Implementação de Modelos de Covariância no arcabouço ToPS

Renato Cordeiro Ferreira

Supervisor: Alan Mitchell Durham

MAC0499 - Trabalho de Conclusão de Curso

Resumo

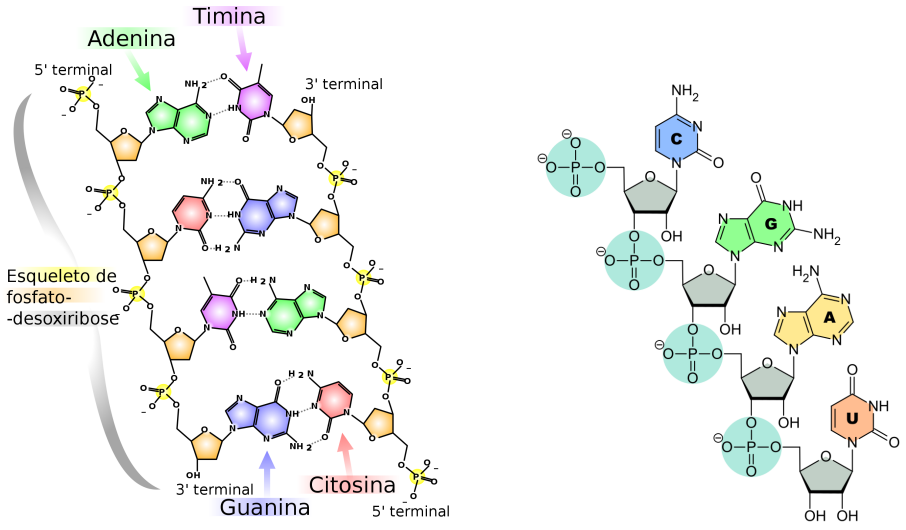
Encontrar a função de RNAs não codificantes (ncRNAs) tornou-se importante para a Biologia Molecular moderna, principalmente após a descoberta dos seus papéis catalíticos e estruturais dentro das células. Neste projeto, pretendemos implementar os Modelos de Covariância, que permitem definir se um ncRNA pertence a uma família cuja função seja conhecida. Adicionaremos o modelo no arcabouço ToPS, que já possui vários algoritmos usados para a descrição do DNA e RNA a partir de suas sequências de símbolos. Para começar esse processo, realizaremos uma refatoração do sistema, padronizando o estilo de código e eliminando inconsistências. Esperamos, assim, facilitar o uso do arcabouço para novos desenvolvedores, e utilizá-lo para criar uma ferramenta de análise e comparação de ncRNAs.

1 Introdução

O RNA (Ácido Ribonucleico) é uma molécula formada por nucleotídeos, que são constituídos de um açúcar ribose, um grupo fostato e um de quatro tipos de bases nitrogenadas: Adenina (A), Uracila (U), Citosina (C) e Guanina (G). Esta composição é parecida com a do DNA (Ácido Desoxirribonucleico), que apresenta no lugar da ribose e da Uracila, o açúcar desoxirribose e a base Timina (T).

Os açúcares pentâmeros ribose e desoxirribose conectam-se entre si com o auxílio dos fosfatos, no sentido do carbono 5' de um para o 3' do seguinte. Os nucleotídeos interagem entre si, ainda, por meio de pontes de hidrogênio, formando pareamentos A-U (com 2 pontes) e C-G (com 3 pontes). No DNA, A e T formam par no lugar de A-U, também com 2 pontes. O conjunto de nucleotídeos que compõe o DNA e o RNA é chamado de **estrutura primária**, e pode ser vista na Figura 1.

