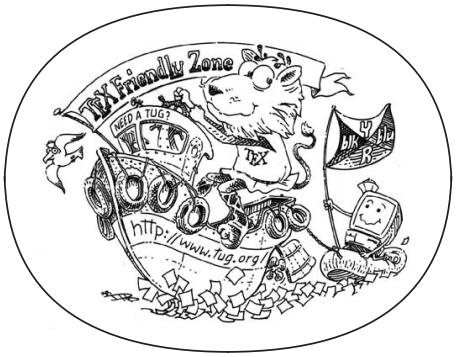
JOSÉ GERALDO DE CARVALHO PEREIRA

AUTÔMATOS CELULARES E PROTEÍNAS

[26 de setembro de 2016 at 16:05 – classicthesis version 4.2]

AUTÔMATOS CELULARES E PROTEÍNAS

JOSÉ GERALDO DE CARVALHO PEREIRA



Exploração de um novo modelo para a predição de estruturas secundárias

Agosto de 2016 – version 4.2

[26 de setembro de 2016 at 16:05 – classicthesis version 4.2]

RESUMO

Short summary of the contents in English... a great guide by Kent Beck how to write good abstracts can be found here:

https://plg.uwaterloo.ca/~migod/research/beckOOPSLA.html

ABSTRACT

Resumo

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTINGS

ACRONYMS

Parte I

FUNDAMENTOS TEÓRICOS

1 INTRODUÇÃO

O problema do enovelamento de proteínas é a questão de como são formadas ou organizadas suas estruturas atômicas tridimensionais. Essa questão surgiu no final da década de 50, logo após a resolução atômica da primeira estrutura proteica por Kendrew e colaboradores [1], trabalho no qual se observou experimentalmente, segundo o próprio autor, uma complexidade maior que as antecipadas pelas teorias da época sobre estruturas proteicas. Posteriormente, Anfinsen [2] realizou experimentos que demonstraram que a ribonuclease poderia ser reversamente desnaturada/renaturada in vitro, e que em condições desnaturantes, tanto a estrutura quanto a função eram perdidas, no entanto, ambas eram recuperadas ao retornarem à condições fisiológicas. A conclusão foi que, apesar da grande complexidade observada, as proteínas se auto-organizavam estruturalmente, assim sendo, apenas a informação contida em sua sequência de aminoácidos seria suficiente para definir sua estrutura e que esta determinaria a sua função. A explicação de Anfinsen para esta auto- organização estrutural foi dada através da hipótese termodinâmica, a qual postula que em condições fisiológicas a população proteica atinge um mínimo de energia livre de Gibbs no seu estado nativo [3].

Dessa forma, devido ao princípio da relação estrutura $\leftrightarrow$ função e a resultados experimentais que demonstraram que a estrutura é determinada pela sequência de aminoácidos, diversos trabalhos buscaram prever a estrutura de uma proteína a partir da sua sequência de resíduos. Alguns dos primeiros trabalhos a discutir uma forma de predição foram publicados por Levinthal [4, 5] os quais o autor menciona que o número de configurações estruturais possíveis para uma cadeia polipeptídica é imenso, sendo impossível explorar todas as conformações possíveis para se determinar qual sua estrutura nativa, ou de menor energia. Apesar disso, as proteínas são capazes de se enovelarem espontaneamente e adotar a conformação nativa rapidamente, numa escala de segundos ou menos. Esta observação ficou popularmente conhecida como Paradoxo de Levinthal. Entretanto, Levinthal não considerou isso como um resultado absurdo, mas baseou-se nessa análise para concluir que um mecanismo aleatório para o enovelamento não seria válido [6]. Segundo Levinthal [5], uma possível explicação para a eficiência observada no processo seria a formação rápida de interações locais que acelerariam e guiariam o enovelamento:

> *We feel that protein folding is speeded and guided by the rapid formation of local interactions, which then determine the further folding of the peptide.*

