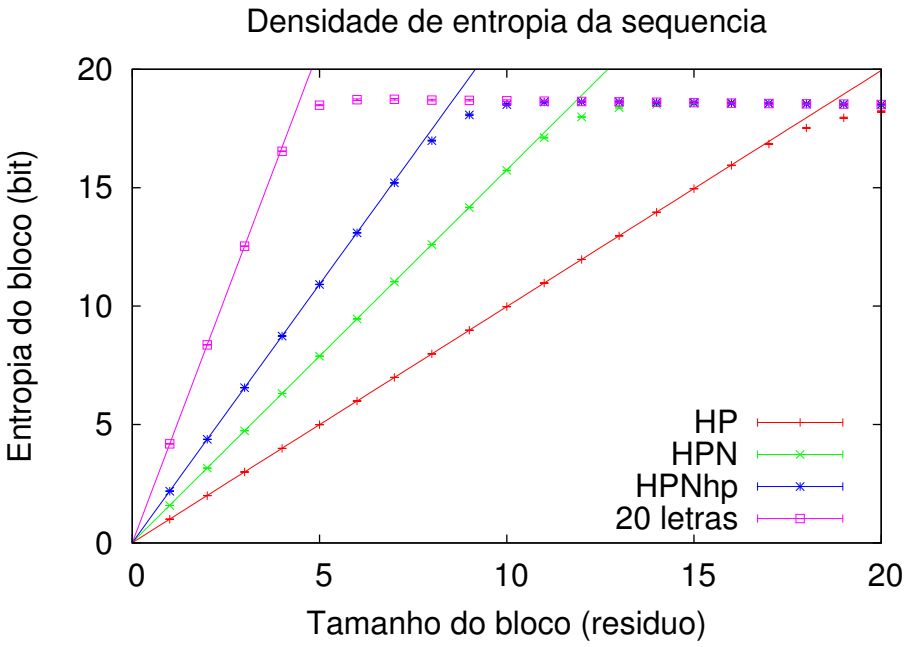


Figura 2: Densidade de entropia da sequência. Os pontos mostram as entropias calculadas para cada um dos alfabetos em cada um dos tamanhos de janela. A atenuação observada nas janelas maiores é resultado da saturação do banco de dados. As linhas foram ajustadas aos pontos para calcular as densidades e os excessos de entropia.

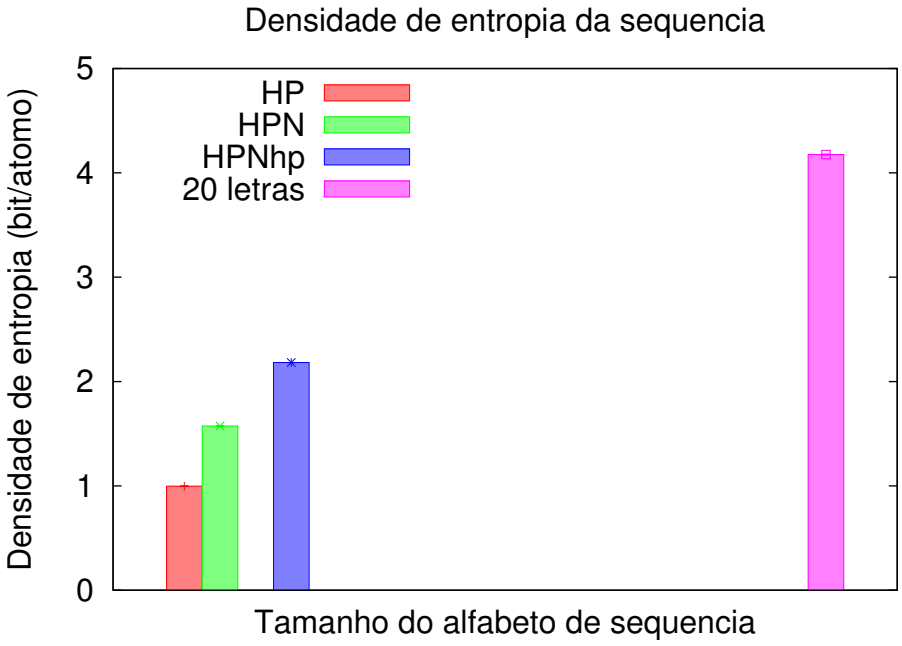


Figura 3: Densidade de entropia da sequência em função do tamanho do alfabeto da mesma.

Os resultados mostram que a densidade de entropia é muito próxima do log_2 do número de estados possíveis, ou seja, é próxima do seu valor teórico máximo. Além disso, o excesso de entropia é aproximadamente igual a zero, de forma que a dúvida sobre o primeiro elemento de um bloco é igual às dúvidas dos elementos seguintes. Esses dois resultados em conjunto evidenciam que os elementos da sequência não possuem correlação local.

Caso a distribuição de probabilidades da estrutura primária se afastasse muito de uma distribuição homogênea, a densidade de entropia calculada não seria próxima de $\log_2 |\mathcal{A}|$, mesmo que não houvesse correlação entre as identidades dos aminoácidos. Por isso é importante considerar a medida do excesso de entropia para avaliar a correlação presente em blocos de sequência. O mesmo deve ser feito em relação à estrutura secundária e aos enterramentos atômicos.

5.2 Entropia da estrutura secundária

A densidade de entropia da estrutura secundária foi calculada da mesma forma que a da estrutura primária (sub-seção 5.1). Entretanto, diferentemente da última, a entropia dos três símbolos constituintes da estrutura secundária não está de acordo com Crooks e Brenner (2004). Por conta disto, foi feito o cálculo da densidade de entropia da estrutura secundária usando-se o mesmo banco de dados de Crooks e Brenner (2004), sendo obtida uma medida muito próxima da que estes autores já haviam encontrado. A Tabela 4 contém os valores obtidos a partir destes ajustes.

Tabela 4: Densidades e excessos de entropia da estrutura secundária calculadas a partir de bancos de dados diferentes.

	Resultado de Crooks e Brenner	Banco de dados de Crooks e Brenner	Banco de dados deste
$h(S)$ (bit/resíduo)	$0,598 \pm 0,001$	$0,5948 \pm 0,0004$	$0,6812 \pm 0,0007$
E_{h_S} (bit)	$0,997 \pm 0,006$	$1,014 \pm 0,003$	$0,943 \pm 0,005$

Uma vez que, ao se utilizar o mesmo banco de dados que Crooks e Brenner (2004) já haviam utilizado, a densidade de entropia corresponde àquela da literatura, conclui-se que a diferença não se deve à técnica. Essa divergência entre a densidade encontrada na literatura e a calculada aqui pode ser explicada pela diferença entre os bancos de dados. Isso porque aqui o banco de dados usado é constituído apenas por proteínas globulares, enquanto o de Crooks e Brenner é formado por todos os tipos de proteínas (2004). Como a estrutura secundária deve variar mais em proteínas globulares do que em proteínas não globulares, por exemplo algumas proteínas filamentosas formadas somente por α-hélices, a entropia do banco de dados deste trabalho é maior.

Foi observada uma relação entre o tipo de estrutura secundária (α-hélice ou folha-β) ou a ausência dela (*Loop*) e o enterramento atômico de seus resíduos. Esta relação

é mostrada na Figura 4. Observa-se uma tendência de as α-hélices se distribuírem uniformemente pelo espaço tridimensional ocupado pela cadeia peptídica, enquanto as folhas-β se localizam mais próximas ao cerne hidrofóbico da proteína. Já os seguimentos sem estrutura secundária regular (*loops*) estão próximos da superfície da proteína. Essa relação sugere que os enterramentos atômicos são bons indicativos de como a proteína deve se organizar em seu estado nativo e serve de justificativa para este trabalho.

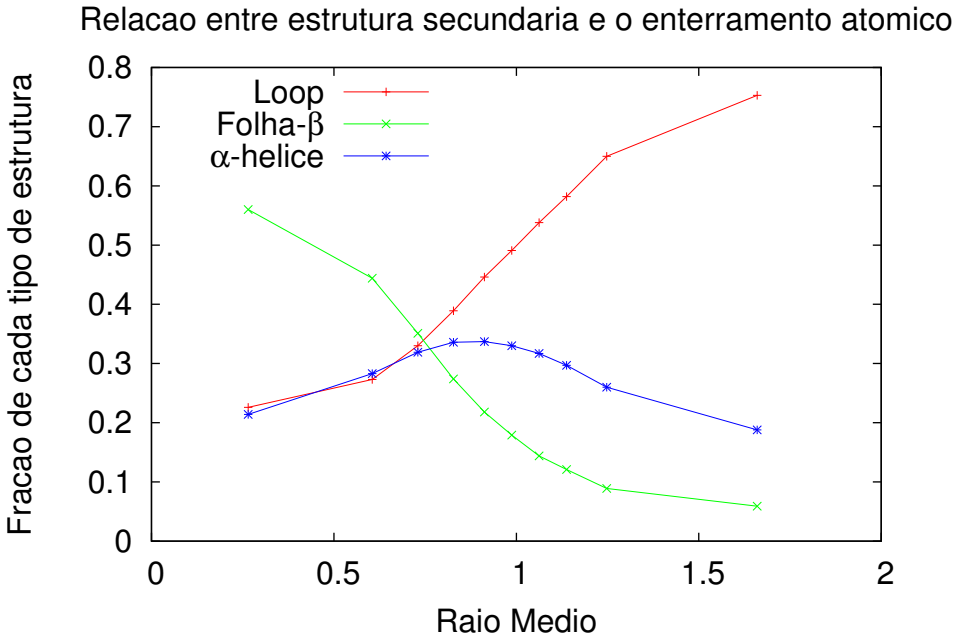


Figura 4: Fração de tipos diferentes em estrutura secundária em função do enterramento atômico do C_α do resíduo correspondente.

5.3 Entropia dos enterramentos atômicos

A densidade de entropia dos enterramentos atômicos foi calculada tanto para carbonos alfa (C_α) quanto para carbonos beta (C_β), bem como para os átomos do esqueleto peptídico (BB), como mostrado pela Tabela 5 e pelas Figuras 5, 6 e 7. A Tab. 5 mostra que a densidade de entropia dos enterramentos atômicos é menor que o máximo possível, o que demonstra uma correlação entre os enterramentos de átomos próximos.

Conforme dito anteriormente, a simples relação entre a medida da densidade de entropia e seu máximo teórico não é suficiente para constatação de presença ou ausência de correlações em determinado ambiente. O fato de a densidade de entropia dos enterramentos atômicos ser menor que o máximo possível pode ser devido a duas possibilidades não exclusivas. A primeira delas é que a distribuição dos enterramentos atômicos não seja equiprovável e a segunda é que haja correlação local entre os elementos da janela.

Sabe-se que o *script* usado para discretizar os enterramentos atômicos (classificaen-

Tabela 5: Entropias e excessos de entropia dos enterramentos atômicos de C_α, C_β e BB. A primeira coluna diz respeito ao número de níveis de enterramento e a última à razão entre a densidade de entropia medida e a máxima densidade de entropia possível.

Níveis	Carbono-α			
	$H(B_{C_\alpha}1)$	$h(B_{C_\alpha})$ (bit/átomo)	E_{h_B} (bit)	$\frac{h(B)}{log_2\lvert\mathcal{A}\rvert}$
2	$0,99978\pm0,00007$	$0,617\pm0,002$	$0,53\pm0,01$	$0,617\pm0,001$
3	$1,5797\pm0,0003$	$0,957\pm0,008$	$0,75\pm0,04$	$0,604\pm0,003$
4	$1,9972\pm0,0002$	$1,23\pm0,01$	$0,86\pm0,05$	$0,615\pm0,002$
5	$2,3110\pm0,0001$	$1,44\pm0,02$	$1,01\pm0,06$	$0,620\pm0,004$
6	$2,5797\pm0,0002$	$1,64\pm0,02$	$1,06\pm0,06$	$0,634\pm0,003$
7	$2,8039\pm0,0002$	$1,81\pm0,02$	$1,14\pm0,08$	$0,645\pm0,003$
8	$2,9940\pm0,0003$	$1,93\pm0,03$	$1,3\pm0,1$	$0,643\pm0,004$
9	$3,1279\pm0,0005$	$2,04\pm0,03$	$1,3\pm0,1$	$0,644\pm0,003$
10	$3,3122\pm0,0004$	$2,16\pm0,04$	$1,4\pm0,1$	$0,650\pm0,004$
Níveis	Carbono-β			
	$H(B_{C_\beta}1)$	$h(B_{C_\beta})$ (bit/átomo)	E_{h_B} (bit)	$\frac{h(B)}{log_2\lvert\mathcal{A}\rvert}$
2	$0,99847\pm0,00008$	$0,725\pm0,004$	$0,36\pm0,02$	$0,725\pm0,004$
3	$1,5800\pm0,0002$	$1,128\pm0,005$	$0,54\pm0,02$	$0,712\pm0,003$
4	$1,9933\pm0,0002$	$1,443\pm0,008$	$0,07\pm0,03$	$0,722\pm0,004$
5	$2,3154\pm0,0002$	$1,693\pm0,008$	$0,76\pm0,03$	$0,729\pm0,003$
6	$2,5786\pm0,0002$	$1,91\pm0,01$	$0,79\pm0,04$	$0,74\pm0,01$
7	$2,7958\pm0,0002$	$2,10\pm0,01$	$0,84\pm0,03$	$0,75\pm0,01$
8	$2,9915\pm0,0002$	$2,26\pm0,01$	$0,88\pm0,04$	$0,75\pm0,01$
9	$3,1264\pm0,0006$	$2,38\pm0,01$	$0,89\pm0,04$	$0,75\pm0,01$
10	$3,3119\pm0,0003$	$2,54\pm0,01$	$0,93\pm0,04$	$0,74\pm0,01$
Níveis	Esqueleto peptídico			
	$H(B_BB1)$	$h(B_{BB})$ (bit/átomo)	E_{h_B} (bit)	$\frac{h(B)}{log_2\lvert\mathcal{A}\rvert}$
2	$0,99991\pm0,00004$	$0,3327\pm0,0005$	$0,686\pm0,004$	$0,3327\pm0,0002$
3	$1,5786\pm0,0003$	$0,5283\pm0,0007$	$1,098\pm0,0005$	$0,3333\pm0,0002$
4	$1,9972\pm0,0002$	$0,684\pm0,002$	$1,40\pm0,01$	$0,3420\pm0,0005$
5	$2,3197\pm0,0002$	$0,807\pm0,002$	$1,65\pm0,02$	$0,3476\pm0,0004$
6	$2,5769\pm0,0004$	$0,903\pm0,004$	$1,86\pm0.02$	$0,3493\pm0,0007$
7	$2,8043\pm0,0002$	$1,018\pm0,003$	$1,95\pm0,02$	$0,3626\pm0,0004$
8	$2,9934\pm0,0004$	$1,096\pm0,005$	$2,09\pm0,03$	$0,3653\pm0,0006$
9	$3,1256\pm0,0005$	$1,176\pm0,005$	$2,12\pm0,02$	$0,3710\pm0,0006$
10	$3,3105\pm0,0004$	$1,267\pm0,003$	$2,19\pm0,01$	$0,3814\pm0,0003$

terramento.awk, detalhado no Apêndice 2) foi criado de forma a classificar os níveis de enterramento de modo que haja o mesmo número de átomos em cada nível, ou seja, o *script* discretiza os enterramentos de maneira equiprovável. Esse fato refuta a primeira hipótese e dá força à segunda. Além disso, o fato de o excesso de entropia ser maior do que zero reforça a hipótese já mencionada.

Ao se comparar as densidades de entropia de C_α, C_β e BB nota-se que a de BB é

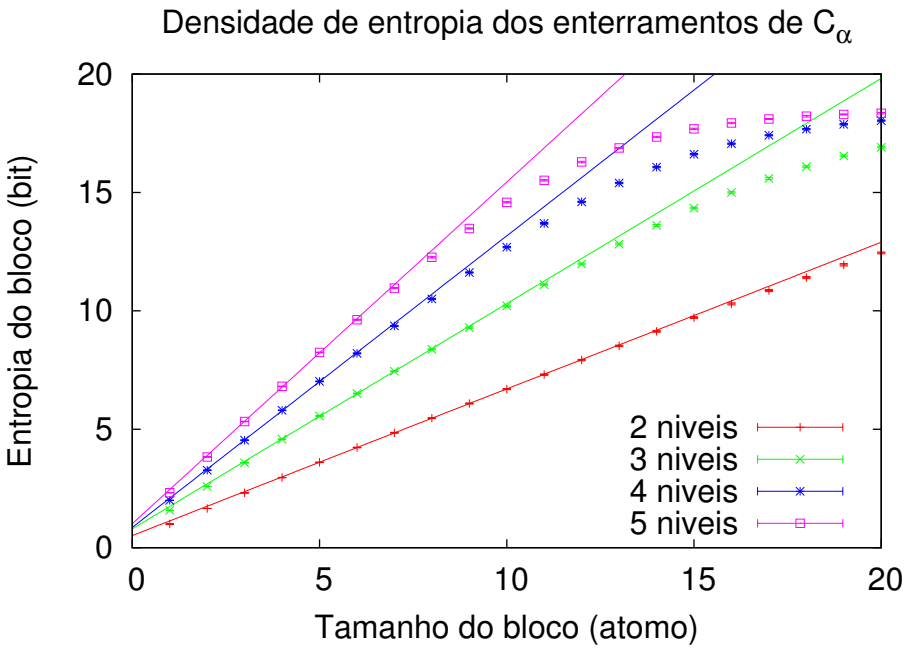


Figura 5: Densidade de entropia dos enterramentos atômicos de C_α quando considerados várias distribuições de níveis de enterramentos. Os pontos mostram as entropias calculadas para diferentes classificações de níveis de enterramento em função do tamanho de janela de enterramentos. As linhas foram ajustadas aos pontos para calcular as densidades e os excessos de entropia.

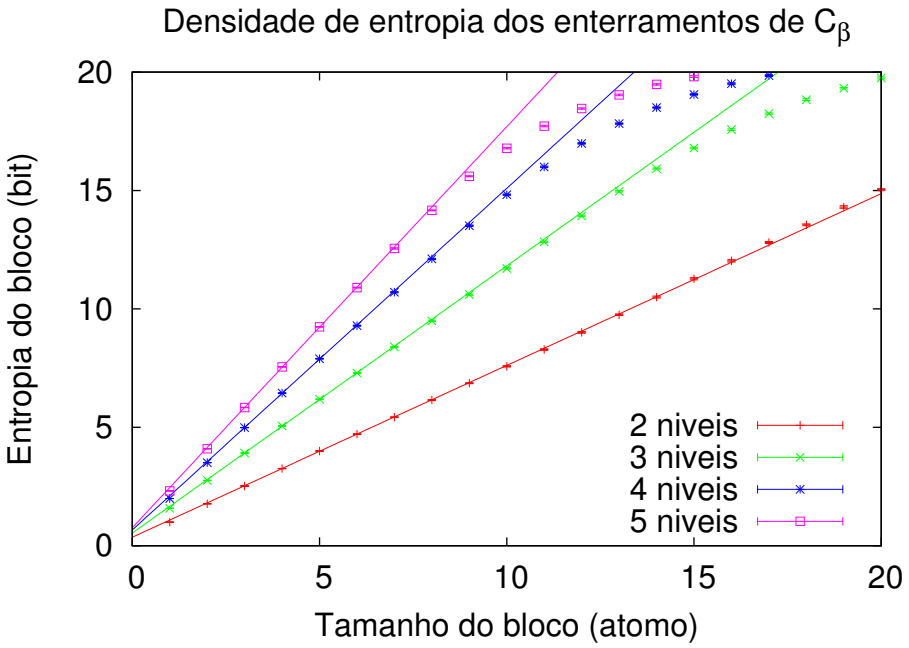


Figura 6: Densidade de entropia dos enterramentos atômicos de C_β quando considerados várias distribuições de níveis de enterramentos. Os pontos mostram as entropias calculadas para diferentes classificações de níveis de enterramento em função do tamanho de janela de enterramentos. As linhas foram ajustadas aos pontos para calcular as densidades e os excessos de entropia.

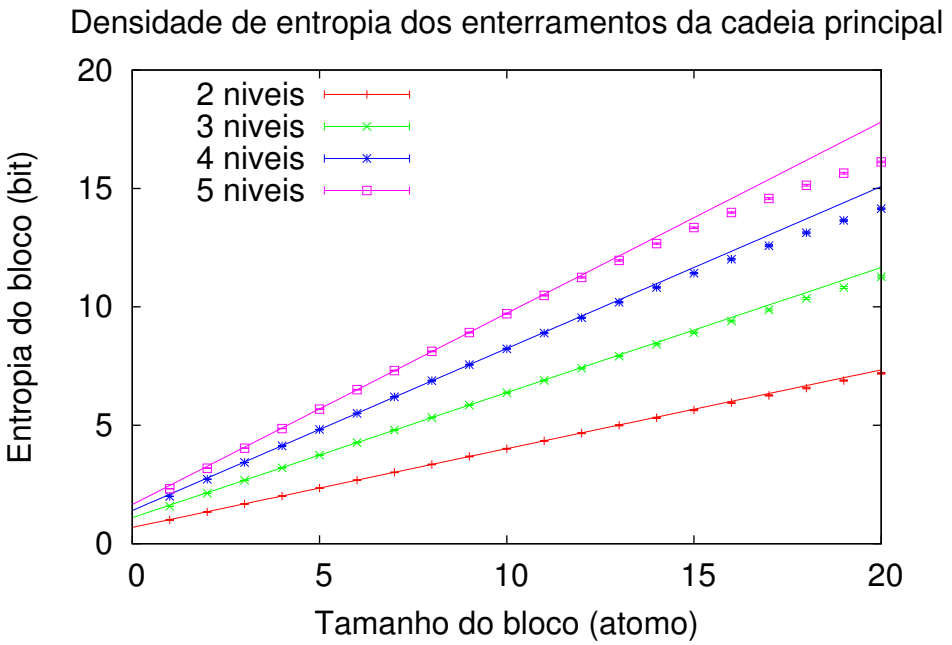


Figura 7: Densidade de entropia dos enterramentos atômicos dos átomos da cadeia principal quando consideradas várias distribuições de níveis de enterramentos. Os pontos mostram as entropias calculadas para diferentes classificações de níveis de enterramento em função do tamanho de janela de enterramentos. As linhas foram ajustadas aos pontos para calcular as densidades e os excessos de entropia.

a menor entre elas, seguida pela de C_α e a de C_β, o que é evidenciado pela Figura 8 e pela Figura 9. Isso ocorre possivelmente por conta das distâncias entre os átomos, pois quanto menor essa distância, maiores são as restrições impostas pelas ligações covalentes. Quando os átomos estão mais próximos entre si o comprimento da ligação restringe os graus de liberdade, o que acontece com menor intensidade quando os átomos estão mais distantes um dos outros.

Comparando-se os resultados obtidos por Crooks e Brenner (2004) acerca da entropia da estrutura secundária com os obtidos aqui acerca da entropia dos enterramentos atômicos de C_α nota-se que a densidade de entropia dos últimos é maior que a dos primeiros. Para efeitos de comparação foram usados apenas três níveis de enterramento, pois este é o número de símbolos presentes na estrutura secundária. A Figura 10 mostra este resultado, cujos valores estão na Tabela 2.

