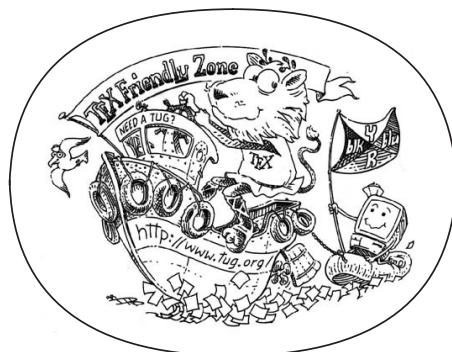


JOSÉ GERALDO DE CARVALHO PEREIRA
AUTÔMATOS CELULARES E PROTEÍNAS

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/beck00PSLA.html>

ABSTRACT

Resumo

SUMÁRIO

I FUNDAMENTOS TEÓRICOS	1
1 INTRODUÇÃO	3
1.1 Métodos computacionais de modelagem estrutural	4
1.2 Hipóteses sobre o enovelamento	6
2 OBJETIVOS	9
2.1 Objetivos específicos	9
3 JUSTIFICATIVA	11
 II DESENVOLVIMENTO	13
4 CONJUNTO DE DADOS	15
4.1 Conjunto de dados de proteínas diversas para treinamento	15
4.2 Proteínas com altíssima identidade sequêncial e diferentes estruturas terciárias e secundárias	16
5 IMPLEMENTAÇÃO	19
5.1 Autômato celular	19
5.1.1 Modelo inicial	19
5.1.2 Modelos extendidos	19
5.2 Busca de regras de transição utilizando Algoritmo de Estimação de Distribuição EDA	20
5.2.1 Função de pontuação dos indivíduos (Função de <i>fitness</i>)	21
5.3 Implementação	22
 III RESULTADOS	23
6 APRENDIZADO DE REGRAS DE TRANSIÇÃO	25
6.1 Autômato celular	25
6.2 EDA	25
6.2.1 Deriva genética	28
6.2.2 Função de fitness	28
6.2.3 Seleção	30
6.3 proteína?	33
7 APRENDIZADO DAS REGRAS GERAIS	35
8 APLICAÇÃO DAS REGRAS DE TRANSIÇÃO	37
8.1 Autômato celular determinístico	37
8.2 Autômato celular probabilístico	37
 IV PERSPECTIVAS FUTURAS	39
9 PERSPECTIVAS FUTURAS	41
9.0.1 Função de fitness	41
9.0.2 CA probabilístico durante a otimização das regras pelo EDA	41

9.0.3 Predição de estados conformacionais dos resíduos 41

V APPENDIX 43

BIBLIOGRAFIA 45

LISTA DE FIGURAS

Figura 1	Figura da sequencia e das estruturas das camaleonicas	17
Figura 2	Figura da sequencia e das estruturas das camaleonicas	17
Figura 3	Esquema da regra simples	19
Figura 4	Figura da sequencia e das estruturas das camaleonicas	26
Figura 5	Figura da sequencia e das estruturas das camaleonicas	26
Figura 6	Figura da sequencia e das estruturas das camaleonicas	27
Figura 7	EDA	28
Figura 8	Deriva genetica	29
Figura 9	Desbalanceamento elementos de estrutura secundária no conjunto de dados	29
Figura 10	q3	30
Figura 11	Histograma de ocorrencia das trincas	31
Figura 12	Probabilidades máximas e mínimas pelo numero de corrência das trincas	31
Figura 13	Relação proporção de ss nas trincas x probabilidade EDA	32
Figura 14	Relação proporção de acertos x probabilidade EDA	32
Figura 15	Relação ocorrências das ss por trincas x probabilidade EDA	33

LISTA DE TABELAS

Tabela 1	Autem timeam deleniti usu id	16
Tabela 2	Correlação (Spearman) entre a proporção de acertos na predição e: (1) proporção das trincas nas estruturas secundárias, (2) número de ocorrências das trincas nas estruturas secundárias	33

LISTINGS

ACRONYMS

Parte I
FUNDAMENTOS TEÓRICOS

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 [KENDREW:1958], 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 [Anfinsen:1968] 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, 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 [Rose:2001].

Dessa forma, devido ao princípio da relação estrutura ↔ 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 previsão foram publicados por Levinthal [Levinthal:1963, Levinthal:]. Nestes trabalhos o autor menciona que o número de configurações estruturais possíveis para uma cadeia polipeptídica é imenso, sendo impossível explorar todas as conformações para encontrar sua estrutura nativa. Apesar disso, as proteínas são capazes de se enovelarem espontaneamente e adotar a conformação nativa 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 [Ben-Naim:2007]. Segundo Levinthal [Levinthal:], 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 há 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 [Moult:2009]. Assim, métodos experimentais, mais especificamente, os métodos de cristalografia de proteínas por difração de raios-X e o de ressonância magnética nuclear, são as principais forma de se obter um modelo estrutural com resolução atômica.

Entretanto, métodos experimentais de resolução da estrutura proteica apresentam diversas dificuldades técnicas. Por exemplo, na cristalografia por difração de raios-X é necessária a obtenção de proteína em alto grau de pureza e a formação de monocrystalis, que muitas vezes são fatores limitantes do processo. Por outro lado, a ressonância magnética nuclear exige altas concentrações de proteína purificada em meios com diferentes isótopos e ainda além de possuir uma limitação quanto ao tamanho da proteína analisada. Essas limitações experimentais são evidenciadas pela disparidade entre o número de estruturas resolvidas experimentalmente (≈ 115 mil depositadas no PDB) e o número de proteínas com sequência de aminoácido conhecidas (≈ 67 milhões depositadas no UniProtKB/TrEMBL – dados de 09/2016).

Consequentemente, 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 com diversas aplicações práticas como o desenvolvimento de novos medicamentos [Baker:1996].

MÉTODOS COMPUTACIONAIS DE MODELAGEM ESTRUTURAL

Os métodos computacionais desenvolvidos e utilizados para construir modelos estruturais das proteínas podem ser classificados em dois tipos: (1) modelagem comparativa e (2) modelagem ab initio ou de novo, sendo considerados ab initio os métodos que não utilizam informações provenientes de proteínas com estruturas similares ao invés de apenas métodos que utilizem informação de caráter exclusivamente físico [Helle:2003]. Dentre os métodos de predição estrutural ab initio, os que tem apresentado melhor desempenho são os que utilizam uma técnica de montagem de fragmentos (*fragment assembly*) como o I-Tasser [Zhang:2003] e o Rosetta [Rohl:1999]. Esses métodos utilizam fragmentos extraídos de proteínas com estrutura

resolvida experimentalmente, os quais posteriormente são reunidos de acordo com a sequência de aminoácidos da proteína a qual se deseja construir um modelo. A utilização de fragmentos tem como objetivo acelerar a busca pelo modelo correto, entretanto, este ainda é um método caro computacionalmente e informações como a predição da estrutura secundária e de contatos não-locais entre resíduos são comumente utilizadas para restringir o número de fragmentos a serem testados, consequentemente reduzindo o espaço de busca e acelerando a modelagem da estrutura [Helle:2003].

