

# Projeto de Pesquisa - Pós-doutorado

## Bioinformática aplicada ao estudo do RNA Estruturoma de *Halobacterium* *salinarum*

Candidato: Dra. Eliane Z. Traldi

Supervisor: Prof. Dr. Ricardo Z. N. Vêncio

Departamento de Computação e Matemática

FFCLRP-USP

Projeto de Pesquisa apresentado à FAPESP para candidatura a bolsa de pós-doutorado vinculado ao Projeto Jovem Pesquisador 2009/09532-0 “Biologia Sistêmica do extremófilo *Halobacterium salinarum*: contribuição dos RNAs não-codificantes ao modelo global de regulação gênica”

Ribeirão Preto, Junho de 2012

# Bioinformática aplicada ao estudo do RNA Estruturoma de *Halobacterium salinarum*

Candidato: Dra. Eliane Z. Traldi

Supervisor: Prof. Dr. Ricardo Z. N. Vêncio  
Departamento de Computação e Matemática  
FFCLRP-USP

## RESUMO

*Estrutura implica função.* No universo das biomoléculas, esse é um paradigma amplamente aceito. Porém, quando o foco é o transcrito, a configuração estrutural é uma camada de informação que ainda é pouco considerada. A partir do reconhecimento das diferentes funções desempenhadas pelos RNAs, é fundamental que sua estrutura seja desvendada para que seja possível a compreensão de como essas biomoléculas atuam em diversos processos cruciais para a existência e manutenção dos sistemas biológicos, como armazenamento, transporte e codificação de informação genética, além de atividades de catálise e regulação. No presente Projeto de Pesquisa de Pós-doutorado, propõe-se realizar a modelagem *in silico* em larga escala da estrutura de todos os RNAs (*RNA Estruturoma*), utilizando como organismo de estudo o extremófilo *Halobacterium salinarum*. A partir da modelagem, será possível classificar essas moléculas em grupos estruturais, além de investigar a correlação entre esses agrupamentos e informações sobre função e regulação do transcrito.

# Bioinformatics applied to *Halobacterium salinarum* RNA Structurome investigation

Candidate: Dra. Eliane Z. Traldi

Supervisor: Prof. Dr. Ricardo Z. N. Vêncio

Departamento de Computação e Matemática

FFCLRP-USP

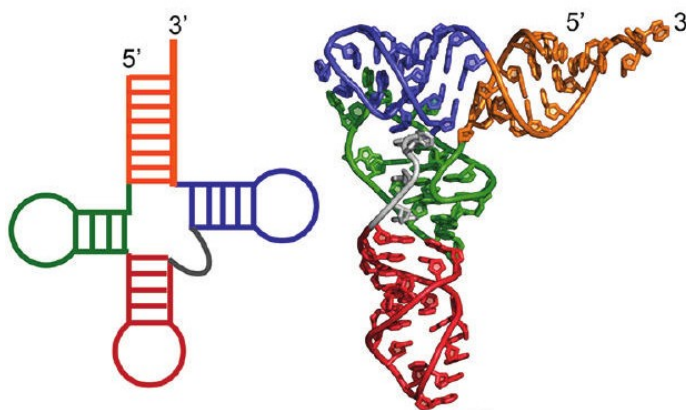
## **ABSTRACT**

*Structure implies function.* In the universe of biomolecules, this is a highly accepted paradigm. However, when the focus is the transcriptome, the structural configuration is a layer of information that is still poorly used. From the recognition of the different functions performed by the RNAs, it is fundamental that their structure is unveiled so that it is possible to understand how these biomolecules act in a variety of processes that are crucial to the existence and maintenance of biological systems, such as storage, transport and codification of genetic information, in addition to catalysis and regulation activities. In this Research Project, large scale modeling *in silico* of the structure of all RNA (RNA structurome) is proposed, using the extremophyle *Halobacterium salinarum* as the study organism. From this modeling it will be possible to classify the molecules in structural groups, in addition to investigate a correlation between these groups and information about function and regulation of the transcriptome.

## 1 – INTRODUÇÃO

No universo das biomoléculas, existe uma relação profunda entre estrutura e função. Essa relação pode ser tão forte a ponto de poder-se afirmar, em geral, que *estrutura implica função*. Este paradigma é bem estabelecido quando se consideram as proteínas, mas ainda não tão difundido quando o foco é o transcrito.

O transcrito de um organismo media o fluxo de informação gênica de diferentes formas, que vão além da simples produção de proteínas. Os RNAs são moléculas extremamente versáteis e possuem propriedades únicas que os permitem armazenar, transportar e codificar informação genética, além de desempenhar diversos tipos de atividades catalíticas e regulatórias. Essas moléculas realizam pareamentos e dobramentos internos, formando estruturas secundárias e terciárias (Figura 1), de modo que a estrutura é uma das características fundamentais para que possam executar corretamente suas mais diversas funções. Dessa forma, a configuração estrutural das moléculas que constituem o transcrito é também essencial, do ponto de vista biomolecular, para a manutenção da vida.



**Figura 1** – Estrutura bidimensional e tridimensional de um RNA transportador. (modificado de Butcher & Pyle, 2011).

Essas biomoléculas são divididas em duas grandes classes funcionais: codificantes, que engloba os RNAs mensageiros (mRNA), e não-codificantes (ncRNAs).

Considerando-se a abundância geral numa célula típica, a maioria esmagadora dos RNAs são ncRNAs. Essas moléculas atuam em diversos processos, como por exemplo a regulação de expressão gênica, o *imprinting* gênico, o *splicing* alternativo, a modificação de outros ncRNAs, entre outros mecanismos intimamente dependentes da sua estrutura para ocorrer (Gorodkin & Hofacker, 2011). Evidentemente, duas subclasses muito conhecidas e importantes de ncRNAs são os RNAs ribossômicos (rRNAs) e transportadores (tRNAs), para os quais a conformação estrutural também é sabidamente chave para o correto funcionamento.

Já os mRNAs, inicialmente vistos como meros intermediários no processo de expressão gênica, tem sido relacionados também a funções estruturais (Kloc, 2008; Kloc, 2009). Recentemente, foi demonstrado que alguns mRNAs poderiam apresentar funções duplas, participando não só da codificação da informação genética mas também auxiliando na manutenção da estrutura citoplasmática e na organização de organelas sub-nucleares (Kloc & Reddy, 2011).

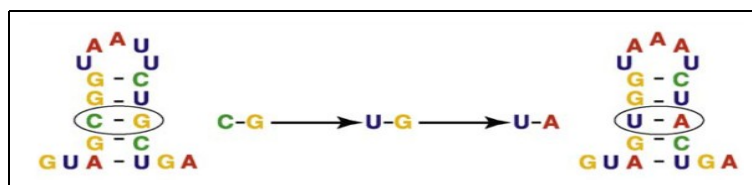
Um exemplo para demonstrar como a própria estrutura de moléculas de mRNA também pode influenciar de forma direta na expressão gênica é a existência de *riboswitches* (Tucker & Breaker, 2005), elementos estruturados localizados geralmente na região 5'UTR de mRNAs, capazes de detectar e responder, através de mudanças conformacionais, a estímulos celulares, como a presença de ligantes específicos.