Apesar da sugestão de Levinthal para explicar um possível mecanismo de enovelamento ter sido publicado a 45 anos, o desafio de se prever as estruturas tridimensionais das proteínas a partir de suas sequências de aminoácidos, mesmo obtendo grande progresso no últimos anos, ainda permanece sem uma solução definitiva [7], sendo os métodos experimentais, mais especificamente, os métodos de cristalografia de proteínas por difração de raios-X e o de ressonância magnética nuclear, ainda a principal forma de se obter um modelo estrutural com resolução atômica. Métodos experimentais de resolução da estrutura proteica apresentam diversas dificuldades técnicas. Para a cristalografia por difração de raios-X é necessária a obtenção de proteína em alto grau de pureza e a obtenção de monocristais, que muitas vezes é o fator limitante do processo. Por outro lado, a ressonância magnética nuclear exige concentrações bastante altas de proteína purificada de meios com diferentes isótopos e ainda, existe limitação quanto ao tamanho da proteína analisada. Essas limitações experimentais são evidenciadas pela disparidade entre o número de estruturas resolvidas experimentalmente (98 mil depositadas no PDB) e o número de proteínas com sequência de aminoácido conhecidas (53 milhões depositadas no UniProtKB/TrEMBL – dados de 02/2014). Dessa forma, a busca por métodos computacionais capazes de prever estruturas proteicas continua uma área de grande interesse científico, tanto como uma forma de se conhecer melhor o mecanismo de enovelamento como também na utilização da informação estrutural para responder diversas questões biológicas e desenvolver novos medicamentos [8].

PROTEÍNAS

Estruturas

Enovelamento

Modelos teóricos para a formação da estrutura secundária

AUTÔMATOS CELULARES

Autômato celular elementar

Outros tipos de autômatos celulares

Problema inverso

2

OBJETIVOS

O objetivo geral deste trabalho será desenvolver um método de reconhecimento de enovelamento capaz de identificar, a partir da sequência de aminoácidos da proteína a qual se deseja modelar, estruturas proteicas resolvidas semelhantes a sua estrutura nativa. Tais estruturas podem ser tanto de proteínas que tenham evoluído divergentemente quanto convergentemente uma vez que não há uma comparação direta entre os resíduos das sequências, mas sim entre estruturas e padrões locais que emergem dessa estrutura primária. Para atingir tal objetivo, delineamos especificamente os objetivos a seguir.

Objetivos específicos 1. Criar um alfabeto com estados discretos e em menor número possível, para representar a conformações dos resíduos na estrutura proteica, buscando minimizar a perda de informação estrutural durante a redução de dimensões (3D Enviroments (Verify3D) onde a redução é feita utilizando estados que combinam a estrutura secundária, a polaridade do ambiente e o grau de exposição ao solvente, totalizando 18 estados; (2) no programa FUGUE, onde há 4 classes para estruturas secundárias, 2 para acessibilidade ao solvente e 8 para ligações de hidrogênio, totalizando 64 estados; (3) no trabalho de Chellapa e Rose (2012) onde são utilizados 11 estados que representam regiões de ângulos diédricos da cadeia principal;

2. Aplicar este alfabeto a domínios proteicos obtidos do CATH e/ou SCOP para criar uma representação das estruturas em dimensões reduzidas – 1D;

3. Utilizar a representação 1D de parte dos domínios obtidos do CATH/SCOP e assim criar um conjunto de treinamento para ser usado na busca de regras de transição para um autômato celular. Essas regras devem ser capazes de guiar a evolução do autômato celular de um estado inicial correspondente a sequência de aminoácidos da proteína até um estado que simbolize a estrutura tridimensional, mas representada unidimensionalmente, utilizando o alfabeto criado;

4. Testar as regras de transição selecionadas por apresentar melhor desempenho em proteínas não incluídas no conjunto de treinamento e assim, analisar a eficácia do autômato celular em obter informação estrutural a partir da sequência de aminoácidos das proteínas;

5. Utilizar o método de autômato celular com a melhor regra de transição para, quando aplicado em sequências de aminoácidos de proteínas sem estrutura conhecida, identificar domínios com enovelamento similar.

