Universidade de São Paulo
Programa de Pós-Graduação em Bioinformática

RECONHECIMENTO DE ENOVELAMENTOS PROTEICOS UTILIZANDO AUTÔMATOS CELULARES

JOSÉ GERALDO DE CARVALHO PEREIRA
ALUNO DE DOUTORADO

DR. PAULO SÉRGIO LOPES DE OLIVEIRA
ORIENTADOR

Relatório anual de bolsista

Novembro de 2014

[5 de janeiro de 2015 at 18:14 – classicthesis version 4.1]

SUMÁRIO

Parte I

RELATÓRIO DE ATIVIDADES

1 INTEGRALIZAÇÃO DOS CRÉDITOS

No primeiro semestre de 2014, foram cursadas duas disciplinas.

1. Introdução ao Aprendizado de Máquina
    - Docente: Dr. Rodrigo Fernandes de Mello
    - Créditos: 12
    - Carga horária: 180 horas
    - Departamento: ICMC-USP, São Carlos
    - Conceito obtido: A
2. Estrutura e Função de Proteínas
    - Docente: Dr. Richard Charles Garratt
    - Créditos: 9
    - Carga horária: 135 horas
    - Departamento: IFSC-USP, São Carlos
    - Conceito obtido: A

E outras duas disciplinas foram cursadas no segundo semestre de 2014.

1. Algoritmos de Estimação de Distribuição
    - Docente: Dr. Alexandre Cláudio Botazzo Delbem
    - Créditos: 6
    - Carga horária: 90 horas
    - Departamento: ICMC-USP, São Carlos
    - Conceito obtido: A
2. Projeto de Inovação com Algoritmos Genéticos
    - Docente: Dr. Alexandre Cláudio Botazzo Delbem
    - Créditos: 6
    - Carga horária: 90 horas
    - Departamento: ICMC-USP, São Carlos
    - Conceito obtido: A

As quatro disciplinas totalizaram 33 créditos, sendo que o mínimo exigido neste programa são 32 créditos.

2

PRODUÇÃO CIENTÍFICA

O aluno enviou um resumo para a VII Escola de Modelagem Molecular em Sistemas Biológicos, no período de 18-22 de agosto, no Laboratório Nacional de Computação Científica - Petrópolis, RJ.

No entanto, apesar do resumo ter sido aceito e do auxílio financeiro ter sido aprovado pelo conselho da pós-graduação, o auxílio foi cancelado devido a mudanças nas regras de requisição para a participação em eventos e por isso, o aluno acabou não participando do evento.

3

PUBLICAÇÕES

No período, foi submetido um artigo em colaboração com o grupo do Dr. Celso Benedetti (LNBio - CNPEM) para a revista Frontiers in Plant Science. Este artigo encontra-se em revisão no momento.

4

ENCONTROS COM O ORIENTADOR

Encontros com o orientador foram realizados frequentemente para discutir, tanto o andamento do projeto, como trabalhos em colaboração com outros grupos e projetos desenvolvidos no laboratório. No início do período, houve uma apresentação do projeto de pesquisa para todos os membros do grupo.

Parte II

ESTÁGIO ATUAL DA PESQUISA

5 ANDAMENTO DO PROJETO

5.1 SUMÁRIO DO PROJETO INICIAL

Neste projeto, propomos o desenvolvimento de um método de reconhecimento de enovelamentos proteicos utilizando autômatos celulares. O reconhecimento do enovelamento é uma etapa crucial para expandir a utilização da modelagem comparativa, pois permite identificar proteínas que tenham estruturas semelhantes, mesmo na ausência de alta similaridade entre suas estruturas primárias. Além do reconhecimento do enovelamento proteico, os autômatos celulares apresentam características que, a princípio, os tornam capazes de fornecer informações sobre a dinâmica do enovelamento, sendo, portanto, um método promissor nesta área, mas até então inédito.