Usualmente,o DNA é encontrado como uma molécula de duas fitas, conectadas entre si pelo pareamento de nucleotídeos. O RNA, por sua vez, forma sequências de fita única, que dobram-se sobre si mesmas pelas forças das pontes de hidrogênio. Pareamentos consecutivos no RNA



(a) DNA (b) RNA

Figura 1: ***Estrutura molecular do DNA e RNA:*** *À esquerda, é possível ver as pontes de hidrogênio entre os pares A-T e C-G. Na maioria dos organismos, o DNA aparece como uma molécula de fita dupla com estrutura helicoidal. O RNA, por sua vez, dobra-se sobre si mesmo, pareando A-U e C-G (Figuras de Ball e Lijealso (2013) e Sponk (2010))*

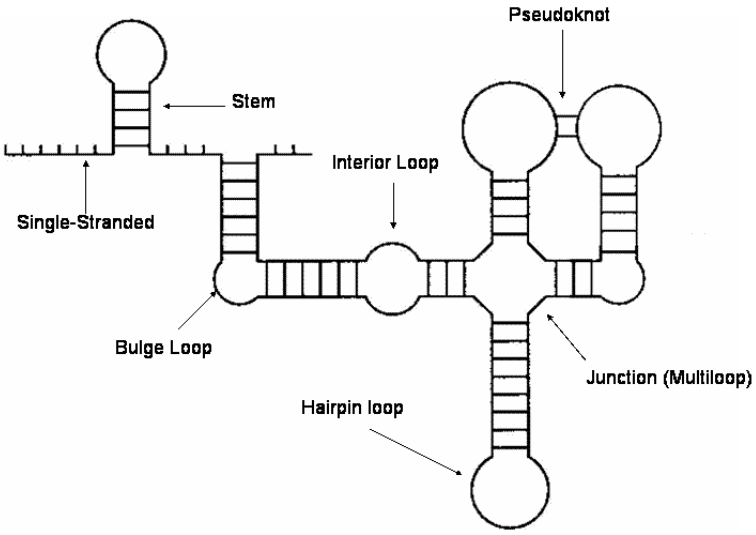


Figura 2: *Representação esquemática da **estrutura secundária** do RNA: fita única (single-stranded), hastes (stem), bojos (bulge loop), laço interior (interior loop), grampos (hairpin loop), junções (junction) e pseudo-nós (pseudo-knots). (Figura de ide , (n.d))*

são chamados de **hastes**, que ao serem interrompidas por nucleotídeos sem par formam **bojos** e **laços**. Laços ligados a somente uma haste são chamados de **grampos**, e interações entre dois grampos ou laços distintos são conhecidas como **pseudo-nós**. Duas ou mais hastes conectadas formam **junções**, que com todos os outros pareamentos compõem a **estrutura secundária** do RNA. As formações citadas estão exemplificadas na Figura 2.

1.1 RNAs não codificantes

De forma geral, os RNAs são lembrados por seu papel fundamental na produção de proteínas - moléculas que exercem funções estruturais, cataliticas (enzimas) e de proteção (anticorpos) nas células. Segundo o Dogma Central da Biologia, o DNA é transcrito em RNA mensageiro (mRNA), cujas trincas de nucleotídeos (códons) são posteriormente traduzidas para os aminoácidos que compõem as proteínas (Alberts *et al.* , 2002). Por servir de repositório intermediário da informação que codifica proteínas, o mRNA é conhecido como **RNA codificante**, .

Entretanto, existem outras variedades de RNA que não realizam esta tarefa: os **RNAs não-codificantes** (ncRNAs), que podem assumir estruturas tridimensionais complexas e exercer funções reguladoras e estruturais dentro das células (Durbin *et al.* , 1998). Um exemplo

importante de ncRNA são os RNAs ribossômicos (rRNAs), que fazem parte da estrutura celular (ribossomos) responsável pela tradução dos RNAs mensageiros.

Devido à sua participação no funcionamento das células, os ncRNAs são tópico de muitas pesquisas no campo da Biologia Molecular moderna. Entretanto, analisar experimentalmente a estrutura de um RNA para descobrir sua função é muito custoso. É desejável, portanto, obter formas de automatizar esse processo, o que envolve resolver três diferentes problemas relacionados às moléculas de RNA:

- **Inferir estruturas secundárias**

  Identificação de estruturas conhecidas, cujos formatos evidenciam a forma tridimensional e função das moléculas.