Dessa forma, foram desenvolvidos diversos métodos experimentais, desde aqueles clássicos de difração de raios-X até os que utilizam agentes químicos ou nucleases, para a determinação detalhada de estrutura molecular tridimensional ou mesmo para uma modelagem mais simplificada com a elucidação dos pareamentos internos em modelos bidimensionais, respectivamente. A partir da determinação da estrutura, os RNAs podem ser classificados com base em características estruturais, que, assim como ocorre no caso das proteínas (Andreeva & Murzin, 2010), poderão ser relacionadas para inferir função, explicar mecanismos de regulação transcricional e pós-transcricional, estudar a dinâmica das estruturas ao longo do tempo ou em resposta à mudança de condições de temperatura, salinidade e na presença de diferentes ligantes (Kertesz *et al.*, 2010).

As abordagens clássicas de cristalografia resolveram cerca de 900 estruturas de RNAs e cerca de 1150 complexos RNA-Proteína (*Protein Data Bank*: <http://www.rcsb.org/pdb/home/home.do>; Rose *et al.*, 2011). Entretanto, apesar de ser a informação mais detalhada possível de ser obtida, a estrutura 3D é custosa e trabalhosa de ser determinada.

Informações estruturais menos detalhadas, como modelos bidimensionais, também podem ser muito úteis para se obter uma visão geral da estrutura das moléculas em questão, além de servirem como base para a modelagem tridimensional, podendo ser obtidas por métodos que utilizam agentes químicos ou nucleases. Entre estes, destacam-se *SHAPE* (*selective 2-hydroxyl acylation analysed by primer extension* - Merino *et al.*, 2005), *SHAMS* (*selective 29-hydroxyl acylation analyzed by mass spectrometry* - Turner *et al.*, 2009), DMS (dimetilsulfato) acoplado a algum método de determinação dos pontos de metilação, como *primer extension*, que permite determinar a estrutura dos RNAs *in vivo* (Wells *et al.*, 2000), entre outros. Entretanto, em sua grande maioria, essas metodologias são focadas em elucidar a estrutura de uma única ou poucas moléculas de RNA por experimento, além de possuírem limitações técnicas relacionadas ao tamanho das moléculas em questão (Underwood *et al.*, 2010).

Assim, se faz necessário o desenvolvimento e aplicação de algoritmos e técnicas computacionais que permitam criar modelos e realizar previsões estruturais *in silico*. Existe uma série de abordagens computacionais para o estudo da estrutura de RNAs. Uma das mais empregadas envolve a comparação de sequências homólogas e a criação de modelos de covariância (Eddy & Durbin, 1994), assumindo-se que essas sequências, apesar de possuírem diferenças, resultam em moléculas com estrutura conservada (Figura 2). Há também metodologias *ab initio* para a determinação da estrutura de moléculas individualmente, com base em propriedades termodinâmicas, em que o objetivo é identificar as estruturas de mínima energia, termodinamicamente mais estáveis (Mathews & Turner, 2006). Métodos híbridos, que combinam os dois tipos de abordagem descritos acima, também foram desenvolvidos (Hofacker *et al.*, 2002; Knight *et al.*, 2004).



**Figura 2** – Exemplo de RNAs com diferentes sequências primárias que conservam a mesma estrutura secundária. (Gorodkin *et al.*, 2009).

Apesar das diversas metodologias existentes, a determinação de estrutura de RNAs ainda possui muitas limitações e ocorre em pequena escala na maioria dos estudos (Wan *et al.*, 2011).

Para que seja possível realizar uma abordagem sistêmica, em que dimensões informacionais diferentes de um mesmo organismo devem ser integradas de forma dinâmica a fim de se obter uma visão global do funcionamento dos seres vivos, necessita-se que esses estudos estruturais possam ser realizados também em larga escala (Beltrao *et al.*, 2007). Esta área de estudo, recentemente denominada de *RNA Estruturômica* (do inglês *Structuromics*), possibilita incorporar a dimensão do RNA estruturômica (Westhof *et al.*, 2010) à emergente Biologia Sistêmica.

Um organismo considerado modelo para estudos de Biologia Sistêmica é *Halobacterium salinarum*, um extremófilo pertencente ao domínio Archea, adaptado a ambientes de alta salinidade, simples e de fácil cultivo em laboratório. O estudo desses seres tem contribuído para a resolução de questões biológicas fundamentais, como a elucidação de mecanismos de regulação transcricional, funcionamento do ciclo celular, reparo de DNA, transporte e degradação de proteínas, entre outros (Soppa, 2006). Além disso, está disponível uma grande quantidade de informações relacionadas a *H. salinarum* nas áreas de genômica, transcritômica, proteômica, entre outras, o que facilita o desenvolvimento e aplicação de métodos de estudo nas mais diversas áreas.

Dessa forma, para que seja possível realizar predições estruturais *in silico* de todos os RNAs de um organismo, como *H. salinarum*, é necessária a adaptação das técnicas computacionais e metodologias estatísticas existentes, incorporando informações biofísicas experimentais disponíveis, para se obter modelos de estruturas mais acurados, sem um consumo de tempo excessivo.

## 2 – OBJETIVOS

### OBJETIVO GERAL

Expandir o entendimento do transcrito para além do estudo dos níveis de expressão, incorporando uma camada de informação adicional, em larga escala, por meio da modelagem do RNA estruturoma.

