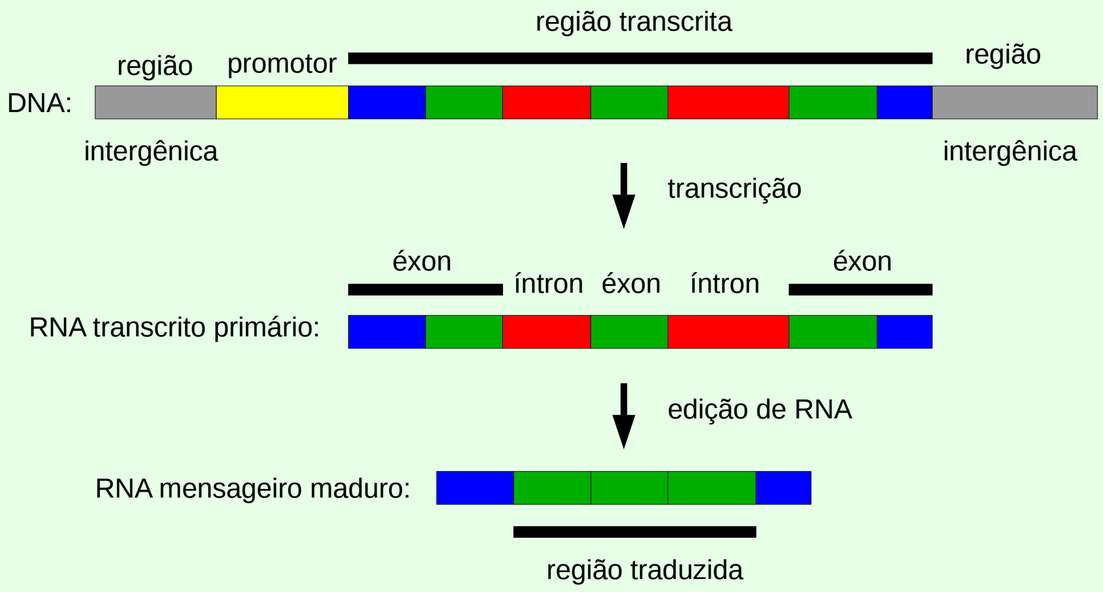
Desenvolvimento de um configurador gráfico para o ToPS

Douglas Vasconcelos Cancherini (aluno), Alan Mitchell Durham (supervisor)
Departamento de Ciência da Computação, IME, USP

Introdução

1. Estrutura dos genes eucarióticos



2. Predição de genes *ab initio*

- Após treinamento do sistema por genes conhecidos, a predição é feita tendo como entrada apenas uma sequência genômica, na qual os genes e seus vários elementos serão reconhecidos
- Boa parte dos preditores gênicos modernos tem uma arquitetura baseada em um modelo oculto de Markov generalizado (GHMM), com um submodelo probabilístico para cada elemento do gene. O comprimento da sequência correspondente ao elemento é dado por um modelo de duração.

3. ToPS

- ToPS é o *Toolkit for Probabilistic Models of Sequence*, previamente desenvolvido pelo grupo do Prof. Alan Durham
- Implementação eficiente de algoritmos para treinamento e decodificação de modelos probabilísticos
- Flexibilidade na combinação de modelos e algoritmos, bem como no número de parâmetros configuráveis
- Está na base do MYOP, um preditor gênico de elevada acurácia também desenvolvido pelo grupo

Motivação

- Os arquivos de configuração dos modelos probabilísticos do ToPS, especialmente os usados no MYOP, são numerosos e guardam relações complexas entre si, dificultando bastante que seja aproveitada na prática a flexibilidade do ToPS