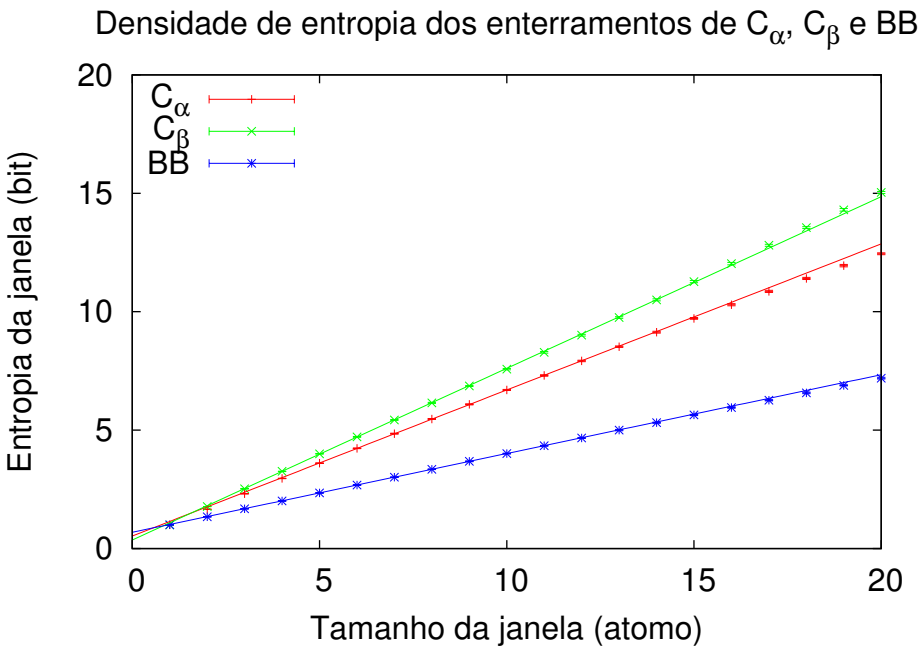


Figura 8: Densidades de entropia dos enterramentos atômicos de C_α, C_β e BB, para dois níveis de enterramento atômico. A densidade de entropia do enterramento de BB deve ser menor que a de C_α e C_β por conta das ligações covalentes que restrigem os graus de liberdade dos átomos.

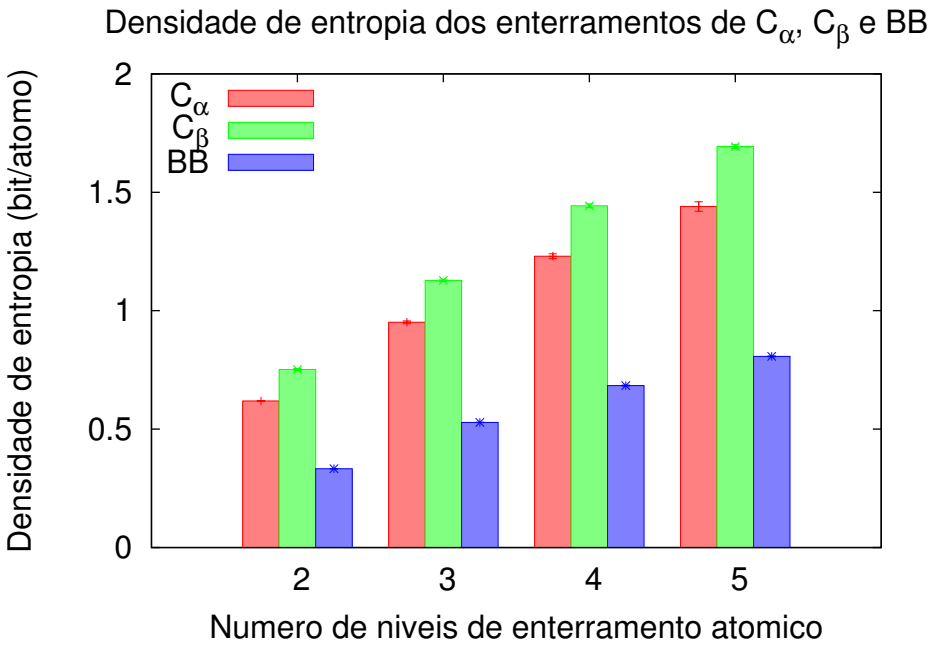


Figura 9: Densidades de entropia dos enterramentos atômicos de C_α, C_β e BB em função no número de níveis de enterramento atômico.

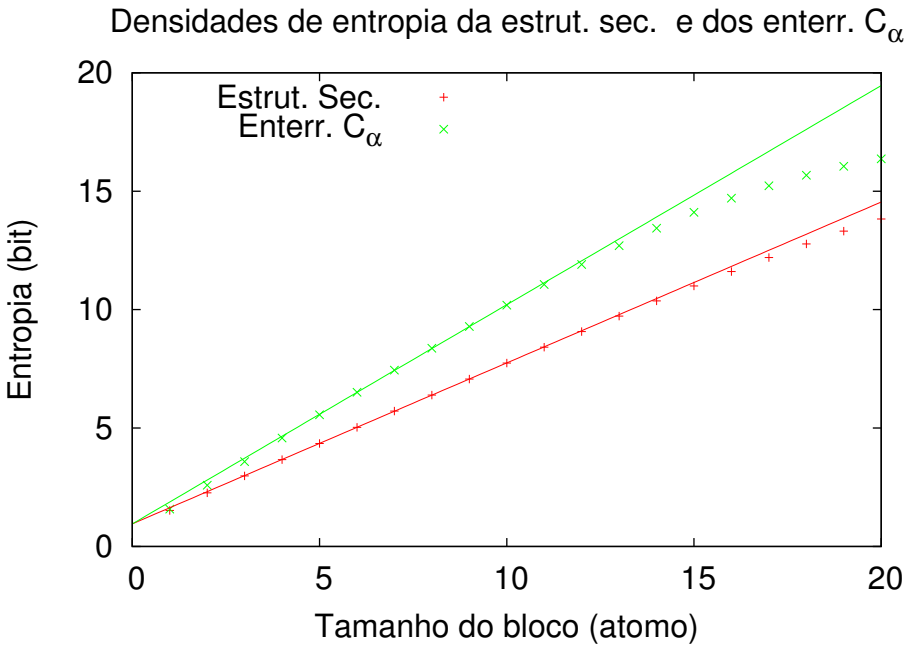


Figura 10: Comparação entre as entropias da estrutura secundária e dos enterramentos atômicos de C_α. "Estrut. sec." significa estrutura secundária e "enterr. C_α" significa enterramentos atômicos de C_α.

5.4 Transinformação entre as identidades dos resíduos centrais e blocos de enterramentos atômicos

Como mostrado nas sub-seções anteriores (5.1, 5.2 e 5.3), a sequência não apresenta correlação local, ao contrário dos enterramentos. Além disso, Crooks e Brenner (2004) observaram que as identidades de resíduos vizinhos, além de serem independentes entre si, são condicionalmente independentes em relação à estrutura secundária, ou seja, a transinformação entre um par de elementos de sequência e um par de elementos de estrutura secundária ($I(Q_i, Q_{i+1}; S_i, S_{i+1})$) é aproximadamente igual à soma das transinformações entre o primeiro elemento de sequência e os dois elementos de estrutura secundária e entre o segundo elemento de sequência e os dois elementos de estrutura secundária ($I(Q_i; S_i, S_{i+1}) + I(Q_{i+1}; S_i, S_{i+1})$). O mesmo pôde ser observado aqui, conforme mostra a Tabela 6, que mostra a média das transinformações entre dois elementos de uma mesma variável (identidade de resíduos de aminoácidos ou níveis de enterramentos atômicos) condicionada a um ou dois elementos da outra variável.

Tabela 6: Transinformação condicional calculada para várias combinações dos dímeros $Q_i; Q_{i+1}$ e $B_i; B_{i+1}$.

Dímeros	$I(Q_i; Q_{i+1})$ (bit)	$I(B_i; B_{i+1})$ (bit)
HP, 2 níveis	$0,00072 \pm 0.00008$	$0,342 \pm 0,003$
HP, 3 níveis	$0,00072 \pm 0.00008$	$0,574 \pm 0,003$
20 lt., 2 níveis	$0,005 \pm 0,007$	$0,342 \pm 0,003$
20 lt., 3 níveis	$0,005 \pm 0,007$	$0,574 \pm 0,003$
Dímeros	$I(Q_i; Q_{i+1}\|B_i)$ (bit)	$I(B_i; B_{i+1}\|Q_i)$ (bit)
HP, 2 níveis	$0,002 \pm 0,002$	$0,334 \pm 0.003$
HP, 3 níveis	$-0,004 \pm 0,006$	$0,567 \pm 0,002$
20 lt., 2 níveis	$0,008 \pm 0,005$	$0,332 \pm 0,009$
20 lt., 3 níveis	$0,008 \pm 0,001$	$0,566 \pm 0,009$
Dímeros	$I(Q_i; Q_{i+1}\|B_i, B_{i+1})$ (bit)	$I(B_i; B_{i+1}\|Q_i, Q_{i+1})$ (bit)
HP, 2 níveis	$0,001 \pm 0,001$	$0,330 \pm 0,003$
HP, 3 níveis	$0,001 \pm 0,001$	$0,570 \pm 0,005$
20 lt., 2 níveis	$0,01 \pm 0,01$	$0,32 \pm 0,02$
20 lt., 3 níveis	$0,01 \pm 0,02$	$0,55 \pm 0,02$

Os resultados mostram que as identidades dos resíduos de aminoácidos da sequência são independentes entre si, e condicionalmente independentes à sequência de enterramentos atômicos. Por outro lado, enterramentos atômicos vizinhos são correlacionados uns com os outros e condicionalmente dependentes da sequência. Por conta da primeira constatação é possível ser utilizada a aproximação descrita na eq. 20. O uso dessa aproximação é interessante devido ao fato de o banco de dados ser limitado, porque, caso ela não pudesse ser usada, seriam necessários blocos crescentes tanto de identidades quanto de enterramentos atômicos, o que causaria a saturação do banco de dados em janelas

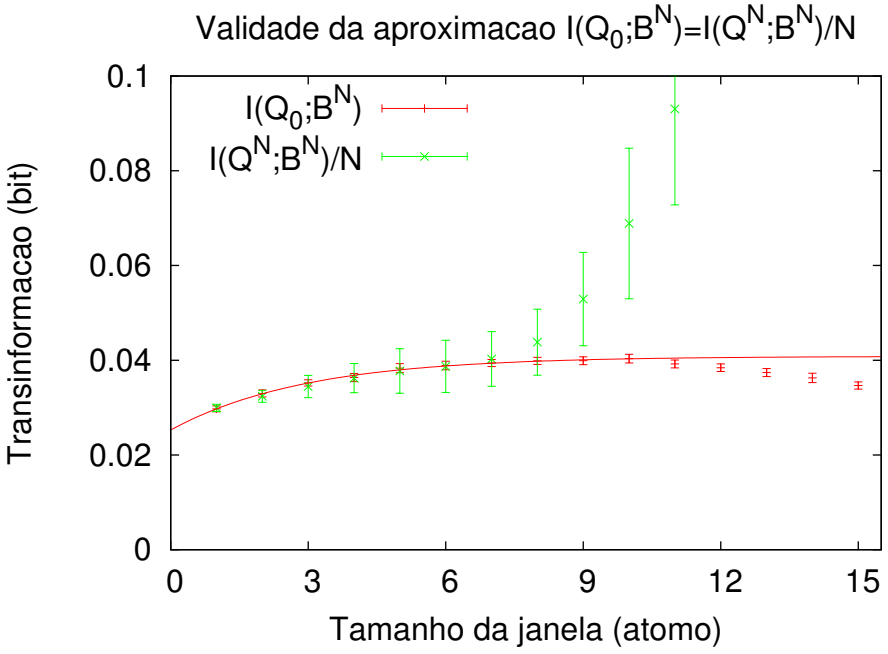


Figura 11: Validade da aproximação $I(Q_0; B^N) = I(Q^N; B^N)/N$. A figura mostra que para janelas pequenas a aproximação da eq. 20 é válida. A diferença observada em janelas grandes é resultado da saturação do banco de dados e a linha contínua é um ajuste aos pontos de $I(Q_0; B^N)$.

ainda pequenas para que fosse possível estimar bem a transinformação (Figura 11).

5.4.1 Cálculo da densidade de transinformação entre estrutura primária e enterramentos atômicos sem o uso de pseudocontagem ou *Bootstrap*

A transinformação entre a sequência e o enterramento de C_α ou de C_β foi calculada variando-se os alfabetos de sequência e o número de camadas de enterramento. Em um primeiro momento, os cálculos foram feitos sem o uso de pseudocontagem e sem a correção do desvio, feita pelo *Bootstrap*. Isso resultou em poucos pontos, o que tornou necessária a utilização de várias equações para se obter estimativas confiáveis. Por conta disso, foram feitos ajustes a partir de três equações diferentes nesses pontos, com o objetivo de se obter a melhor estimativa. Os ajustes mencionados foram feitos a partir das equações 26, 28 e 30 e chegaram a resultados parecidos, conforme as Tabelas 7, 8 e 9. As Figuras 12 e 13 mostram os ajustes de onde foram extraídos os dados da Tab. 9.

A densidade de transinformação não deve depender do ponto inicial ($I(Q^1; B^1)$), mas somente do número de níveis de enterramento atômico e do tamanho do alfabeto. Assim, não deve ser correto multiplicar o valor do primeiro ponto pela exponencial, pois este exerceria influência sobre o formato da curva. Por isso, as eq. 26 e 30 seriam as mais corretas. Entretanto, o fato de o banco de dados ser limitado precisa ser levado em consideração e, assim, uma equação com menor número de parâmetros é mais adequada. Como a eq. 26 possui três parâmetros a serem ajustados, enquanto a eq. 30 possui apenas

Tabela 7: Densidades de transinformação calculadas com a equação $I(Q;B^N) = a - b \times e^{(-N/c)}$ (eq. 26).

	Parâmetros			Densidade de
	a	b	c	Transiformação (bit/átomo)
C_α, 2 níveis	$0,044 \pm 0,005$	$0,018 \pm 0,004$	4 ± 2	$0,044 \pm 0,005$
C_α, 3 níveis	$0,062 \pm 0,006$	$0,037 \pm 0,004$	3 ± 1	$0,062 \pm 0,006$
C_β, 2 níveis	$0,065 \pm 0,001$	$0,023 \pm 0,001$	$4,1 \pm 0,5$	$0,065 \pm 0,001$
C_β, 3 níveis	$0,086 \pm 0,002$	$0,053 \pm 0,002$	$2,0 \pm 0,3$	$0,086 \pm 0,002$

Tabela 8: Densidades de transinformação calculadas com a equação $I(Q;B^N) = \frac{\beta}{\alpha} + (I(Q^1;B^1) - \frac{\beta}{\alpha}) \times e^{\alpha(N-1)}$ (eq. 28).

	$I(Q^1;B^1)$ (bit)	Parâmetros		Densidade de
		α	β	Transiformação (bit/átomo)
C_α, 2 níveis	0,0299	$0,25 \pm 0,08$	$0,011 \pm 0,003$	$0,044 \pm 0,018$
C_α, 3 níveis	0,0356	$0,4 \pm 0,1$	$0,023 \pm 0,005$	$0,06 \pm 0,02$
C_β, 2 níveis	0,0470	$0,25 \pm 0,02$	$0,016 \pm 0,001$	$0,064 \pm 0,006$
C_β, 3 níveis	0,0545	$0,51 \pm 0,09$	$0,044 \pm 0,006$	$0,086 \pm 0,019$

Tabela 9: Densidades de transinformação calculadas com a equação $I(Q;B^N) = I(Q^1;B^1) + \frac{\beta}{\alpha}(1 - e^{\alpha(N-1)})$ (eq. 30).

	$I(Q^1;B^1)$ (bit)	Parâmetros		Densidade de
		α	β	Transiformação (bit/átomo)
C_α, 2 níveis	0,0299	$0,25 \pm 0,08$	$0,0034 \pm 0,0004$	$0,044 \pm 0,004$
C_α, 3 níveis	0,0356	$0,4 \pm 0,1$	$0,010 \pm 0,001$	$0,06 \pm 0,01$
C_β, 2 níveis	0,0470	$0,25 \pm 0,02$	$0,0045 \pm 0,0001$	$0,065 \pm 0,001$
C_β, 3 níveis	0,0545	$0,50 \pm 0,05$	$0,0159 \pm 0,0009$	$0,086 \pm 0,004$

dois, e como o banco de dados não fornece estatística suficiente para janelas maiores que cinco ou seis resíduos, a eq. 30 foi escolhida para todos os outros cálculos de densidade de transinformação desta fase do trabalho.

As Figuras 14 e 15 são resultado dos ajustes da eq. 30 à transinformação entre sequência, levando em conta cada um dos alfabetos, e dois ou três níveis de enterramento atômico de C_α, respectivamente. Estes ajustes não fornecem medida do erro porque foram ajustados somente para os dois primeiros pontos da curva e representam o limite superior da densidade de transiformação. As linhas pontilhadas possuem os mesmos parâmetros que o ajuste para HP, com exceção de $I(Q^1;B^1)$. Elas foram simplesmente translocadas para o ponto inicial de cada um dos alfabetos e representam o limite inferior da densidade de transinformação. Os valores dos ajustes para dois e três níveis de enterramento se encontram nas Tabelas 10 e 11.

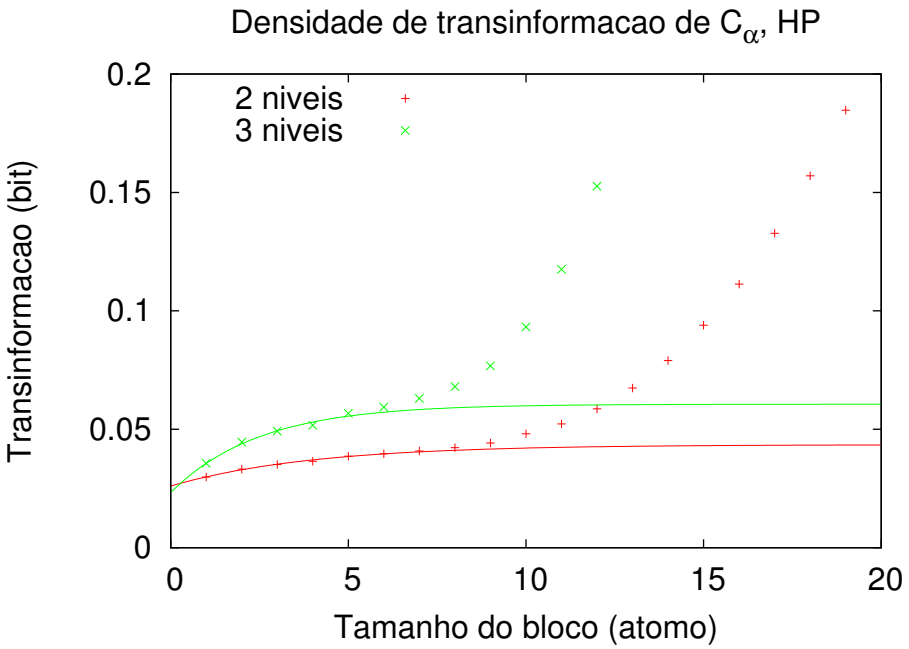


Figura 12: Transinformação entre sequência, considerando-se o alfabeto HP, e dois e três níveis de enterramento atômico de C_α. As linhas mostram os ajustes feitos com a eq. 30 aos pontos de mesma cor.

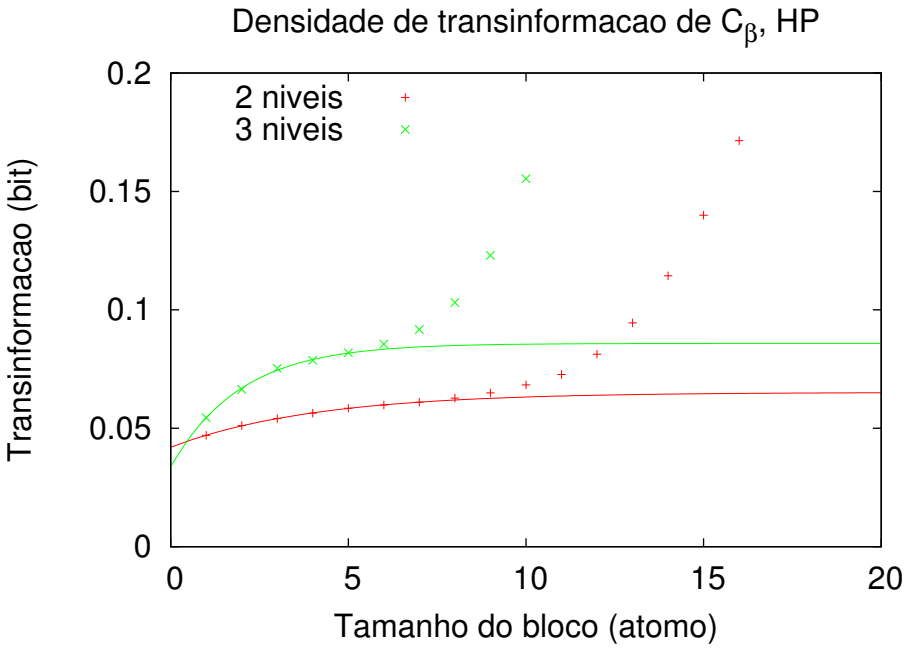


Figura 13: Transinformação entre sequência, considerando-se o alfabeto HP, e dois e três níveis de enterramento atômico de C_β. As linhas mostram os ajustes feitos com a eq. 30 aos pontos de mesma cor.

Tentou-se calcular a densidade de transinformação com outros alfabetos também para C_β, mas não houve estatística suficiente. A Figura 16 mostra que mesmo com o alfabeto

Tabela 10: Valores obtidos do ajuste da eq. 30 aos pontos da transinformação entre cada um dos alfabetos e dois níveis de enterramento atômico de C_α.

Alfabeto	$I(Q^1;B^1)$ (bit)	Parâmetros		$i(Q;B)$ máxima (bit/átomo)	$i(Q:B)$ mínima (bit/átomo)
		α	β		
HP	$0,0299$	$0,25\pm0,08$	$0,0034\pm0,0004$	$0,45\pm0,004$	$0,044\pm0,004$
HPN	$0,0419$	$0,19$	$0,0037$	$0,061$	$0,056$
HPNhp	$0,0431$	$0,11$	$0,0036$	$0,0831$	$0,0572$
20 letras	$0,0458$	$-0,10$	$0,041$	$0,3642$	$0,0599$

Tabela 11: Valores obtidos do ajuste da eq. 30 aos pontos da transinformação entre cada um dos alfabetos e três níveis de enterramento atômico de C_α.