### OBJETIVOS ESPECÍFICOS

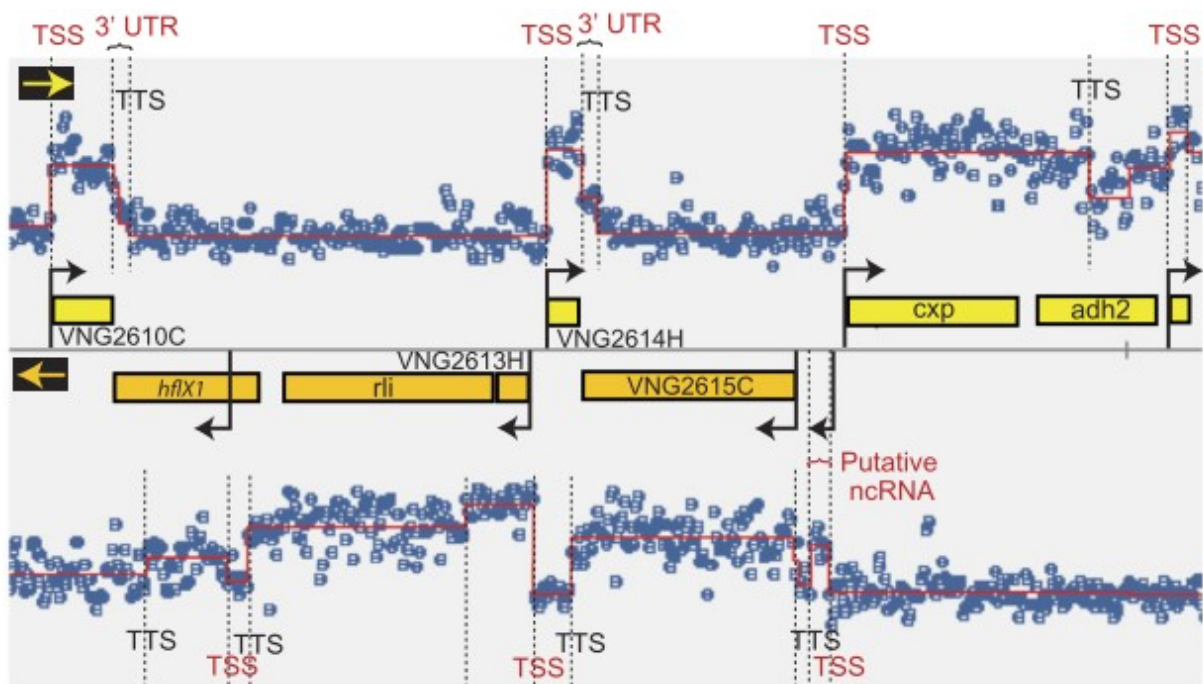
1. Auxiliar na identificação todos os potenciais transcritos codificantes e não-codificantes, em *cis*, *trans* ou internos de *Halobacterium salinarum*;
2. Adaptar metodologias computacionais de predição estrutural de RNAs para que sejam capazes de lidar com quantidade massiva de dados, como o transcrito completo de um organismo;
3. Predizer a estrutura, e seu *ensemble*, de todos os transcritos (RNA estruturoma) de *H. salinarum*;
4. Classificar os transcritos em famílias/grupos estruturais, com ou sem similaridade de sequência;
5. Investigar a correlação entre os agrupamentos estruturais e informações sobre função e/ou regulação do transcrito.

### 3 – MATERIAIS E MÉTODOS

#### 3.1. IDENTIFICAÇÃO DE TODOS OS POTENCIAIS TRANSCRITOS DE *HALOBACTERIUM SALINARUM*.

A base para um esforço de estabelecimento do RNA Estruturoma de um organismo é a determinação dos sítios de início (TSS - *transcript start site*) e término de transcrição (TTS - *transcript termination site*) de cada transcrito já observado experimentalmente, sejam estes codificantes, não-codificantes, alternativos internos a transcritos conhecidos, etc, enfim, do transcriptoma completo.

O transcriptoma completo de *H. salinarum*, em diversas condições fisiológicas, foi mapeado experimentalmente a partir de um de *tiling array* (Koide *et. al*, 2009) realizado previamente pelo grupo colaborador, coordenado pela Profa Koide (Depto de Bioquímica FMRP-USP) e ao qual o presente Projeto de Pesquisa de Pós-doutorado está associado (Projeto Jovem Pesquisador FAPESP 2009/09532-0). O *array* continha sondas de 60-mer de tamanho, com 40-mer de sobreposição entre elas, representando todo o genoma de *H. salinarum*. O sinal proveniente da hibridização no *tiling array* foi modelado com base na metodologia de Árvores de Regressão e foram identificados os sítios de início (TSS) e término de transcrição (TTS), utilizados para determinar os mRNAs, tRNAs, rRNAs e também ncRNAs de *H. salinarum* (Figura 3).



**Figura 3** – Estrutura de um trecho do transcriptoma de *H. Salinarum*, em que TSS e TTS representam os sítios de início e término de transcrição, respectivamente (Koide *et. al*, 2009).



Apesar destes resultados do grupo fornecerem um catálogo preliminar de todos os transcritos para iniciar o estabelecimento do RNA Estruturoma, a resolução dessa determinação possui limitações devido ao espaçamento das sondas desenhadas para o *array*.

Atualmente, nosso grupo de pesquisa vem trabalhando, em conjunto com o Laboratório de Biologia Sistêmica de Microorganismos (LaBiSisMi) coordenado pela Profa. Koide em estreita colaboração com o doutorando Felipe ten Caten (projeto FAPESP doutorado 2011/14455-4), na determinação precisa dos transcritos de *H. salinarum* através de técnicas de modelagem de sinal aplicadas a dados provenientes de *RNA-Seq*. Estes podem fornecer uma melhor resolução na identificação do início e término dos transcritos, uma etapa prévia importante para a correta determinação estrutural dessas moléculas. Dessa forma, auxiliar na identificação mais precisa dos transcritos é o passo inicial do presente Projeto de Pesquisa de Pós-doutorado.

Metodologias provenientes da área de Processamento de Sinais, como *Hidden Markov Model* (HMM) e *Change-point Model*, comumente utilizadas para a segmentação de sinais provenientes de arquivos de áudio, dados de econometria, entre outros, já estão sendo empregadas para a análise de dados biológicos (Du *et al.*, 2006; Day *et al.*, 2007; Winchester *et al.*, 2009; Li *et al.*, 2011; Wang *et al.*, 2011) e serão utilizadas nessa etapa.

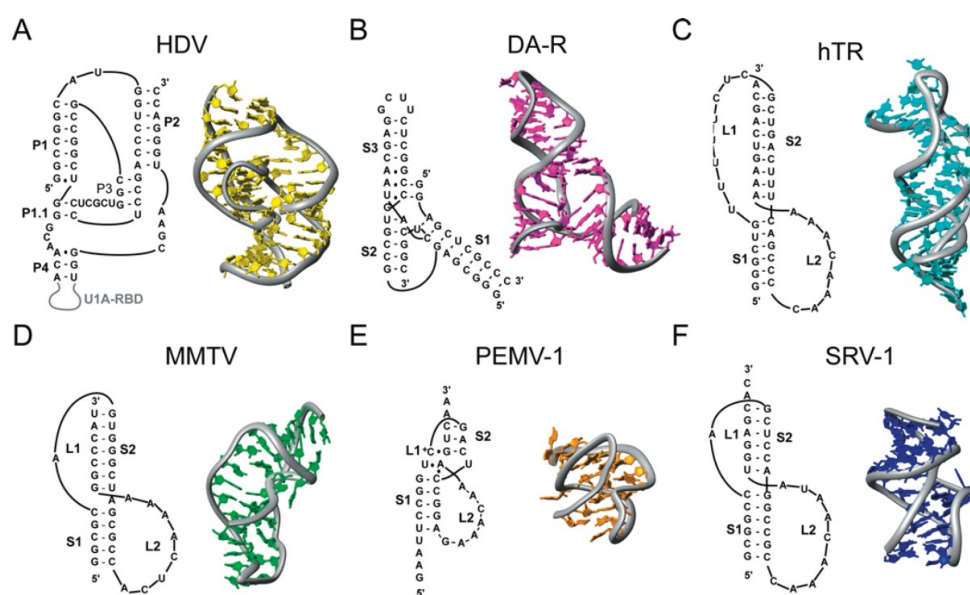
### 3.2. ADAPTAÇÃO DAS METODOLOGIAS DE DETERMINAÇÃO DA ESTRUTURA DE RNAs PARA UTILIZAÇÃO EM LARGA ESCALA.