A outra categoria de métodos de modelagem, a modelagem comparativa, necessita que estruturas similares à da proteína que se deseja modelar tenham sido previamente resolvidas experimentalmente. Os métodos de modelagem comparativa baseiam-se no princípio que, em proteínas homólogas, a estrutura é mais conservada do que a sequência de aminoácidos. Sendo assim, proteínas que tenham uma identidade entre as sequências maior que 30%, apresentando assim uma evidência de homologia, podem ser modeladas caso uma delas tenha estrutura resolvida [Marti-Renom:1995]. Isso não significa que proteínas com identidade sequencial menor que 30% não possam apresentar estruturas tridimensionais similares. Entretanto, a identificação dessas proteínas homólogas e o alinhamento entre as sequências, ambos passos essenciais durante a modelagem comparativa, tornam-se mais suscetíveis a erros [Marti-Renom:1995].

Na tentativa de contornar a deficiência da construção do alinhamento para a modelagem comparativa foram desenvolvidos métodos que buscam identificar proteínas com estruturas similares, mas baixa identidade sequencial (< 30%). Esse métodos, conhecidos como métodos de reconhecimento de enovelamento, englobam métodos de comparação sequência-estrutura e métodos de alinhavamento (*threading*) [Dunbrack:2001]. O diferencial desses métodos em relação ao simples alinhamento entre sequências primárias está na utilização de informações estruturais como por exemplo, estrutura secundária, exposição ao solvente, entre outros, para descrever o ambiente em que cada resíduo se encontra na proteína. Esses ambientes alteram os padrões de substituições de aminoácidos como demonstrado por Overington e colaboradores [Overington:1984]. Consequentemente, a utilização desses ambientes na construção de matrizes de substituição ou nos métodos de alinhavamento, permite uma estimativa da estrutura tridimensional que melhor acomoda a sequência de aminoácidos da proteína que se deseja modelar.

Outro argumento que justifica a aplicação do método de modelagem comparativa é a existência aparente de um número finito de enovelamentos adotados pelas proteínas, o qual alguns autores estimam ser entre 1.000 e 10.000 [Chothia:1987, Coulson:1997]. Portanto, com o aumento do número de estruturas resolvidas, acredita-se que futuramente esses enovelamentos estarão completamente representa-

dos nos bancos de dados, possibilitando a modelagem de um número cada vez maior de proteínas [Kolodny:2008].

Entretanto os métodos de modelagem comparativa não fornecem informações sobre o caminho, ou mesmo o mecanismo, de enovelamento da proteína, pois baseiam-se apenas na estrutura nativa para a construção do modelo [Helle:2003]. Essas informações sobre o caminho de enovelamento podem ser importantes para melhorar a previsão de estruturas terciárias como pode ser observado no trabalho de Giri e colaboradores [Giri:2007]. Neste trabalho os autores analisaram experimentalmente o enovelamento de proteínas com alta identidade entre as sequências (30%, 77% e 88%), mas que, mesmo com esta alta identidade, possuem topologias diferentes e notaram que as diferenças entre as topologias surgem logo no início do processo de enovelamento. Proteínas como essa, caso fossem modeladas comparativamente, provavelmente resultariam em modelos estruturais incorretos, o que a princípio poderia ser evitado com alguns métodos ab initio, como por exemplo os baseados em dinâmica molecular, ou por algum outro método capaz de obter informações sobre estágios intermediários do enovelamento a partir da sequência.

HIPÓTESES SOBRE O ENOVELAMENTO

A hipótese termodinâmica, apesar de explicar o enovelamento, não fornece informações sobre qual o mecanismo de enovelamento adotado pelas cadeias polipeptídicas na transição entre os estados desenovelado e nativo [Rose:2001]. Consequentemente, diversos mecanismos de enovelamento foram propostos [DilleChan:1997, Dill:2008; DilleMacallum:2012]. De maneira geral, esses mecanismos podem se distinguir em dois tipos: hierárquico e não-hierárquico. Num mecanismo hierárquico de enovelamento, acredita-se que o processo se inicia com estruturas que são formadas localmente na sequência e que comumente apresentam baixa estabilidade. A interação entre essas estruturas locais produziriam estruturas intermediárias, com complexidade crescente e maior estabilidade, até atingir a conformação nativa. Diferentemente, num mecanismo de enovelamento não-hierárquico, as interações não-locais não apenas estabilizariam as estruturas locais, mas seriam as responsáveis por determiná-las [BaldwinRose:1999].

Não há na literatura um consenso sobre qual tipo de mecanismo – hierárquico ou não hierárquico – descreve com maior fidelidade o enovelamento proteico, uma vez que há evidências experimentais e de simulação computacional que, ora sustentam um mecanismo, ora outro [BaldwinRose:1999, DaggetFersht:2003]. Devido a essas evidências, alguns autores propõem que, na realidade, ambos os mecanismos possam ocorrer, sendo portanto, não apenas a estrutura proteica, mas também o processo de enovelamento, determinados pela sequência de aminoácidos [DaggetFersht:2003].

Dentre as evidências que apontam para um mecanismo de enovelamento hierárquico estão estudos de proteínas como a α lactalbumina a apo-mioglobina, a RNase H, a barnase e o citocromo c, onde análises do processo de enovelamento indicam que ocorre uma rápida formação de estrutura secundária semelhante à observada na proteína nativa (native-like) e que a mesma é estabilizada em estruturas intermediárias (molten globules), ou seja, antes da proteína atingir sua conformação nativa [BaldwineRose:1999a]. Outras evidências que sugerem a existência de um mecanismo hierárquico são os padrões na sequência de aminoácidos que ocorrem imediatamente após as extremidades N e C terminal de hélices α [HarpereRose:1993, AuroraSrinivasaneRose:1994, Aurora:1997, BaldwineRose:1999a] e fitas β [ColloceCohen:1991], os quais acredita-se que atuem como sinal de término ou “parada” desses elementos de estrutura secundária. Outros trabalhos ainda demonstram que há preferência de algumas trincas de aminoácidos por determinadas conformações e estruturas secundárias [BetancourteSkolnick:2004, Otaki:2010], sendo que algumas trincas, interessantemente, não foram observadas uma única vez em alguns tipos de estruturas secundárias das proteínas analisadas [Otaki:2010].

Alguns trabalhos de simulação computacional [AbagyanTrotov:1994, PedersenMoult:1997] também identificaram que algumas sequências peptídicas, correspondentes a pequenas regiões de proteínas, apresentam maior propensão a adotar uma estrutura secundária, ou mesmo uma conformação, similar a observada na estrutura nativa da proteína original. Posteriormente, Srinivasan e Rose [SrinivasanRose:1999] realizaram simulações para demonstrar que essas propensões por estruturas secundárias surgiam de impedimentos estéricos entre os átomos de resíduos consecutivos na sequência e que essas estruturas correspondem a estrutura secundária de estados intermediários da proteína, sendo por vezes conservada na estrutura nativa e em outras, alterada devido a interações não locais.

Esses resultados observados na literatura sugerem ser possível um algoritmo para o reconhecimento de enovelamentos proteicos construído a partir de um método que recapitule a formação de estruturas locais durante o processo de enovelamento das proteínas. A princípio, um método com estas características permitiria não apenas identificar proteínas que tenham uma conformação nativa similar, mas uma identidade sequencial baixa, assim como os métodos atuais de reconhecimento de enovelamentos proteicos, mas possivelmente apresentar sensibilidade suficiente para evitar que proteínas com considerável identidade sequencial e entretanto baixa similaridade estrutural, como as observadas no trabalho de Giri e colaboradores [Giri:2012], sejam incorretamente modeladas por comparação [Helles:2008]. Acreditamos que um método com tais características possa ser desenvolvido utilizando autômatos celulares.