Alfabeto	$I(Q^1;B^1)$ (bit)	Parâmetros		$i(Q:B)$ máxima (bit/átomo)	$i(Q:B)$ mínima (bit/átomo)
		α	β		
HP	$0,0356$	$0,4\pm0,1$	$0,010\pm0,001$	$0,06\pm0,01$	$0,06\pm0,01$
HPN	$0,0517$	$0,4$	$0,012$	$0,0817$	$0,0761$
HPNhp	$0,0532$	$0,3$	$0,012$	$0,0932$	$0,0776$
20 letras	$0,0572$	$0,1$	$0,012$	$0,1772$	$0,816$

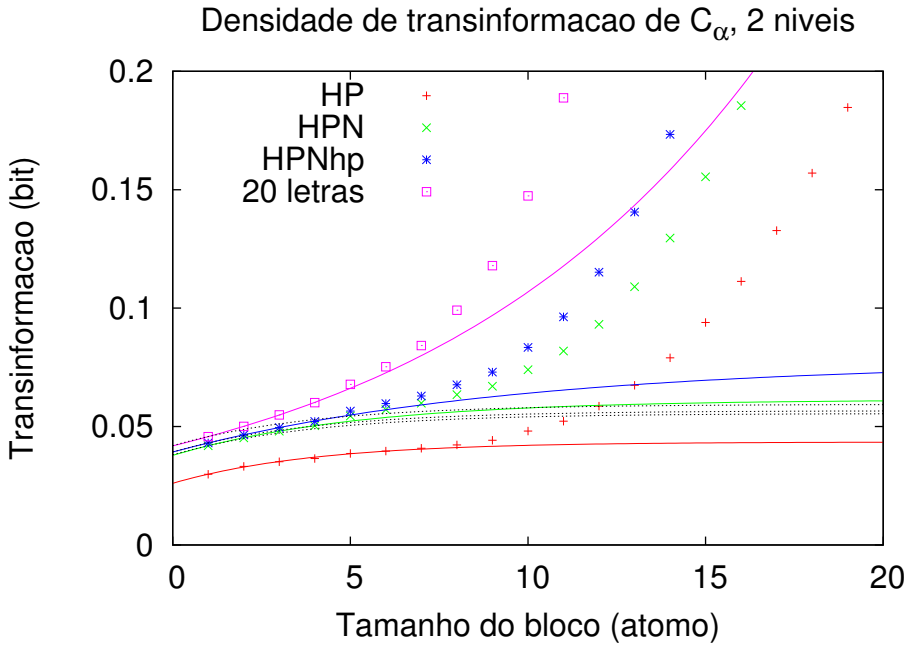


Figura 14: Transinformação entre sequência, considerando-se cada um dos alfabetos, e dois níveis de enterramento atômico de C_α. A convexidade da curva em cor-de-rosa apenas evidencia que para 20 letras o banco de dados satura já nos primeiros pontos e, por isso, eles não são confiáveis. As linhas pontilhadas represetam a translocação da curva que contém os parâmetros α e β de HP para os pontos iniciais ($I(Q^1;B^1)$) de cada alfabeto.

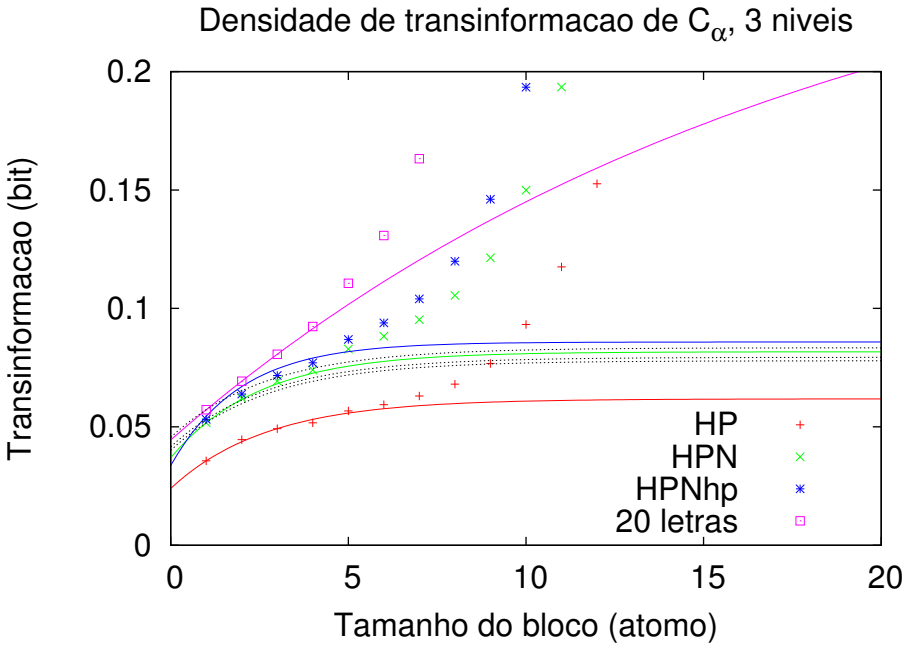


Figura 15: Transinformação entre sequência, considerando-se cada um dos alfabetos, e três níveis de enterramento atômico de C_α. A enorme diferença de densidade de transinformação entre a curva em cor-de-rosa e as outras apenas evidencia que para 20 letras o banco de dados satura já nos primeiros pontos e, por isso, eles não são confiáveis. As linhas pontilhadas represetam a translocação da curva que contém os parâmetros α e β de HP para os pontos iniciais $(I(Q^1;N^1))$ de cada alfabeto.

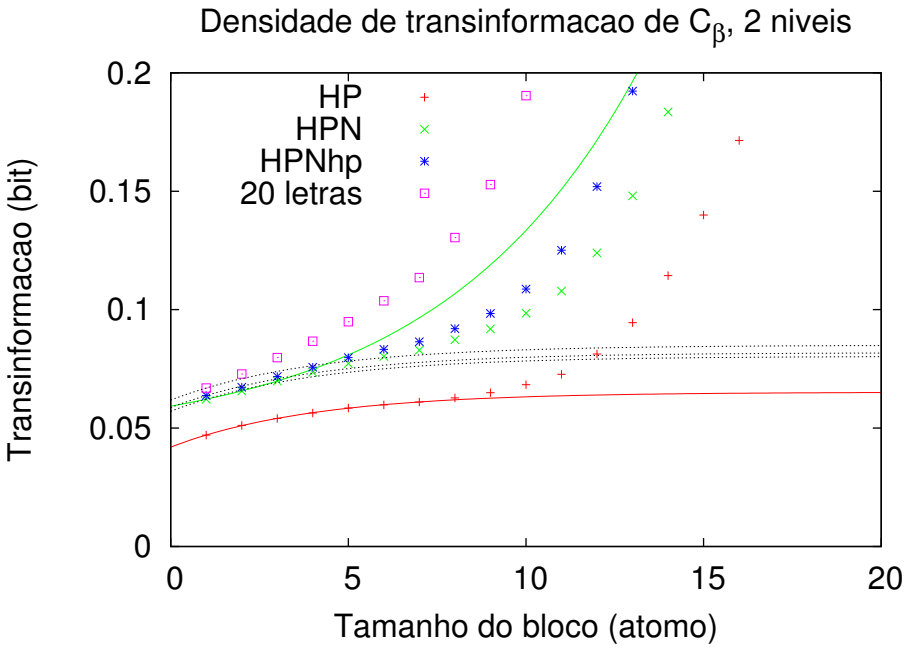


Figura 16: Transiformação entre sequência, considerando-se cada um dos alfabetos, e dois níveis de enterramento atômico de C_β. A convexidade dos pontos do alfabeto HPN (pontos em verde) apenas evidencia que para três letras o banco de dados satura já nos primeiros pontos e, por isso, eles não são confiáveis. As linhas pontilhadas represetam a translocação da curva que contém os parâmetros α e β de HP para os pontos iniciais $(I(Q^1;N^1))$ de cada alfabeto.

HPN e dois níveis de enterramento os pontos evidenciam a saturação do banco de dados.

Por conta da impossibilidade se obter os valores de densidade de transinformação para várias combinações de alfabetos foi introduzido na metodologia de estimativa de frequência a pseudocontagem. Aliado a isso, o desvio do banco de dados foi corrigido a partir de um *Bootstrap* de 50 repetições.

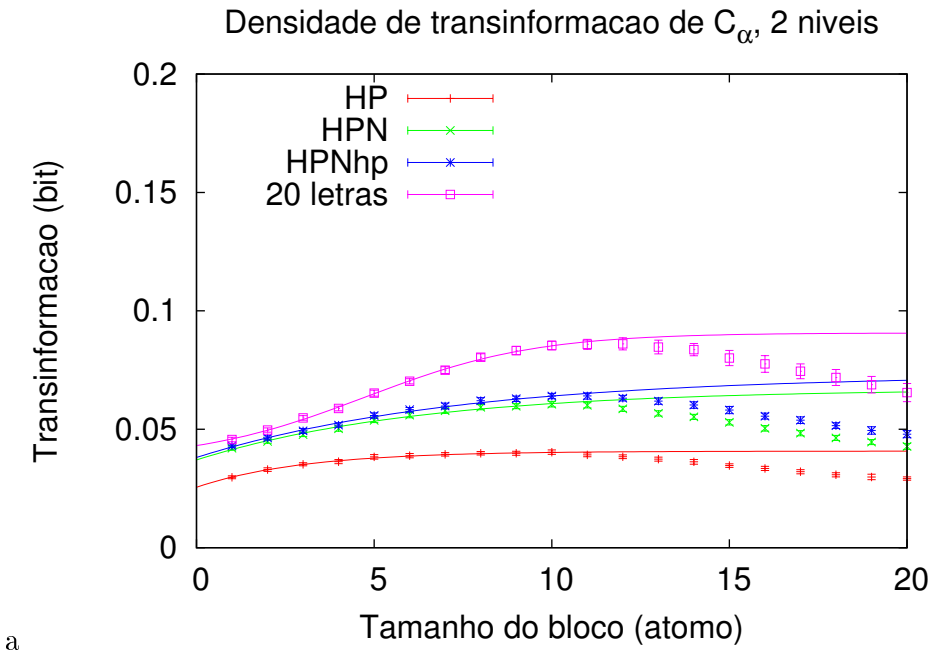
5.4.2 Cálculo da densidade de transinformação entre estrutura primária e enterramentos atômicos com o uso de pseudocontagem e de *Bootstrap*

Da mesma forma que na seção anterior, a transinformação entre a sequência e o enterramento de C_α ou de C_β foi calculada variando-se os alfabetos de sequência e o número de camadas de enterramento. Entretanto, nesta fase do trabalho, o cálculo da transinformação foi feito com pseudocontagem e *Bootstrap*, o que permitiu que os ajustes fossem feitos a partir de mais pontos e eliminou a necessidade de se utilizar equações com um menor número de variáveis. Por esse motivo, aqui foram usadas duas equações para os ajustes, uma exponencial (eq. 26) e uma sigmóide (eq. 32), de acordo com o conjunto de pontos obtidos. A estimativa cujo resultado apresentou o menor erro relativo foi considerada. O uso do *Bootstrap* e da pseudocontagem também permitiu o cálculo da densidade de transinformação para um número maior de níveis de enterramento atômico. Os resultados estão na Tabela 12 e nas Figuras 17, 18, 19 e 20. As Fig. 19 e 20 contém linhas pontilhadas no alfabetos HPNhp e 20 letras porque os pontos de HPNhp permitem inferir que o banco de dados está saturado. O motivo para tal constatação é o fato de que a transinformação não pode diminuir com o tamanho do alfabeto, devendo sempre ser igual ou maior que a dos alfabetos menores. Como o ajuste nos pontos calculados com HPNhp apresenta uma estimativa para a densidade de transinformação menor que a obtida com HPN, necessariamente a primeira não está correta, o que permite inferir que a estimativa para 20 letras está igualmente incorreta.

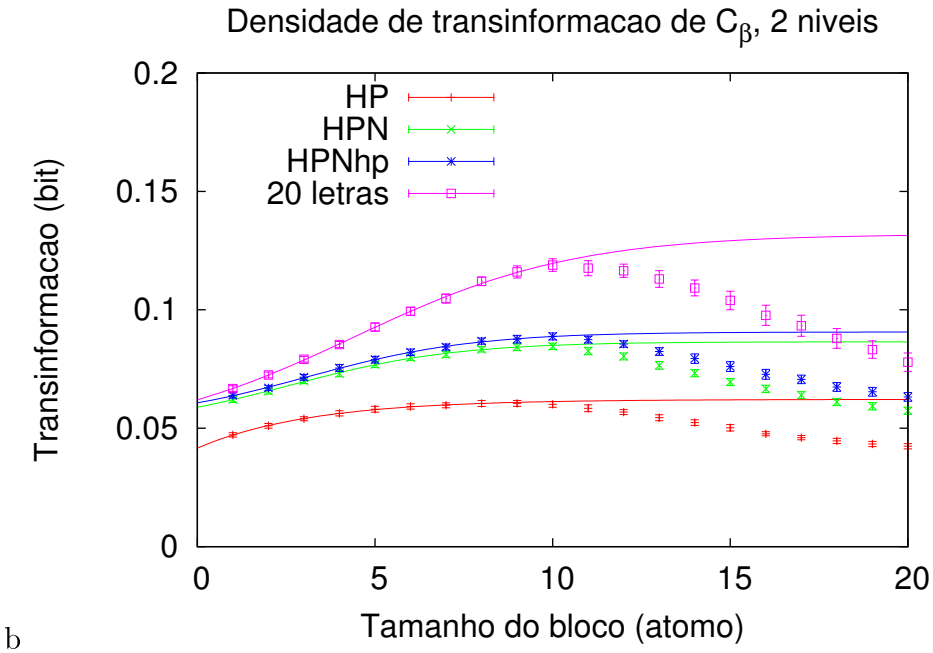
Os pontos calculados usando-se pseudocontagem e *Bootstrap* são mais exatos que os anteriores pois tem seu desvio e a influência da saturação do banco de dados corrigidos. Além disso, a utilização dessas duas técnicas permitiu que as curvas fossem ajustadas a um maior número de pontos e que o erro de cada um dos pontos fosse considerado para tal, o que aumenta a confiabilidade dos resultados. De acordo com o exposto, é natural que esses últimos dados apresentados sejam considerados como a melhor estimativa da densidade de transinformação entre a sequência de resíduos de aminoácidos e os enterramentos atômicos de C_α e C_β. Uma vez que esses resultados são a já mencionada estimativa produzida por este trabalho, a Tabela 13 apresenta a melhor estimativa da densidade de transinformação considerada e a fração da densidade de entropia do enterramento respondida por ela.

Tabela 12: Densidade de transinformação entre sequência e enterramentos atômicos de C_α e C_β para vários alfabetos de sequência e de níveis de enterramento calculada com pseudocontador e *Booststrap*. As colunas denotadas por $i(Q;B)$ mostram os resultados dos ajustes como uma extrapolação dos pontos obtidos, e as colunas denotadas por $i(Q;B)^-$ são o ponto mais alto obtido, que é uma medida inferior da densidade de transinformação mais conservadora que a apresentada na seção anterior (chamada de $i(X;Y)$ mínima), por isso chamada de $i(Q;B)^-$.

Alfabeto	C_α		C_β	
	$i(Q;B)$ (bit/átomo)	$i(Q;B)^-$ (bit/átomo)	$i(Q;B)$ (bit/átomo)	$i(Q;B)^-$ (bit/átomo)
2 níveis				
HP	$0,0408 \pm 0,0002$	$0,0403 \pm 0,0009$	$0,0622 \pm 0,0003$	$0,060 \pm 0,001$
HPN	$0,068 \pm 0,003$	$0,061 \pm 0,001$	$0,087 \pm 0,009$	$0,085 \pm 0,001$
HPNhp	$0,075 \pm 0,003$	$0,064 \pm 0,001$	$0,100 \pm 0,003$	$0,089 \pm 0,001$
20 letras	$0,091 \pm 0,007$	$0,086 \pm 0,003$	$0,13 \pm 0,01$	$0,119 \pm 0,003$
3 níveis				
HP	$0,058 \pm 0,001$	$0,056 \pm 0,001$	$0,081 \pm 0,001$	$0,080 \pm 0,001$
HPN	$0,091 \pm 0,004$	$0,084 \pm 0,002$	$0,117 \pm 0,003$	$0,111 \pm 0,002$
HPNhp	$0,096 \pm 0,003$	$0,088 \pm 0,002$	$0,123 \pm 0,003$	$0,116 \pm 0,002$
20 letras	$0,130 \pm 0,006$	$0,118 \pm 0,002$	$0,176 \pm 0,006$	$0,153 \pm 0,003$
4 níveis				
HP	$0,069 \pm 0,002$	$0,068 \pm 0,001$	$0,119 \pm 0,005$	$0,104 \pm 0,003$
HPN	$0,102 \pm 0,003$	$0,098 \pm 0,002$	$0,16 \pm 0,01$	$0,132 \pm 0,003$
HPNhp	$0,108 \pm 0,003$	$0,102 \pm 0,002$	$0,134 \pm 0,007$	$0,127 \pm 0,002$
20 letras	$0,154 \pm 0,007$	$0,132 \pm 0,003$	$0,19 \pm 0,01$	$0,163 \pm 0,003$
5 níveis				
HP	$0,076 \pm 0,003$	$0,074 \pm 0,008$	$0,123 \pm 0,004$	$0,107 \pm 0,003$
HPN	$0,112 \pm 0,005$	$0,106 \pm 0,001$	$0,16 \pm 0,04$	$0,138 \pm 0,003$
HPNhp	$0,12 \pm 0,01$	$0,111 \pm 0,002$	$0,15 \pm 0,02$	$0,131 \pm 0,002$
20 letras	$0,16 \pm 0,01$	$0,139 \pm 0,003$	$0,20 \pm 0,02$	$0,168 \pm 0,003$

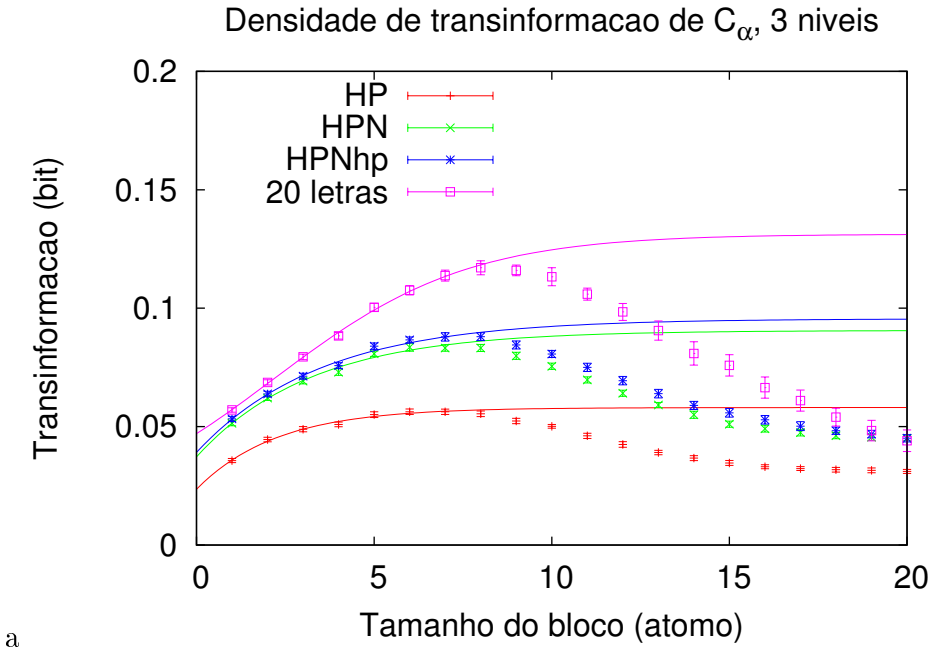


a

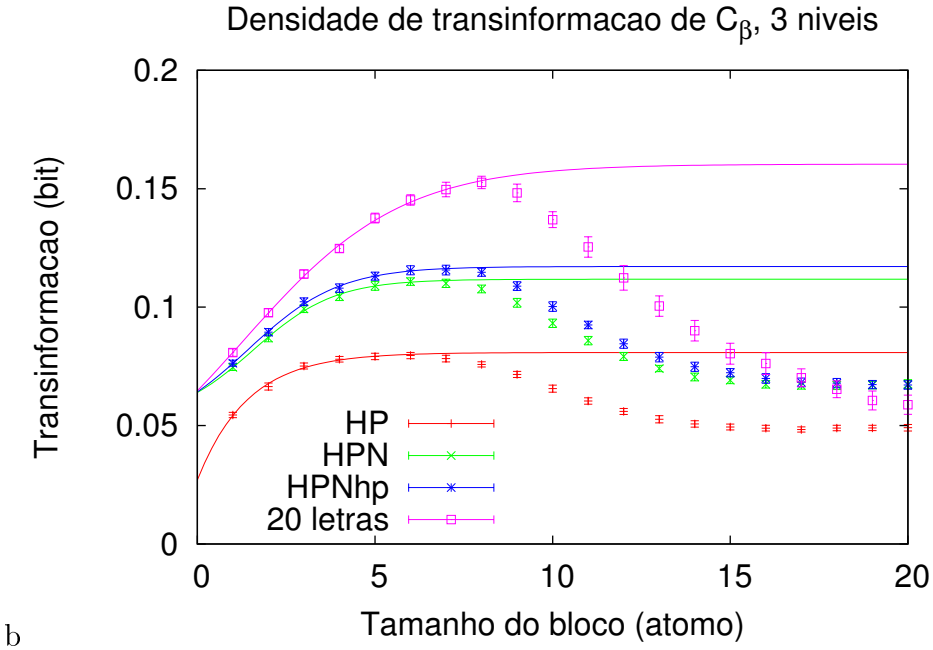


b

Figura 17: Ajustes para a estimativa da densidade de transinformação entre vários alfabetos de sequência e 2 níveis de enterramento atômico de C_α e C_β, respectivamente. Nestes gráficos, os pontos da transinformação foram obtidos com o uso de *Bootstrap* e de pseudo-contador.