Atualmente, existem dezenas de metodologias para a predição estrutural de RNAs, porém a grande maioria delas se limita apenas à predição de algumas classes específicas de RNAs, como os não-codificantes por exemplo, e focam em predições em pequena escala. Além disso, apenas um pequeno número de metodologias de determinação estrutural é capaz de prever estruturas contendo os chamados pseudo-nós (Sato *et al.*, 2011). Os pseudo-nós (Figura 4) são um tipo de motivo estrutural, com diversas topologias possíveis, que desempenham uma série de funções biológicas distintas, como *splicing* de *introns* (Staple *et al.*, 2005) e deslocamento do quadro de leitura, o chamado *frameshifting*, através do bloqueio dos ribossomos (Tholstrup, 2011). O banco de dados *PseudoBase++* (Taufers *et al.*, 2009) contém mais de 250 estruturas de pseudo-nós determinadas por diversas técnicas experimentais e computacionais.

A predição de estruturas de moléculas de RNA que contém esses pseudo-nós é uma tarefa árdua do ponto de vista computacional e a boa parte dos algoritmos existentes para realizar essa tarefa apresentam um consumo de tempo elevado (Sato *et al.*, 2011), tornando inviável sua aplicação para análises em larga escala.

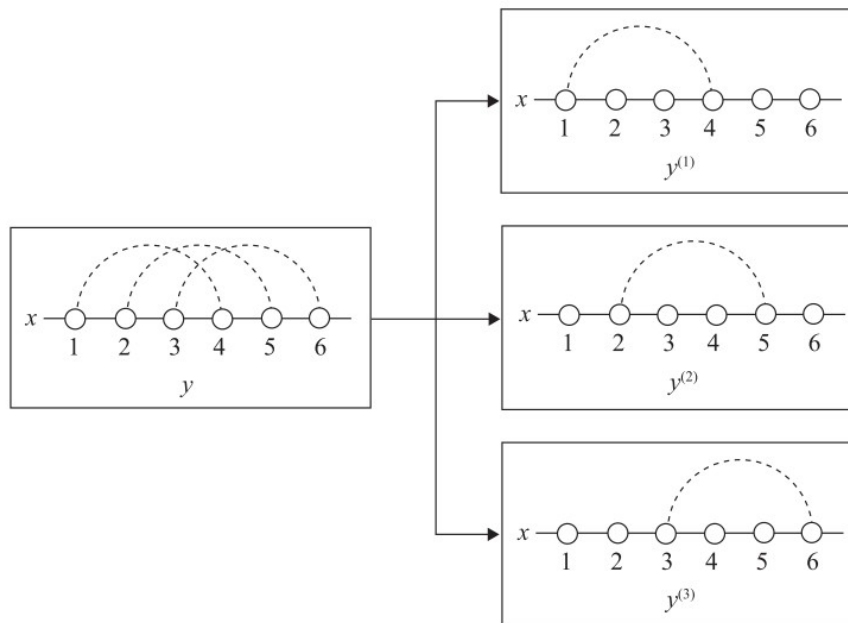
Devido aos problemas apresentados acima, a escolha de metodologias para a determinação estrutural dos RNAs fica restrita, pois para realizar essa predição em larga escala, conforme proposto no presente projeto, é necessário que o algoritmo escolhido seja capaz de determinar todos os tipos de estruturas de RNA, incluindo as que contêm pseudo-nós, de forma acurada e sem um consumo excessivo de tempo.

Uma metodologia proposta para a predição de estruturas de RNAs, contendo pseudo-nós, de maneira rápida e acurada é o *IPknot* (Sato *et al.*, 2011). Por essas características, rapidez e acurácia, além da capacidade de lidar com sequências longas de RNA, esse método será utilizado como ponto de partida para a adaptação, a fim de realizar a predição estrutural em larga escala.



**Figura 4** – Estruturas bidimensionais e tridimensionais de diferentes tipos de pseudo-nós (Staple *et al.*, 2005).

O *IPKnot* é baseado na abordagem de *Integer Programming* (*IP* - Programação Inteira), utilizada em problemas de otimização. Em linhas gerais, o *IPknot* realiza a decomposição da estrutura que contém o pseudo-nó em um conjunto de sub-estruturas sem esse motivo estrutural (Figura 5) e realiza uma aproximação da distribuição de probabilidades dos pareamentos internos das bases que formam a molécula contendo o pseudo-nó. É proposto ainda um algoritmo heurístico para refinar as probabilidades de pareamento e assim melhorar a acurácia da predição (Sato *et al.*, 2011).



**Figura 5** – Decomposição de uma estrutura contendo pseudo-nós em um conjunto de sub-estruturas livres de pseudo-nós (Sato *et al.*, 2011).

Apesar da escolha inicial desse método, ao longo do estudo aprofundado das metodologias existentes, novos algoritmos poderão também ser incorporados para a realização dessa etapa.

Ainda, a possibilidade da incorporação de informações experimentais disponíveis também é um dos passos desejáveis durante a adaptação dos métodos disponíveis, para melhorar ainda mais a acurácia da modelagem. Existe uma intenção do LaBiSisMi (pós-graduando Vicente Gomes Filho) de desenvolver uma metodologia em larga-escala para isto e em sendo bem sucedida deverá ser incorporada às metodologias estudadas no Projeto de Pesquisa de Pós-doutorado.

### 3.3. PREDIÇÃO DA ESTRUTURA 2D DE TODOS OS POTENCIAIS TRANSCRITOS DE *HALOBACTERIUM SALINARUM*.

*H. salinarum* é organismo extremófilo, modelo em biologia sistêmica, sobre o qual há uma gama de informações genômicas, transcritômicas, proteômicas, entre outras, disponíveis, o que torna possível o desenvolvimento e aplicação de metodologias *in silico* de estudo nas mais diversas áreas, incluindo a predição estrutural de RNAs, permitindo também a posterior integração dessas informações para criação de um modelo global multi-escala de funcionamento desses seres (Bonneau *et al.*, 2007).

Assim, após a adaptação dos algoritmos de predição estrutural, a ser realizada na segunda etapa do presente Projeto de Pesquisa de Pós-doutorado, as metodologias resultantes serão aplicadas para a modelagem da estrutura de todos os potenciais transcritos de *H. salinarum*.

### 3.4. AGRUPAMENTO EM BUSCA DE PADRÕES ESTRUTURAIS PARA CLASSIFICAÇÃO DOS GRUPOS DE RNAs.

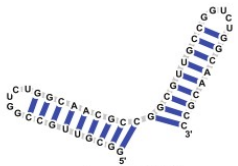


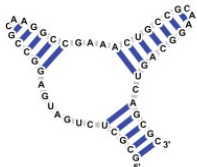
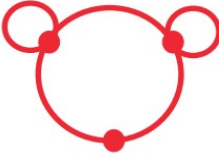
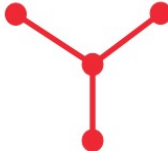
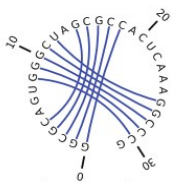

Após a predição da estrutura dos transcritos de *H. salinarum*, essas moléculas serão agrupadas e classificadas de acordo com os padrões estruturais encontrados, relacionados ou não às sequências primárias dos respectivos RNAs. Esses padrões geralmente são mais conservados do que as sequências, e poderão ser utilizados para inferir a função dos RNAs estudados.