Os autômatos celulares foram inventados na década de 40 por John von Neumann baseando-se em sugestões de seu colega, o matemático Stanislaw Ulam [Mitchell:2009]. Autômatos celulares são modelos matemáticos para representar sistemas complexos e consistem num conjunto de células discretas espacialmente que apresentam um estado dentre um conjunto finito de estados possíveis. Os autômatos celulares evoluem paralelamente, ou seja, o estado de cada célula evolui de maneira síncrona em passos discretos de tempo e de acordo com regras simples e determinísticas gerando uma complexidade a partir do efeito cooperativo de elementos simples - as regras e as células - tratando-se portanto de uma complexidade emergente, que surge globalmente no sistema a partir de regras simples, locais e determinísticas [Wolfram:1984].

Autômatos celulares tem sido utilizados em diversos campos de pesquisa como por exemplo na modelagem de sistemas: (1) biológicos, desde eventos intracelulares, como redes de interação proteicas, até estudo de populações; (2) químicos, na modelagem cinética de sistemas moleculares e no crescimento de cristais; (3) físicos, para o estudos sistemas dinâmicos, desde a interação entre partículas até o agrupamento de galáxias (Ganguly et al., 2003). No entanto, não há na literatura artigos que mencionem a sua utilização na predição de enovelamentos proteicos. Apesar disso, acreditamos tratar-se de um modelo promissor.

2

OBJETIVOS

O objetivo principal deste trabalho é desenvolver um método de previsão de estruturas secundárias utilizando autômatos celulares. Diferentemente dos métodos atuais, onde a predição é realizada em apenas dois estados, sequência -> estrutura secundária, o método proposto é iterativo, representando uma dinâmica de formação das estruturas secundárias.

Neste trabalho, optamos por utilizar o modelo de autômatos celulares como o método iterativo. Entretanto, métodos recentes de aprendizado profundo poderiam ser utilizados como alternativa ou complemento do modelo proposto, desde que mantivessem implicitamente ou explicitamente o caráter iterativo.

OBJETIVOS ESPECÍFICOS

O objetivo principal envolve os seguintes objetivos específicos:

1. Preparação de um conjunto de treinamento composto da sequência de aminoácidos e das estruturas secundárias atribuídas por diferentes métodos a partir de uma grande variedade de estruturas proteicas;
2. Implementação de autômatos celulares com estados que possibilitem representar a sequência de aminoácidos e os elementos de estrutura secundária. Assim, o estado inicial desses autômatos celulares representariam a sequência de aminoácidos e durante a evolução do autômato celular, os elementos de estrutura secundária surgiriam e se organizariam para formar a estrutura secundária proteica;
3. Implementação de um método de otimização das regras de transição dos autômatos celulares como o objetivo de maximizar a acurácia do método de predição;
4. Aplicação do método, análise e comparação dos resultados com outros métodos de predição.

3

JUSTIFICATIVA

Apesar de métodos de predição de estruturas secundárias serem estudados a mais de ?? anos e, ao longo do tempo, terem atingido uma acurácia em torno de 80%, o método proposto por nós contém características não encontradas em outros métodos de predição de estrutura secundária.

Algumas dessas características são a utilização da sequência de aminoácidos completa ao invés do particionamento da sequência em janelas, como comumente utilizado em redes neurais. A característica iterativa do autômato celular, a qual proporciona a emergência de padrões complexos originados a partir de padrões simples. O uso somente da sequência de aminoácidos da proteína, ao invés de matrizes de substituição específica por posição (PSSM) ou descritores físico-químicos. E a simplicidade na compreensão da informação contida nas regras de transição de um autômato celular capaz de predizer a formação de estruturas secundárias.

Acreditamos que a possibilidade de obter informações que outros métodos de predição de estrutura secundária não fornecem tornam esse trabalho relevante cientificamente. Por exemplo, algumas possibilidades seriam a obtenção de informações sobre a dinâmica de formação das estruturas secundárias, ou até mesmo, sobre o enovelamento. Assim como aplicações práticas como a análise de perturbações causadas por mutações pontuais, a utilização no design de proteínas e a contribuição para métodos ab initio de predição de estruturas terciárias.

Parte II
DESENVOLVIMENTO

4

CONJUNTO DE DADOS

Neste trabalho foram utilizados dois conjuntos de dados compostos da sequência de aminoácidos de proteínas com estruturas resolvidas experimentalmente e da estrutura secundária atribuída aos seus resíduos por quatro diferentes algoritmos: DSSP [Kabsch:1978], Stride [Frishman:1990], Kaksi [Martin:2000] e PROSS[Srinivasan:1994].

O primeiro conjunto selecionado é formado por estruturas de alta qualidade e tem como finalidade ser utilizado na busca de regras de transição para o autômato celular capazes de reproduzir os padrões de estruturas secundárias presentes nas proteínas. Essas regras de transição 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 celular para o universo de proteínas existentes.

O segundo conjunto selecionado é composto de quatro proteínas com altíssima identidade sequencial, tendo no máximo 3 resíduos mutados entre as sequências de 56 aminoácidos, e diferentes estruturas terciárias e secundárias.

CONJUNTO DE DADOS DE PROTEÍNAS DIVERSAS PARA TREINAMENTO

O conjunto de proteínas diversas utilizado para o treinamento do autômato celular foi obtido do banco de dados “Top8000” (versão de 2015). Esse banco de dados é organizado pelo Richardson Lab da Universidade de Duke (disponível em github.com/rlabduke/reference_data). As estruturas de proteínas, separadas em cadeias quando há mais de uma por estrutura, atendem aos seguintes critérios:

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

As proteínas selecionadas pelos critérios acima são subagrupadas de acordo com o grau de identidade sequencial: < 50%, <70% e <95%.

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 possuíam resíduos indeterminados.

Proteínas que apresentassem resíduos indeterminados na estrutura ou que tivessem 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.

PROTEÍNAS COM ALTÍSSIMA IDENTIDADE SEQUÊNCIAL E DIFERENTES ESTRUTURAS TERCIÁRIAS E SECUNDÁRIAS

Em 2007, Alexander e colaboradores [17609385], desenharam um experimento onde obtiveram dois enovelamentos com topologias diferentes para sequências com mais de 88% de identidade sequencial. O ponto de partida do experimento foram dois domínios chamados G_A e G_B com 56 aminoácidos. O domínio G_A possui um feixe de 3 hélices α (β - α helix bundle) enquanto o domínio G_B apresenta a enovelamento $4\beta+\alpha$, ou seja, 4 fitas β mais uma hélice α (Figura 1).

Posteriormente, uma série de outros estudos sobre esses dois domínios demonstrou ser possível obter os dois enovelamentos com identidade sequencial ainda maiores [10.1073/pnas.0805857105, 10.1073/pnas.0906408106], até que em 2012, He e colaboradores [10.1016/j.str.2011.11.018], obtiveram mutações pontuais capazes de alterar a estrutura entre os dois enovelamentos (Figura 2). As estruturas resolvidas por RMN das quatro proteínas foram as utilizadas para compor esse conjunto de dados (PDB IDs: 2LHC, 2LHD, 2LHE, 2LHG) utilizando os 10 primeiros modelos de cada estruturas com o objetivo de manter os dados平衡ados.