Autômatos celulares são modelos computacionais que consistem num conjunto de células discretas espacialmente, as quais encontram-se em estado discreto num tempo também discreto. O estado dessas células podem se alterar ao longo do tempo para outros estados pertencentes à um conjunto finito de estados possíveis. A alteração desses estados ao longo do tempo é chamada de evolução e é definida por regras de transição simples e locais, onde o estado atual da célula e de seus vizinhos definirá para qual estado a célula irá evoluir. Esta evolução, muitas vezes, produz padrões complexos a partir do efeito cooperativo desses elementos simples - as regras e as células. Essa complexidade que surge globalmente no sistema a partir de regras simples, locais e determinísticas é conhecida como Emergência. A figura 1 exemplifica um autômato celular elementar.

Visando o objetivo principal de desenvolver um método de reconhecimento de enovelamentos proteicos baseado em autômato celulares foram estipuladas algumas etapas de menor complexidade e custo computacional. Essas etapas permitirão avaliar progressivamente o potencial do método e assim direcionar o desenvolvimento e contornar possíveis dificuldades.

Uma dessas etapas é a utilização de autômatos celulares para a predição de estruturas secundárias proteicas. As estruturas secundárias são estruturas locais e diversas hipóteses sugerem que durante o enovelamento elas se originam em determinadas regiões da sequência de aminoácidos e propagam-se, formando uma estrutura secundária semelhante a encontrada na proteína em sua conformação nativa. Essa característica de iniciar localmente e propagar-se formando a estrutura secundária sugere que este é um evento promissor para ser modelado por autômatos celulares. A figura 2 representa um modelo de

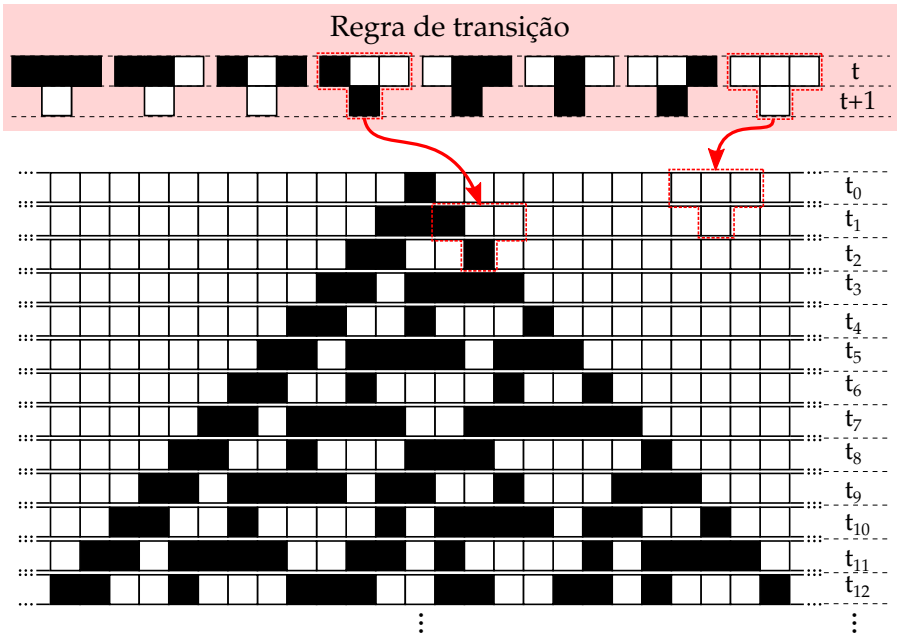

Figura 1: Exemplo de um autômato celular elementar. Este é o tipo de autômato celular mais simples, pois é unidimensional, possui apenas dois estados (0 ou 1) e possui vizinhança 1 (uma célula a esquerda e uma a direita). O autômato inicia com uma linha de células (t_0) e evolui através da aplicação de uma regra de transição que irá determinar o estado de cada uma das células na geração seguinte (t_1). Este processo ocorre iterativamente produzindo, muitas vezes, uma complexidade global que emerge da aplicação das regras de transição locais.

autômato celular unidimensional adaptado a predição de estruturas secundárias proteicas.