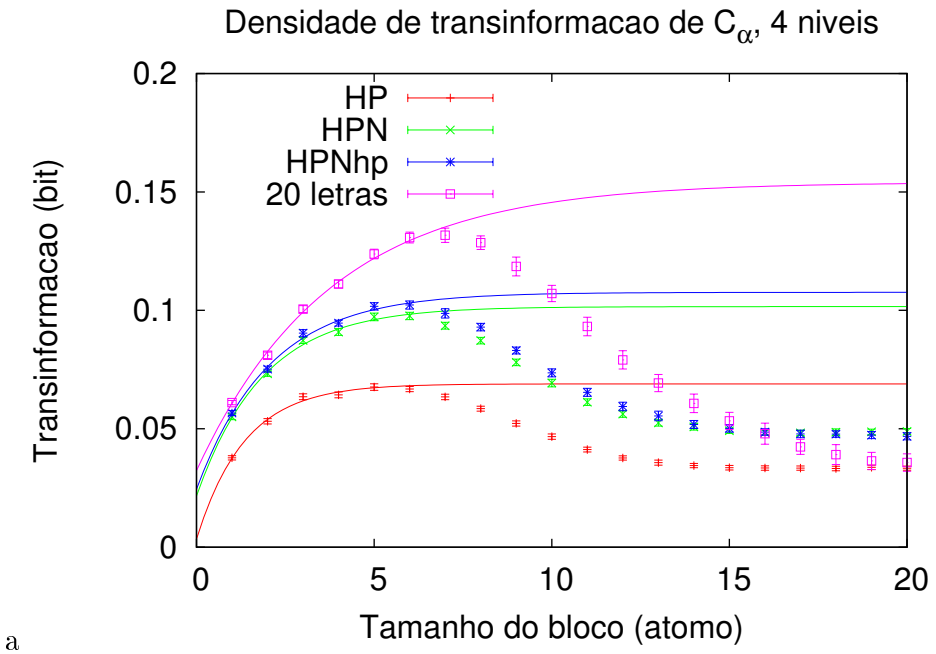


a

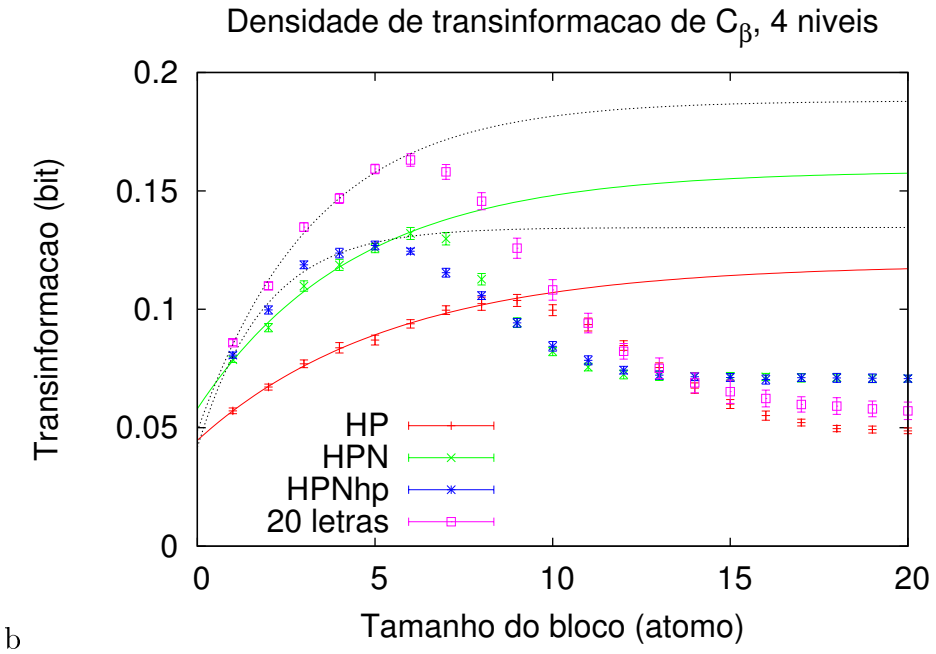


b

Figura 18: Ajustes para a estimativa da densidade de transinformação entre vários alfabetos de sequência e 3 níveis de enterramento atômico de C_α e C_β, respectivamente. Nestes gráficos, os pontos da transinformação foram obtidos com o uso de *Bootstrap* e de pseudo-contador.

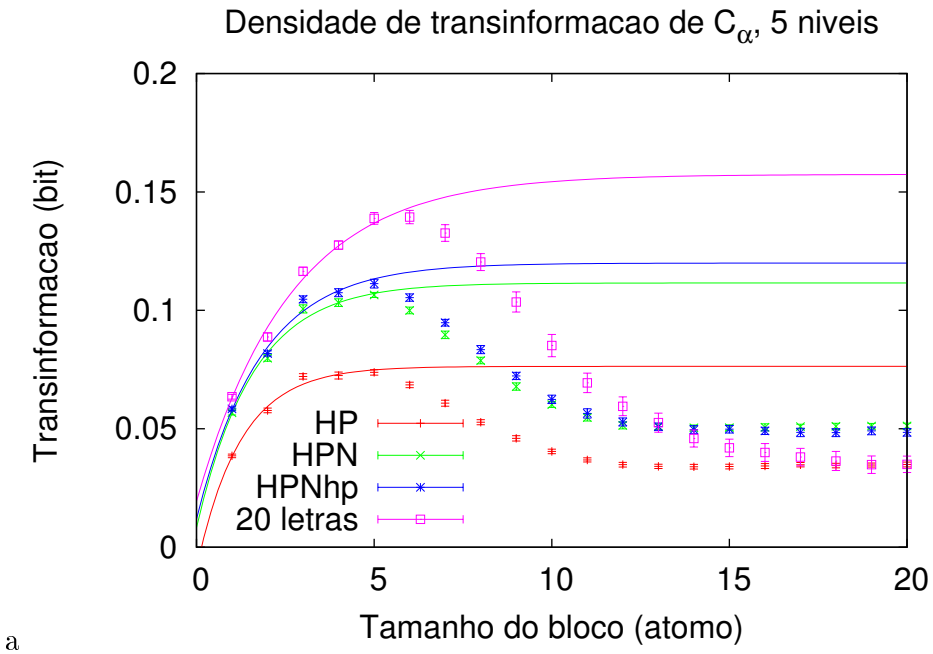


a

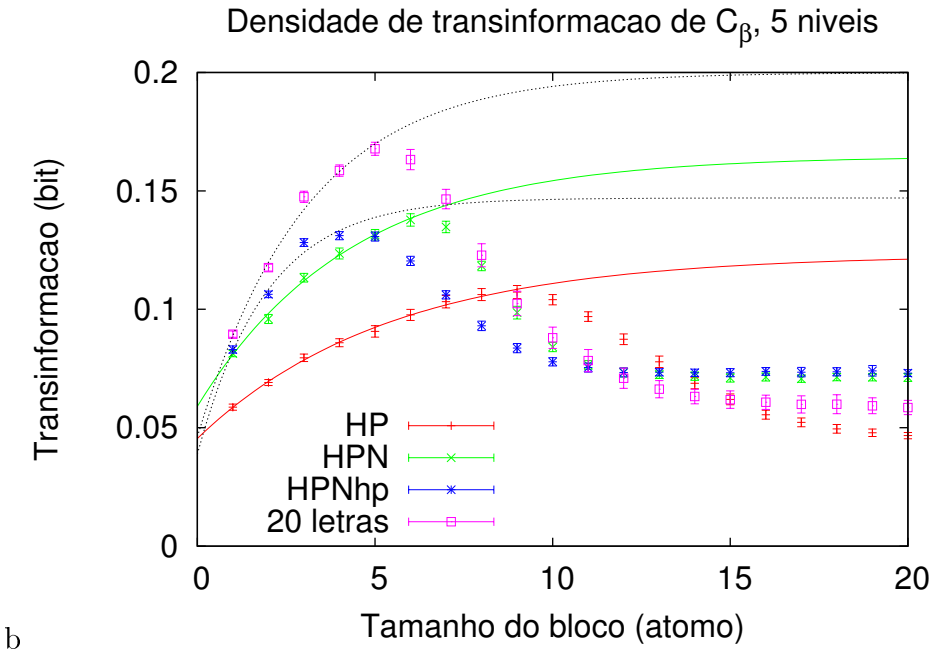


b

Figura 19: Ajustes para a estimativa da densidade de transinformação entre vários alfabetos de sequência e 4 níveis de enterramento atômico de C_α e C_β, respectivamente. Nestes gráficos, os pontos da transinformação foram obtidos com o uso de *Bootstrap* e de pseudo-contador. As linhas pontilhadas evidenciam que os ajustes para os alfabetos HPNhp e 20 letras não são estimativas razoáveis da densidade de transinformação porque, como o ajuste de HPNhp sugere uma densidade de transinformação menor que a calculada para HPN, o resultado permite inferir que o banco de dados está saturado.



a



b

Figura 20: Ajustes para a estimativa da densidade de transinformação entre vários alfabetos de sequência e 5 níveis de enterramento atômico de C_α e C_β, respectivamente. Nestes gráficos, os pontos da transinformação foram obtidos com o uso de *Bootstrap* e de pseudo-contador. As linhas pontilhadas evidenciam que os ajustes para os alfabetos HPNhp e 20 letras não são estimativas razoáveis da densidade de transinformação porque, como o ajuste de HPNhp sugere uma densidade de transinformação menor que a calculada para HPN, o resultado permite inferir que o banco de dados está saturado.

Tabela 13: Densidade de transinformação ($iQ;B$)) e fração da densidade de entropia do enterramento respondida por ela ($\frac{i(Q;B)}{h(B)}$).

Alfabeto	C_α		C_β	
	$i(Q;B)$ (bit/átomo)	$\frac{i(Q;B)}{h(B)}$	$i(Q;B)$ (bit/átomo)	$\frac{i(Q;B)}{h(B)}$
2 níveis				
HP	$0,0408 \pm 0,0002$	$0,0661 \pm 0,0004$	$0,0622 \pm 0,0003$	$0,086 \pm 0,001$
HPN	$0,068 \pm 0,003$	$0,110 \pm 0,005$	$0,087 \pm 0,009$	$0,12 \pm 0,01$
HPNhp	$0,075 \pm 0,003$	$0,122 \pm 0,005$	$0,100 \pm 0,003$	$0,138 \pm 0,004$
20 letras	$0,091 \pm 0,007$	$0,15 \pm 0,01$	$0,13 \pm 0,01$	$0,18 \pm 0,01$
3 níveis				
HP	$0,058 \pm 0,001$	$0,089 \pm 0,001$	$0,081 \pm 0,001$	$0,071 \pm 0,001$
HPN	$0,091 \pm 0,004$	$0,095 \pm 0,004$	$0,117 \pm 0,003$	$0,104 \pm 0,003$
HPNhp	$0,096 \pm 0,003$	$0,100 \pm 0,003$	$0,123 \pm 0,003$	$0,109 \pm 0,003$
20 letras	$0,130 \pm 0,006$	$0,136 \pm 0,006$	$0,176 \pm 0,006$	$0,156 \pm 0,005$
4 níveis				
HP	$0,069 \pm 0,002$	$0,056 \pm 0,002$	$0,119 \pm 0,005$	$0,082 \pm 0,003$
HPN	$0,102 \pm 0,003$	$0,083 \pm 0,003$	$0,16 \pm 0,01$	$0,111 \pm 0,007$
HPNhp	$0,108 \pm 0,003$	$0,088 \pm 0,003$	$0,134 \pm 0,007$	$0,093 \pm 0,005$
20 letras	$0,154 \pm 0,007$	$0,125 \pm 0,006$	$0,19 \pm 0,01$	$0,13 \pm 0,01$
5 níveis				
HP	$0,076 \pm 0,003$	$0,053 \pm 0,002$	$0,123 \pm 0,004$	$0,07 \pm 0,01$
HPN	$0,112 \pm 0,005$	$0,078 \pm 0,004$	$0,16 \pm 0,04$	$0,09 \pm 0,02$
HPNhp	$0,12 \pm 0,01$	$0,083 \pm 0,007$	$0,15 \pm 0,02$	$0,09 \pm 0,01$
20 letras	$0,16 \pm 0,01$	$0,11 \pm 0,01$	$0,20 \pm 0,02$	$0,12 \pm 0,01$

5.5 Transinformação entre blocos de identidades de resíduos e enterramentos de um dos átomos do resíduo central

Apesar de as identidades dos resíduos de aminoácidos não serem correlacionadas localmente, elas aparentemente apresentam correlação quando condicionadas ao enterramento de um dos átomos do resíduo central, de acordo com a eq. 23. Isso ficou evidente ao se comparar a transinformação entre um bloco de identidades de aminoácidos e o enterramento do C_α ou do C_β do resíduo central ($I(Q^N; B_0)$) e a soma das transinformações dos N resíduos do bloco e o enterramento do C_α ou do C_β do resíduo central ($\sum I(Q_n; B_0)$).

A Figura 21 mostra cada um dos pontos $I(Q_n; B_0)$ que foram somados para se obter a estimativa $\sum I(Q_n; B_0)$ e a Figura 22 mostra o valor de $\sum I(Q_n; B_0)$ para tamanhos crescentes de janelas. Se forem considerados pares de identidades e enterramentos atômicos separados por 15 ou mais resíduos, essas transinformações tendem a zero e, por isso, as janelas consideradas na fig. 22 tem tamanho menor ou igual a 31 resíduos. A Figura 23 compara $I(Q^N; B_0)$ e $\sum I(Q_n; B_0)$ e evidencia que a diferença entre essas duas estimativas existe mesmo que para janelas pequenas. A tabela 14 apresenta os dados referentes às diferenças mencionadas acima. Resultados considerando-se o alfabeto de 20 letras na sequência não estão presentes porque não há estatística suficiente para este cálculo a partir de blocos de sequência formados por um alfabeto tão diverso.

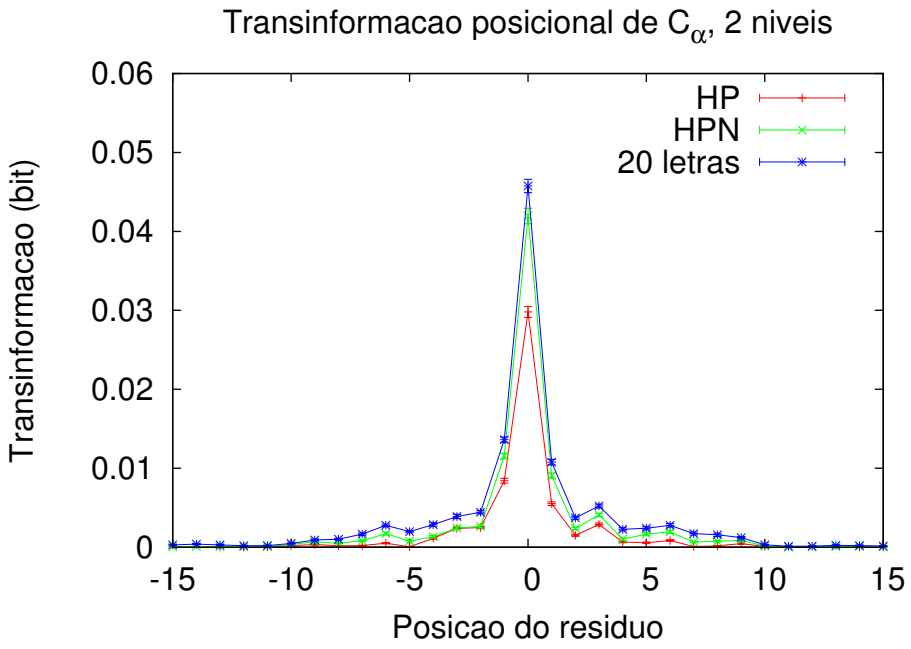


Figura 21: Transinformação posicional entre três diferentes alfabetos de sequência e dois níveis de enterramentos atômicos de C_α.

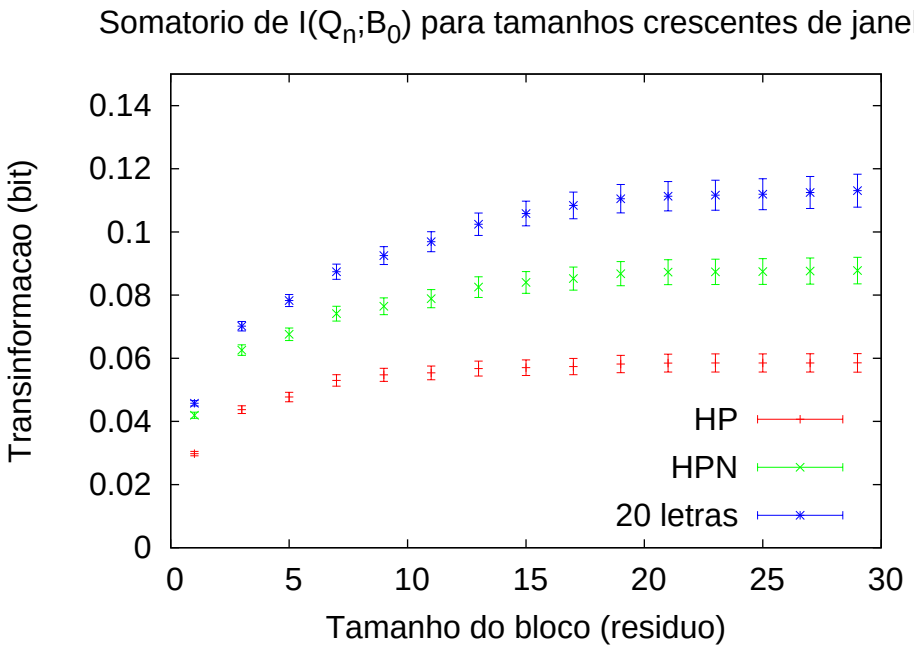


Figura 22: Somatório de $I(Q_n; B_0)$ para tamanhos crescentes de janelas. A transinformação entre sequência e enterramento atoômico referente a cada par de posições Q_n e Q_{-n} foi somado à janela anterior, resultando em janelas de tamanho cada vez maiores, porém cuja $\sum I(Q_n; B_0)$ não necessariamente é maior.

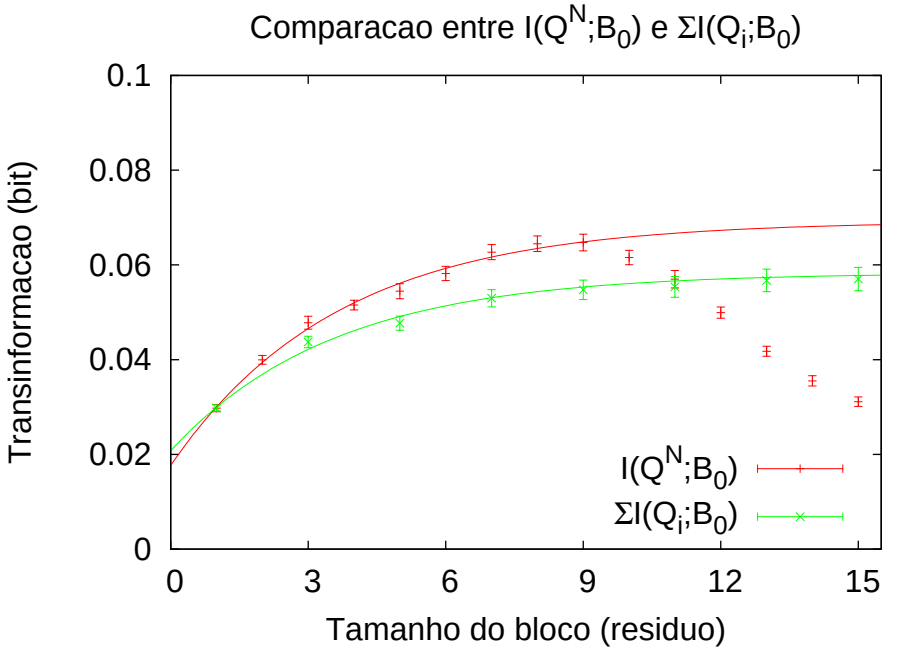


Figura 23: Comparação entre a transinformação de um bloco de identidades de aminoácidos e o enterramento do C_α ou do C_β do resíduo central $I(Q^N; B_0)$ e a soma das transinformações dos N resíduos do bloco e o enterramento do C_α ou do C_β do resíduo central $\sum I(Q_n; B_0)$.

Tabela 14: Comparação entre a transinformação calculada para um bloco de identidades de aminoácidos e o enterramento do C_α ou do C_β do resíduo central $I(Q^N;B_0)$ e a soma das transinformações dos N resíduos do bloco e o enterramento do C_α ou do C_β do resíduo central $\sum I(Q_n;B_0)$.

Alfabeto	C_α		C_β	
	$I(Q^N;B_0)$ (bit)	$\sum I(Q_n;B_0)$ (bit)	$I(Q^N;B_0)$ (bit)	$\sum I(Q_n;B_0)$ (bit)
2 níveis				
HP	$0,065\pm0,002$	$0,059\pm0,003$	$0,077\pm0,002$	$0,070\pm0,003$
HPN	$0,083\pm0,002$	$0,075\pm0,004$	$0,096\pm0,003$	$0,086\pm0,004$
3 níveis				
HP	$0,085\pm0,002$	$0,075\pm0,004$	$0,096\pm0,003$	$0,086\pm0,004$
HPN	$0,110\pm0,002$	$0,0114\pm0,005$	$0,122\pm0,002$	$0,125\pm0,005$

Porque os enterramentos atômicos são correlacionados entre si a aproximação da eq. 20 não pode ser feita para as identidades dos resíduos de aminoácidos, diferentemente dos enterramentos atômicos (conforme visto na seção anterior), porque o limite tem que ser feito em relação à variável correlacionada. A aproximação da eq. 20 pode ser calculada, mas não terá como resultado a densidade de transinformação. Entretanto, o estudo de $I(Q^N;B_0)$ e de $\sum I(Q_n;B_0)$ é importante porque ele diz respeito às informações que podem ser extraídas a partir de blocos de identidades de resíduos, que é o princípio de funcionamento dos algoritmos de predição de enterramentos atômicos a partir da sequência de resíduos de aminoácidos.

5.6 Entropia condicional dos átomos da cadeia lateral

Para cada átomo pertencente à cadeia lateral dos resíduos de aminoácidos foi calculada a entropia de seu enterramento condicionada ao enterramento do C_α e à identidade do resíduo, ou seja, foi medida a dúvida do enterramento de um determinado átomo da cadeia lateral uma vez que são conhecidas a identidade do resíduo e o enterramento do C_α correspondente, considerando-se para tal 2 níveis de enterramento atômico. Foi observado que esta entropia é menor que a entropia de C_α ou de C_β, de forma que esta redução é efeito do condicionamento a outras variáveis, e é essa característica que permite ser possível a predição dos enterramentos atômicos, conforme discutido adiante. A Tabela 15 mostra o número de ligações covalentes que separam o átomo em questão do C_α do mesmo resíduo, e a Tabela 16 apresenta os valores calculados para as entropias condicionais mencionadas, de forma que combinadas a informações apresentadas nas duas tabelas é possível saber a entropia condicional de cada um dos átomos das cadeias laterais. A Figura 24 mostra as densidades de entropia condicional e não condicional para cada um dos átomos da cadeia lateral do aminoácido Lisina.

A Tab. 16 mostra que entropia dos enterramentos dos átomos da cadeia lateral condicionada à identidade do resíduo e ao enterramento do C_α correspondente é maior conforme o número de ligações covalentes que separam o átomo em questão de seu C_α,

Tabela 15: Distância entre os átomos das cadeias laterais e o C_α do mesmo resíduo. A distância é o número de ligações covalentes que separam o átomo em questão do C_α do mesmo resíduo.

Resíduo	Distância (número de ligações covalentes)					
	1	2	3	4	5	6
A	C_β					
R	C_β	C_γ	C_δ	N_ϵ	C_ζ	
N	C_β	C_γ	$O_{\delta 1}$ e $N_{\delta 2}$			
D	C_β	C_γ	$O_{\delta 1}$ e $O_{\delta 2}$			
C	C_β	S_γ				
F	C_β	C_γ	$C_{\delta 1}$ e $C_{\delta 2}$	$C_{\epsilon 1}$ e $C_{\epsilon 2}$	C_ζ	
E	C_β	C_γ	C_δ	$O_{\epsilon 1}$ e $O_{\epsilon 2}$		
Q	C_β	C_γ	C_δ	$O_{\epsilon 1}$ e $N_{\epsilon 2}$		
H	C_β	C_γ	$N_{\delta 1}$ e $C_{\delta 2}$	$C_{\epsilon 1}$ e $N_{\epsilon 2}$		
I	C_β	$C_{\gamma 1}$ e $C_{\gamma 2}$	$C_{\delta 1}$			
L	C_β	C_γ	$C_{\delta 1}$ e $C_{\delta 2}$			
K	C_β	C_γ	C_δ	C_ϵ	N_ζ	
M	C_β	C_γ	S_δ	C_ϵ		
P	C_β	C_γ e C_δ				
S	C_β	O_γ				
Y	C_β	C_γ	$C_{\delta 1}$ e $C_{\delta 2}$	$C_{\epsilon 1}$ e $C_{\epsilon 2}$	C_ζ	O_η
T	C_β	$O_{\gamma 1}$ e $C_{\gamma 2}$				
W	C_β	C_γ	$C_{\delta 1}$ e $C_{\delta 2}$	$N_{\epsilon 1}$ e $C_{\epsilon 2}$ e $C_{\epsilon 3}$	$C_{\zeta 2}$ e $C_{\zeta 3}$	C_η
V	C_β	$C_{\gamma 1}$ e $C_{\gamma 2}$				

o que era esperado. Entretanto, uma outra abordagem para o problema poderia, em vez de condicionar os enterramentos de cada um dos átomos da cadeia lateral ao C_α do resíduo, condicioná-los ao átomo vizinho imediatamente anterior em relação ao C_α, ou seja, calcular $H(C_\beta|C_\alpha, Q)$, $H(C_\gamma|C_\beta, Q)$, $H(C_\delta|C_\gamma, Q)$ e assim sucessivamente. Esses valores foram calculados para a Lisina e estão na Tabela 17 e na Figura 25. A Tab. 17 e na Fig. 25 evidenciam que a entropia é reduzida com o condicionamento ao enterramento do átomo imediatamente anterior e não se altera substancialmente com o condicionamento aos outros átomos da cadeia lateral, de modo que não é necessário o conhecimento dos enterramentos atômicos destes.