4.2 PROTEÍNAS COM ALTÍSSIMA IDENTIDADE SEQUÊNCIAL E DIFERENTES ESTRUTURAS TERCIÁRIAS E SEQUÊNCIAIS

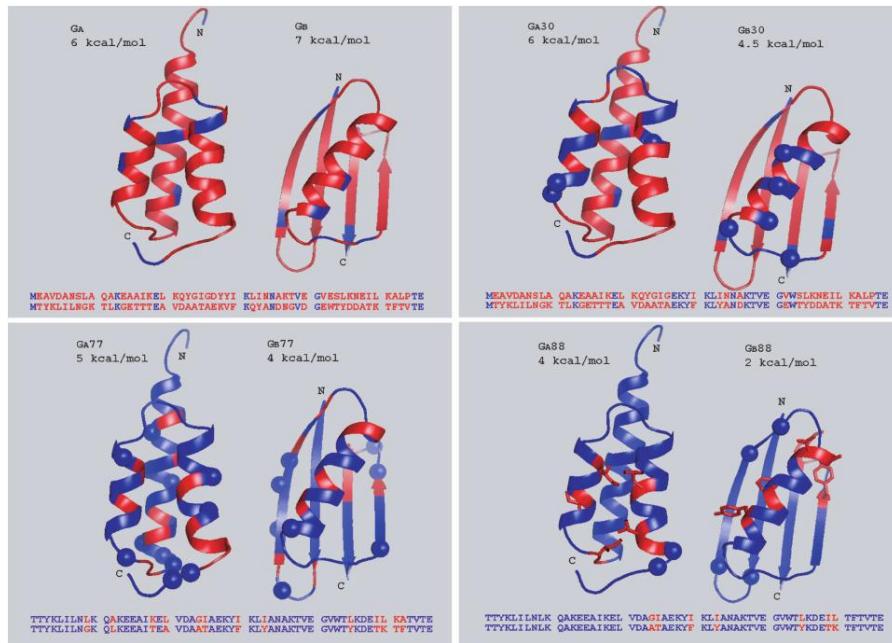


Figura 1: Figura da sequencia e das estruturas das camaleonicas

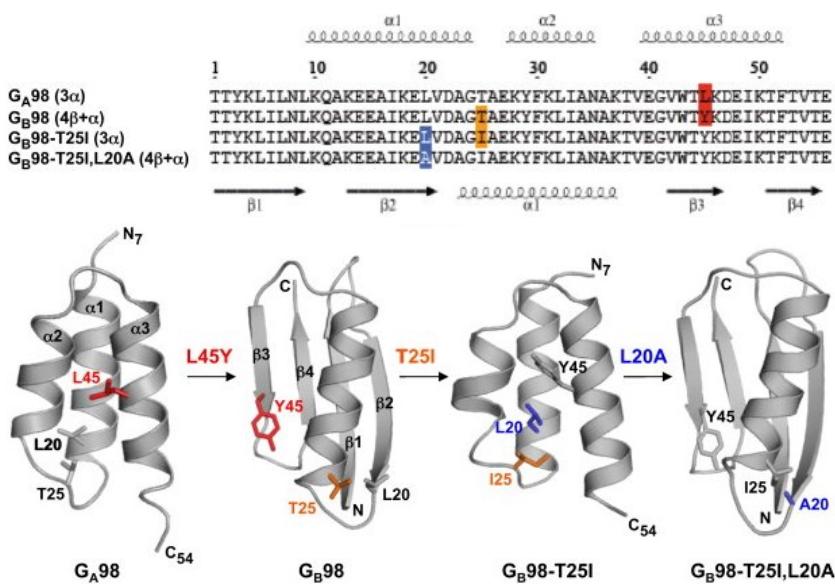


Figura 2: Figura da sequencia e das estruturas das camaleonicas

5

IMPLEMENTAÇÃO

AUTÔMATO CELULAR

Modelo inicial

O autômato celular inicialmente proposto possui 24 estados discretos. Esses estados correspondem aos 20 aminoácidos, aos 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, sendo 3 estados que representam os elementos de estrutura secundária e um que representa o resíduo presente naquela posição da cadeia polipeptídica.

Logo, temos que o total de elementos na regra desse autômato é 24^3 ou 13824, dos quais 24 são elementos estáticos. Esses elementos estáticos correspondem a células no estado # que não transitam para estados diferentes dele mesmo, permanecendo nesse estado durante toda 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 correspondentes aos aminoácidos para estados que representam elementos de estruturas secundárias. 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

$$\begin{array}{c} t_i \quad \boxed{\square \quad \square \quad \square} \\ \quad \quad \quad \boxed{\square} \\ t_{i+1} \end{array} = 24 \times 24 \times 24 \\ 24 \rightarrow \{20 \text{ AAs, H, E, C, }\#\\ = 4 \\ 4 \rightarrow \{H, E, C, ?\} \text{ onde ? representa o estado anterior}$$

Figura 3: Esquema da regra simples

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. Esse aumento de complexidade nos levou a descartar, pelo menos até o momento, este modelo.

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 2 estados para cada um dos 3 elementos 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 adicionando estados para diferenciar glicinas e prolínas, e outro acrescentando estados para diferenciar resíduos com cargas positivas e negativas assim como glicinas e prolínas. 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.

A motivação para o uso da hidrofobicidade dos resíduos foi influenciada por trabalhos de Hecht e colaboradores [Xiong07, West1995] que examinaram a influência de padrões periódicos de hidrofobicidade nas sequências proteicas e sua relação com elementos de estruturas secundárias, concluindo que alguns padrões apresentam preferência por hélices α enquanto outros padrões apresentam preferência por fitas β [West1995].

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.

BUSCA DE REGRAS DE TRANSIÇÃO UTILIZANDO ALGORITMO DE ESTIMAÇÃO DE DISTRIBUIÇÃO 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 têmpera simulada (*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 menor 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 de transição. A probabilidade é distribuída pelo nó mestre para os nós escravos. Os nós escravos utilizam a probabilidade recebida para gerar $c \geq 2$ 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. Em seguida, um torneio entre as regras candidatas geradas no nó escravo e a regra com maior fitness é enviada ao nó mestre. Ao receber as n/c regras vencedoras, onde n é o tamanho da população do EDA, o nó mestre atualiza a probabilidade e começa a distribuí-la para os nós escravos, iniciando assim a geração $T + 1$ do EDA. A otimização termina após um número específicos de gerações ou quando as probabilidades convergem.

Função de pontuação dos indivíduos (Função de fitness)

A função de fitness utilizada pelo EDA baseia-se na porcentagem de estados corretos durante a evolução do autômato celular ($t_1 \rightarrow t_{final}$) onde os estados corretos são os elementos de estrutura secundária idênticos ao concenso obtido entre os quatro métodos de atribuição de estruturas secundária. Quando não há concenso entre os métodos de atribuição de estrutura secundária, a posição é descartada pela função de fitness. A função de fitness é, portanto, equivalente a acurácia.

$$ACC = \frac{TP + TN}{P + N} \quad (1)$$

$$\text{fitness} = \frac{TH + TE + TC}{H + E + C} \quad (2)$$

Onde TP e TN são verdadeiros positivos e verdadeiros negativos, ou seja, elementos preditos corretamente. P e N correspondem ao número de elementos positivos e negativos, logo, $P + N$ resulta no número total de elementos do conjunto. Na função de *fitness*, TH, TE e TC correspondem ao elementos de estrutura secundária preditos

corretamente como hélices, fitas e coils ao longo da evolução do autômato celular. H + E + C representam o número de resíduos que compõem cada um dos elementos multiplicado pelo número de passos que o autômato celular irá evoluir.