Existem iniciativas para o agrupamento estrutural dessas moléculas, que serão utilizadas como base para essa etapa do projeto. Uma delas é o *Rfam*, um banco de dados de famílias de RNA que contém alinhamentos múltiplos, com estrutura secundária consenso de RNAs e suas respectivas famílias, além de modelos de covariância (Griffiths-Jones *et al.*, 2003; Gardner *et al.*, 2009; Gardner *et al.*, 2011). Modelos de covariância são generalizações de *Hidden Markov Models* (HMMs), modelos probabilísticos que tem sido usados na modelagem de muitos processos biológicos. Os HMMs, por sua vez, são similares as famosas Cadeias de Markov usuais mas com os estados da cadeia inacessíveis diretamente. Cada estado gera um conjunto de estados observáveis com uma certa probabilidade de emissão. Tendo acesso a série temporal destes estados observáveis, faz-se inferência a cerca dos estados (ocultos) da cadeia propriamente dita. Modelos de covariância para RNAs são baseados em uma árvore ordenada e os algoritmos para usar modelos de covariância na análise de sequências correspondem aos algoritmos para HMMs estendidos à estruturas de árvores. A partir desses modelos de covariância é possível determinar escores que combinam não apenas as sequências consenso de RNAs, mas também estruturas secundárias consenso, o que permite a identificação de RNAs homólogos que possuem estrutura mais conservada que a sequência primária (Eddy & Durbin, 1994). A versão atual do *Rfam*, 10.1, possui 1973 famílias de RNA catalogadas (<http://rfam.sanger.ac.uk/>).

Há também o *SCOR*, um outro exemplo de banco de dados focado na classificação estrutural das moléculas de RNA (Klosterman *et al.*, 2002; Tamura *et al.*, 2004) e o *RNA Strand*, que contém estruturas secundárias de RNAs de uma série de organismos, além de diversos tipos de informações estatísticas de cada uma dessas estruturas (Andronescu *et al.*, 2008).

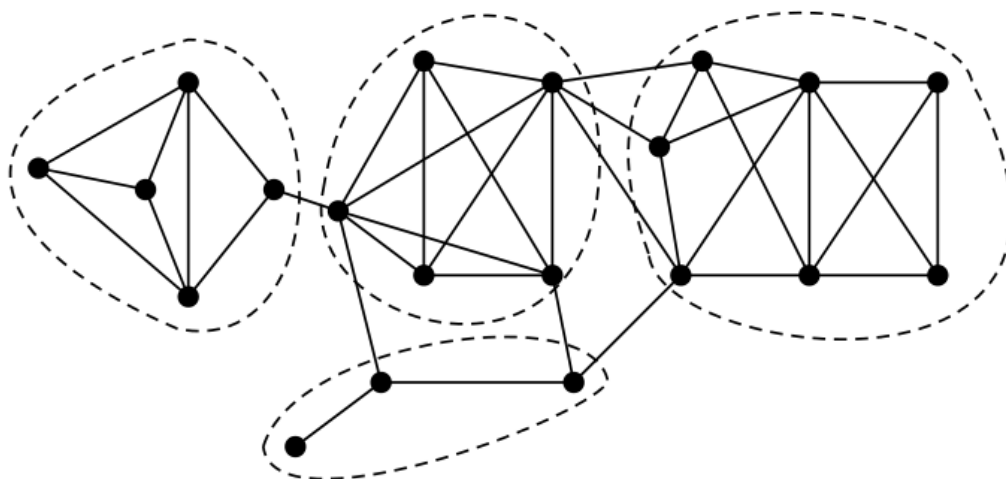
Outro recurso interessante que poderá ser utilizado nessa etapa é o *RAG* (*RNA-As-Graphs* - Fera *et al.*, 2004; Izzo *et al.*, 2011), que permite classificar, catalogar e prever motivos de estrutura secundária de RNAs, utilizando para as representações e classificações elementos da teoria dos

grafos, como *dual graphs* (grafos duais) e *tree graphs* (grafos árvore) (Figura 6). A partir da representação das estruturas de RNAs em grafos, elas poderão ser agrupadas e funções específicas poderão ser relacionadas aos grupos formados. Para esse agrupamento será realizada a adaptação do algoritmo MCL (*Markov Cluster Algorithm* – Dongen, 2000), capaz de encontrar agrupamentos em grafos.

Structure	Dual Graph	Tree Graph
 <p>Secis RNA (PDB_01270)</p>	 <p>Bridge</p>	
 <p>Hammerhead Ribozyme (PDB_00693)</p>	 <p>Tree</p>	
 <p>Ribosomal RNA (PDB_00133)</p>	 <p>Pseudoknot</p>	<p>None</p>

**Figura 6** – Representação de estrutura de RNAs na forma de grafos (Izzo *et al.*, 2011).

As aplicações clássicas de metodologias de agrupamento consideram que as entidades a serem estudadas são representadas como vetores, em que são descritas suas características. É calculada então uma medida de distância entre esses vetores, e essa medida é utilizada para a realização dos agrupamentos. Já no caso de grafos, que possuem características intrínsecas distintas dos vetores, como grau (número de vizinhos), caminho, ciclos, conectividade, entre outros, esse tipo de procedimento não pode ser aplicado diretamente. O MCL foi desenvolvido então com o objetivo de encontrar agrupamentos em estruturas de grafos simples (Figura 7), onde a dissimilaridade entre os vértices pode ser definida de forma implícita através das características das relações de conectividade do grafo (Dongen, 2000).



**Figura 7** – Representação de agrupamentos em grafo utilizando MCL (Dongen, 2000).

Essa técnica já foi aplicada com sucesso para a investigação de dados biológicos, por exemplo na classificação de proteínas em famílias (Enright *et al.*, 2002) e na identificação de grupos de ortólogos, através do *OrthoMCL* (Li *et al.*, 2003), podendo então, em teoria, ser adaptada para o problema em questão de realizar agrupamento de estruturas de RNAs representadas por grafos.