3

JUSTIFICATIVA

A aplicação de autômatos celulares no desenvolvimento de um novo método de reconhecimento de enovelamentos não visa ser apenas uma alternativa aos métodos atuais, mas também tem como objetivo a criação de um modelo que seja capaz de fornecer informações sobre a dinâmica do enovelamento de proteínas, ao permitir a análise de padrões que indicam pontos de início do enovelamento, de formação de estruturas locais e da propagação global dessas estruturas locais ao longo da sequência. O método poderá auxiliar também no estudo de como mutações podem afetar o enovelamento proteico, assim como no design de proteínas e contribuir também para a predição ab initio de estruturas proteicas.

Parte II

DESENVOLVIMENTO

4

DADOS

Neste trabalho foram utilizados dois conjuntos de dados compostos de proteínas com estruturas resolvidas experimentalmente e da estrutura secundária atribuída aos seus resíduos por quatro diferentes algoritmos: DSSP, Stride, Kaksi e Pross.

O primeiro conjunto selecionado é formado por um grande número de estruturas de alta qualidade e tem como finalidade ser utilizado na busca de regras gerais para o autômato celular. Essas regras gerais são um dos elementos mais importantes desse trabalho, pois permitem avaliar a generalização do autômato celular, isto é, qual o grau de sucesso da aplicação do autômato para o universo de proteínas existentes.

O segundo conjunto selecionado é composto de quatro proteínas denominadas de camaleônicas. Esse conjunto foi selecionado por ser, possivelmente, o exemplo experimental mais desafiador para os métodos de predição de estrutura secundária. Como discutiremos ao longo do texto, todos os métodos de predição de estrutura secundária, assim como os de modelagem comparativa, tendem a falhar nesse conjunto devido à limitações teóricas dos métodos.

PROTEÍNAS CAMALEÔNICAS

PROTEÍNAS DIVERSAS

O conjunto de proteínas diversas utilizado para o treinamento do autômato foi obtido do banco de dados "Top8000" (versão de 2015). Esse banco de dados foi organizado pelo Richardson Lab da Universidade de Duke (disponível em github.com/rlabduke/reference_data). As cadeias selecionadas atendem aos seguintes critérios:

- Resolução < 2.0 Å
- MolProbity score < 2.0
- $\leqslant$ 5% dos resíduos apresentando comprimentos de ligação anormais ($> 4\sigma$)
- $\leqslant$ 5% dos resíduos apresentando ângulos de ligação anormais ($> 4\sigma$)
- $\leqslant$ 5% dos resíduos com desvios anormais do C_β (> 0.25 Å)

As cadeias selecionadas pelos critérios acima são subagrupadas de acordo com o grau de identidade sequencial (homologia): < 50%,