- **Realizar alinhamentos múltiplos**

  Geração de perfis com posições e pareamentos conservados numa família de RNAs.

- **Fazer buscas por similaridade**

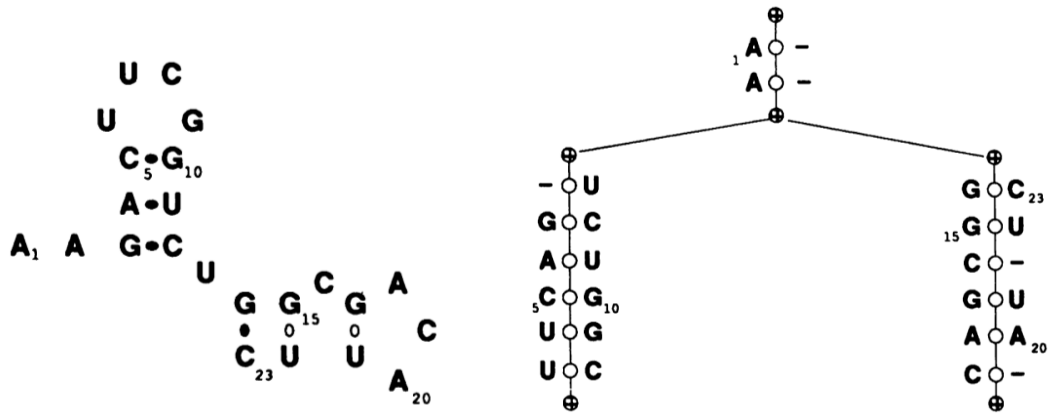
  Identificação da proximidade de uma molécula com familias de RNAs, mostrando as diferenças presentes nas moléculas de organismos distintos.

Na próxima seção, apresentaremos o **Modelo de Covariância**, que provê algoritmos que utilizam apenas as sequências de RNA para realizar essas tarefas.

1.2 Modelos de Covariância

A maioria dos ncRNAs é encontrada em **famílias**: conjunto de moléculas com provável ancestral comum, cujas diferenças são originadas por mutações ocorridas ao longo da evolução das espécies. Em geral, estas mudanças são selecionadas de modo a preservar a estrutura secundária das moléculas, que determina o formato tridimensional (e, com isso, a função) dos ncRNAs. Além de substituições de nucleotídeos simples, são comuns comuns eventos de **covariância**, nos quais há troca de um pareamento de hastes e grampos.

Uma das formas de analisar o DNA, o RNA e as proteínas é utilizar modelos probabilísticos que descrevam suas sequências de símbolos. Para o primeiro e terceiro caso, uma análise



(a) Exemplo de RNA (b) Representação em árvore

Figura 3: ***Representação da estrutura secundária do RNA:*** *A árvore binária contém nós extras para demarcar o início, fim e bifurcação, além de nós para representar nucleotídeos simples e pareamentos presentes na sequência. (Figuras de Eddy e Durbin , 1994)*

linear das sequências costuma ser suficiente para obter informações significativas sobre os dados estudados (Durbin *et al.* , 1998). Modelos baseados em Cadeias Ocultas de Markov (HMMs) permitem criar programas preditores de genes (Kashiwabara , 2011) e ferramentas que definem se uma proteína pertence a uma dada família. Entretanto, as variantes tradicionais dos HMMs não são suficientes para descrever as dependências de longa distância presentes na estrutura secundária dos ncRNAs.

Para representar um RNA, é possível utilizar uma árvore binária ordenada, que é capaz de descrever os nucleotídeos e os pareamentos presentes na estrutura secundária da sequência (conforme ilustrado na Figura 3). Esta modelagem, porém, não permite representar interações tridimensionais entre as bases, como pareamentos triplos e pseudo-nós (Eddy e Durbin , 1994). Também é limitada a apenas uma sequência, o que a impede de modelar variações ocorridas entre diferentes membros de uma família.