### 3.5. CORRELAÇÃO ENTRE AGRUPAMENTOS ESTRUTURAIIS E INFORMAÇÕES SOBRE FUNÇÃO E REGULAÇÃO DO TRANSCRITOMA.

A relação entre estrutura e função em proteínas é inquestionável. Recentes descobertas revelaram a versatilidade surpreendente dos RNAs e sua importância em uma variedade de funções. Não é mais surpresa que características estruturais de RNA tenham grande papel na sua função biológica, já que apenas sua sequência não contém informações funcionais suficientes (Laing & Schlick, 2010). RNA *switches*, ou seja, RNAs que mudam drasticamente sua estrutura, são elementos regulatórios importantes (Sullenger, 2004). RNA *switches* não são meras exceções de moléculas de RNA apresentando comportamento não-usual e sim, por outro lado, representantes de vários outros exemplos que vem sendo descobertos atualmente (Flamm *et al.*, 2000). O uso de duas conformações competitivas de RNA permite que eventos moleculares, como o ligamento de uma molécula por uma proteína, possa ser usado para regular a expressão gênica, quando essas configurações alternativas mutuamente exclusivas correspondem a uma configuração ativa e uma configuração inativa do transcrito (Merino & Yanofsky, 2002). Os *riboswitches* são o exemplo mais

conhecido desse comportamento (Vitreschak *et al.*, 2004). Alguns vírus com genoma do tipo RNA (ou RNA vírus) também fazem uso de transições entre domínios estruturais para realizar funções diferentes. Assim, um dos objetivos do estudo da estrutura de RNA é obter o entendimento de como a estrutura e a dinâmica resultam em funções específicas desempenhadas pelos RNAs.

A partir do agrupamento estrutural obtidos, pretendemos então estudar a correlação entre os grupos de RNA e informações sobre função e regulação do transcrito. Pode ser muito difícil inferir algum tipo de funcionalidade apenas através da verificação direta dos grupos com o conhecimento já existente sobre famílias ou motivos estruturais, já que estes não são tão numerosos por ser esta uma área emergente de pesquisa com acúmulo de informação ainda limitado quando comparada, por exemplo, ao que se conhece de famílias proteicas. Um diferencial da nossa proposta é a forte cooperação com um grupo de pesquisa que já quantificou o transcrito da arquea *Halobacterium salinarum* em várias condições experimentais.

Nosso grupo já tem acesso a um compêndio de dados de expressão gênica em larga-escala obtidos por RNAseq e/ou microarray, em diversas condições ambientais e ao longo de séries-temporais, introduzindo o desejado fator dinâmico nas análises. O grupo colaborador já tem disponível em seus bancos-de-dados internos o transcrito da curva de crescimento em condições padrão e em algumas condições de estresse, de mutantes para proteínas que sabidamente interagem com RNA, e de uma biblioteca RNA-seq enriquecida para RNAs pequenos. Usando quantificação por tiling array, várias condições de crescimento foram obtidas e um conjunto muito grande de dados já está disponível na literatura, uma vez que *Halobacterium salinarum* é um organismo modelo em biologia sistêmica e os líderes de pesquisa mundial nesta área são parceiros dos grupos de pesquisa envolvidos no presente projeto. A pertinência de certos membros num dado agrupamento pode correlacionar com comportamentos dinâmicos observados no transcrito e, assim, fornecer pistas de potenciais papéis funcionais que dificilmente seriam indicados de outra forma. Essa é uma das principais motivações para uma abordagem multi-escala e multidisciplinar.

#### **4 – EQUIPE COLABORADORA E INFRA- ESTRUTURA**

Este projeto de pós-doutorado será de responsabilidade principal da Dra. Eliane Z. Traldi, com supervisão próxima do Prof Dr. Vêncio e co-supervisão da Profa. Dra. Koide.

A Dra. Traldi é formada em Matemática pela USP e doutora em Biologia Computacional e Biofísica Molecular pela Rutgers University. Em seu doutorado trabalhou com a modelagem matemática da transição de fase lisogênica/lítica em vírus e na modelagem matemática da classe de RNAs abortivos (aRNAs), ambos sob a orientação do Prof. Konstantin Mischaikow.

Este projeto conta com a colaboração de Felipe ten Caten, aluno de doutorado (Processo FAPESP: 2011/14455-4) no Programa Interunidades de Pós-graduação em Bioinformática USP, José Vicente Gomes Filho e Livia Zaramela (Processo FAPESP: 2011/07487-7), alunos de mestrado e doutorado, respectivamente, do Programa de Pós-graduação em Bioquímica, todos orientados pela Profa. Dra. Tie Koide. Conta ainda com a colaboração dos estudantes de doutorado Marcos Abraão Fonseca (Processo FAPESP: 2012/02896-9) e Diego Martinez (Processo FAPESP: 2011/08104-4), orientados pelo Prof. Dr. Vêncio no doutorado em Bioinformática. O projeto 2011/14455-4 irá fornecer a identificação das sequências de RNAs a serem investigadas no presente projeto ao passo que os projetos 2011/07487-7 e 2012/02896-9 irão fornecer as evidências experimentais e computacionais para posterior correlação com os agrupamentos de famílias a serem encontrados na presente proposta. O projeto 2011/08104-4 deve fornecer o apoio de software necessário para manipulação prática do compêndio de dados já disponíveis para os quais busca-se correlação com estrutura no presente projeto. O envolvimento tangencial de diversos estudantes de pós-graduação, com diferentes *backgrounds*, no presente projeto de pesquisa caracteriza um verdadeiro grupo multidisciplinar e demonstra a inserção desta proposta num contexto maior e ciência colaborativa, típica e necessária para o enfrentar questões complexas de Biologia Sistêmica.

Este projeto é supervisionado pelo físico Prof. Dr. Ricardo Z. N. Vêncio. Após obter seu Doutorado em Bioinformática pela USP, ele passou dois anos em Seattle realizando pós-doutorado no Institute for Systems Biology (<http://systemsbiology.org>) e retornou ao Brasil como Professor Doutor na USP. Sua experiência em Bioinformática e Biologia Sistêmica Computacional estão bem documentadas em seu histórico de publicações envolvendo assuntos que vão desde Expressão Gênica (microarranjos e sequenciamento) até métodos Bayesianos em Bioinformática.