IMPLEMENTAÇÃO

Tanto o autômato celular quanto o EDA foram implementados na linguagem de programação Go. A estrada do nó mestre é um arquivo de configuração no formato TOML que contém os parâmetros para o autômato celular, o EDA e os dados a serem utilizados. Os dados, ou seja as sequências de aminoácidos das proteínas e suas respectivas estruturas secundárias, são armazenados em um banco de dados chave/valor. A comunicação entre os nós escravo e o nó mestre é feita utilizando chamadas remotas de procedimento (RPC). O código fonte está acessível publicamente no GitHub (github.com/jgcarvalho/zeca-search, [zeca-search-master](https://github.com/jgcarvalho/zeca-search-master) e [zeca-search-slave](https://github.com/jgcarvalho/zeca-search-slave))

Parte III
RESULTADOS

6

APRENDIZADO DE REGRAS DE TRANSIÇÃO

Nas seções seguintes mostraremos resultados da capacidade de autômatos celulares reproduzirem padrões de estruturas secundárias a partir da sequência de aminoácidos e da utilização do EDA para buscar regras de transição. Inicialmente, discutiremos os autômatos celulares e as diferentes opções de estados utilizados. Em seguida, discutiremos os resultados da otimização das regras de transição utilizando um EDA distribuído.

AUTÔMATO CELULAR

Os modelos de autômatos celulares foram testados primeiramente, e até o momento, apenas no conjunto de proteínas com alta identidade sequencial. A escolha desse conjunto para os testes dos autômatos celulares baseou-se na facilidade de observar e analisar a formação e propagação dos elementos de estruturas secundárias nessas proteínas.

O modelo idealizado para o autômato celular deveria ter a capacidade de propagar sinais locais ao longo da sequência, e assim, resultar na formação de padrões globais. Tal capacidade está relacionada aos estados que ocorrem durante a evolução do autômato celular, sendo dependentes do número de estados possíveis e também do tamanho da vizinhança utilizada. Como todos os autômatos celulares propostos tem vizinhança 1 por questões de complexidade (ver ref métodos), a capacidade de formar e propagar os sinais será dependente apenas dos estados possíveis do autômato celular.

Entre os quatro modelos testados, o número de estados para os elementos de estrutura secundária demonstrou relação com a acurácia do modelo. Assim, CAs com estados de elementos de estrutura secundária que conservam mais características dos resíduos mostraram-se mais promissores (Figura 4).

EDA

O aprendizado das regras realizado através de um algoritmo de EDA distribuído demonstrou-se eficiente dado a complexidade do problema. Utilizando sete nós (448 núcleos) no cluster de computação de alto desempenho EMU-2 (adquirido pelo "Programa de Equipamento Multiusuário da FAPESP-2009", 2009/53853-5, localizado no Centro Internacional de Pesquisa e Ensino do Hospital A. C. Camargo em

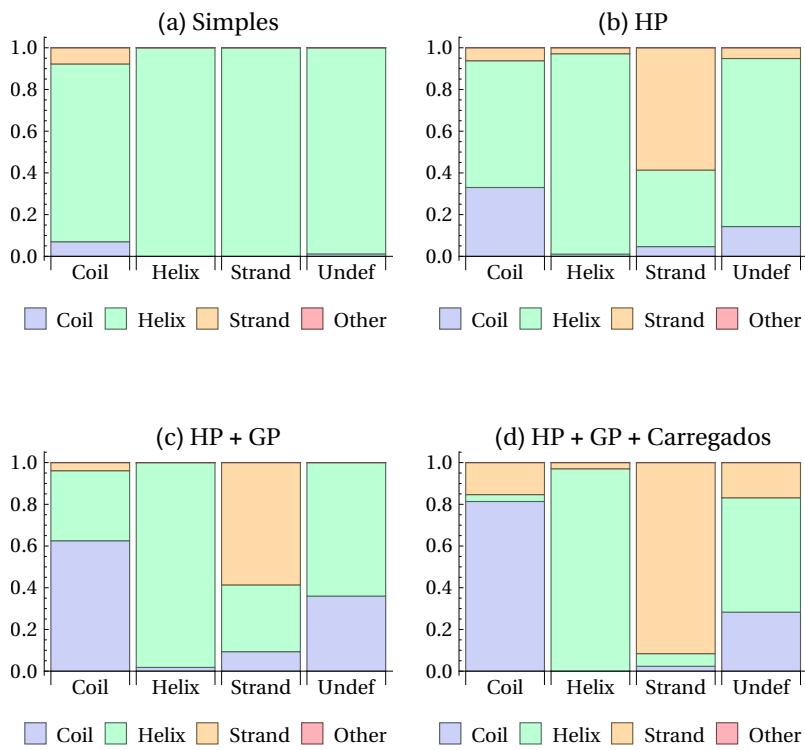


Figura 4: Figura da sequencia e das estruturas das camaleonicas

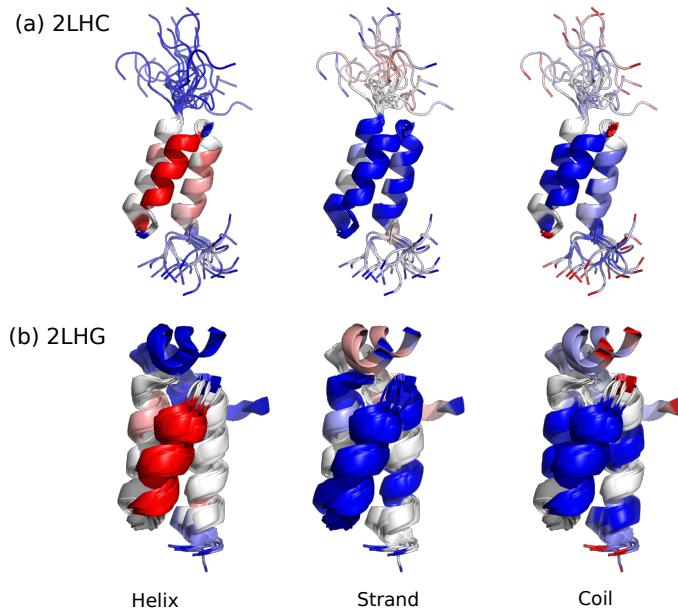


Figura 5: Figura da sequencia e das estruturas das camaleonicas

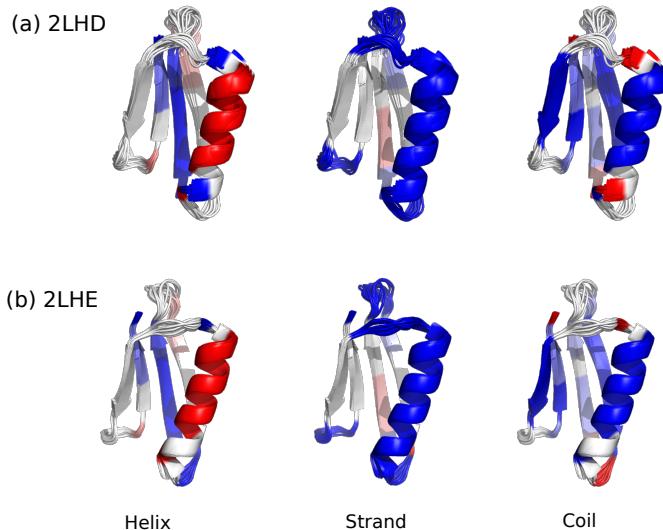


Figura 6: Figura da sequencia e das estruturas das camaleonicas

São Paulo) foi possível evoluir o EDA por 1000 gerações com 10000 indivíduos em pouco menos de duas semanas.