O **Modelo de Covariância** (CM, Eddy e Durbin (1994)) expande os nós da árvore que descreve as sequências. Cada nó representa uma sequência ou pareamento consenso (conservados). Dentro do CM, os nós tornam-se conjuntos de estados, que permitem inserções e remoções de nucleotídeos. Arcos dentro e dentre os nós descrevem as variações válidas sobre o perfil que gerou a árvore. O conjunto de possíveis nós, estados e arcos está ilustrado na Figura 4.

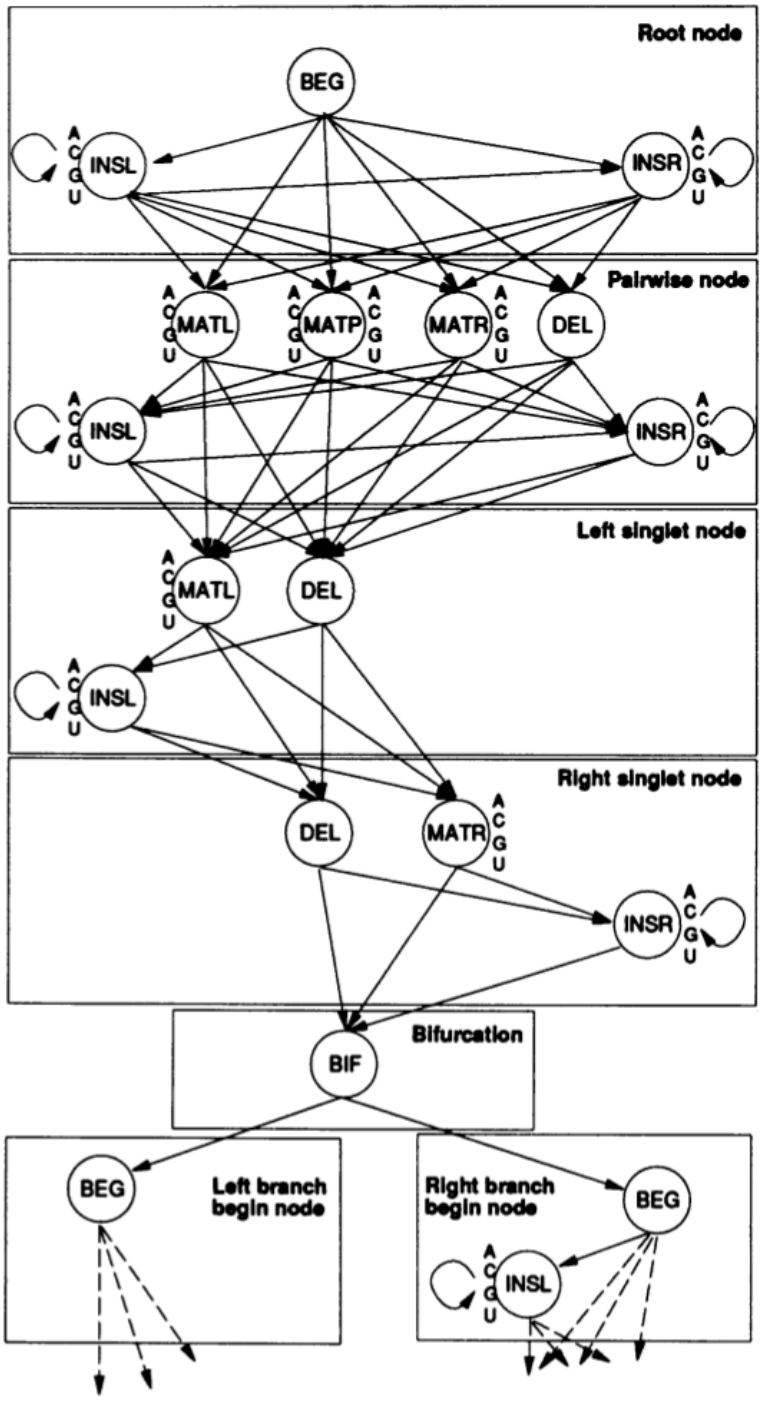


Figura 4: ***Modelo de Covariância:*** *Cada conjunto de estados é equivalente à um tipo de nó na árvore binária rígida (Figura 3): raiz, pareamento, nucleotídeo único à esquerda, nucleotídeo único à direita, bifurcação e ramos da bifurcação.* `INS` *representa inserções à esquerda (*`INSL`*) e à direita (*`INSR`*),* `DEL` *representa remoções e* `MAT` *representa correspondências com as sequências consenso à esquerda (*`MATL`*), à direita (*`MATR`*) e no par (*`MATP`*). Os estados* `BEG`*,* `END` *e* `BIF` *são artificiais, para completar o diagrama de estados e simplificar sua estrutura. Arcos representam transições válidas entre estados. (Figura de Eddy e Durbin , 1994)*

O Modelo de Covariância pode ser visto como um gerador de sequências (Eddy e Durbin , 1994). Estados de correspondência (`MAT`) e de inserção (`INS`) têm associados a si uma probabilidade de emissão de nucleotídeos ou pareamentos. Estados de remoção (`DEL`) não emitem símbolos, mas representam a perda de um nucleotídeo em uma posição consenso. Outros estados sem emissão (`BEG`, `END` e `BIF`) modelam a estrutura da árvore. Os arcos representam probabilidades de transição entre cada estado. O conjunto de todas as probabilidades cria um modelo probabilístico, que permite gerar qualquer sequência de RNA com uma probabilidade associada. Paralelamente, qualquer sequência existente pode ter uma probabilidade associada a si, dado um CM previamente conhecido.

1.3 Algoritmos