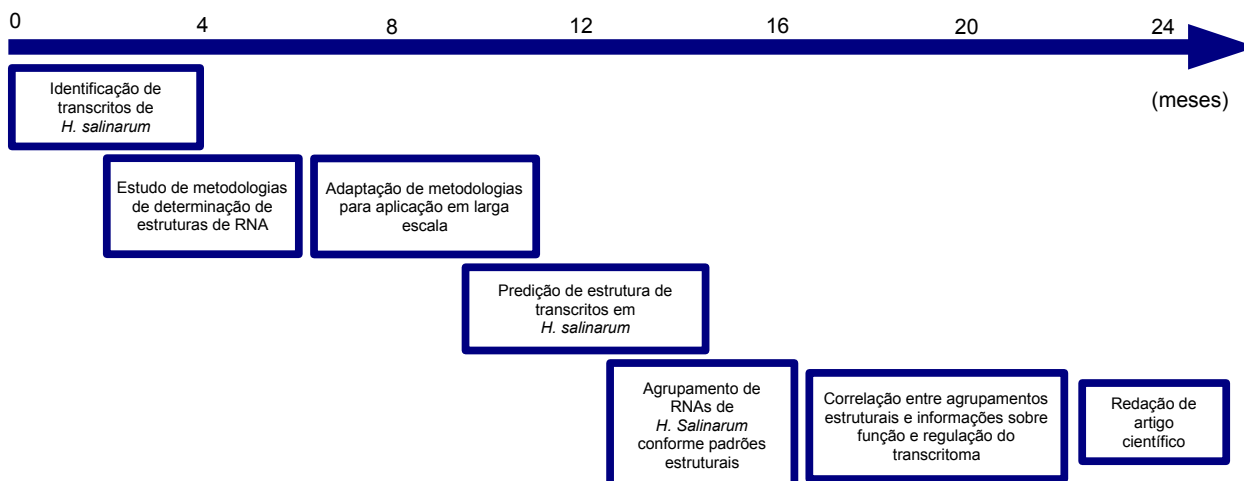
O presente projeto está associado oficialmente ao Projeto Jovem Pesquisador FAPESP 2009/09532-0 “Biologia Sistêmica do extremófilo *Halobacterium salinarum*: contribuição dos RNAs não-codificantes ao modelo global de regulação gênica”, coordenado pela Profa. Dra. Tie Koide, do Departamento de Bioquímica e Imunologia FMRP-USP, do qual o Prof. Dr. Vêncio é formalmente Pesquisador Associado. Ainda, o projeto está associado não-formalmente ao projeto FAPESP 2010/50425-0 “Mecanismos moleculares fundamentais responsáveis pela complexação de proteínas”, coordenado pelo físico Prof. Dr. Fernando Luis Barroso da Silva, do Departamento de Física e Química FCFRP-USP. O Prof. Barroso é especialista em Biofísica Estrutural e pretende cooperar com o presente projeto como consultor em questões envolvendo termodinâmica das estruturas biomoleculares envolvidas no presente projeto, bem como irá ceder tempo de máquina no supercomputador SGI Altrix XE1300 adquirido pelo projeto FAPESP supracitado e instalado no Departamento de Computação e Matemática com parte de uma série de projetos cooperativos entre seu grupo e o LabPIB.



Toda infra-estrutura necessária está disponível no Laboratório de Processamento de Informação Biológica (LabPIB <http://labpib.openwetware.org/>) do Departamento de Computação e Matemática - FFCLRP, que tem como professor responsável o Prof. Dr. Ricardo Z. N. Vêncio, e no Laboratório de Biologia Sistemática de Microorganismos (LaBiSisMi <http://labisismi.fmrp.usp.br/>), que tem como responsável a Profa. Dra. Tie Koide, do Departamento de Bioquímica e Imunologia FMRP.

## 5 – CRONOGRAMA APROXIMADO

As atividades propostas neste projeto de pesquisa deverão se desenvolver aproximadamente conforme o cronograma proposto abaixo:



## 6 – REFERÊNCIAS BIBLIOGRÁFICAS

Andreeva, A., & Murzin, A. G. (2010). Structural classification of proteins and structural genomics: new insights into protein folding and evolution. *Acta crystallographica. Section F, Structural biology and crystallization communications*, 66(Pt 10), 1190-7. International Union of Crystallography. doi:10.1107/S1744309110007177

Andronescu, M., Bereg, V., Hoos, H. H., & Condon, A. (2008). RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC bioinformatics*, 9, 340. doi:10.1186/1471-2105-9-340

Beltrao, P., Kiel, C., & Serrano, L. (2007). Structures in systems biology. *Current opinion in structural biology*, 17(3), 378-84. doi:10.1016/j.sbi.2007.05.005

- Bonneau, R., Facciotti, M. T., Reiss, D. J., Schmid, A. K., Pan, M., Kaur, A., Thorsson, V., et al. (2007). A predictive model for transcriptional control of physiology in a free living cell. *Cell*, *131*(7), 1354-65. doi:10.1016/j.cell.2007.10.053
- Butcher, S. E., & Pyle, A. M. (2011). The Molecular Interactions That Stabilize RNA Tertiary Structure: RNA Motifs, Patterns, and Networks. *Accounts of chemical research*. doi:10.1021/ar200098t
- Day, N., Hemmaplardh, A., Thurman, R. E., Stamatoyannopoulos, J. a, & Noble, W. S. (2007). Unsupervised segmentation of continuous genomic data. *Bioinformatics (Oxford, England)*, *23*(11), 1424-6. doi:10.1093/bioinformatics/btm096
- Du, J., Rozowsky, J. S., Korb, J. O., Zhang, Z. D., Royce, T. E., Schultz, M. H., Snyder, M., et al. (2006). A supervised hidden markov model framework for efficiently segmenting tiling array data in transcriptional and chIP-chip experiments: systematically incorporating validated biological knowledge. *Bioinformatics (Oxford, England)*, *22*(24), 3016-24. doi:10.1093/bioinformatics/btl515
- Eddy, S. R., & Durbin, R. (1994). RNA analysis using covariance models, *22*(11).
- Enright, a J., Van Dongen, S., & Ouzounis, C. a. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, *30*(7), 1575-84.
- Fera, D., Kim, N., Shiffeldrim, N., Zorn, J., Laserson, U., Gan, H. H., & Schlick, T. (2004). RAG: RNA-As-Graphs web resource. *BMC bioinformatics*, *5*, 88. doi:10.1186/1471-2105-5-88.
- Flamm, C., Hofacker, I. L., Maurer-Stroh, S., Stadler, P. F., & Zehl, M. (2000). Design of multi-stable RNA molecules. *RNA*, *7*, 254-265.
- Gardner, P. P., Daub, J., Tate, J. G., Nawrocki, E. P., Kolbe, D. L., Lindgreen, S., Wilkinson, A. C., et al. (2009). Rfam: updates to the RNA families database. *Nucleic acids research*, *37*(Database issue), D136-40. doi:10.1093/nar/gkn766
- Gardner, P. P., Daub, J., Tate, J., Moore, B. L., Osuch, I. H., Griffiths-Jones, S., Finn, R. D., et al. (2011). Rfam: Wikipedia, clans and the “decimal” release. *Nucleic acids research*, *39*(Database issue), D141-5. doi:10.1093/nar/gkq1129
- Gilbert, W. (1986). Origin of life: The RNA world. *Nature* *319* (6055): 618–618. Bibcode Nature.319..618G. doi:10.1038/319618a0
- Gorodkin, J., & Hofacker, I. L. (2011). From Structure Prediction to Genomic Screens for Novel Non-Coding RNAs. (M. Levitt, Ed.) *PLoS Computational Biology*, *7*(8), e1002100. doi:10.1371/journal.pcbi.1002100
- Gorodkin, J., Hofacker, I. L., Torarinsson, E., Yao, Z., Havgaard, J. H., & Ruzzo, W. L. (2010). De novo prediction of structured RNAs from genomic sequences. *Trends in biotechnology*, *28*(1), 9-19. doi:10.1016/j.tibtech.2009.09.006
- Griffiths-Jones, S. (2003). Rfam: an RNA family database. *Nucleic Acids Research*, *31*(1), 439-441. doi:10.1093/nar/gkg006
- Izzo, J. a, Kim, N., Elmetwaly, S., & Schlick, T. (2011). RAG: An Update to the RNA-As-Graphs Resource. *BMC bioinformatics*, *12*(1), 219. BioMed Central Ltd. doi:10.1186/1471-2105-12-219
- Kertesz, M., Wan, Y., Mazor, E., Rinn, J. L., Nutter, R. C., Chang, H. Y., & Segal, E. (2010). Genome-wide measurement of RNA secondary structure in yeast. *Nature*, *467*(7311), 103-7. Nature Publishing Group. doi:10.1038/nature09322
- Kloc, M. (2008). Emerging novel functions of RNAs, and binary phenotype? *Developmental biology*, *317*(2), 401-4. doi:10.1016/j.ydbio.2008.03.003