No entanto, a principal dificuldade do projeto consiste em encontrar regras de transição para os autômatos celulares capazes de reproduzir o evento desejado. Essa dificuldade ocorre devido ao imenso espaço de regras possíveis. Por exemplo, considerando apenas regras que contenham os estados correspondentes aos aminoácidos e a elementos de estrutura secundária, como na figura 2, teríamos 4^{13225} combinações, ou regras, possíveis que poderiam predizer a formação de estrutura secundária a partir da sequência de aminoácidos.

$$\begin{aligned} \mathtt{nElementos} &= (\mathtt{nAA} + \mathtt{nSS} + \mathtt{n\#})^3 \\ &= (20 + 3 + 1)^3 \\ &= 13824 \end{aligned}$$

Onde nAA é o número de aminoácidos (20), nSS é o número de elementos da estrutura secundária (3 = hélice, fita, alça) e n# é o sinal de início/término da cadeia polipeptídica. O expoente 3 é resultado do número de células que determinam o estado seguinte, nesse caso,

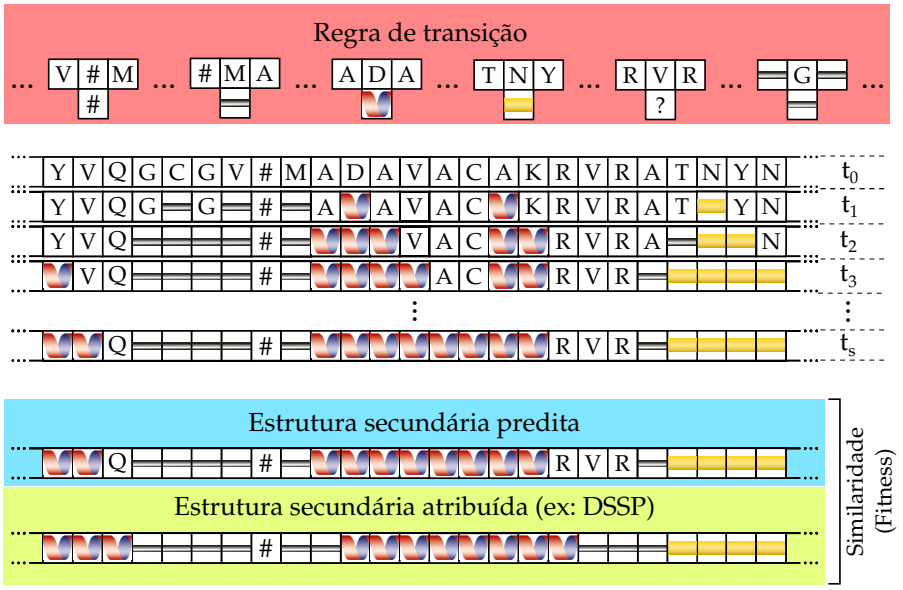

Figura 2: Esquema de um autômato celular para a predição de estruturas secundárias. A regra de transição do autômato é composta de 13824 elementos. O estado inicial do autômato (t_0) é formado pela sequência de aminoácidos da proteína. O estado # indica o início ou fim da cadeia polipeptídica e pode ser utilizado para concatenar múltiplas proteínas, funcionando como uma barreira de influência entre elas. O estado final do autômato celular t_s, onde s é um número finito, é a estrutura secundária predita e a sua similaridade com a estrutura secundária atribuída pode ser considerada uma medida do acurácia da regra de transição.

o estado da própria célula e os estados dos vizinhos, um a esquerda e um a direita.

Cada um desses 13824 elementos da regra de transição podem evoluir pra quatro estados possíveis. Três deles representam elementos de estrutura secundária e um deles representa um estado indefinido o qual indica que a célula assumirá um estado igual ao seu estado inicial (estado em t_0).

Desses 13824 elementos, 576 deles possuem o sinal de início/término # na célula central. Esses elementos evoluem sempre para o estado # para conservar o sinal de início/término da cadeia polipeptídica. Outros 23 elementos possuem ambos os vizinhos com o sinal de início/término # e permanecerão constantemente no estado inicial t_0. Assim, somente os 13225 elementos restantes poderão assumir um dos quatro estados possíveis mencionados anteriormente. Nosso problema resume-se, portanto, em encontrar a melhor combinação desses 13225 elementos capaz de predizer a estrutura secundária proteica.