Cada nó de trabalho apresentou um uso de memória de 75% (48 GB), e manteve o processamento próximo a 100% por núcleo (1.6 GHz). O mecanismo de comunicação por RPC entre o nó mestre e os nós de trabalho não sobrecarregou a rede (1Gbs). Isso nos permite concluir que o algoritmo é escalável clusters com maior número de nós e processadores de maior desempenho.

A opção por realizar o torneio entre soluções candidatas nos nós de trabalho, permite apenas a opção de realizar o torneio entre as k últimas soluções geradas no próprio nó. Consequentemente, as $k-1$ soluções perdedoras são descartadas, havendo portanto, a remoção das soluções perdedoras em cada torneio.

Usualmente, o método de seleção por torneio utilizado em algoritmos genéticos não distribuídos acumula as solução candidatas até atingir o tamanho máximo populacional, quando então, é realizado o torneio. Isso permite que o torneio seja feito sem a eliminação dos perdedores, possibilitando que a seleção destes ocorram em outros combates.

Entretanto, no EDA distribuído, para aplicarmos um torneio sem eliminação dos perdedores seria necessário:

1. o envio de todas as soluções candidatas para o nó mestre, o que resultaria em maior consumo de rede;
2. o acúmulo de todas as soluções candidatas até atingir o tamanho máximo da população, gerando uma limitação da memória disponível;

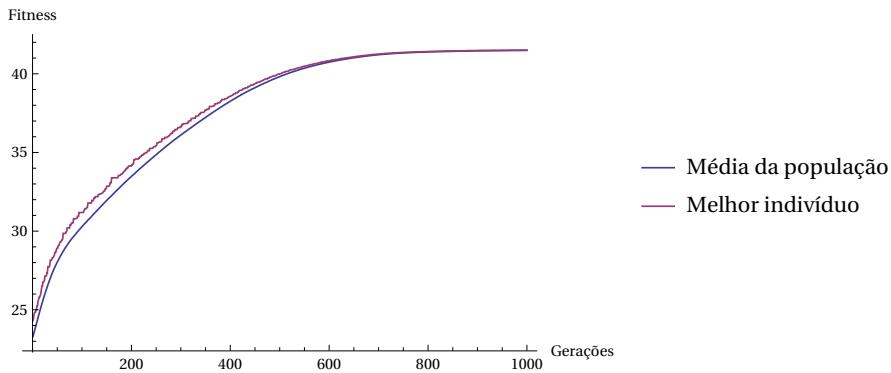


Figura 7: EDA

3. a espera até a realização do torneio para iniciar o cálculo das probabilidades, resultando no aumento do tempo ocioso nos nós de trabalho igual ao intervalo de tempo do envio da última solução até o cálculo das probabilidades.

A escolha em realizar o torneio nos nós de trabalho demonstrou ser escalável, manter a variabilidade das soluções candidatas ao longo da evolução e convergência (Figura 7).

Deriva genética

A evolução do EDA por 1000 gerações com torneio de dois ($k=2$) e população de 10000 indivíduos, demonstrou sinais de deriva genética a partir de 564 gerações. Tais sinais podem ser detectados observando-se as probabilidades dos 38 elementos de regra que não ocorrem no conjunto de proteínas. Esses elementos são do tipo $[#][x][#]$.

O elementos $[#][x][#]$, onde o x corresponde a qualquer estado exceto $[#]$, não apresentam probabilidade fixa, mas também, por estar ausente nas proteínas, não sofrem pressão seletiva. Logo, suas variações são aleatórias.

Na geração 564 um desses 38 elementos apresentou probabilidade zero ($p = 0$), de transitar para um dos quatro estados possíveis, indicando a eliminação de um gene da população por deriva genética. Ao final das 1000 gerações 11 dos 38 elementos apresentavam probabilidade zero para uma das transições (Figura 8).

Função de fitness

A simplicidade da equação de fitness (Equação 1) demonstrou problemas que acreditamos serem solucionáveis na continuação deste trabalho. Um desses problemas é ocasionado pelo desbalanceamento dos dados de treinamento (Figura 9).

Os resultados obtidos indicam uma maior acurácia para os elementos de estrutura secundária mais frequentes no conjunto de treina-

Elementos da regra

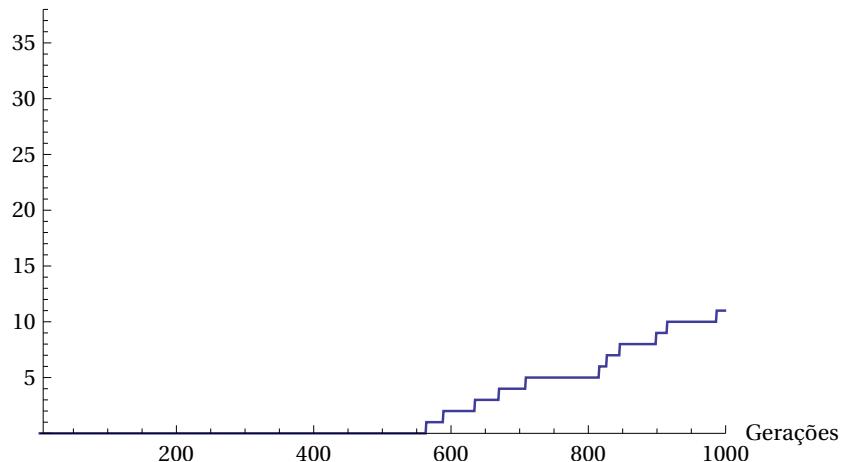


Figura 8: Deriva genética

Ocorrências

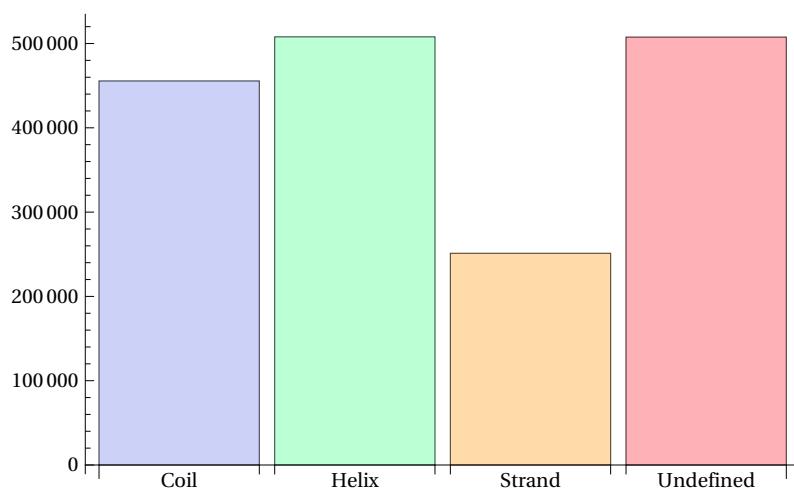


Figura 9: Desbalanceamento elementos de estrutura secundária no conjunto de dados