Kloc, M. (2009). Teachings from the egg: new and unexpected functions of RNAs. *Molecular reproduction and development*, 76(10), 922-32. doi:10.1002/mrd.21043

Kloc, M., Foreman, V., & Reddy, S. a. (2011). Binary function of mRNA. *Biochimie*, 1-7. Elsevier Masson SAS. doi:10.1016/j.biochi.2011.07.008

Klosterman, P. S., Tamura, M., Holbrook, S. R., & Brenner, S. E. (2002). SCOR: a Structural Classification of RNA database. *Nucleic acids research*, 30(1), 392-4.

Knight, R., Birmingham, A., & Yarus, M. (2004). BayesFold: rational 2 degrees folds that combine thermodynamic, covariation, and chemical data for aligned RNA sequences. *RNA (New York, N.Y.)*, 10(9), 1323-36. doi:10.1261/rna.5168504

Koide, T., Reiss D. J., Bare, J. C. et al. (2009). Prevalence of transcription promoters within archaeal operons and coding sequences. *Molecular systems biology*, 5(285):285.

Laing, C., Schlick, T. (2010). Computational approaches to 3D modeling of RNA. *Journal of Physics: Condensed Matter*, 22, 283101(18pp). doi:10.1088/0953-8984/22/28/283101

Li, A., Liu, Z., Lezon-Geyda, K., Sarkar, S., Lannin, D., Schulz, V., Krop, I., et al. (2011). GPHMM: an integrated hidden Markov model for identification of copy number alteration and loss of heterozygosity in complex tumor samples using whole genome SNP arrays. *Nucleic acids research*, 39(12), 4928-41. doi:10.1093/nar/gkr014

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research*, 13(9), 2178-89. doi:10.1101/gr.1224503

Mathews, D. H., & Turner, D. H. (2006). Prediction of RNA secondary structure by free energy minimization. *Current opinion in structural biology*, 16(3), 270-8. doi:10.1016/j.sbi.2006.05.010

Merino, E. J., Wilkinson, K. a, Coughlan, J. L., & Weeks, K. M. (2005). RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *Journal of the American Chemical Society*, 127(12), 4223-31. doi:10.1021/ja043822v

Merino, E. & Yanofsky, C. (2002). Regulation by termination-antitermination: a genomic approach. In A. L. Sonenshein, J. A. Hock, and R. Losick, editor, *Bacillus subtilis and its closest relatives: From Genes to Cells*. ASM Press, Washington D.C., 323-336.

Meyer, F., Kurtz, S., Backofen, R., Will, S., & Beckstette, M. (2011). Structator: fast index-based search for RNA sequence-structure patterns. *BMC bioinformatics*, 12(1), 214. BioMed Central Ltd. doi:10.1186/1471-2105-12-214

*Protein Data Bank homepage*: <<http://www.rcsb.org/pdb/home/home.do>>. Acesso em dezembro de 2011.

*Rfam homepage*: <<http://rfam.sanger.ac.uk/>>. Acesso em dezembro de 2011.

Rose, P. W., Beran, B., Bi, C., Bluhm, W. F., Dimitropoulos, D., Goodsell, D. S., Prlic, A., et al. (2011). The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic acids research*, 39(Database issue), D392-401.

Sato, K., Kato, Y., Hamada, M., Akutsu, T., & Asai, K. (2011). IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics (Oxford, England)*, 27(13), i85-i93. Doi:10.1093/bioinformatics/btr215

- Soppa, J. (2006). From genomes to function: haloarchaea as model organisms. *Microbiology (Reading, England)*, 152(Pt 3), 585-90. doi:10.1099/mic.0.28504-0
- Stijn van Dongen, *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, May 2000.
- Tamura, M., Hendrix, D. K., Klosterman, P. S., Schimmelman, N. R. B., Brenner, S. E., & Holbrook, S. R. (2004). SCOR: Structural Classification of RNA, version 2.0. *Nucleic acids research*, 32(Database issue), D182-4. doi:10.1093/nar/gkh080
- Sullenger, B. A. (2004). Riboswitches – to kill or save the messenger. *New England journal of Medicine*, 351, 2759-2760.
- Tholstrup, J., Oddershede, L. B., & Sorensen, M. a. (2011). mRNA pseudoknot structures can act as ribosomal roadblocks. *Nucleic Acids Research*, 1-11. doi:10.1093/nar/gkr686
- Tucker, B. J., & Breaker, R. R. (2005). Riboswitches as versatile gene control elements. *Current opinion in structural biology*, 15(3), 342-8. doi:10.1016/j.sbi.2005.05.003
- Turner, K. B., Yi-Brunozzi, H. Y., Brinson, R. G., Marino, J. P., Fabris, D., & Le Grice, S. F. J. (2009). SHAMS: combining chemical modification of RNA with mass spectrometry to examine polypurine tract-containing RNA/DNA hybrids. *RNA (New York, N.Y.)*, 15(8), 1605-13. doi:10.1261/rna.1615409
- Underwood, J. G., Uzilov, A. V., Katzman, S., Onodera, C. S., Mainzer, J. E., Mathews, D. H., Lowe, T. M., et al. (2010). FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nature methods*, 7(12), 995-1001. doi:10.1038/nmeth.1529
- Vitreschak, A.G., Rodionov, D.A., Mironov, A.A., Gelfand, M.S.(2004). Roboswitches: the oldest mechanism for regulation of gene expression? *Trends in Genetics*,20(1), 44-50.
- Wan, Y., Kertesz, M., Spitale, R. C., Segal, E., & Chang, H. Y. (2011). Understanding the transcriptome through RNA structure. *Nature Reviews Genetics*, 12(9), 641-655. Nature Publishing Group. doi:10.1038/nrg3049
- Wang, Y., Wu, C., Ji, Z., Wang, B., & Liang, Y. (2011). Non-parametric change-point method for differential gene expression detection. *PloS one*, 6(5), e20060. doi:10.1371/journal.pone.0020060
- Wells, S. E., Hughes, J. M. X., Igel, A. H., & Ares, M. (2000). Use of dimethylsulfate to probe RNA structure in vivo. *Methods in Enzymology*, 318, 479-493.
- Westhof, E., & Romby, P. (2010). The RNA structurome: high-throughput probing. *Nature methods*, 7(12), 965-7. Nature Publishing Group. doi:10.1038/nmeth1210-965
- Winchester, L., Yau, C., & Ragoussis, J. (2009). Comparing CNV detection methods for SNP arrays. *Briefings in functional genomics & proteomics*, 8(5), 353-66. doi:10.1093/bfpg/elp017