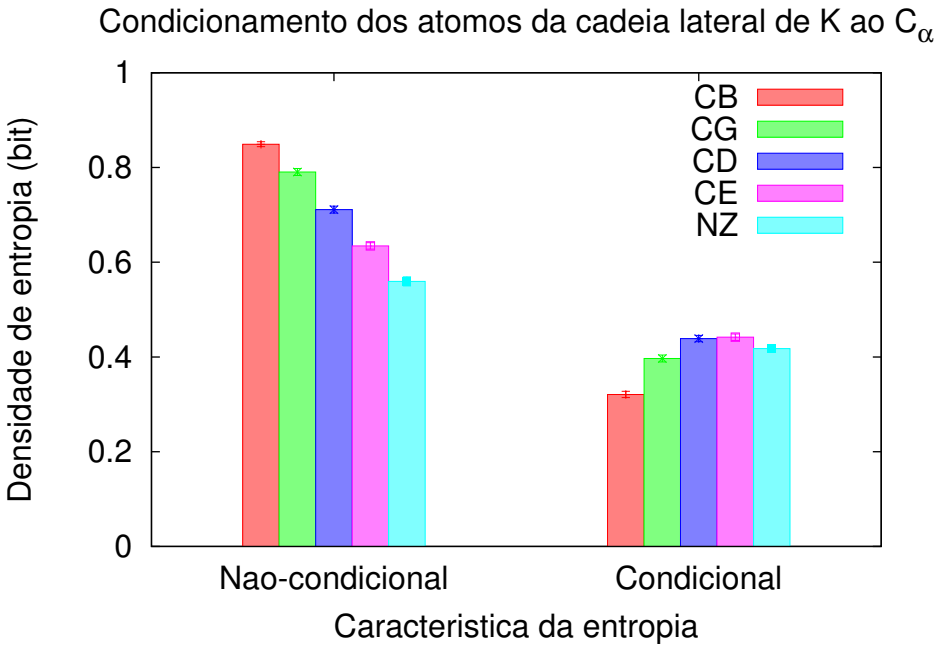


Figura 24: Efeito do condicionamento dos atomos da cadeia lateral do aminoácido Lisina ao C_α da mesma.

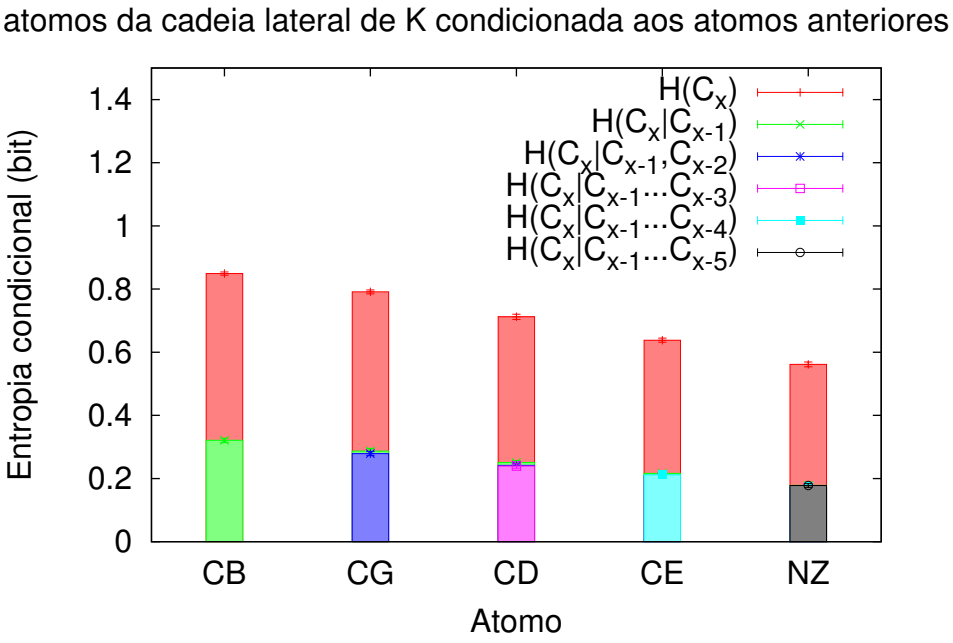


Figura 25: Entropia dos enterramentos dos átomos da cadeia lateral da Lisina condicionada aos átomos anteriores

Tabela 16: Entropia do enterramento de cada um dos átomos pertencentes às cadeias laterais dos resíduos condicional ao enterramento do C_α e à identidade do resíduo. Os números mostrados no topo da tabela representam o número de ligações covalentes que separam o átomo em questão do C_α do mesmo resíduo.

Resíduo	$H(B_i\|B_{C_\alpha},Q)$ (bit)					
	1	2		3		
A	$0,33\pm0,04$					
R	$0,36\pm0,09$	$0,5\pm0,1$		$0,58\pm0,09$		
N	$0,32\pm0,05$	$0,39\pm0,08$		$0,44\pm0,07$	$0,45\pm0,09$	
D	$0,30\pm0,05$	$0,39\pm0,08$		$0,40\pm0,03$	$0,42\pm0,09$	
C	$0,29\pm0,08$	$0,5\pm0,1$				
F	$0,30\pm0,07$	$0,42\pm0,07$		$0,3\pm0,1$		
E	$0,28\pm0,04$	$0,4\pm0,1$		$0,43\pm0,09$		
Q	$0,32\pm0,04$	$0,43\pm0,06$		$0,49\pm0,06$		
H	$0,32\pm0,05$	$0,45\pm0,04$		$0,51\pm0,05$	$0,53\pm0,05$	
I	$0,32\pm0,05$	$0,40\pm0,07$	$0,40\pm0,04$	$0,5\pm0,1$		
L	$0,33\pm0,04$	$0,4\pm0,1$		$0,5\pm0,$		
K	$0,30\pm0,04$	$0,4\pm0,1$		$0,4\pm0,1$		
M	$0,35\pm0,03$	$0,46\pm0,01$		$0,6\pm0,2$		
P	$0,28\pm0,03$	$0,36\pm0,03$	$0,324\pm0,005$			
S	$0,31\pm0,05$	$0,40\pm0,07$				
Y	$0,35\pm0,07$	$0,49\pm0,09$		$0,56\pm0,09$		
T	$0,32\pm0,05$	$0,4\pm0,1$	$0,4\pm0,1$			
W	$0,34\pm0,06$	$0,5\pm0,1$		$0,55\pm0,04$	$0,6\pm0,1$	
V	$0,31\pm0,04$	$0,42\pm0,05$				

Resíduo	$H(B_i\|B_{C_\alpha},Q)$ (bit)					
	4			5		6
R	$0,6\pm0,2$			$0,6\pm0,2$		
F	$0,6\pm0,1$			$0,7\pm0,1$		
E	$0,4\pm0,2$	$0,4\pm0,1$				
Q	$0,5\pm0,1$	$0,5\pm0,1$				
H	$0,60\pm0,05$	$0,62\pm0,05$				
K	$0,4\pm0,4$			$0,4\pm0,1$		
M	$0,63\pm0,07$					
Y	$0,7\pm0,1$			$0,7\pm0,1$		$0,8\pm0,2$
W	$0,7\pm0,1$	$0,66\pm0,08$	$0,6\pm0,1$	$0,8\pm0,1$	$0,7\pm0,1$	$0,77\pm0,09$

Tabela 17: Entropia dos átomos da cadeia lateral da Lisina condicionada aos átomos anteriores da mesma cadeia. C_x é o átomo em questão, mostrado na primeira coluna, e C_{x-i} é o átomo da i-ésima coluna à esquerda da coluna de C_x na tabela, ou seja, ao i-ésimo átomo anterior.

	C_β	C_γ	C_δ	C_ϵ	C_ζ
$H(C_x)$ (bit)	$0,849 \pm 0,005$	$0,791 \pm 0,006$	$0,712 \pm 0,008$	$0,638 \pm 0,007$	$0,561 \pm 0,008$
$H(C_x\|C_{x-1})$ (bit)	$0,321 \pm 0,007$	$0,287 \pm 0,008$	$0,250 \pm 0,006$	$0,216 \pm 0,006$	$0,178 \pm 0,005$
$H(C_x\|C_{x-1}, C_{x-2})$ (bit)		$0,279 \pm 0,006$	$0,241 \pm 0,006$	$0,207 \pm 0,005$	$0,172 \pm 0,005$
$H(C_x\|C_{x-1}...C_{x-3})$ (bit)			$0,240 \pm 0,007$	$0,203 \pm 0,006$	$0,170 \pm 0,006$
$H(C_x\|C_{x-1}...C_{x-4})$ (bit)				$0,211 \pm 0,005$	$0,175 \pm 0,005$
$H(C_x\|C_{x-1}...C_{x-5})$ (bit)					$0,178 \pm 0,006$

6 Discussão

Primeiramente, foi mostrado que a estrutura primária não possui correlação local (Tab. 3 e Fig. 2), ou seja, a identidade de um resíduo de aminoácido é independente das identidades dos seus vizinhos. Essa característica propicia que uma grande quantidade de informação seja armazenada na sequência de aminoácidos, informação esta que pode determinar as camadas dos enterramentos atômicos. Entretando, é intuitivo pensar que os elementos da sequência precisam ser correlacionados, por conta da formação de domínios. Assim, é possível que não haja correlação local, mas exista um grau de correlação não local entre as identidades dos resíduos de aminoácidos que não foi medido neste trabalho.

Por outro lado, a estrutura secundária e os enterramentos atômicos de C_α, C_β ou BB são correlacionados localmente (Tab. 4-5 e Fig. 5-7). No caso da estrutura secundária isso acontece porque α-hélices e folhas-β são arranjos locais, assim, se um resíduo de aminoácido faz parte de uma α-hélice, é muito provável que seus vizinhos também assumam a mesma estrutura secundária. Já no caso dos enterramentos atômicos, esse fato é observado porque os átomos que compõe a cadeia principal, assim como os das cadeias laterais, estão fisicamente ligados uns aos outros, o que restringe a liberdade dos mesmos, por exemplo, se um atómo está muito enterrado o seu vizinho não poderá estar completamente exposto. É aí que reside a correlação entre os enterramentos atômicos. Isso também explica a relação entre a medida da densidade de entropia dos átomos de BB, C_α e C_β, na qual a primeira é a menor e a última é a maior entre elas (Fig. 8)

O fato de a entropia da sequência ser maior que a do enterramento atômico torna teoricamente possível que aquela guarde a informação necessária para a determinação destes. Todavia, a simples comparação das densidades de entropia não é suficiente para se afirmar que esta informação está de fato presente na sequência. Essa relação pode ser medida pela transinformação, ou informação mútua.

Antes de tudo foi necessário provar a independência condicional entre os enterramentos atômicos locais e uma pequena janela de sequência local, para que fosse possível se utilizar a aproximação da Eq. 20 (Tab. 6 e Fig. 11). O uso dessa aproximação é essencial para o trabalho porque o tamanho limitado do banco de dados inviabiliza que sejam analizados simultaneamente blocos crescentes de enterramentos atômicos e de identidades de aminoácidos. Uma vez demonstrada a independência condicional entre esses elementos, o trabalho pôde prosseguir.

A transinformação foi calculada, primeiramente, a partir de três equações diferentes devido ao número de pontos passíveis de serem usados para os ajustes ser pequeno (Tab. 26-30 e Fig. 12-13). Destas três equações, uma foi eleita por sua coerência com o problema e por ter o menor número de parâmetros, característica desejada pois o banco de dados impôs uma limitação a janelas maiores que cinco ou seis resíduos (Tab. 10-11 e Fig. 14-15). Vale ressaltar que na eq. 26, c é relacionado ao tamanho de janela no qual o

enterramento de um resíduo ainda pode ser sentido por outro, b representa a importância dessa influência e a é a densidade de transinformação entre a identidade do resíduo central e os enterramentos de uma janela com tamanho tendendo ao infinito. Comparando-se a eq. 26 e a eq. 30 pode-se perceber que c é igual a $\frac{1}{\alpha}$, b é igual a $-\frac{\beta}{\alpha}$ e a é igual a $I(X;Y^1) + \frac{\beta}{\alpha}$ ou $I(X;Y^1) - b$. Ou seja, as duas equações são formas diferentes de se escrever o mesmo enunciado.

Foram feitas tentativas de se ajustar as equações acima nos pontos relativos à transinformação entre o enterramento de C_β e a sequência, mas não foi obtido sucesso (Fig. 16). Assim, houve a necessidade de se tratar o banco de dados com o objetivo de melhorar a estimativa dos pontos e, com isso, tornar possível uma melhor estimativa da densidade de transinformação entre enterramentos atômicos e identidade de aminoácidos.

Por isso, em um segundo momento, os pontos passaram a ser obtidos com o uso de pseudocontagm e *Bootstrap*. A combinação dessas duas técnicas possibilitou que as equações fossem ajustadas a mais pontos porque o *Bootstrap* corrigiu o vício dos pontos e a pseudocontagem tornou mais evidente a saturação do banco de dados uma vez que os valores para a transinformação diminuem quando a saturação é importante (Tab. 12-13 e Fig. 17-20).

De acordo com os resultados apresentados na Tab. 13, a transinformação entre os enterramentos atômicos de C_α ou de C_β e a sequência local de aminoácidos não se mostra maior que 15% da dúvida do enterramento. A fração $\frac{i(Q;B)}{h(B)}$, ainda que aparentemente pequena, deve ser suficiente para a predição dos enterramentos dos átomos mencionados pois muitas das conformações são restringidas pela própria cadeia por meio dos ângulos das ligações químicas. Conforme apresentado na Introdução, a ligação peptídica é uma ligação química simples com caráter de ligação dupla, o que a torna planar. Esta planaridade reduz as conformações possíveis da cadeia peptídica e cria pares de ângulos Φ e Ψ com maior probabilidade de ocorrência, que dão origem às estrutura secundárias ou aos *loops*, e pares que são proibidos (Ramachandran *et al.*, 1963). Além do mais, diferentemente da estrutura secundária, que é um arranjo puramente local, os enterramentos atômicos são determinados informações locais que dependem do tipo de resíduo de aminoácido e de sua vizinhança, e por informações não locais relacionadas à estrutura global da proteína e que não foram medidas neste trabalho.

Dessa forma, a estrutura nativa de uma proteína depende do diálogo entre informações locais e não locais a respeito das interações entre resíduos de aminoácidos e solvente. A importância da informação não local pode ser inferida através do pequeno número de conformações proteicas possíveis, que alguns estimam como sendo da ordem de 10^3 ou 10^4 (Govindarajan *et al.*,1999; Kooning *et al.*, 2002). Partindo do pressuposto que o número de estruturas possíveis $|\mathcal{A}_X|$ é igual a 10^4, de acordo com as eq. 1 e 2,

$$\begin{aligned} H(X) &\geq \log_2 |\mathcal{A}_X| \\ &\geq \log_2 10^4 \approx 13 \text{ bit.} \end{aligned}$$

Para uma proteína de 260 resíduos, isto representa uma dúvida da estrutura igual a $0,05$ bit/resíduo que é respondida pela sequência, ou seja, a densidade de transinformação estimada aqui deve ser maior ou igual a este valor. A Tab. 13 mostra que para o alfabeto de 20 letras isso é verdade.

Se forem considerados dois níveis de enterramento atômico para átomos de C_α, $0,95$ bit da dúvida devem ser respondidos por informações independentes da sequência. Uma vez que a densidade de entropia do enterramento desse tipo de átomo é aproximadamente $0,6$ bit/átomo, infere-se que $0,4$ bit (dos $0,95$ bit mencionados) é resolvido pela informação local dos enterramentos atômicos, ainda restando $0,55$ bit para ser respondido por informação não local. Para a entropia inicial de $1,5$ bit correspondente a três níveis de enterramento de C_β, $0,05$ bit da dúvida é resolvido pela sequência, $0,4$ bit é respondido por informação local independente da sequência e $1,05$ bits devem ser resolvidos por informações não locais.

Vale ressaltar que o valor de $0,05$ bit/resíduo como a medida da dúvida que é respondida pela sequência é apenas uma estimativa grosseira da realidade. Um dos aspectos que deve ser levado em consideração é que essa estimativa foi obtida simplesmente dividindo $H(X)$, dada por $\log_2 |\mathcal{A}_X|$, pelo número de resíduos de aminoácidos de uma proteína hipotética, assumindo que a dúvida do enterramento não depende da identidade do resíduo. Outro ponto é que os diferentes átomos que compõem os resíduos, em especial as cadeias laterais, tem entropias de enterramento diferentes entre si e essa dúvida é reduzida pelo condicionamento ao enterramento dos átomos vizinhos, conforme mostra a seção 5.6. Uma análise mais detalhada feita para o aminoácido Lisina e mostra que o condicionamento a apenas um átomo, sendo ele covalentemente ligado ao átomo em questão e mais próximo ao C_α do que este, é suficiente para reduzir consideravelmente a entropia do enterramento do mesmo (Tab. 17 e Fig. 25).

Ainda em relação à estimativa de $0,05$ bit/resíduo de informação respondida pela sequência, é fundamental para que a predição dos enterramentos atômicos seja possível que a densidade de transinformação entre sequência e enterramentos atômicos seja maior ou igual a essa estimativa. Isso porque é necessário que a informação esteja na sequência para que o algoritmo de predição seja capaz de recuperá-la. Além do mais, deve-se considerar que o algoritmo dificilmente conseguirá extrair toda a informação presente no ambiente, de forma que a quantidade na qual a transinformação supera a referida estimativa se torna uma margem de segurança que ainda mantém possível a predição dos enterramentos atômicos a partir da sequência.

Embora a aproximação da eq. 20 não seja válida para blocos de identidades de resíduos e que não seja possível calcular a densidade de transinformação entre sequência e enterramentos atômicos a partir de $I(Q^N; B_0)$, essa análise é importante porque o mais comum é que os algoritmos de predição utilizem fragmentos de sequência local para fazer sua predição. Dessa forma, essa estimativa serve de base para a avaliação do desempenho de algoritmos de predição.

Foi encontrada correlação entre estrutura secundária e enterramento atômico, o que pode ser de grande valia para a predição de estruturas terciárias de proteínas, pois o enterramento atômico é função da localização espacial do átomo em questão, e esta depende da estrutura secundária do resíduo de aminoácido ao qual este átomo pertence. Certas sequências de resíduos de aminoácidos são conhecidas por gerarem determinado tipo de estrutura secundária, o que levaria a uma forma indireta de se predizer os enterramentos atômicos. Contudo, chegar à estrutura nativa de uma proteína a partir da predição dos enterramentos atômicos baseada em estruturas secundárias preditas necessita de um algoritmo bastante robusto, capaz de ignorar o ruído do sistema e fornecer resultados razoáveis, e obter um algoritmo com essas características não é uma tarefa trivial.

Uma vez que Pereira de Araújo e Onuchic (2009) mostraram que é possível, em uma simulação de dinâmica molecular, se chegar à estrutura nativa a partir da informação das camadas de enterramento atômico, os enterramentos se mostraram como alternativa viável para a predição de estruturas terciárias de proteínas. Os resultados deste trabalho mostram que aproximadamente 15% da dúvida acerca das camadas de enterramento atômico é possível se resolvida a partir da sequência. Conforme discutido anteriormente, este valor aparentemente pequeno aparenta ser maior do que o que seria necessário para a correta predição das camadas em questão para C_α e C_β. Além disso, o estudo sobre a entropia condicional dos átomos das cadeias laterais mostrou que não é necessária a predição dos enterramentos de cada um dos átomos da proteína uma vez que o conhecimento das posições de alguns átomos determina o conhecimento das posições de outros átomos.

Essas informações combinadas, somadas a trabalhos paralelos desenvolvidos no Laboratório de Biologia Teórica e Computacional da Universidade de Brasília (LBTC-UnB), mostram a possibilidade real de se utilizar enterramentos atômicos para a predição de estruturas terciárias de proteínas e evidenciam que algoritmos que já estão sendo desenvolvidos no LBTC-UnB são capazes de extrair quase 100% da informação sobre os enterramentos atômicos presente na sequência de aminoácidos.

7 Conclusão

A totalidade das perspectivas traçadas no exame de qualificação foi concluída pois foram utilizadas as técnicas de pseudocontagem e *Bootstrap* para melhorar a estimativa dos pontos relativos às transinformações, foi provada a independência condicional das entropias e transinformações calculadas, e foi medida a entropia para cada um dos átomos dos diferentes resíduos ao longo da cadeia lateral condicionada ao enterramento do C_α correspondente. Também foi feita uma análise mais detalhada a respeito dos átomos da cadeia lateral do aminoácido Lisina e foi mostrado que o conhecimento do enterramento do átomo vizinho mais próximo ao C_α do mesmo resíduo reduz a entropia do enterramento do átomo em questão e que o condicionamento desta entropia aos outros átomos da cadeia lateral, além do vizinho, não a altera.

Ao longo deste trabalho foram investigadas as densidades de entropia das estruturas primárias, secundárias e enterramentos atômicos e a densidade de transinformação entre sequência e enterramentos aômicos. Foi observado que a sequência pode armazenar a informação para a determinação de camadas de enterramentos atômicos, mas que a transinformação entre essas duas grandezas não responde mais que 20% que a dúvida do último. Mesmo parecendo pouca, essa informação deve ser suficiente para que a predição dos enterramentos atômicos a partir da sequência de aminoácidos de uma proteína seja acurada o suficiente para levar à correta estrutura nativa. Por conta da última afirmação é que este trabalho é importante, pois a densidade de transinformação entre sequência e enterramentos atômicos provê um máximo teórico para a quantidade de informação que os algoritmos de predição são capazes de extrair da sequência.

Referências

ABKEVICH, Victor I.; GUTIN, Alexander M.; SHAKHNOVICH, Eugene I. Specific nucleus as the transition state for protein folding: evidence from lattice model. **Biochemistry**, Washington, v.33, p.10026-10036. 1994.

ALBERTS, Bruce *et al.* **Molecular Biology of The Cell**. Nova York: Garland Science, 2002.

