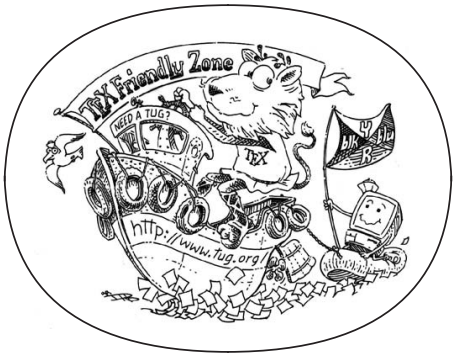
JOSÉ GERALDO DE CARVALHO PEREIRA

AUTÔMATOS CELULARES E PROTEÍNAS

[22 de setembro de 2016 at 10:47 – 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

José Geraldo de Carvalho Pereira: *Autômatos celulares e proteínas,* Exploração de um novo modelo para a predição de estruturas secundárias, © Agosto de 2016

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

[22 de setembro de 2016 at 10:47 – classicthesis version 4.2]

LISTA DE FIGURAS

LISTA DE TABELAS

LISTINGS

ACRONYMS

Parte I

FUNDAMENTOS TEÓRICOS

1 INTRODUÇÃO

PROTEÍNAS

Estruturas

Enovelamento

Modelos teóricos para a formação da estrutura secundária

AUTÔMATOS CELULARES

Autômato celular elementar

Outros tipos de autômatos celulares

Problema inverso

[22 de setembro de 2016 at 10:47 – classicthesis version 4.2]

2

OBJETIVOS

[22 de setembro de 2016 at 10:47 – classicthesis version 4.2]

3 JUSTIFICATIVA

[22 de setembro de 2016 at 10:47 – classicthesis version 4.2]

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%,

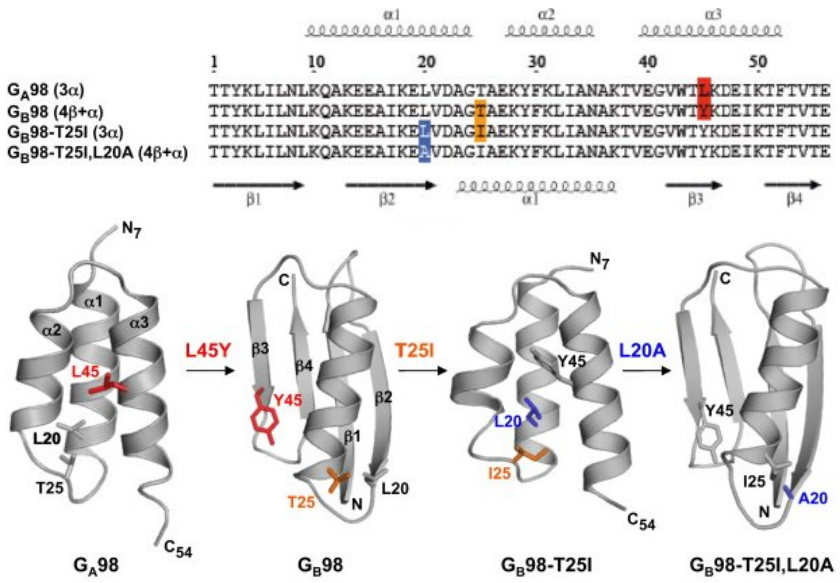


Figura 1: Figura da sequencia e das estruturas das camaleonicas

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,

[22 de setembro de 2016 at 10:47 – classicthesis version 4.2]

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

[22 de setembro de 2016 at 10:47 – classicthesis version 4.2]

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.

[22 de setembro de 2016 at 10:47 – classicthesis version 4.2]

8

APRENDIZADO DAS REGRAS GERAIS

[22 de setembro de 2016 at 10:47 – classicthesis version 4.2]

9

ANÁLISE DAS REGRAS GERAIS

[22 de setembro de 2016 at 10:47 – classicthesis version 4.2]

Parte IV

PERSPECTIVAS FUTURAS

10 DESAFIOS FUTUROS

[22 de setembro de 2016 at 10:47 – classicthesis version 4.2]

11

ALTERNATIVAS EM ANÁLISE

[22 de setembro de 2016 at 10:47 – classicthesis version 4.2]

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;
```

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/