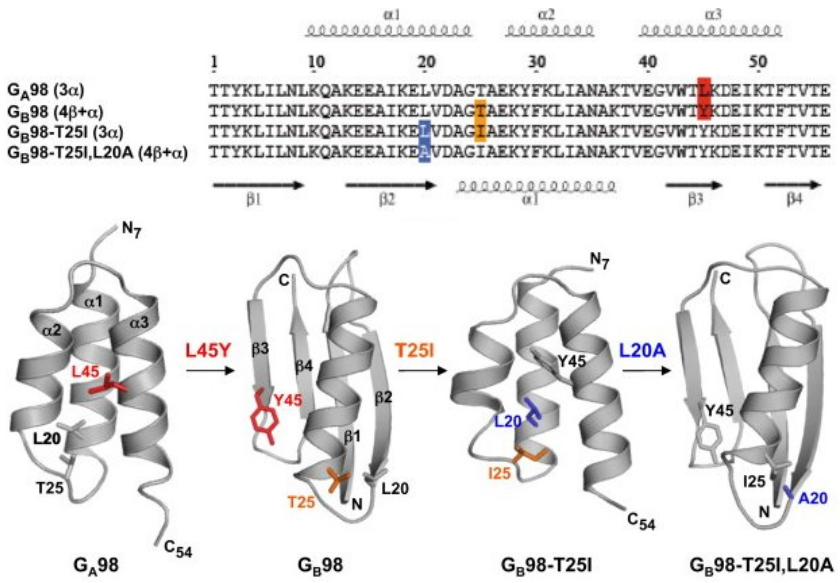


Figura 1: Figura da sequencia e das estruturas das camaleonicas

CONJUNTO	# ORIGINAL	# UTILIZADAS
Top8000-hom50	7233	6749
Top8000-hom70	7958	7435
Top8000-hom95	8826	8227

Tabela 1: Número de cadeias presentes no banco de dados Top8000 (Richardson Lab) e número de cadeias utilizadas neste trabalho após a exclusão de cadeias que apresentaram algum problema durante a atribuição da estrutura secundária ou que possuiam resíduos indeterminados.

<70% e <95%. Cadeias que apresentavam resíduos indeterminados na estrutura ou que apresentaram algum erro durante a atribuição da estrutura secundária por algum dos quatro métodos foram removidos do conjunto. A tabela 1 mostra o número de cadeias utilizadas.

5 IMPLEMENTAÇÃO

AUTÔMATO CELULAR

Modelo inicial

O autômato celular inicialmente proposto possui 24 estados discretos. Esses estados correspondem aos 20 aminoácidos, a 3 elementos de estruturas secundárias (hélice, fita e random coil) e mais um estado que indica o início/fim da cadeia polipeptídica (*estado*=#). A vizinhança deste autômato celular é igual a 1 (r=1), o que indica que as regras de transição são dependentes dos dois vizinhos mais próximos, um a esquerda e um a direita. Cada transição pode ocorrer para apenas quatro estados, ou um dos 3 estados que representam os elementos de estrutura secundária ou para o resíduo presente naquela posição da cadeia polipeptídica.

Logo, temos que o total de elementos na regra desse autômato é 24^3 ou 13824, das quais 24 são elementos estáticos, pois células no estado # sempre permanecerão nesse estado durante a evolução do autômato. Assim temos 4^{24^3-24} regras possíveis para esse autômato celular.

Modelos extendidos

Uma das limitações do modelo proposto inicialmente é a perda de informação que ocorre durante a evolução do autômato celular quando as células transitam de estados correpondentes aos aminoácidos para estados de elementos de estrutura secundária. Por exemplo, quando uma lisina evolui para uma hélice, o estado de hélice não possui mais a informação de qual aminoácido havia naquela posição. Acreditamos que essa perda de informação possa ser um fator crítico para o modelo. Consequentemente, avaliamos modelos alternativos que pudessem manter essa informação.

Uma possibilidade seria manter a informação do resíduo juntamente com o elemento de estrutura secundária. Esse modelo teria 20 estados para os aminoácidos, 20 estados para hélices (um estado diferente para cada aminoácido), 20 estados para fitas e 20 estados para random coils, além do estado de início/fim da cadeia polipeptídica, totalizando 81 estados. Cada regra para esse autômato celular teria 81^3 ou 531441 elementos, o que seria aproximadamente 38 vezes maior que uma regra do modelo proposto inicialmente, resultando em um aumento significativo da complexidade e, consequentemente,

da dificuldade na busca por regras que reproduzam o padrão desejado.

Assim, a alternativa escolhida foi utilizar características dos aminoácidos que mantivessem parcialmente a informação do resíduo durante a evolução do autômato celular, mas sem resultar em um aumento tão elevado do número de regras em relação ao modelo inicial. O primeiro modelo concebido que atende esses requisitos utiliza as características de hidrofobicidade dos aminoácidos. Isso resulta em modelo com 27 estados, sendo dois estados para cada um dos 3 elemento de estrutura secundária, mais os 20 aminoácidos e o início/fim da cadeia polipeptídica. No total, a regra deste autômato celular é formada por 27^3, ou 19683, elementos, sendo aproximadamente 1,42 vezes maior que a regra do modelo inicial.