ALON, Uri. **An Introduction to Systems Biology:** Design Principles of Biological Circuits. Londres: CRC Press, Taylor & Francis Group, 2007.

ANFINSEN, Christian B. Principles that govern the folding of proteins. **Science**, v.181, p.223-230. 1973.

BERMAN,Helen M. *et al.* The Protein Data Bank. **Nucleic Acids Research**, Oxford, v.28(1), p.235-242. 2000.

BONNEAU, Richard; Baker, David. Ab initio protein structure prediction: progress and prospects. **Annual Review of Biophysics and Biomolecular Structure**, Palo Alto, v.30, p.173-189. 2001.

CHITI, Fabrizio; DOBSON, Christopher M. Protein misfolding, functional amyloid and human disease. **Annual Review of Biochemistry**, Palo Alto, v.75, p.333-366. 2006.

COVER, Thomas M.; THOMAS, Joy A. **Elements of Information Theory**. 2. ed. Hoboken, Wiley-Interscience, 2006. 748p.

CROOKS, Gavin E.; BRENNER, Steven E. Protein secondary structure: entropy, correlations and prediction. **Bioinformatics**, Oxford, v.20(10), p.1603-1611. 2004.

CRUTCHFIELD, James A.; FELDMAN, David P. Regularities Unseen, Randomness Observed: Levels of Entropy Convergence. **Chaos**, Santa Fe, v.13(25), p.25-54. 2003.

DILL, Ken A.; CHAN, Hue S. From levinthal to pathways to funnels. **Nature Structural Biology**, v.4, p.10-19. 1997.

DILL, Ken A. *et al.* The protein folding problem: when will it be resolved?. **Current Opinion in Structural Biology**, v.17, p.342-346. 2007.

DILL, Ken A. *et al.* The protein folding problem. **Annual Review of Biophysics**, v.37. p.289-316. 2008.

DOLGIKH, D. A. *et al.* α-lactalbumin: compact state with fluctuating terciary structure? **FEBS letters**, v.136(2), p.311-315. 1981

DOOLITTLE, Russell F. Redundancies in protein sequences. In: FASMAN, Gerald D. **Prediction of protein structure and the principles of protein conformation**. Nova York e Londres: Plenum Press, 1989. cap.14, p.599-624.

EISENBERG, David; CROTHERS, Donald. **Physical chemistry with applications to the life sciences**. Teh Benjamin/Cummins Publishing Company, 1979. 868p.

FASMAN, Gerald D. The development of the prediction of protein structure. In:____. **Prediction of protein structure and the principles of protein conformation**. Nova York e Londres: Plenum Press, 1989. cap.6, p.193-316.

GOMES, Antonio L.C. **Transmissão de Informação entre Seqência Primária, Enterramentos Atômicos e Estrutura Tridimensional de Proteínas Globulares.** Brasília: Universidade de Brasília, 2007. Dissertação (Mestrado em Biologia Molecular) - Programa de Pós-Graduação em Biologia Molecular, Instituto de Biologia, Universidade de Brasília, Brasília, 2007.

GOMES, Antonio L. C. *et al.* Description of atomic burials in compact globular proteins by Fermi-Dirac probability distributions. *Proteins: Structure, Function and Bioinformatics*, v.66, p.304-320, 2007.

GOVINDARAJAN, Sridhar; RECABARREN, Ruben; GOLDSTEIN, Richard A. Estimating the total number of protein folds. *Proteins: Structure, Function and Genetics*, v.35, p.408-4414, 1999.

GRIEP, Sven; HOBOHM, Uwe. PDBselect 1992-2009 and PDBfilter-select. **Nucleic Acids Research**, Oxford, v. 38, p.D318-D319. 2010.

GUTIN, A. M.; ABKEVICH, V. I.; SHAKHNOVICH, E. I. Is burst hydrophobic collapse necessary for protein folding? *Biochemistry*, v.34, p.3066-3076. 1995.

HARDIN, Corey; POGORELOV, Taras V.; LUTHEY-SCHULTEN, Z. *Ab initio* protein structure prediction. *Current Opinion in Structural Biology*, v.12, p.176-181. 2002.

KAMTEKAR, Satwik *et al.* Hecht. Protein design by binary patterning of polar and nonpolar amino acids. *Science*, v.262, p.1680-1685. 1993.

KARANICOLAS, John; BROOKS, Charles L. An evolution of minimalist models for protein folding: from the behavior of protein-like polymers to protein function. *Drug Discovery Today: BIOSILICO*, v.2(3), p.127-133. 2004.

KOONING, Eugene V. *et al.* The structure of the protein universe and genome evolution. *Nature*, v.420, p.218-223. 2002.

LEACH, Andrew R. **Molecular Modelling: Principles and Applications**. Prentice Hall, 2001. 744p.

LEVINTHAL, Cyrus. **How to Fold Graciously**. In: MŐSSBAUN SECTROSCOPY IN BIOLOGICAL SYSTEMS. 1969, Monticello, Illinois. Illinois: University of Illinois Press, 1969. p.22-24.

LODISH, Harvey *et al.* **Molecular cell biology**. W. H. Freeman, 2000. 973p.

MATHEWS, Christopher K.; VAN HOLDE, Kensal E.; AHERN, Kevin G. **Biochemistry**. 3. ed. Redwood city: Benjamin Cummings, 1999. 1186p.

NELSON, David L.; COX, Michael M. **Lehninger Principles of Biochemistry**. 5. ed. Nova York: Worth, 2008. 1158p.

OGUSHI, Mikio; WADA, Akiyoshi. 'molten-globule state': a compact form of globular proteins with mobile side-chains. *FEBS*, v.164, p.21-24. 1983.

ONUCHIC, José N. O.; LUTHEY-SCHULTEN, Z.; WOLYNES, Peter G. Theory of protein folding: The energy landscape perspective. *Annual Review of Physical Chemistry*, v.48, p.545-600. 1997.

PEREIRA DE ARAÚJO, Antônio F.; ONUHIC, José N. A sequence-compatible amount of native burial information is sufficient for determining the structure of small globular proteins. *Proceedings of the National Academy of Sciences*, Washington, v.106(45), p.19001-19004. 2009.

PEREIRA DE ARAÚJO, Antônio F. *et al.* Native atomic burials, supplemented by physically motivated hydrogen bond constraints, contain sufficient information to determine the tertiary structure of small globular proteins. **Proteins**, v.70, p.971-983. 2008.

RAMACHANDRAN, G. N.; RAMAKRISHNAN, R.; SASISAKHARAN, V. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, v.7, p.95-99. 1963.

SHANNON, Claude E. A mathematical theory of communication.**Bell System Technical Journal**. Illinois, v.27, p.379-423 e p.623-656. 1949.

So much more to know. [Editorial] *Science*, v.309, p. 78-102. 2005.

TREPTOW, Werner L. *et al.* Non-native Interactions, Effective Contact Order, and Protein Folding: A Mutational Investigation With the Energetically Frustrated Hydrophobic Model. *Proteins: Structure, Function, and Genetics*, v.49, p.167-180. 2002.

RCSB Protein Data Bank. Disponível em: <http://www.rcsb.org/pdb/home/home.do>. Acesso em: 15 jan 2012.

APÊNDICE A — *Scripts* utilizados

O formato geral de entrada dos arquivos do banco de dados é constituído por duas colunas, a primeira correspondendo ao código de três letras dos aminoácidos e a segunda correspondendo seu enterramento atômico. Quando há quebras na cadeia, são colocados "-" para marcá-las e as proteínas são separadas por linhas em branco ou por "b". Esses arquivos entram nos dois próximos *scripts*, que chamam os vários seguintes. A seguir é calculada a média das repetições do *Bootstraps* e o desvio é corrigido. Os pontos resultantes são os que deram origem a todos os dados deste trabalho.

Script geral para os cálculos sem pseudocontagem e sem *Bootstrap*

O *script* a seguir foi usado para fazer os cálculos dos pontos de entropia e transinformação sem o uso de pseudocontagem e *Bootstrap*. Este é o *script* geral, usado para chamar os que tratam de partes específicas do cálculo. seq2HP.awk pode ser substituído por seq2HPN.awk, seq2HPNhp.awk ou seq2letra.awk, de acordo com o alfabeto de interesse.

```
for ((niv=2; niv<=5; niv++)); do
for ((tam1=1; tam1<=21; tam1++)); do
#for ((tam2=1; tam2<=21; tam2++)); do
echo $i `awk '! nres {nres=$4}; $2=="N" || $2=="CA" || $2=="C"
{if($4-nres>1 || nres-$4>1) print "- -"; print $3, $6; nres=$4}' |
awk -v niveis=$niv -f classificaenterramento.awk |
awk -f seq2HP.awk |
awk -f fazjanelaXY.awk -v tj1=$tam1 -v tj2=$tam2 | awk '$2 !~ /-/' |
awk -f transinf.awk | awk '{print $1, $2, $3, $4, $5}'`;
done;
done
```

Vale deixar claro que as variáveis *tam1* e *tam2* são as que determinam, respectivamente, os tamanhos da janelas de sequência e de enterramentos atômicos e que elas não são usadas simultaneamente, por conta da já mencionada saturação do banco de dados. Depois de passar por cada um dos *scripts* que tratam partes de específicas do cálculo (como a classificação das identidades dos aminoácidos em alfabetos reduzidos ou a classificação dos enterramentos atômicos em diferentes números de camadas), o banco de dados passa pelo *script* que faz o cálculo de cada uma das grandezas de interesse. No caso dos pontos calculados sem pseudocontagem e *Bootstrap*, o *script* responsável por isso recebe o nome de transinf.awk e é detalhado mais adiante.

Script geral para os cálculos com pseudocontagem e sem *Bootstrap*

O *script* a seguir foi usado para fazer os cálculos dos pontos de entropia e transinformação com o uso de pseudocontagem e sem o uso de *Bootstrap*. Esses cálculos são feitos porque são esses resultados que têm seu desvio corrigido pela média dos *Bootstraps*, ou seja, os pontos corrigidos são resultado da diferença entre a média dos *Bootstraps* e os pontos resultantes desse *script*.

Este é o *script* geral, usado para chamar os que tratam de partes específicas do cálculo. seq2HP.awk pode ser substituído por seq2HPN.awk, seq2HPNhp.awk ou seq2letra.awk, de acordo com o alfabeto de interesse.

```
for ((niv=2; niv<=5; niv++)); do
for ((tam1=1; tam1<=21; tam++)); do
#for ((tam2=1; tam2<=21; tam++)); do
awk -v niveis=$niv -f classificaenterramento.awk |
awk  -f seq2HP.awk |
awk -v tj1=$tam1 -v tj2=$tam2 -f fazjanelaXY.awk |
awk '$0!~/b/ && $0!~/-/ {print $1, $2, "1"}' |
echo $tam `awk -v tjy=$tam -f ~/bin/prob.awk23` ;
done;
done
```

Vale deixar claro que, assim como no caso anterior, as variáveis *tam1* e *tam2* são as que determinam, respectivamente, os tamanhos da janelas de sequência e de enterramentos atômicos e que elas não são usadas simultaneamente, por conta da já mencionada saturação do banco de dados.

Script geral para os cálculos com pseudocontagem e com *Bootstrap*

O *script* a seguir foi usado para fazer os cálculos dos pontos de entropia e transinformação com o uso de pseudocontagem e *Bootstrap*. Este é o *script* geral, usado para chamar os que tratam de partes específicas do cálculo. Da mesma forma que no *script* anterior, seq2HP.awk pode ser substituído por seq2HPN.awk, seq2HPNhp.awk ou seq2letra.awk, de acordo com o alfabeto de interesse.

```
for ((niv=2; niv<=5:niv++)); do
for ((b=1; b<=50; b++)); do
for ((tam=1; tam<=21; tam++)); do
#for ((tam2=1; tam2<=21; tam++)); do
awk -v niveis=$niv -f classificaenterramento.awk |
awk  -f seq2HP.awk |
awk -v tj1=$tam1 -v tj2=$tam2 -f fazjanelaXY.awk |
bash core.sh |
awk '$3!=0' |
echo $b $tam `awk -v tjy=$tam -f prob.awk23` ;
done;
done;
done
```

Vale deixar claro que, assim como nos casos anteriores, as variáveis *tam1* e *tam2* são as que determinam, respectivamente, os tamanhos da janelas de sequência e de enterramentos atômicos e que elas não são usadas simultaneamente, por conta da já mencionada saturação do banco de dados.

classificaenterramento.awk

O *script* a seguir tem como entrada os valores contínuos dos enterramentos atômicos (medidos em unidades de raio de giro) e os classifica em diferentes níveis de enterramento atômico. O número de níveis é uma variável definida pelo usuário no momento da classificação ou em *script* acessório. Neste caso, o número de níveis de enterramento atômico consta nos dois primeiros *script* deste apêndice.

```
BEGIN {
    if (!niveis) niveis=2;
}
{
    #if ($NF=="b") nivel="b";
    if ($NF=="-" || $NF=="b") nivel=$NF;
    else if (niveis==2) {
        if ($NF<0.95) nivel=0;
        else nivel=1;
    }
    else if (niveis==3) {
        if ($NF<0.8) nivel=0;
        else if ($NF<1.075) nivel=1;
        else nivel=2;
    }
    else if (niveis==4) {
        if ($NF<0.73) nivel=0;
        else if ($NF<0.95) nivel=1;
        else if ($NF<1.13) nivel=2;
        else nivel=3;
    }
    else if (niveis==5) {
        if ($NF<0.68) nivel=0;
        else if ($NF<0.875) nivel=1;
        else if ($NF<1.025) nivel=2;
        else if ($NF<1.175) nivel=3;
        else nivel=4;
    }
    else if (niveis==6) {
        if ($NF<0.63) nivel=0;
        else if ($NF<0.8) nivel=1;
        else if ($NF<0.95) nivel=2;
        else if ($NF<1.075) nivel=3;
        else if ($NF<1.22) nivel=4;
        else nivel=5;
    }
    else if (niveis==7) {
        if ($NF<0.6) nivel=0;
        else if ($NF<0.775) nivel=1;
        else if ($NF<0.9) nivel=2;
        else if ($NF<1.0) nivel=3;
        else if ($NF<1.1) nivel=4;
        else if ($NF<1.225) nivel=5;
        else nivel=6;
    }
    else if (niveis==8) {
        if ($NF<0.575) nivel=0;
        else if ($NF<0.73) nivel=1;
```

```
        else if ($NF<0.85) nivel=2;
        else if ($NF<0.95) nivel=3;
        else if ($NF<1.03) nivel=4;
        else if ($NF<1.13) nivel=5;
        else if ($NF<1.27) nivel=6;
        else nivel=7;
    }
    else if (niveis==9) {
        if ($NF<0.55) nivel=0;
        else if ($NF<0.72) nivel=1;
        else if ($NF<0.775) nivel=2;
        else if ($NF<0.9) nivel=3;
        else if ($NF<0.98) nivel=4;
        else if ($NF<1.075) nivel=5;
        else if ($NF<1.17) nivel=6;
        else if ($NF<1.275) nivel=7;
        else nivel=8;
    }
    else if (niveis==10) {
        if ($NF<0.53) nivel=0;
        else if ($NF<0.68) nivel=1;
        else if ($NF<0.78) nivel=2;
        else if ($NF<0.875) nivel=3;
        else if ($NF<0.95) nivel=4;
        else if ($NF<1.025) nivel=5;
        else if ($NF<1.1) nivel=6;
        else if ($NF<1.175) nivel=7;
        else if ($NF<1.32) nivel=8;
        else nivel=9;
    }
    $NF=nivel;
    print $0;
}
```

seq2HP.awk

Script responsável por classificar os tipos de resíduos de aminoácidos em hidrofóbicos ou polares, ou seja, este *script* faz o alfabeto HP.

```
BEGIN {
    t["ALA"]=t["CYS"]=t["PHE"]=t["GLY"]=t["ILE"]=t["LEU"]=t["MET"]=t["VAL"]=t["TRP"
    t["ASP"]=t["GLU"]=t["HIS"]=t["LYS"]=t["ASN"]=t["PRO"]=t["GLN"]=t["ARG"]=t["SER"
}
{$1 = t[$1] ? t[$1] : $1; print $0}
```

seq2HPN.awk

Script responsável por classificar os tipos de resíduos de aminoácidos em hidrofóbicos, polares ou neutros, ou seja, este *script* faz o alfabeto HPN.

```
BEGIN {
    t["CYS"]=t["PHE"]=t["ILE"]=t["LEU"]=t["MET"]=t["VAL"]=t["TRP"]=t["TYR"]="H";
    t["ASP"]=t["GLU"]=t["LYS"]=t["ASN"]=t["PRO"]=t["GLN"]=t["ARG"]="P";
    t["ALA"]=t["GLY"]=t["HIS"]=t["SER"]=t["THR"]="N";
}
{$1 = t[$1] ? t[$1] : $1; print $0}
```

seq2HPNhp.awk

Script responsável por classificar os tipos de resíduos de aminoácidos em muito hidrofóbicos, puco hidrofóbicos, muito polares, pouco polares ou neutros, ou seja, este *script* faz o alfabeto HPNhp.

```
BEGIN {
    t["PHE"]=t["ILE"]=t["LEU"]=t["VAL"]=t["TRP"]="H";
    t["ASP"]=t["GLU"]=t["LYS"]="P";
    t["ALA"]=t["GLY"]=t["HIS"]=t["SER"]=t["THR"]="N";
    t["CYS"]=t["MET"]=t["TYR"]="h";
    t["ASN"]=t["PRO"]=t["GLN"]=t["ARG"]="p";
}
{$1 = t[$1] ? t[$1] : $1; print $0}
```

seq2letra.awk

Script responsável por transformar o código de três letras das identidades dos aminoácidos no código de uma letra, ou seja, este *script* faz o alfabeto de 20 letras.

```
BEGIN {
    t["CYS"]="C";
    t["ILE"]="I";
    t["LEU"]="L";
    t["MET"]="M";
    t["VAL"]="V";
    t["ASN"]="N";
    t["PRO"]="P";
    t["GLN"]="Q";
    t["ALA"]="A";
    t["GLY"]="G";
    t["HIS"]="H";
    t["SER"]="S";
    t["THR"]="T";
    t["PHE"]="F";
    t["TRP"]="W";
    t["TYR"]="Y";
    t["ASP"]="D";
    t["GLU"]="E";
    t["LYS"]="K";
    t["ARG"]="R";
}
{$1 = t[$1] ? t[$1] : $1; print $0}
```

fazjanelaXY.awk

Este *script* forma as janelas crescentes de identidades de resíduos e/ou de enterramentos atômicos. O elemento central é sempre o da linha em questão e as janelas crescem à medida que são adicionados elementos das linhas anteriores e posteriores alternadamente. Quando não há elementos para serem adicionados (no caso das extremidades das cadeias), são colocados -"e depois essas linhas são excluídas pelos *scripts* acessórios (dois primeiros *scripts* deste apêndice).

```
BEGIN {if(!tj1) tj1=1; if(!tj2) tj2=1}
{x[NR]=$1; y[NR]=$2}
END{
    for(i=1; i<=NR; i++){
        X=Y="";
        for(j=i-int(tj1/2); j<i-int(tj1/2) + tj1; j++)
            X = X ((j>=1 && j <=NR) ? x[j]:"-");
        for(j=i-int(tj2/2); j<i-int(tj2/2)+tj2; j++)
            Y = Y ((j>=1 && j <=NR) ? y[j]:"-");
        print X, Y
    }
}
```

transinf.awk

Este é o *script* responsável por fazer a contagem das janelas sem a utilização de pseudocontagem e sem os pesos atribuídos pelo *Bootstrap*.

```
BEGIN{SUBSEP=" "}
{X[$1]++; Y[$2]++; XY[$1,$2]++; n++; x[$1,$2]=$1; y[$1,$2]=$2;}
END{
    for(i in X)
    {
        Px[i]=(X[i]/n); infx[i]= -log(Px[i])/log(2); Tx+=Px[i]*infx[i];
    }
    for(i in Y)
    {
        Py[i]=(Y[i]/n); infy[i]= -log(Py[i])/(log(2)); Ty+=Py[i]*infy[i];
    }
    for(i in XY)
    {
        Pxy[i]=(XY[i]/n); infxy[i]= -log(Pxy[i])/(log(2)); Txy+=Pxy[i]*infxy[i]
    }
    print "TOTAL",  Tx,Ty,Txy, Tx+Ty-Txy, n, Txy-Ty;
}
```

core.sh

Este é o *script* responsável por preparar a entrada do scripts a seguir, que é responsável por fazer as repetições do *Bootstrap*.

```
awk '{if ($1~/b/) print "   "; else if ($2!~/b/) print}' |
awk -f ~/bin/resample.awk2 |
awk '$0!~/b/ && $0!~/-/ && $0!~/X/'
```

resample.awk

Conforme dito anteriormente, este é o *script* responsável por fazer as repetições do *Bootstrap*. A cada quebra na cadeia, ele gera um número aleatório que passa a ser o peso da proteína em questão. O *Bootstrap* foi feito assim para que não fosse necessário serem feitas 50 cópias do banco de dados.

```
function poisson(lambda,       k, max_k, p, P, sum) {
    k=0;
    max_k=1000;
    p=rand();
    P=exp(-lambda);
    sum=P;
    if (sum>=p) return 0;
    for (k=1; k<max_k; ++k) {
        P*=lambda/k;
        sum+=P;
        if (sum>=p) break;
    }
    return k;
}

BEGIN{srand(); weight=poisson(1)}
NF {print $0, weight}
!NF {weight=poisson(1)}

#const int PoissonRandomNumber(const double lambda)
#{
#  int k=0;                          //Counter
#  const int max_k = 1000;           //k upper limit
#  double p = UniformRandomNumber(); //uniform random number
#  double P = exp(-lambda);          //probability
#  double sum=P;                     //cumulant
#  if (sum>=p) return 0;             //done allready
#  for (k=1; k<max_k; ++k) {         //Loop over all k:s
#    P*=lambda/(double)k;            //Calc next prob
#    sum+=P;                         //Increase cumulant
#    if (sum>=p) break;              //Leave loop
#  }
#
#  return k;                         //return random number
#}
```

prob.awk23

Script que realiza a contagem das janelas considerando os pesos atribuídos a elas pelo *script* anterior ou considerando o peso igual a um no caso da repetição sem os pesos aleatórios.

```
function zog(x) {if(x==0){return 0} else {return log(x)/log(2)}}

BEGIN{SUBSEP=" "}
{
n+=$3;
nx[$1]+=$3;
ny[$2]+=$3;
nxy[$1,$2]+=$3;
X[$1,$2]=$1
Y[$1,$2]=$2

ny0[substr($2, int(length($2)/2)+1, 1)]+=$3;
nxy0[$1, substr($2, int(length($2)/2)+1, 1)]+=$3;
tj2=length($2);
}

END{
for(x in nx){
        px[x]=nx[x]/n;
        hx-=px[x]*zog(px[x]);
        }
for(y in ny){
        py[y]=ny[y]/n;
        hy-=py[y]*zog(py[y]);
        }

for(xy in nxy){
        pxy[xy]=nxy[xy]/n;
        pxgy[xy]=nxy[xy]/ny[Y[xy]];

        hxy-=pxy[xy]*zog(pxy[xy]);
        hxgy-=pxy[xy]*zog(pxgy[xy]);
        }

for(y0 in ny0){
        py0[y0]=ny0[y0]/n
        for(x in nx){
                p1[x,y0]=nxy0[x,y0]/n;
                p0[x,y0]=nxy0[x,y0]/ny0[y0];
                }
        }

for(y in ny){
```

```
        for(x in nx){
                ly=substr(y, int(length(y)/2)+1, 1)
                if(!nxy[x,y]){nxy[x,y]=0}
                pxy0[x,y]=(nxy[x,y] + (20*p1[x,ly]))/(n+20)
                pxgy0[x,y]=(nxy[x,y] + (20*p0[x,ly]))/(ny[y]+20)

                hxy0-=pxy0[x,y]*(zog(pxy0[x,y]))
                hxgy0-=pxy0[x,y]*(zog(pxgy0[x,y]))

                S[y]-=pxgy0[x,y]*(zog(pxgy0[x,y]))
                }
        h0xgy+=S[y]*py[y]
        }


printf("%s H(X)= %1.4f H(Y)= %1.4f H(X,Y)= %1.4f H'(X|Y)= %1.4f
I(X;Y)= %1.5f I'(X;Y)= %1.5f\n", n, hx, hy, hxy, h0xgy, hx+hy-hxy,
hx-h0xgy)
}
```

BIOINFORMATICS

Vol. 00 no. 00 2011
Pages 1–7

Information-theoretic analysis of atomic burials in globular proteins: On the search for an informational intermediate between sequence and structure

Juliana R. Rocha, Diogo C. Ferreira, Paulo H. Azevêdo and Antônio F. Pereira de Araújo*

Laboratório de Biologia Teórica e Computacional, Departamento de Biologia Celular, Universidade de Brasília, Brasília-DF 70910-900, Brazil

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: It has been recently suggested that atomic burials, as expressed by molecular central distances, contain sufficient information to determine the tertiary structure of small globular proteins. A possible approach to structural determination from sequence could therefore involve a sequence-to-burial intermediate prediction step whose accuracy, however, is theoretically limited by the mutual information between these two variables.