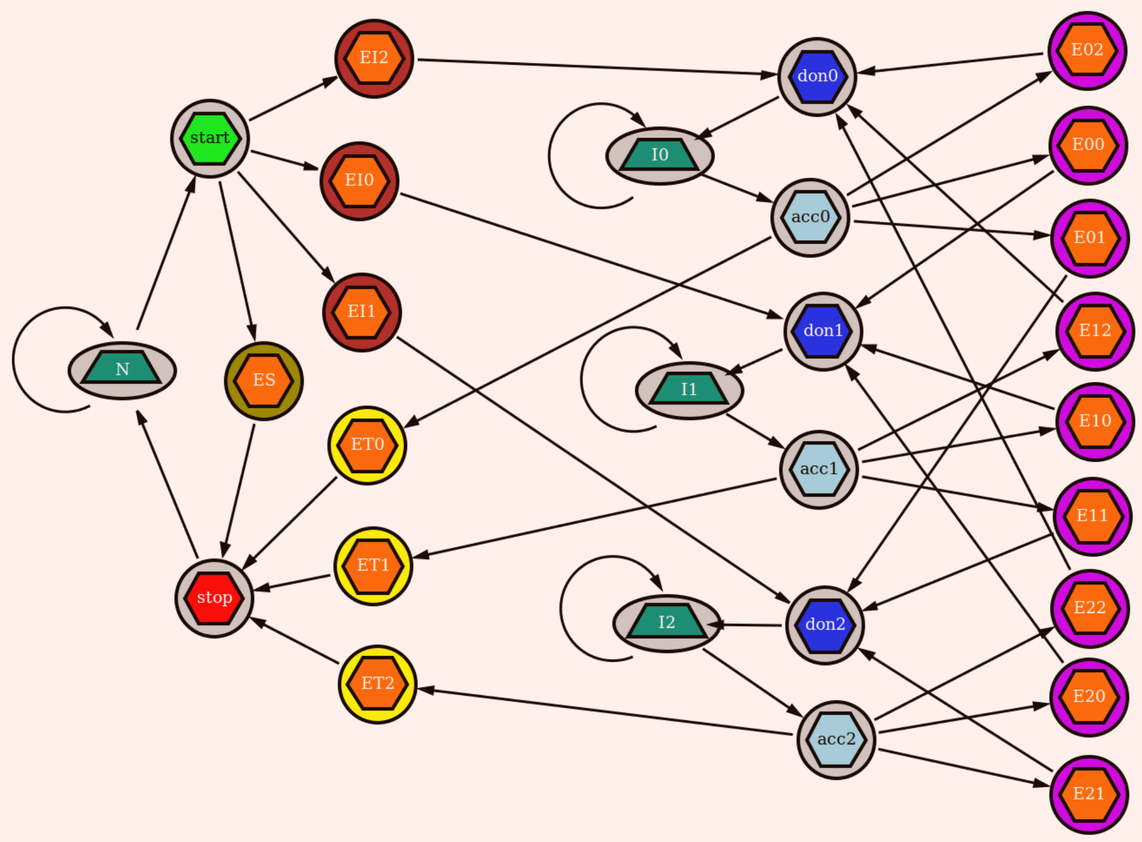
Objetivos

Nossa intenção foi permitir, a usuários do ToPS e do MYOP em uso de um modelo baseado em GHMM:

- compreender melhor a arquitetura do GHMM, bem como obter visualizações convenientes de tal arquitetura e ter a possibilidade de modificá-la
- poder com facilidade editar os parâmetros dos submodelos do GHMM e/ou os parâmetros para seu treinamento, bem como os parâmetros dos modelos de duração correspondentes

Implementação

- Cliente em HTML5 e JavaScript para navegadores, com uso de elementos SVG
- Interface intuitiva, com submodelos probabilísticos do GHMM como vértices de um grafo e arestas dirigidas representando a probabilidade de transição entre eles
- Cada classe de submodelo estatístico tem seus vértices com uma certa forma geométrica
- Facilidade em mostrar o compartilhamento de parâmetros de configuração entre submodelos por meio de cores iguais de preenchimento
- Elipse colorida ao redor de cada submodelo representa o modelo de duração correspondente
- Pequeno volume de dados permite seu armazenamento no próprio DOM da página, com funções acionadas por eventos na página editando diretamente o DOM
- Restrições de segurança nos navegadores exigem que treinamento de modelos e predições gênicas por meio do ToPS sejam feitas por meio de um servidor (futuro)
- Funcionalidades futuras a serem disponibilizadas por meio do servidor: treinamento, predição e simulação
- Funcionalidade adicional da interface (em implementação): escolha dos parâmetros de treinamento



Representação, na interface do sistema que implementamos, do GHMM para um preditor gênico simplificado. Hexágonos e trapézios representam, respectivamente, cadeias de Markov inomogêneas e cadeias de Markov de alcance variável. Polígonos e elipses de mesma coloração indicam compartilhamento de, respectivamente, submodelos de sequência e modelos de duração. É o caso, por exemplo, dos submodelos para éxons (ES, EIk, ETk, Ejk, j, k = 0, 1 ou 2, em laranja), íntrons (Ik, em verde escuro), sítios doadores (donk, em azul escuro) e aceptores de íntrons (acck, em azul claro), ou dos modelos de duração para éxons internos (Ejk, em magenta), iniciais (Eik, em bordô) ou terminais (Etk, em amarelo). N = região não codificante, start (stop) = sítio de início (parada) da tradução, ES = éxon único.

Referências bibliográficas

- AY Kashiwabara. MYOP/ToPS/SGEval: um ambiente computacional para estudo sistemático de predição de genes. Tese de Doutorado, Departamento de Ciência da Computação, Universidade de São Paulo, Brasil, 2012
- AY Kashiwabara, I Bonadio, V Onuchic, F Amado, R Mathias e AM Durham. Tops: a framework to manipulate probabilistic models of sequence data. PLoS Comput Biol , 9(10):e1003234, 2013
- WH Majoros. Methods for computational gene prediction. Cambridge University Press, 2007