Além deste modelo extendido, dois outros modelos foram utilizados. Um deles acrescentando estados para diferenciar glicinas e prolinas, e outro acrescentando estados para diferencia resíduos com cargas positivas e negativas assim como glicinas e prolinas. Ambos utilizam também a hidrofobicidade dos demais resíduos. As regras para esses modelos apresentam respectivamente 33^3 e 39^3 elementos, o que corresponde a um aumento aproximado de 2,6 e 4,3 vezes em relação ao modelo inicial.

Em todos os modelos extendidos cada elemento da regra continua com a possibilidade de transitar para apenas 4 estados, ou um dos 3 elementos de estrutura secundária ou o resíduo encontrado naquela posição da cadeia polipeptídica.

EDA

A busca por regras de um autômato celular que reproduzam um padrão específico, conhecido como problema inverso, é um problema de otimização. Na literatura, esse problema é normalmente abordado utilizando metaheurísticas como algoritmos genéticos ou anelamento simulado (*simulated annealing*). Neste trabalho optamos por utilizar o Algoritmo de Estimação de Distribuição (EDA). Os fatores que determinaram a utilização desse algoritmo foram a facilidade de implementação do EDA de forma distribuída e o pequeno número de parâmetros em relação à algoritmos genéticos.

No EDA distribuído implementado neste trabalho cada elemento da regra do autômato celular, com excessão dos elementos onde a célula apresenta o estado início/fim da cadeia polipeptídica (*estado*=#), tem a mesma probabilidade inicial ($p = 0,25$) para cada um dos 4 estados transição. A probabilidade é distribuída pelo nó mestre para os nós escravos. Os nós escravos utilizam a probabilidade recebida para gerar $c \geqslant 2$ regras candidatas. As regras candidatas são então utilizadas para evoluir o autômato celular por t passos. Após a evolução, um valor de fitness é atribuído a cada regra. Após um torneio

entre as regras candidatas geradas no nó escravo, a regra com maior fitness é enviada ao nó mestre. Após

Função de fitness

Parte III

RESULTADOS

6 ANÁLISE DOS DADOS

7

APRENDIZADO DE REGRAS

A proteína Ga98 e seus mutantes, os quais sofrem alterações globais na estrutura secundária, são casos interessantes para o teste de novas metodologias de predição de estrutura secundária. Nas metodologias atuais, que comumente utilizam redes neurais, a predição é feita utilizando uma janela de resíduos, em geral com comprimentos de 9, 11 ou 13 resíduos, onde o resíduo central da janela é classificado pela rede neural. Como a predição nas demais janelas presentes na sequência polipeptídica não influencia na classificação da janela, o método apresenta a limitação de responder apenas localmente às variações dos dados de entrada.

Por outro lado, os autômatos celulares, apesar de evoluirem de acordo com regras locais, tem a capacidade de propagar as variações locais e influenciar o surgimento ou alteração de padrões globais, distantes do ponto de origem da variação.

Para avaliar a capacidade dos modelos propostos e da eficácia do método de otimização em encontrar regras capazes de reproduzir o padrão correspondente às estruturas secundárias, testamos a nossa metodologia nessas quatro proteínas.

8

APRENDIZADO DAS REGRAS GERAIS

9

ANÁLISE DAS REGRAS GERAIS

Parte IV

PERSPECTIVAS FUTURAS

10 DESAFIOS FUTUROS

11

ALTERNATIVAS EM ANÁLISE

Parte V

APPENDIX

A

APPENDIX TEST

Lorem ipsum at nusquam appellantur his, ut eos erant homero concludaturque. Albucius appellantur deterruisset id eam, vivendum partiendo dissentiet ei ius. Vis melius facilisis ea, sea id convenire referrentur, takimata adolescens ex duo. Ei harum argumentum per. Eam vidit exerci appetere ad, ut vel zzril intellegam interpretaris.