Methods: We use a non-redundant set of globular protein structures to estimate the mutual information between local amino acid sequence and atomic burials. Representing amino acids with various reduced alphabets and discretizing central distances of C_α or C_β atoms in equiprobable burial levels, we estimate relevant mutual information measures to which actual burial prediction algorithms should be compared.

Results: Mutual information density for 20 amino acid letters was estimated to be around 15% of the unconditional burial entropy density, ranging from approximately 0.09 out of 0.6 bit/atom for C_α atoms and $L = 2$ burial levels to around 0.2 out of 1.7 bits/atom for C_β atoms and $L = 5$ burial levels. Lower estimates for the mutual information between single residue burial and local amino acid sequence, obtained under the approximation of independence between amino acids conditional to single burial, were found to be consistent with corresponding densities although tending to display slightly lower values, particularly for C_α atoms.

Contact: aaraujo@unb.br

1 INTRODUCTION

It has become a common statement in biology that amino acid sequences contain sufficient information to determine protein tertiary structures. Fulfillment of the implied possibility of structure prediction from sequence is actually considered one of the most important unsolved problems of molecular biophysics, as recently reviewed by different groups (Dill *et al.*, 2008; Shakhnovich, 2006; Onuchic and Wolynes, 2004). Such an intrinsically informational assertion, however, has not been as extensively investigated within the context of Shannon's information theory. Although informational concepts have been used in algorithms for secondary structure prediction from local sequence since the seventies (Garnier *et al.*, 1978), for example, the limit imposed on any prediction by the mutual information between these two quantities was estimated only a few years ago (Crooks and Brenner, 2004). Incidentally, an informational analysis of backbone dihedral angles has also exposed the unfeasibility of tertiary structure determination from an even perfect three-state secondary structure prediction (Solis and Rackovsky, 2004). More generally, the recurrent utilization of statistical potentials in computational biology has only recently been interpreted explicitly in informational terms (Solis and Rackovsky, 2007). Other recent applications of information theory to studies of protein structure include the analysis of pairwise contact potentials (Cline *et al.*, 2002), which revealed a rather modest mutual information between the identity of contact partners, and the investigation of general distance constraints in minimalist protein models (Sullivan *et al.*, 2003).

Contrasting with secondary structure, atomic burials appear to encode sufficient information for structural determination. Contrasting with pairwise contacts, they have a much better chance of being adequately estimated from sequence information. Monte Carlo simulations of geometrically realistic protein models using native burial information, as expressed by atomic distances from the molecular center, have successfully recovered the tertiary structure of small globular proteins (Pereira de Araújo *et al.*, 2008). A simple computational experiment combining Molecular Dynamics of similar models with discretized burial levels has additionally provided an upper bound for the amount of required burial information. It actually turned out to be comparable to, and therefore encodable by, the information (entropy) of local protein sequences (Pereira de Araújo and Onuchic, 2009). The suggested qualitative distinction can be rationalized intuitively. Pairwise contacts provide a nonlocal representation, even if simplified, of nonlocal information while secondary structure, on the other extreme, is a local representation of purely local information. Burials, however, include nonlocal information in a local representation, as is evident from the requirement of knowledge about the whole tertiary structure for assigning burials, but not secondary structure, of any short protein fragment. The possibility of structural determination from

*to whom correspondence should be addressed

sequence-dependent burial information, when combined to appropriate sequence-independent constraints, is also consistent with the perceptible previous success in native fold recognition from the arrangement of hydrophobic and polar residues (Huang *et al.*, 1995). It has been further supported more recently by a purely analytical model which was able to recover native-like burial traces from sequence hydrophobicity information combined to a simple size constraint intended to approximate the overall effect of excluded volume (England, 2011).

A potential approach to tertiary structure prediction could therefore involve a sequence-to-burial intermediate prediction step. It must be noted, nevertheless, that theoretical encodability, as provided by entropy compatibility, is necessary but not sufficient to demonstrate actual encoding. The accuracy of any burial prediction from sequence must be further limited by the observed correlation between burials and sequences, as conveniently quantified by the mutual information between these two quantities. In the present study we estimate the mutual information between burials and local amino acid sequence in globular proteins. The resulting fraction of sequence entropy actually involved in burial encoding provides theoretical limits to which prediction algorithms should be compared.

2 THEORETICAL BACKGROUND

A fundamental concept of Shannon's information theory is the entropy, $H(X)$, of a discrete random variable X assuming values in an "alphabet" $\mathcal{X} = \{x_1, \cdots, x_L\}$ with probabilities $p(X) = \{p(x_1), \cdots, p(x_L)\}$, which is given, in bits, by

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x). \tag{1}$$

As discussed in standard text-books on information theory, e.g. (Reza, 1961; Cover and Thomas, 2006), entropy is an appropriate measure of the uncertainty associated to the particular random variable, varying itself continuously from the minimal value of zero, when $p(\tilde{x}) = 1$ for a single $x = \tilde{x}$ and $p(x) = 0$ for all $x \neq \tilde{x}$, to the maximal value of $\log_2 L$, when all possible values are equally probable, or $p(x) = 1/L$ for all x. The concept is naturally extended to joint and conditional probabilities, resulting in joint and conditional entropies, $H(X,Y)$ and $H(X|Y)$, respectively. The mutual information between two random variables, $I(X;Y) = I(Y;X)$, is a symmetric measure of the amount of information about any one of the variables contained in the second variable, and is given simply by the decrease in the uncertainty of one variable when the other variable is known,

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y). \end{aligned} \tag{2}$$

The entropy, $H(X^N)$, associated with a given set of $M = L^N$ possible sequences of N symbols or "letters" or, in other words, the entropy of the random variable obtained by concatenation of N random variables, $X^N = X_1X_2\cdots X_N$, with individual symbols chosen from the same alphabet of L possible letters, $\mathcal{X} = \{x_1, \cdots, x_L\}$, and associated probabilities $p(X^N) = \{p(x^N)\}$, is therefore given by

$$H(X^N) = -\sum_{x^N\in\mathcal{X}^N} p(x^N)\log_2 p(x^N). \tag{3}$$

Due to unavoidable finite size effects, however, it is clear that these probabilities tipycally *cannot* be directly estimated from corresponding frequencies observed in any real data set of representative examples. The situation is greatly simplified if the sequences can be considered to be "statistically homogeneous", by which it is meant that

$$p(x_i, \cdots, x_{i+m-1}) = p(x_1, \cdots, x_m) \tag{4}$$

for any integers i, m with $i + m - 1 \leq N$, and furthermore if they are "m-order markovian", i.e.

$$p(x_i|x_{i-1}, \cdots, x_1) = p(x_i|x_{i-1}, \cdots, x_{i-m}) \tag{5}$$

with $m \leq i \leq N$, or in other words, if different sequences can be regarded as different realizations of a single stationary and m-order markovian stochastic process. In this case it is not dificult to show that the entropy increases linearly with N, for $N > m$, as

$$H(X^N) = Nh(X) + E(m), \tag{6}$$

where

$$h(X) = H(X_{m+1}|X_m, \cdots, X_1) \tag{7}$$

is the "entropy density", or "entropy rate", and

$$E(m) = \sum_{n=1}^{m-1} [H(X_n|X_{n-1}, \cdots, X_1) - h(X)] \tag{8}$$

is the non-extensive "excess entropy" which depends on m but not on N. Note that $E(0) = 0$, when adjacent letters are statistically independent, while $E(1) = H(X_1) - h(X)$, and $E(2) = H(X_1) + H(X_2|X_1) - 2h(X)$, and so on.

From the frequencies observed in a data set of representative examples, the entropy density $h(X)$, order of markovicity m, and excess entropy $E(m)$, can be estimated from the size dependence of the entropy of blocks formed by N adjacent letters as far as $m < N < m^*$, where m^* is the value above which frequencies of individual blocks in the data set cease to resemble their probabilities, a situation we refer to as "saturation". A simple upper estimate for m^* can be obtained from considering that for sufficiently large block sizes, $N' > m^*$, frequencies of individual blocks in any finite data set become either 1 or 0, depending on wether or not the block happens to be present in the set, resulting in an (incorrect) entropy estimate of $H(X^{N'}) = \log_2 C$, where C is the sequence-independent number of N'-blocks in the data set. The onset of this anomalous but predictable behavior can therefore be used to determine an upper estimate for m^*.

For the mutual information between two types of sequences, X^N and Y^N, again assuming homogeneity and markovicity, we have

$$\begin{aligned} I(X^N;Y^N) &= H(X^N) + H(Y^N) - H(X^N,Y^N) \\ &= N\,[h(X) + h(Y) - h(X,Y)] + \\ &\quad (E_X + E_Y - E_{X,Y}) \\ &= Ni(X;Y) + E_{X;Y}, \end{aligned} \tag{9}$$

or

$$i(X;Y) = \lim_{N\to\infty} \frac{I(X^N;Y^N)}{N}, \tag{10}$$

and an estimate for the corresponding mutual information density, $i(X;Y)$, a quantity of much interest that imposes an upper limit on any possible prediction of the local sequence of one variable from the local sequence of the other variable, can again be obtained from N-block entropy estimates. In this case, however, because the number of blocks now increases with block size as $L_{\mathcal{X}}^N L_{\mathcal{Y}}^N$, saturation should occur at a much shorter block length. The situation is somewhat improved if adjacent letters in one of the sequences, e.g. X^N, can be considered as statistically independent both unconditionally and conditionally to the other sequence, Y^N, in which case the mutual information density turns out to be expressible as the limit

$$i(X;Y) = \lim_{N\to\infty} I(X;Y^N) = I(X;Y^\infty), \tag{11}$$

and therefore obtainable by extrapolation from the behavior of blocks whose number increases only as $L_{\mathcal{X}} L_{\mathcal{Y}}^N$. This procedure was used by Crooks and Brenner, 2004, to estimate the mutual information density between sequences of amino acid residues and corresponding sequences of secondary structure assignments.

It should be stressed that $I(X^\infty;Y) \neq I(X;Y^\infty)$ and the limit in eq. 11 must be taken on the correlated variable Y and not on the uncorrelated variable X. We note, however, that $I(X^\infty;Y)$ is also interesting since it provides a limit for the prediction of individual Y values given the local sequence of X. Saturation might again become a problem when $L_{\mathcal{X}}$ is large, in which case it might be useful to consider the following lower bound,

$$\begin{aligned} \sum_{i=1}^{N} I(X_i;Y) &= \sum_{i=1}^{N} [H(X_i) - H(X_i|Y)] \\ &\leq H(X^N) - H(X^N|Y) \\ &= I(X^N;Y). \end{aligned} \tag{12}$$

In this last equation, each of the N "positional" mutual information terms between X_i and Y, with i indicating position within X^N, is computed from just $L_{\mathcal{X}} \times L_{\mathcal{Y}}$ possible combinations independently of N. The inequality disappears in case of statistical independence between elements of the sequence X^N not only unconditionally, as already assumed, but also conditionally to single Y values.

3 METHODS

In the present study we estimated probabilities from frequencies observed in a data set of representative globular structures derived from PDBSELECT (Hobohm and Sander, 1994). From the list made available in Nov. 2009, we selected structures determined by X-ray chrystalography with resolution better than 2.5Å and excluded chains not satisfying the globularity criterion given by the expected relation between radius of gyration and the number of residues, $R_g \leq 2.9 N_r^{1/3}$Å (Gomes *et al.*, 2007). Membrane proteins were also excluded, simply by removing PDB files containing the word "MEMBRANE". The resulting collection, from now on simply referred to as the data bank, is composed of 1499 chains, with a total of approximately 263000 residues. Statistical errors on computed probabilities and entropies were estimated, and systematic biases corrected for, by a bootstrap procedure using 50 randomly generated replicas of the data bank. In addition to the complete alphabet of 20 amino acid identities, we have also used the reduced alphabets HP, HPN and HhPpN. Hyrophopbic and polar residues are grouped in the HP alphabet as $\mathrm{H} = \{\mathrm{A,C,F,G,I,L,M,V,W,Y}\}$ and $\mathrm{P} = \{\mathrm{D,E,H,K,N,P,Q,R,S,T}\}$, respectively. In HPN a third, "neutral", class includes residues from both HP groups, $\mathrm{N} = \{\mathrm{A,G,H,S,T}\}$. Remaining hydrophobic and polar residues are further subdivided in HhPpN. Burials were obtained from the atomic distances from the molecular center of C_α or C_β atoms, normalized by the radius of gyration, and grouped in approximately equiprobable burial levels, resulting in a collection of burial alphabets $\{\chi L\}$, where χ is either α or β, representing the atomic type for which burials are defined, and L is the number of burial layers. We usually use superscripts to indicate block size and integer subscripts to indicate position within the block, with "0" representing the central block position by convention. If necessary, however, we also indicate particular alphabets as subscripts in our notation, such as $H(Q_{\mathrm{HP}}^N)$, $h(B_{\beta 5})$, $I(Q_{20}^N; B_{\alpha 2}^N)$, etc.

N-block entropies for residue identities, $H(Q^N)$, and burials, $H(B^N)$, were computed according to eq. 3, using individual block probabilities estimated from normalized single counts, e.g. for N-blocks of identities

$$p(q^N) = \frac{n(q^N)}{n_t}, \tag{13}$$

where $n(q^N)$ is the number of occurrences of the given N-block within the n_t occurrences of all N-blocks in the data set, and analogously for $p(b^N)$. Entropy densities, $h(Q)$ and $h(B)$, were then estimated according to eq. 6 from the linear region in the dependence of $H(Q^N)$ and $H(B^N)$ on N. We estimated the mutual information between N-blocks of burials and single amino acid identity at the central position in the block,

$$I(Q_0; B^N) = H(Q_0) - H(Q_0|B^N), \tag{14}$$

where $H(Q_0) = H(Q^1)$ is the single identity entropy obtained with probabilities estimated from single counts as explained above, while $H(Q_0|B^N)$ is the conditional entropy of central residue identity conditional to burial block, obtained from conditional probabilities estimated as

$$p(Q_0|B^N) = \frac{n(Q_0, B^N) + (20 \times p(Q_0|B_0))}{n(B^N) + 20}. \tag{15}$$

The use of 20 "pseudo-counts" is intended to mitigate artifacts resulting from low frequency events as N increases and $n(Q_0, B^N)$ for some identities might become too small for reliable estimates, in which case counting is dominated by the single pair conditional probability. Conditional probabilities were computed only for observed burial blocks and all amino acid identities. The estimated mutual information turns out to be increasingly smaller than its actual value as N becomes large. While the actual mutual information must increase monotonically with N, its estimate will decrease for large N, providing again a simple signature of data bank saturation. Since $i(Q;B) \approx I(Q_0; B^\infty)$ according to eq. 11, the mutual information density was obtained by extrapolation from the observed behavior before saturation of estimated $I(Q_0; B^N)$ values as a function of block size N, as previously done for secondary structure by Crooks and Brenner, 2004. In addition to fitting the data,

before saturation, to a single exponential $f(x) = a - b\exp(x/c)$, with limiting behavior provided by adjusted parameter a, we have also used a sigmoid $f(x) = \frac{a}{1-\exp(-b(x-c))} + d$, with limiting behavior provided by $a + d$. An analogous procedure was used to estimate $I(Q^\infty; B_0)$ for the tractable HP and HPN representations but saturation becomes a problem for more detailed alphabets. The lower bound $I(Q^\infty; B_0)^- \equiv \sum_{i=1}^{\infty} I(Q_i; B_0)$ was estimated for all alphabets.

4 RESULTS

Fig. 1 illustrates the statistical behavior of local sequences of amino acid identities and burials, as determined here from C_α central distances normalized by radius of gyration. N-block entropy for sequences of amino acid identities, $H(Q^N)$, and burials, $H(B^N)$, are shown in (a) and (b), respectively. Different curves correspond to different alphabets, ranging from 2 to 20 amino acid identities and from 2 to 5 equally probable burial levels. It is apparent that $H(Q^N)$ increases linearly from the origin for all identity alphabets, being consistent with zero order markovicity, $m = 0$, or equivalently, statistical independence between amino acid identities along the sequence. Deviation from linearity for large N results from saturation of the data bank as all curves converge to the same alphabet-independent saturated limit behavior. The plot of N-block entropy for burials, $H(B^N)$, on the other hand, deviates from linearity also for small N and, more perceptively, displays a positive intercept with the ordinate axis due to correlations between adjacent burial levels. Low-order markovicity, with m not higher than 2 or 3, is indicated in this case by the linear region for intermediate N. Accordingly, as shown in Table 1, residue entropy density $h(Q)$ is very close to the single letter entropy, $H(Q^1)$, increasing from essentially 1 for HP sequences, $h(Q_{\rm HP}) \approx H(Q^1_{\rm HP}) \approx 1$ bit/residue, to $h(Q_{20}) \approx H(Q^1_{20}) \approx 4.18$ bits/residue for 20 amino acid letters while mutual information between adjacent identities is in the order of milibits and excess entropy around 1 centibit. Entropy densities of correlated burials, however, are significantly lower than corresponding single burial entropies, with a positive mutual information between adjacent burials, such as $h(B_{\alpha 2}) \approx 0.62 < H(B_{\alpha 2}) \approx 1$ bit/residue and $I(B_i; B_{i+1}) \approx 0.34$ bit for two C_α burial levels. The behavior of C_β burials was found to be qualitatively similar but consistently displaying larger entropy densities, such as $h(B_{\beta 2}) \approx 0.73$ and $h(B_{\beta 3}) \approx 1.1$ bits/residue for two and three burial levels, respectively, to be compared to $h(B_{\alpha 2}) \approx 0.62$ and $h(B_{\alpha 3}) \approx 0.95$ bit/residue for C_α burials.

The approximations used in our estimates for the relevant mutual information measures, $i(Q; B)$ and $I(Q^\infty; B_0)^-$, are illustrated in Fig. 2 for the tractable alphabet combination of HP sequences and 2 levels of C_α burials. Eq. 11 actually provides a good approximation for the mutual information density between amino acid identities and burials, since direct computations of $I(Q^N; B^N)/N$ and $I(Q_0; B^N)$ result in the same value, within sampling error, before saturation (Fig. 2a). As expected, saturation for $I(Q^N; B^N)/N$ occurs at smaller N, as indicated by a steep increase at $N \approx 8$ for the present data set, when compared to saturation of $I(Q_0; B^N)$ which occurs at $N \approx 11$ as indicated by the maximum position in the curve. This observation is consistent with statistical independence between identities not only unconditionally, as suggested by the previous figure, but also conditionally to the sequence of burials, B^N. We assume this approximation to be valid also for more detailed alphabets. The mutual information between local sequence and central burial, on the other hand, is not as accurately reproduced by the sum of positional contributions and a strict inequality should be assumed in eq. 12, since $\sum_{i=1}^{N} I(Q_i; B_0) < I(Q^N; B^0)$ even for N as small as 3 (Fig. 2b). The difference in the extrapolated limits, $I(Q^\infty; B_0)$ and $I(Q^\infty; B_0)^-$, is close to $(0.07 - 0.06) = 0.01$ bit, above sampling error and typical for the tractable alphabet combinations under consideration, as shown in Table 2, but actually comparable to the small excess entropy observed in amino acid sequences described above. It is indicated, therefore, that small correlations between amino acid identities in local sequences, although usually negligible (Crooks and Brenner, 2004; Weiss *et al.*, 2000), at least when compared to stronger correlations on burials or secondary structure, might be perceptible in burial prediction schemes, particularly upon conditioning to a single residue burial, B_0. As a practical consequence previously discussed in the context of accessible surface area prediction (Thompson and Goldstein, 1996), prediction algorithms that account for these correlations are expected to perform better than algorithms in which they are neglected.

The dependence on N of the estimates for mutual information between N-blocks of burials and central residue identities, $I(Q_0; B^N)$, is shown in Fig. 3 for different combinations of identity and burial alphabets. Mutual information density, $i(Q; B) \approx I(Q_0; B^\infty)$, was obtained by extrapolation from exponential or sigmoidal fits to the points before saturation, as indicated by solid lines and shown Table 2. Mutual information density is always larger for C_β burials when compared to C_α burials with the same alphabet combination, such as $i(Q_{20}; B_{\alpha 2}) \approx 0.09 < i(Q_{20}; B_{\beta 2}) \approx 1.13$ bits/residue. As could be anticipated, it tends to increase with alphabet size either of amino acid identities or burials such as, in the case of C_β atoms, from $i(Q_{HP}; B_{\beta 2}) \approx 0.06$ bit/residue for the HP alphabet and 2 burial layers, to $i(Q_{20}; B_{\beta 3}) \approx 0.18$ bit/residue, for 20 amino acid letters and 3 layers. Fig. 4 shows that mutual information density for 20 amino acid letters further increases with the number of burial layers at least up to 5 layers, where it approaches 0.16 and 0.20 bit/residue for C_α and C_β burials, respectively. No reliable estimate was possible for 6 layers.