5.2 ANÁLISE DO PERÍODO

Neste período, o código fonte inicial foi quase totalmente refeito para permitir maior flexibilidade. Isso foi necessário, pois o código desenvolvido inicialmente e que foi utilizado em testes preliminares, abrangia apenas a predição da estruturas secundárias das proteínas e tinha como objetivo avaliar a viabilidade do projeto.

As modificações feitas no código permitirão criar autômatos celulares com outros estados além dos já utilizados, que correspondem aos aminoácidos e elementos de estrutura secundária. Isso possibilitará testarmos tanto códigos simplificados que representem, por exemplo, as características físico-químicas dos aminoácidos, como carga, polaridade, entre outros, assim como códigos mais complexos que representem, por exemplo, diversos elementos de estrutura secundária juntamente com a exposição ao solvente.

Diversos

Com a modificação dos objetivos do projeto para a criação de um método de reconhecimento de enovelamentos proteicos

5.3 DISCUSSÕES E CONCLUSÕES PARCIAIS

Discussões

5.4 PERSPECTIVAS FUTURAS

No próximo periodo

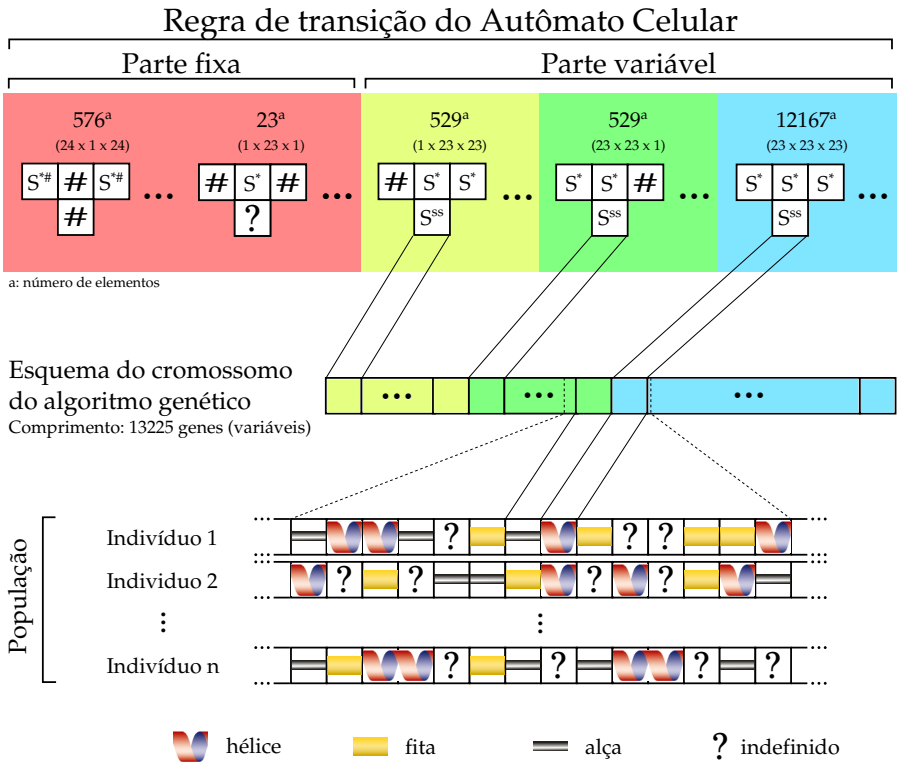

Figura 3: Representação da utilização de um algoritmo genético na busca por uma regra de transição capaz de predizer estruturas secundárias proteicas. Os cromossomos possuem 13225 genes, ou variáveis, que podem assumir quatro estados distintos. Cada indivíduo apresenta uma combinação desses estados. Essa combinação corresponde a uma regra de transição de um autômato celular que irá evoluir por um número definido de passos. A comparação do resultado dessa evolução com a estrutura secundária atribuída ("real") fornece uma medida do fitness da regra. Utilizando-se um algoritmo genético competente esperasse que ocorra uma convergência para a regra com maior fitness, ou seja, a que melhor prediz a estrutura secundária da proteína.

6 APRECIAÇÃO DO ORIENTADOR

O orientador deverá emitir parecer sobre o desempenho acadêmico e os resultados de pesquisa apresentados pelo bolsista.