Dada uma sequência $x = x_1..x_L$ e um Modelo de Covariância com M estados $W_1..W_M$ (W_1 estado inicial), é possível definir uma série de algoritmos de programação dinâmica que associam a uma subsequência $x_i..x_j$ de x as seguintes probabilidades (Durbin *et al.* , 1998):

- ***Inside*:** $\alpha(i, j, k)$, a probabilidade da subsequência $x_i..x_j$ ser gerada por uma árvore com raiz no estado W_k.
- ***Outside*:** $\beta(i, j, k)$, a probabilidade de gerar a sequência x sem ter utilizado (ainda) a subsequência $x_i..x_j$, que será gerada por uma árvore com raiz no estado W_k.
- **CYK (Cocke–Younger–Kasami):** $\gamma(i, j, k)$, a probabilidade da melhor árvore que gera a subsequência $x_i..x_j$ com raiz em W_k .

Por meio destas probabilidades, é possível inferir a árvore base de um Modelo de Covariância, utilizando apenas um conjunto de sequências de RNA (Durbin *et al.* , 1998). Esse processo simula a análise comparativa de sequências, utilizada para criar manualmente perfis de famílias de RNA. O processo começa com um alinhamento aleatório entre as sequências, a partir do qual usa-se alguma heurística que gere a árvore base, então expandida para um Modelo de Covariância. Uma destas heurísticas é descrita em Eddy e Durbin (1994). Com a primeira versão do Modelo de Covariância, o processo repete dois passos básicos:

(a) Construir um alinhamento múltiplo dado o Modelo de Covariância atual.

(b) Construir um Modelo de Covariância ótimo ou subótimo dado o alinhamento atual;

Para o item (a), é possível aplicar o algoritmo CYK sobre cada sequência, uma vez que o alinhamento múltiplo é equivalente a encontrar o melhor alinhamento de cada membro sobre o perfil do conjunto (Durbin *et al.* , 1998). Para o item (b), aplica-se o algoritmo de maximização da esperança **inside-outside**, que utiliza as probabilidades $\alpha(i, j, k)$ e $\beta(i, j, k)$ para reestimar as probabilidades de emissão e transição. Os passos são repetidos até que as diferenças na estrutura da árvore e nas probabilidades não sejam significativas. O processo completo está ilustrado na Figura 5:

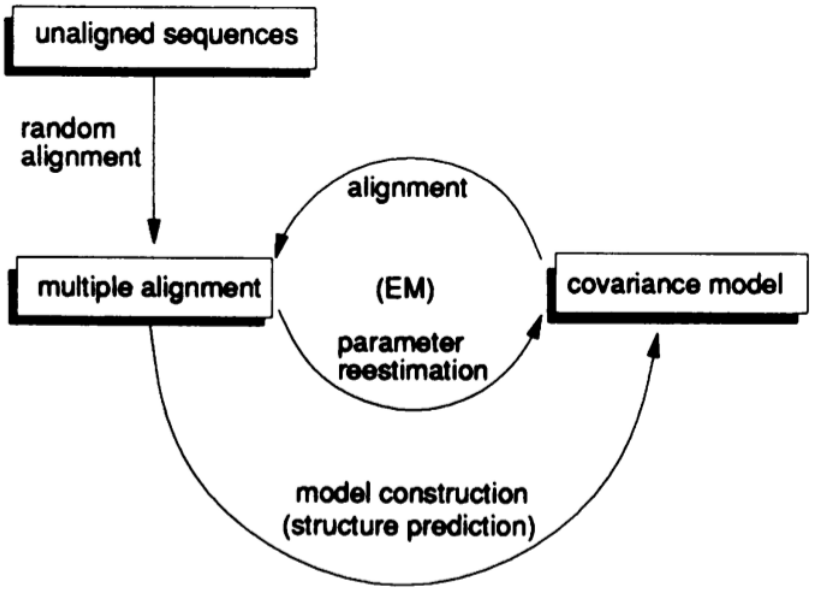


Figura 5: ***Análise Comparativa de Sequências Automatizada:*** *Utilizando um processo iterativo de dois passos, é possível gerar um Modelo de Covariância que represente um conjunto de sequências, bem como estimar as probabilidades de emissão e transição do modelo. Para obter a primeira versão do CM, utiliza-se uma heurística (Figura de Eddy e Durbin , 1994)*

Com a análise comparativa, podemos inferir a estrutura secundária e alinhar múltiplas sequências de RNA. Com o algoritmo CYK, podemos fazer buscas em bancos de dados de CMs, verificando se uma molécula de RNA pertence a uma dada família. Em conjunto, os 4 principais algoritmos de Modelos de Covariância permitem resolver os problemas de análise automatizada de RNAs indicados na Subseção 1.1.

1.4 ToPS

Para implementar os algoritmos de Modelos de Covariância (apresentados na Subseção 1.3), utilizaremos o arcabouço de modelos probabilísticos ToPS (Kashiwabara *et al.* , 2013), desenvolvido pelo grupo de pesquisa do Professor Alan Durham. Em sua publicação, o arcabouço disponibiliza 8 modelos utilizados em predição de genes e estudo de proteínas. A principal hierarquia do sistema associa os modelos por meio de um padrão *composite* (Gamma *et al.* , 1994), conforme ilustrado na Figura 6:

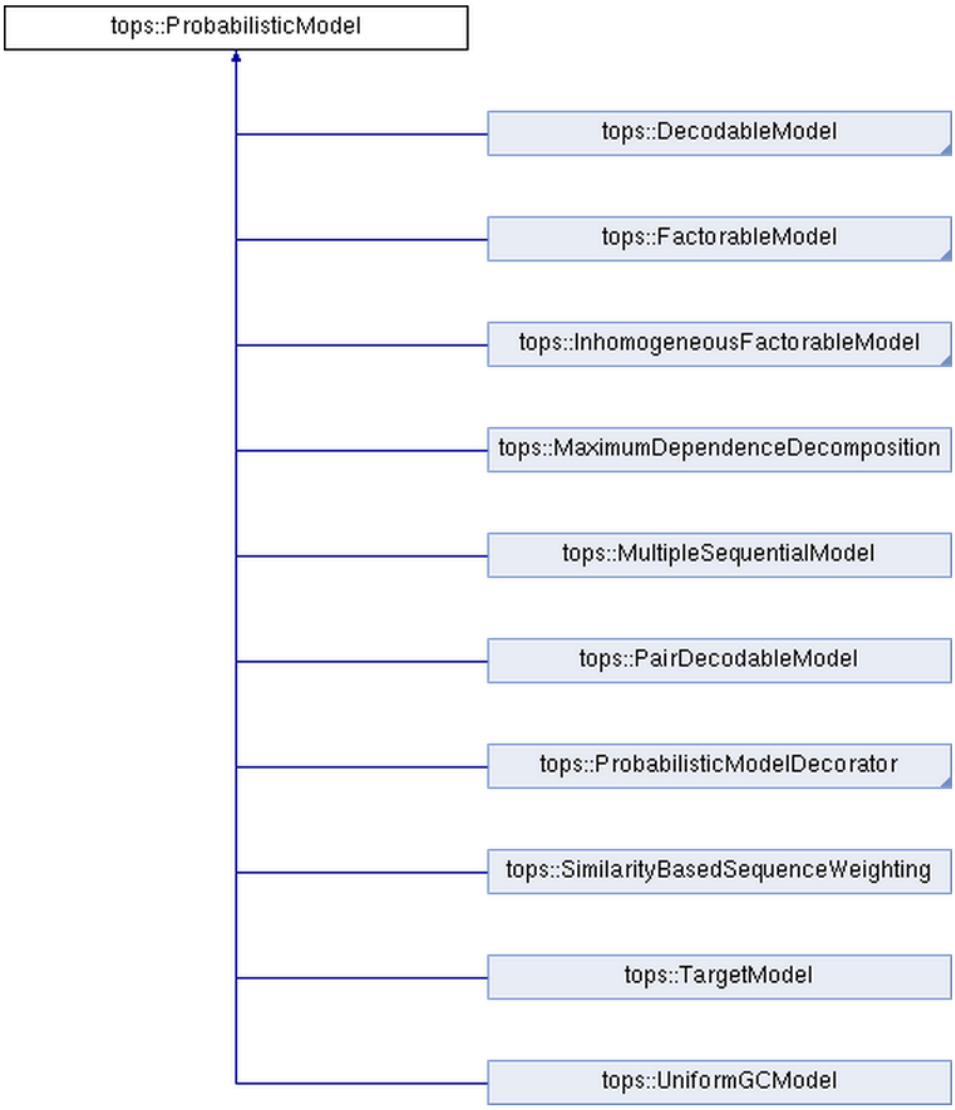


Figura 6: ***Hierarquia principal do ToPS:*** *Todos os modelos são filhos da classe ProbabilisticModel, que define a interface geral para utilização de um modelo probabilístico. Os modelos integradores (como o GHMM) usam esta interface para incluir modelos dentro de seus estados (Kashiwabara , 2011)*

Por conta desta estrutura, o ToPS apresenta uma série de vantagens que podem simplificar a implementação dos algoritmos, tais como funções básicas de probabilidade já implementadas, programas aplicativos (construídos sobre o arcabouço) para inferência de probabilidades e uma linguagem pré-definida para especificação dos modelos.

Contudo, diversas modificações e extensões feitas no sistema abrem campo para aperfeiçoamentos, que permitiriam simplificar a hierarquia de classes e melhorar a sua utilização. Para explorar esse potencial e facilitar ainda mais a integração do Modelo de Covariância ao ToPS, começaremos este trabalho fazendo uma refatoração (Fowler *et al.* , 1999) do arcabouço.

2 Objetivos

Neste trabalho, temos por objetivo implementar os algoritmos do Modelo de Covariância, de modo a permitir estudos sobre ncRNAs. Para tanto, estenderemos o arcabouço ToPS (Kashiwabara *et al.* , 2013), adicionando um novo modelo à sua hierarquia de classes.

Para iniciar o projeto, realizaremos uma refatoração do arcabouço, portando-o para um novo repositório, padronizando o estilo de código e modificando a hierarquia de classes, de modo a simplificar a inserção de novos modelos. Esta fase facilitará o entendimento da biblioteca do ToPS, com especial atenção à sua estrutura.