Positional mutual information values, $I(Q_i; B_0)$, are shown in Fig. 5a for 20 amino acid letters and different numbers of burial levels of C_β atoms. Positional mutual information is essentially 0 for burial and identity pairs separated by more than 15 residues. We use therefore the sum $\sum_{i=1}^{N} I(Q_i; B_0)$ with $N = 31$ as a reasonable approximation of $I(Q^\infty; B_0)^- \equiv \sum_{i=1}^{\infty} I(Q_i; B_0)$ which, as discussed above, is a strictly lower bound for $I(Q^\infty; B_0)$. Accordingly, positional mutual information partial sums are already essentially constant at $N \approx 30$ for 2 burial levels and various amino acid alphabets, as shown in Fig. 5(b), and also for 20 amino acid letters and various numbers of burial levels, as shown in Fig. 5(c). Since the increase in the number of burial levels is not accompanied by data bank saturation in the present approximation, we are able to explore the effect of many burial levels on $I(Q^\infty; B_0)^-$ while the behavior of $i(Q; B)$ could not be observed beyond the increase up to 5 burial layers discussed above. As shown in Fig. 5(d), $I(Q^\infty; B_0)^-$ for C_β also increases significantly from 2 layers to 5 layers, approximately from 0.13 to 0.18 bit, but only slightly for additional layers with asymptotic limit close to 0.2 bit. Qualitatively similar results were obtained for C_α atoms but mutual information

between single burials and local sequence tends again to be smaller in this case when compared to C_β atoms, although the difference is smaller than for mutual information density, as also seen in Table 2. We also show for comparison in the same table the mutual information between single letters, $I(Q;B)$, as well as $I(Q^\infty;B_0)$ for a few tractable alphabet combinations.

5 DISCUSSION

In the present study we estimate by extrapolation, neglecting long range correlations, the mutual information density between local sequence of amino acid identities and corresponding burials, $i(Q;B) \approx I(Q_0;B^\infty)$. It must be noted that the underlying probability distributions, estimated from local block statistics, are much simpler than distributions of whole amino acid sequences and tertiary structures. In particular they are consistent with markovicity and a linear dependence of entropy, and mutual information, on block length, as shown in Fig. 1. Meaningful densities of entropy and mutual information can be estimated for this simplified statistical scheme with different reduced alphabets. Additionally, and most importantly, resulting estimates for $i(Q;B)$ provide upper limits for the quality of any prediction associating local sequences of burials and identities, a clearly attemptable task with established learning algorithms. Predictions of single burial values from local sequence, on the other hand, should be limited simply by the mutual information between local sequence and single burial, $I(Q^\infty;B_0)$, which is difficult to estimate for 20 amino acid letters due to data bank saturation. We provide therefore a lower bound, $I(Q^\infty;B_0)^- < I(Q^\infty;B_0)$, further neglecting local correlations between amino acids conditional to single burial.

As shown in Table 2, comparison between $I(Q;B)$ and $I(Q^\infty;B_0)$ (or $I(Q^\infty;B_0)^-$) indicates that reduction in single burial uncertainty provided by knowledge of single residue identity is typically around half the reduction provided by the whole local sequence of identities, such as 0.05 and 0.11 bits, respectively, for 2 C_α burial level and 20 amino acids. Furthermore, it is apparent from the difference between $I(Q^\infty;B_0)$ and $I(Q^\infty;B_0)^-$, for the few tractable alphabet combinations, that local correlations in identities might account for a couple of centibits in mutual information. Our estimates for the mutual information density, $i(Q;B)$, indicate that the uncertainty about burials resolvable from local sequence, already considering the reduction provided by sequence-independent burial local correlations, can be smaller than 0.1 bit/residue, as for 2 levels of C_α burials, and probably not much larger than 0.2 bit/residue, as observed for 5 levels of C_β burials. These values are comparable to estimates involving secondary structure (0.16 bits/residue (Crooks and Brenner, 2004)), although no more than around 15% of the corresponding burial entropy density.

As has been previously noted (Crooks and Brenner, 2004), small mutual information between local sequence and structural descriptors indicates that local structure, as reflected in secondary structure or burials, must be largely determined by non-local information. It is useful, however, to distinguish between sequence-dependent and sequence-independent non-local information. After all, a large amount of structure-determining non-local information might be provided by sequence-independent constraints, analogous to grammatical rules of human languages (Pereira de Araújo and Onuchic, 2009). The information to be obtained from sequences, corresponding in the same analogy to the actual literature codified in written texts, could actually be much smaller. Note that this distinction is already apparent and easily quantified locally from the comparison between single letter entropies and corresponding densities of entropy, on one hand, or, on the other hand, mutual information with sequence. The uncertainty of 1 bit for two burial levels of a single C_α atom, for example, diminishes to 0.6 bit due to sequence-independent local information, or a reduction of 0.4 bit, while only around 0.09 or 0.11 bit is resolvable from sequence-dependent local information, depending on wether sequence-independent local information is previously assumed.

A large amount of sequence-independent non-local structural information is actually inferred from the small expected total number of protein shapes, Ω_s, which has been estimated by different groups to be in the order of several thousands (Chotia, 1992; Zhang and DeLisi, 1998; Govindarajan *et al.*, 1999; Koonin *et al.*, 2002). If Ω_s is assumed to be 10000, for example, the corresponding entropy would be limited from above by $\log_2 \Omega_s$ and could not be more than around 13 bits per structure, or only 0.05 bit/residue for a putative typical length of 260 residues (0.1 bit/residue for 130 residues). This would be the uncertainty about whole structures, and therefore burials, to be resolved from sequence. The large remaining single burial uncertainty, e.g. $\approx (1 - 0.05 = 0.95)$ bits/residue for 2 C_α burial levels, must therefore be resolved by sequence-independent information, both local ($\approx$ 0.4 bits/residue, as discussed above) and non-local ($\approx$ 0.55 bits/residue, as a consequence). Note that even if the total effective number of structures turns out to be larger or smaller by up to two orders of magnitude the estimated amount of sequence-dependent structural information could not change by more than a couple of centibits/residue.

This small amount of sequence-dependent information (literature) when compared to the large amount of sequence-independent constraints (grammar) is an unavoidable consequence of a modest total number of structures when compared to possible sequences. It is also clearly consistent with the sound elusiveness of possible solutions for the problem of *ab initio* protein structure prediction, contrasting to significant success in homology modeling. Note that the entropy of whole amino acid sequences must indeed be much larger than structural entropy since many sequences fold to each single structure (Larson *et al.*, 2002; Koehl and Levitt, 2002), although smaller, and less trivial, than estimated from local statistics. Long range sequence correlations have been detected (Pande *et al.*, 1994) and must produce deviations from markovicity, contributing not only to reduce the entropy but also to destroy its linear dependence on chain length. Any possible additional folding restrictions are likely to intensify this effect. In particular, the entropy associated to whole protein sequences should NOT be assumed to rise simply as $Nh(Q)$, where $h(Q) \sim 4.2$ bits/residue is the entropy density estimated from local sequence.

It must also be noted that our present estimates for the mutual information density between amino acid sequence and burials of single atomic types, either C_α or C_β, appears to be smaller than our previous upper estimate between 3 and 4 bits/residue for the minimal amount of required burial information from the combination of all atomic types (Pereira de Araújo and Onuchic, 2009). Assuming 0.2 bit/atom as representative and simply multiplying by the average number of atoms in each residue we arrive at $(0.2 \times 7.8 \approx 1.6)$ bits/residue. In any case, our present mutual information estimates

provide parameters to which actual prediction algorithms should be compared. A parallel study shows that most of this burial information shared by local sequences is easily captured by simple statistical prediction schemes based on Hidden Markov Models (HMM) or even Naive Bayesian Classifiers (NBC) (Van der Linden *et al.*, 2012). Interestingly, this parallel study indicates that $I(Q^{\infty};B_0)^{-}$ is approached by the simplest NBC algorithms, which neglect identity correlations conditional to single burials, and are usually surpassed by HMM algorithms, in which case such correlations are accounted for. Furthermore, near optimal prediction for HMM algorithms is indicated by the corresponding mutual information density approaching our present estimate for $i(Q;B)$.

6 CONCLUSION

We have estimated relevant mutual information measures between local amino acid sequences and corresponding burials in globular proteins, providing parameters to evaluate the quality of algorithms for burial prediction from local sequence. This is a crucial step preliminary to actual burial prediction, as described in a parallel study. It is therefore fundamental for our goal of determining the possibility of recovering the native conformation of globular proteins using atomic burials as informational intermediates between local sequence and tertiary structure.

REFERENCES

Chotia, C. (1992). One thousand families for the molecular biologist. *Nature*, **357**, 543–544.

Cline, M., Karplus, K., Lathrop, R., Smith, T., Rogers, R., and Haussler, D. (2002). Information-theoretic dissection of pairwise contact potentials. *Proteins: Struct., Funct. and Genet.*, **49**, 7–14.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*, chapter 2. Wiley-Interscience.

Crooks, G. E. and Brenner, S. E. (2004). Protein structure prediction: entropy, correlations and prediction. *Bioinformatics*, **20**, 1603–1611.

Dill, K. A., Ozcan, S. B., Shell, M. S., and Weikl, T. R. (2008). The protein folding problem. *Annu. Rev. Biophys.*, **37**, 289–316.

England, J. (2011). Allostery in protein domains reflects a balance of steric and hydrophobic effects. *Structure*, **19**, 967–975.

Garnier, J., Osguthorpe, D. J., and Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J.Mol.Biol.*, **120**, 97–120.

Gomes, A. L. C., de Rezende, J. R., Pereira de Araújo, A. F., and Shakhnovich, E. I. (2007). Description of atomic burials in compact globular proteins by Fermi-Dirac probability distributions. *Proteins: Struct., Funct. and Bioinf.*, **66**, 304–320.

Govindarajan, S., Racabarren, R., and Goldstein, R. A. (1999). Estimating the total number of protein folds. *Prot. Sci.*, **35**, 408–414.

Hobohm, U. and Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.*, **3**, 522–524.

Huang, E. S., Subbiah, S., and Levitt, M. (1995). Recognizing native folds by the arrangement of hydrophobic and polar residues. *J. Mol. Biol.*, **252**, 709–720.

Koehl, P. and Levitt, M. (2002). Protein topology and stability define the space of allowed sequences. *PNAS*, **99**, 1280–1285.

Koonin, E. V., Wolf, Y. I., and Karev, G. P. (2002). The structure of protein universe and genome evolution. *Nature*, **420**, 218–223.

Larson, S. M., England, J. L., Desjarlais, J. R., and Pande, V. S. (2002). Throughly sampling sequence space: Large-scale protein design of structural ensembles. *Prot. Sci.*, **11**, 2804–2813.

Onuchic, J. N. and Wolynes, P. G. (2004). Theory of protein folding. *Curr. Opin. Struct. Biol.*, **14**, 70–75.

Pande, V. S., Grosberg, A. Y., and Tanaka, T. (1994). Nonrandomness in protein sequences: Evidence for a physically driven stage of evolution? *Proc. Natl. Acad. Sci. USA*, **91**, 12972–12975.

Pereira de Araújo, A. F. and Onuchic, J. N. (2009). A sequence-compatible amount of native burial information is sufficient for determining the structure of small globular proteins. *Proc. Natl. Acad. Sci. USA*, **106**, 19001–19004.

Pereira de Araújo, A. F., Gomes, A. L. C., Bursztyn, A. A., and Shakhnovich, E. I. (2008). Native atomic burials, supplemented by physically motivated hydrogen bond constraints, contain sufficient information to determine the tertiary structure of small globular proteins. *Proteins: Struct., Funct. and Bioinf.*, **70**, 971–983.

Reza, F. M. (1994 (1961)). *An introduction to information theory*, chapter 3. Dover Publications, INC.

Shakhnovich, E. (2006). Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. *Chem. Rev.*, **106**, 1559–1588.

Solis, A. and Rackovsky, S. (2004). On the use of secondary structure in protein structure prediction: a bioinformatic analysis. *Polymer*, **45**(2), 525–546.

Solis, A. D. and Rackovsky, S. (2007). Property-based sequence representations do not adequately encode local protein folding information. *Proteins:Struct., Funct. and Bioinf.*, **67**(4), 785–788.

Sullivan, D. C., Aynechi, T., Voelz, V. A., and Kuntz, I. D. (2003). Information content of molecular structures. *Biophys. J.*, **85**, 174–190.

Thompson, M. J. and Goldstein, R. A. (1996). Predicting solvent accessibility: Higher accuracy using Bayesian statistics and optimized residue substitution classes. *Proteins: Struct. Funct. and Genet.*, **25**, 38–47.

Van der Linden, M. G., Ferreira, D. C., Rocha, J. R., and Pereira de Araújo, A. F. (2012). Simple statistical schemes for the discrete prediction from sequence of atomic burials in globular proteins. *submitted to Bioinformatics*, **00**, 00.

Weiss, O., Jimenez-Montano, M., and Herzel, H. (2000). Information content of protein sequences. *J. Theor. Biol.*, **206**, 379–386.

Zhang, C. and DeLisi, C. (1998). Estimating the total number of protein folds. *J. Mol. Biol.*, **284**, 1301–1305.

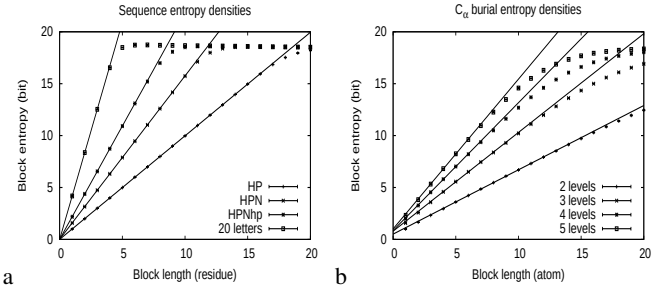


Fig. 1. N-block sequence entropy estimates as a function of block size N for different alphabets of amino acid identities (a) and C_α burial levels (b). Straight lines represent linear fits to the data from which the entropy density (inclination) and excess entropy (intersect with the ordinates) are obtained.

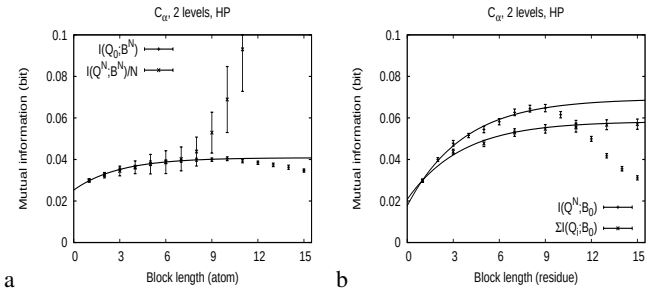


Fig. 2. Comparison between $I(Q_0; B^N)$ and $I(Q^N; N^N)/N$, for the HP alphabet and 2 levels of C_α burials, reveals virtually coincident values before saturation (a), suggesting that eq. 11 is a good approximation. Comparison between $I(Q^N; B_0)$ and $\sum I(Q_i; B_0)$, on the other hand, reveals discrepancies even for small blocks (b), indicating that a strict inequality should be assumed in eq. 12. Curves represent single exponential fits to the data before saturation.

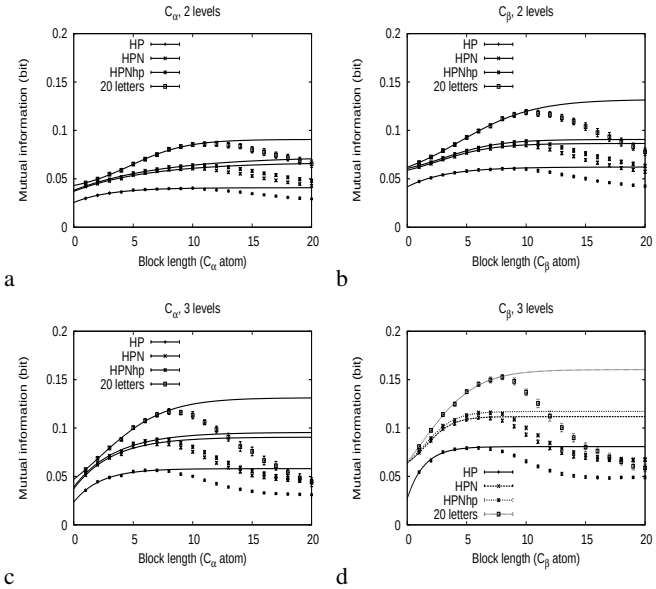


Fig. 3. Estimates for the mutual information, $I(Q_0; B^N)$, between a single central amino acid identity, Q_0, and N-blocks of burials, B^N, as a function of block size N, for 2 and 3 levels of C_α (a and c) and C_β (b and d) burials. Different sets of points correspond to different alphabets of amino acid identities. Lines represent exponential or sigmoidal fits to the data before saturation from which limiting values $i(Q; B) \approx I(Q_0; B^\infty)$ are obtained.

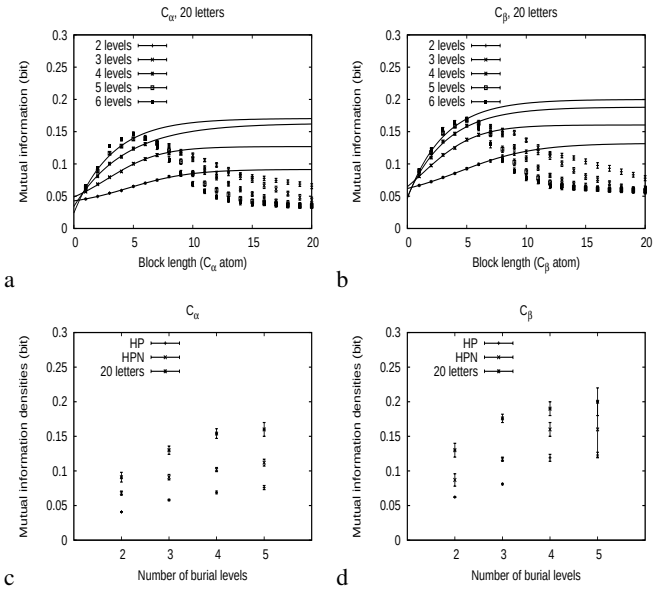


Fig. 4. Estimates for the mutual information, $I(Q_0; B^N)$, between a single central amino acid identity, Q_0, and N-blocks of burials, B^N, as a function of block size N, for the alphabet of 20 amino acids and various levels of C_α (a) and C_β (b) burials, from $L = 2$ to $L = 6$, as represented by different sets of points. Lines represent exponential or sigmoidal fits the data before saturation, except for $L = 6$. Limiting values obtained from fitted curves, $i(Q; B) \approx I(Q_0; B^\infty)$, are plotted in (c) and (d) as a function of the number of burial levels, L, for C_α and C_β atoms, respectively, together with analogous values obtained for the simplified alphabets of amino acid identities, HP and HPN.

Table 1. Single sequence analysis. Single letter entropy, $H(X)$, and mutual information between adjacent letters, $I(X_i; I_{i+1})$, are in bits. Entropy density h(X), in bits/letter, and corresponding excess entropy E_X, in bits, were obtained from data fits shown in Fig. 1. Each line corresponds to a different alphabet of amino acid identities or burials, as indicated in the first column. Error in the last significant digit is shown in parentheses.

	$H(X)$	$I(X_i; X_{i+1})$	$h(X)$	E_X
HP	1.00000(9)	0.00072(8)	0.9969(2)	0.0096(7)
HPN	1.5806(4)	0.0009(1)	1.5734(7)	0.018(3)
20	4.185(2)	0.005(7)	4.176(4)	0.010(5)
$\alpha 2$	0.99978(7)	0.342(3)	0.619(1)	0.513(9)
$\alpha 3$	1.5796(3)	0.574(3)	0.951(6)	0.79(3)
$\alpha 5$	2.3200(2)	0.808(5)	1.44(2)	1.01(6)
$\beta 2$	0.99847(8)	0.211(3)	0.725(4)	0.36(2)
$\beta 3$	1.5800(2)	0.377(4)	1.128(5)	0.54(2)
$\beta 5$	2.3154(2)	0.508(5)	1.693(8)	0.76(3)

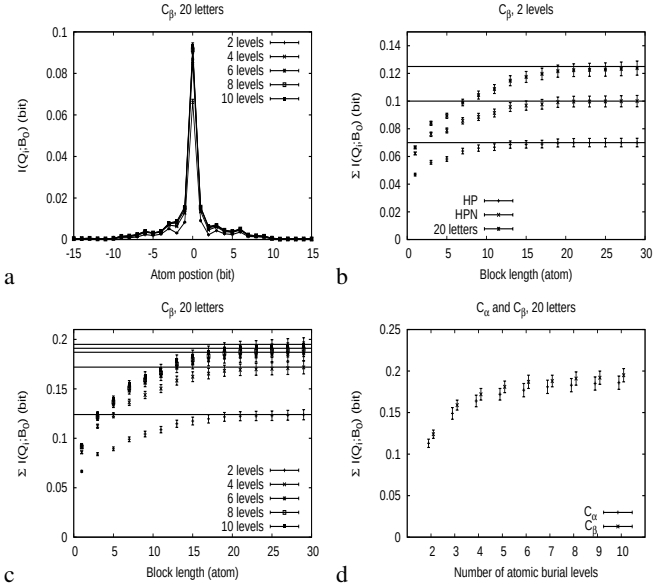


Fig. 5. (a) Positional mutual information $I(Q_i; B_0)$ between amino acid identity at position i, Q_i, within the N-block of identities Q^N, and central C_β burial, B_0, as a function of i for 20 amino acid letters and various numbers of burial levels (a). Sums of positional mutual information terms, $\sum_N I(Q_i; B_0)$, as a function of block size N, for 2 C_β burial levels with various alphabets of amino acid identities (b) and for 20 amino acid letters with various numbers of C_β burial levels (c). Horizontal lines represent the limiting behavior obtained for $N = 31$. Limiting behavior for sums of positional mutual information terms, obtained with fixed block size $N = 31$, as a function of the number of burial levels for C_α and C_β atoms (d).

Table 2. Inter-sequence analysis. Mutual information between single letters, $I(Q; B)$ in bits, mutual information density, $i(Q; B)$ in bits/pair, as obtained in Figs. 3 and 4, $I(Q^\infty; B_0)$ in bits, for tractable alphabet combinations, and $I(Q^\infty; B_0)^-$ in bits, as obtained in Fig. 5, for C_α and C_β atoms are shown for different combinations of identity alphabet and number of burial layers, as indicated in the first two columns. Error in the last significant digit is shown in parentheses.

		$I(Q;B)$		$i(Q;B)$		$I(Q^\infty;B_0)$		$I(Q^\infty;B_0)^-$	
L		C_α	C_β	C_α	C_β	C_α	C_β	C_α	C_β
2	HP	0.0297(6)	0.0472(9)	0.0408(2)	0.0622(3)	0.069(2)	0.088(5)	0.059(3)	0.070(3)
	HPN	0.0420(9)	0.062(1)	0.068(3)	0.087(9)	0.098(7)	0.109(9)	0.089(7)	0.100(4)
	20	0.046(1)	0.067(1)	0.091(7)	0.13(1)	–	–	0.113(5)	0.124(5)
3	HP	0.0357(9)	0.054(1)	0.058(1)	0.081(1)	0.091(2)	0.109(4)	0.075(4)	0.086(4)
	HPN	0.051(1)	0.075(1)	0.091(4)	0.117(3)	0.125(5)	0.14(1)	0.114(5)	0.125(5)
	20	0.0570(9)	0.081(2)	0.130(6)	0.176(6)	–	–	0.149(7)	0.159(6)
5	HP	0.0386(8)	0.059(1)	0.076(3)	0.123(4)	–	–	0.083(4)	0.095(4)
	HPN	0.057(1)	0.081(1)	0.112(5)	0.16(4)	–	–	0.130(5)	0.141(5)
	20	0.064(1)	0.090(1)	0.16(1)	0.20(2)	–	–	0.172(7)	0.181(7)