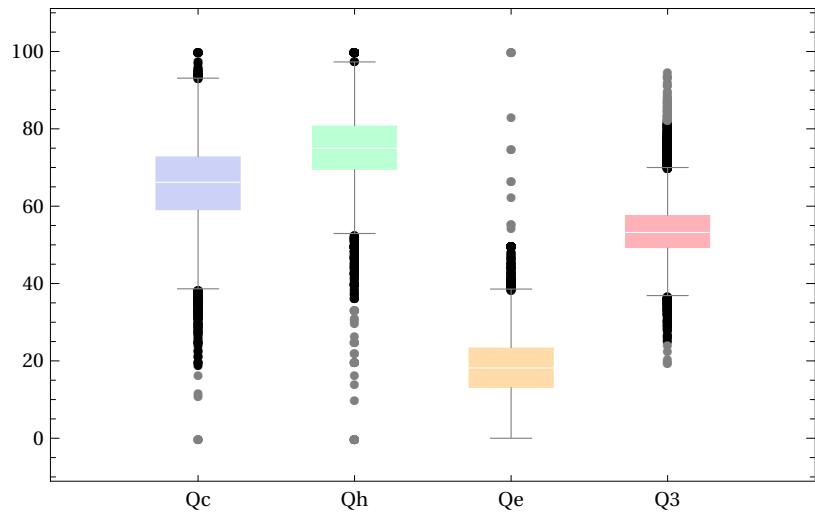


Figura 10: q3

mento. Esse é um problema recorrente no aprendizado com classes desbalanceadas e costuma ser tratado na função de fitness (Figura 10), por exemplo, com funções que fazem uma média da acurácia por classes. As alterações na função de fitness para tratar este problema são preferíveis em relação à modificações no conjunto de dados para equilibrar as classes, uma vez que modificações no conjunto de dados envolveriam a retirada de informação de classes mais frequentes ou a duplicação de informação das classes menos frequentes, ambas produzindo consequências distorcivas na informação presente no conjunto de treinamento.

Seleção

As ocorrência de trincas no conjunto de proteínas apresentou grande variação (Figura 11). Para avaliar a influência do número de ocorrências das trincas na seleção do EDA nós procuramos por correlações entre fatores que poderiam influenciar na seleção.

AAA 2 BBB 1000

A figura 12 representa a distribuição das probabilidades máximas e mínimas das dos elementos da regra em relação a frequência de ocorrência das trincas no conjunto. O gráfico e o valor de correlação (Spearman = 0.09) entre as probabilidades e a frequência de ocorrência das trincas, indica não haver uma grande influência da frequência das trincas durante a seleção. Essa influência também não foi encontrada durante a evolução do EDA (Anexo).

Outros fatores que poderiam influenciar a evolução do EDA seriam: (1) a probabilidade da trinca ser observada em uma determinada estrutura secundária; (2) o número de ocorrências da trinca em determinada estrutura secundária; (3) a taxa de acertos por trincas para determinada estrutura secundária.

Elementos da regra

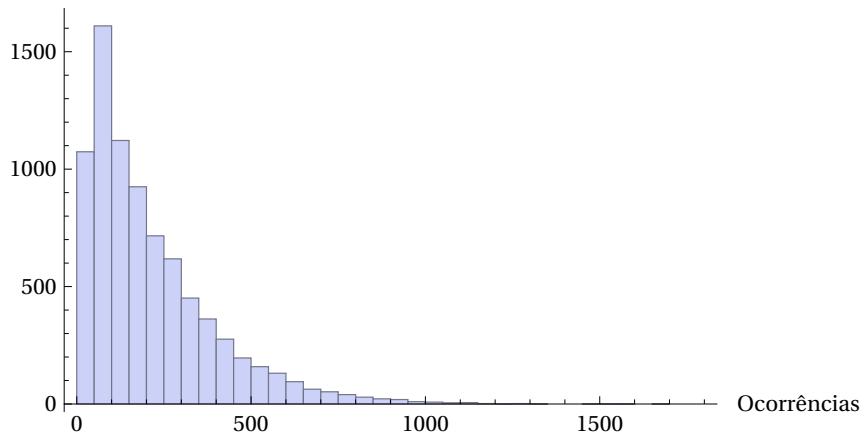


Figura 11: Histograma de ocorrência das trincas

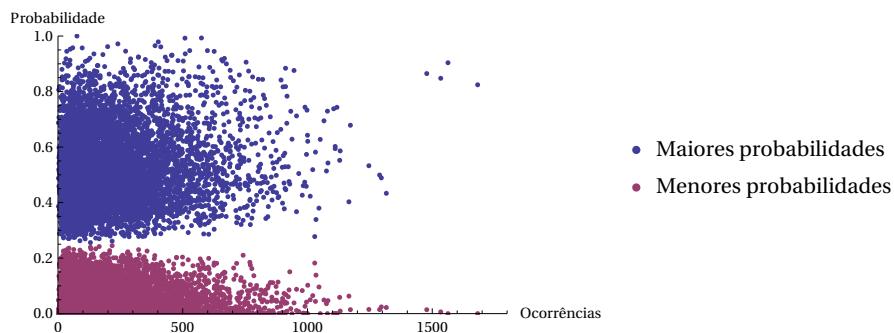


Figura 12: Probabilidades máximas e mínimas pelo número de ocorrência das trincas

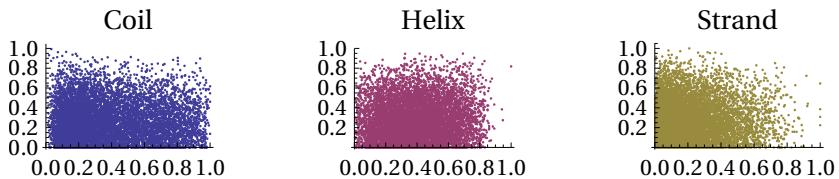


Figura 13: Relação proporção de ss nas trincas x probabilidade EDA

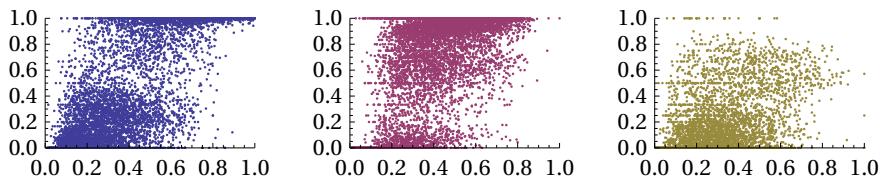


Figura 14: Relação proporção de acertos x probabilidade EDA

As probabilidades de transição dos elementos da regra, não apresentam aparente relação com as proporções encontradas para cada estrutura secundária nas trincas (Figuras H E C). Assim como não apresentaram relação com a frequência de ocorrências de determinada estrutura para uma trinca e com a proporção de trincas corretamente classificadas por estrutura secundária. Essas duas últimas relações não eram mesmo esperadas (Anexo Figuras H E C e figuras H E C). Contudo, foi interessante observar que a primeira relação aparentemente não ocorria.