Para a implementação, reutilizaremos parte do código do Modelo Oculto de Markov Sensível ao Contexto (CSHMM, Agarwal *et al.* (2010)), que está sendo desenvolvido por Rafael Mathias em seu mestrado. Os CSHMMs também podem ser aplicados na análise da estrutura secundária de RNAs, e por esse motivo apresentam similaridades (em termos de algoritmos) com os Modelos de Covariância.

Por fim, uma vez que a implementação seja bem sucedida, coletaremos estatísticas de sensibilidade e precisão dos algoritmos, aplicando-os sobre sequências cujo modelo de covariância seja conhecido. Estudaremos as ferramentas já existentes, comparando o desempenho delas com a implementação do ToPS.

3 Plano de trabalho

Dividimos a realização deste trabalho em 6 etapas:

1. Criar novo repositório para o ToPS, portando o código para uma nova biblioteca.
2. Realizar refatorações na hierarquia de classes. principal do arcabouço, facilitando a reutilização de código e a inserção de novos modelos.
3. Implementar algoritmos de modelo de covariância (*inside*, *outside*, CYK, *inside-outside* e análise comparativa automatizada)
4. Realizar testes com outras ferramentas que utilizam modelos de covariância.
5. Escrever monografia
6. Montar apresentação e pôster

	mar	abr	mai	jun	jul	ago	set	out	nov
1	X	X							
2	X	X	X	X					
3					X	X			
4						X	X	X	X
5						X	X	X	X
6								X	X

Tabela 1: *Cronograma do plano de trabalho*

Referências

Agarwal *et al.* **(2010)** Sumeet Agarwal, Candida Vaz, Alok Bhattacharya e Ashwin Srinivasan. Prediction of novel precursor mirnas using a context-sensitive hidden markov model (cshmm). *BMC bioinformatics*, 11(Suppl 1):S29. Citado na pág. 9

Alberts *et al.* **(2002)** Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts e Peter Walter. *Molecular Biology of the Cell.* Garland Science, 4th ed. Citado na pág. 2

Ball e Lijealso (2013) Madeleine Price Ball e Lijealso. Dna chemical structure.svg, 2013. URL http://pt.wikipedia.org/wiki/Ficheiro:DNA_chemical_structure_pt.svg. Citado na pág. 1

Durbin *et al.* **(1998)** Richard Durbin, Sean R. Eddy, Anders Krogh e Graeme Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1st ed. Citado na pág. 2, 4, 6, 7

Eddy e Durbin (1994) Sean R Eddy e Richard Durbin. Rna sequence analysis using covariance models. *Nucleic acids research*, 22(11):2079–2088. Citado na pág. 4, 5, 6, 7

Fowler *et al.* **(1999)** Martin Fowler, Kent Beck, J Brant, William Opdyke e Don Roberts. *Refactoring: Improving the design of existing programs*. Addison-Wesley Reading. Citado na pág. 9

Gamma *et al.* **(1994)** Erich Gamma, Richard Helm, Ralph Johnson e John Vlissides. *Design patterns: elements of reusable object-oriented software*. Pearson Education. Citado na pág. 8

ide () ide. Rna_sec_struct2.gif. URL http://kelder.zeus.ugent.be/. Citado na pág. 2

Kashiwabara *et al.* **(2013)** André Yoshiaki Kashiwabara, Ígor Bonadio, Vitor Onuchic, Felipe Amado, Rafael Mathias e Alan Mitchell Durham. Tops: A framework to manipulate probabilistic models of sequence data. *PLoS computational biology*, 9(10):e1003234. Citado na pág. 8, 9

Kashiwabara (2011) André Y. Kashiwabara. *MYOP/ToPS/SGEval: Um ambiente computacional para estudo sistemático de predição de genes*. Tese de Doutorado, Instituto de Matemática e Estatística, Universidade de São Paulo, Brasil. Citado na pág. 4, 8

Sponk (2010) Sponk. Rna-nucleobases.svg, 2010. URL http://commons.wikimedia.org/wiki/File:RNA-Nucleobases.svg. Citado na pág. 1