More dummy text.

APPENDIX SECTION TEST

Test: Tabela 2 (This reference should have a lowercase, small caps A if the option `floatperchapter` is activated, just as in the table itself → however, this does not work at the moment.)

LABITUR BONORUM PRI NO	QUE VISTA	HUMAN
fastidii ea ius	germano	demonstratea
suscipit instructior	titulo	personas
quaestio philosophia	facto	demonstrated

Tabela 2: Autem usu id.

ANOTHER APPENDIX SECTION TEST

Equidem detraxit cu nam, vix eu delenit periculis. Eos ut vero constituto, no vidit propriae complectitur sea. Diceret nonummy in has, no qui eligendi recteque consetetur. Mel eu dictas suscipiantur, et sed placerat oporteat. At ipsum electram mei, ad aeque atomorum mea. There is also a useless Pascal listing below: Listing 1.

Listing 1: A floating example (`listings` manual)

```
for i:=maxint downto 0 do
begin
{ do nothing }
end;
```

BIBLIOGRAFIA

[1] C. J. KENDREW, G. BODO, M. H. DINTZIS, G. R. PARRISH, H. WYCKOFF e C. D. PHILLIPS. "A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis". Em: *Nature* (1952), pp. 662–666. ISSN: 00280836. DOI: 10.1038/181662a0. URL: http://dx.doi.org/10.1038/181662a0.

[2] B. C. Anfinsen. "Principles that govern the folding of protein chains". Em: *Science (New York, N.Y.)* (1967), pp. 223–30. ISSN: 00368075. URL: http://www.ncbi.nlm.nih.gov/pubmed/4124164.

[3] D. George Rose, J. Patrick Fleming, R. Jayanth Banavar e Amos Maritan. "A backbone-based theory of protein folding". Em: *Proceedings of the National Academy of Sciences of the United States of America* (2000), pp. 16623–33. ISSN: 00278424. DOI: 10.1073/pnas.0606843103. URL: http://www.pnas.org/content/103/45/16623.long.

[4] C. Levinthal. "Are there pathways for protein folding?" Em: *Journal de Chimie Physique et de Physico-Chimie Biologique* 65 (1962), pp. 44–45. ISSN: 00217689. URL: http://www.biochem.wisc.edu/courses/biochem704/Reading/Levinthal1968.pdf.

[5] *How to fold graciously*. Vol. 24. Mössbaun Spectroscopy in Biological Systems Proceedings, pp. 22–24. URL: http://www.cc.gatech.edu/~turk/bio_sim/articles/proteins_levinthal_1969.pdf.

[6] Arieh Ben-Naim. "Levinthal's question revisited, and answered". Em: *Journal of biomolecular structure & dynamics* (2006), pp. 113–24. ISSN: 15380254. DOI: 10.1080/07391102.2012.674286. URL: http://www.ncbi.nlm.nih.gov/pubmed/22571437.

[7] John Moult, Krzysztof Fidelis, Andriy Kryshtafovych, Torsten Schwede e Anna Tramontano. "Critical assessment of methods of protein structure prediction (CASP)–round x". Em: *Proteins* 82 Suppl 2 (2008), pp. 1–6. ISSN: 10970134. DOI: 10.1002/prot.24452. URL: http://www.ncbi.nlm.nih.gov/pubmed/24344053.

[8] D. Baker e A. Sali. "Protein structure prediction and structural genomics". Em: *Science (New York, N.Y.)* (1995), pp. 93–6. ISSN: 00368075. DOI: 10.1126/science.1065659. URL: http://www.ncbi.nlm.nih.gov/pubmed/11588250.

DECLARATION

Put your declaration here.

Campinas, Agosto de 2016

José Geraldo de Carvalho Pereira

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

https://bitbucket.org/amiede/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

http://postcards.miede.de/