A ausência da primeira relação pode indicar que as probabilidades de transição dos demais elementos da regra e sua aplicação ao longo da evolução do CA estão propagando as probabilidades e sofrendo influência da vizinhança na busca de produzir estruturas secundárias mais próximas as reais.

Entretanto, observamos uma relação entre a proporção de elementos de estruturas secundárias para um trinca e a proporção de acertos da trinca para a mesma estrutura secundária. Acreditamos que isso seja um indício que o aprendizado, ou otimização da regra, precisa ser melhorado para que elementos de estruturas secundárias menos comuns a determinadas trincas possam ser corretamente preditos (Figura 14 e Tabela ??).

Há também uma relação menos influente entre o número de ocorrências de uma determinada estrutura secundária nas trincas e a proporção de acertos dessa trinca para a mesma estrutura secundária (Figura 15 e Tabela ??).

Ambas as relações encontradas eram esperadas pois indicam uma tendência do método a privilegiar o aprendizado, ou otimização, de elementos capazes de influenciar com maior intensidade a função de fitness. Isso mostra a importância de manter um variabilidade alta na população e reduzir efeitos de deriva genética para que a otimização escape de mínimos locais e consiga aprender, também, as pro-

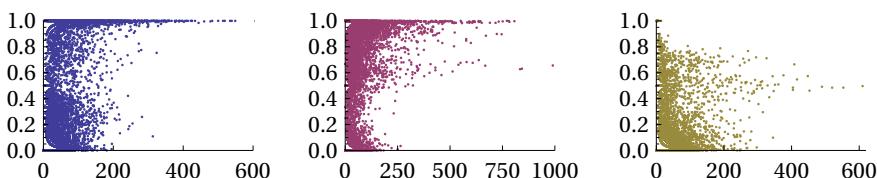


Figura 15: Relação ocorrências das ss por trincas x probabilidade EDA

CORRELAÇÃO	COIL	HÉLICES	FITAS
Proporção das trincas	0,75	0,60	0,60
Ocorrências das trincas	0,60	0,24	0,38

Tabela 2: Correlação (Spearman) entre a proporção de acertos na predição e: (1) proporção das trincas nas estruturas secundárias, (2) número de ocorrências das trincas nas estruturas secundárias

babilidades para trincas menos frequentes nas proteínas e/ou com proporções pequenas para determinadas estruturas secundárias.

PROTEINA?

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.

7

APRENDIZADO DAS REGRAS GERAIS

8

APLICAÇÃO DAS REGRAS DE TRANSIÇÃO

Após a otimização das regras de transição utilizando o EDA, nós avaliamos o desempenho das regras em classificar corretamente os elementos de estrutura secundária. A regra de transição que apresentou a melhor acurácia ao longo da evolução do EDA foi utilizada em um autômato celular determinístico, onde cada elemento da regra apresenta sempre a mesma transição.

Além do autômato celular determinístico, nós utilizamos as probabilidades da última geração do EDA como uma regra de transição para um autômato celular probabilístico.

AUTÔMATO CELULAR DETERMINÍSTICO

Q₃
curva Roc

AUTÔMATO CELULAR PROBABILÍSTICO

Q₃
curva Roc

Parte IV
PERSPECTIVAS FUTURAS

9

PERSPECTIVAS FUTURAS

Neste capítulo discutiremos alterações que estão sendo planejadas para melhorar o método que estamos desenvolvendo. Essas alterações foram pensadas baseando-se nos resultados obtidos até o momento e nas possibilidades para melhorá-los.

Função de fitness

Os resultados obtidos até o momento indicam que a função de fitness precisa de modificada para conseguirmos melhorar a acurácia da predição. A função utilizada até o momento não mostrou-se eficaz ao lidar com classes desbalanceadas como podemos notar pela menor acurácia obtida em fitas quando comparada a hélices e coils.

A ineficácia ao lidar com classes desbalanceadas, não prejudica somente a predição das classes menos numerosas, mas também a das classes mais numerosas, uma vez que essas tendem a ser superestimadas.

Entre as funções de fitness que planejamos testar estão a CBA e a MCC.

CA probabilístico durante a otimização das regras pelo EDA

Como descrito anteriormente, durante a otimização, o EDA envia probabilidades para os nós de trabalho que irão gerar regras de transição candidatas que finalmente disputarão os torneios. Atualmente, estas regras de transição geradas nos nós de trabalho são determinísticas.

Uma alternativa que planejamos implementar na continuidade deste trabalho é a modificação dos autômatos celulares utilizados durante a otimização das regras para autômatos celulares do tipo probabilísticos. Com isso, as probabilidades do EDA não representariam apenas as probabilidades das transições na população do EDA, mas sim a probabilidade de transição do elemento em si. Acreditamos que essa modificação aproximaria as regras de transição de um modelo físico, onde as probabilidades de transição teriam relação com a energia livre de uma transição da trinca de estados.

Predição de estados conformacionais dos resíduos

Uma possibilidade que está sendo avaliada é modificação do objetivo da predição. A predição de elementos da estrutura secundária por resíduo não apresenta uma correspondente física uma vez que um

resíduo raramente compõe isoladamente uma estrutura secundária. Por exemplo, hélices e fitas são formadas pela repetição de resíduos que se encontram em estados conformacionais específicos. Logo, um resíduo isolado no estado conformacional de uma hélice, poderia fazer fazer parte outro elemento de estrutura secundária como um coil ou volta.

Por sua vez, uma estrutura secundária poderia ser classificada pelo estado conformacional dos resíduos, sendo o estado conformacional representado por regiões de ângulos phi e psi. Assim, ao invés do autômato celular predizer os elementos de estrutura secundária ele poderia predizer estados conformacionais referentes a regiões dos ângulos.

Esta representação, combinada com o uso de um CA probabilístico, representaria as probabilidades dos estados conformacionais, fornecendo também as alterações de probabilidades que ocorrem ao longo da evolução do autômato celular. Assim, a evolução desse CA probabilístico resultaria na predição das probabilidades de cada aminoácido estar em cada estado conformacional.

Isso traria a vantagem de, além de ser possível descrever os elementos de estrutura secundária utilizando os estados conformacionais de cada resíduo, seria possível obter também informações da estrutura tridimensional das proteínas. Mesmo que essa abordagem diminuísse a acurácia na predição de elementos de estrutura secundária, o ganho de informação que poderíamos na conformação tridimensional, por exemplo, em coils, poderia compensar a mudança para a adoção dos estados conformacionais na predição.

Por outro lado, o número de estados conformacionais depende da forma como regiões de ângulos phi e psi serão discretizadas uma vez que o CA exige estados discretos. A princípio, poderíamos optar por poucos estados como a região de hélices, região de fitas e outras regiões, sendo essa última, todas as não hélices e não fitas. A princípio, um número maior de estados conformacionais poderia ser usado permitindo identificar, por exemplo, regiões de volta. No entanto, isso aumentaria exponencialmente o tamanho da regra.

Parte V
APPENDIX

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 L^AT_EX 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/>