MARCOS ABRAÃO DE SOUZA FONSECA

Identificação *in silico* de ncRNAs no organismo modelo *Halobacterium salinarum* NRC-1

Ribeirão Preto 2016

MARCOS ABRAÃO DE SOUZA FONSECA

Identificação *in silico* de ncRNAs no organismo modelo *Halobacterium salinarum* NRC-1

Tese de doutorado apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática da Universidade de São Paulo para obtenção do título de Doutor em Ciências.

Área de Concentração: Bioinformática.

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

Ribeirão Preto 2016

Agradecimentos

Aproveito esse momento, o mais descontraído, para expressar meus sinceros agradecimentos a todos que permearam este tão importante período de minha formação.

A minha família por compreender minhas fases de ausência e ainda oferecer todo apoio e amizade possível. Agradeço em especial à minha avó Dona Lourdes que apesar de ter frequentado muito pouco a escola sempre percebeu minha busca pelo conhecimento e contribuiu nessa minha jornada com todo seu valor de vida.

A minha mãe Aldeti por estar presente e também buscar me entender e apoiar. Participou, a sua maneira, das principais etapas e fez o seu melhor para me acompanhar.

Ao meu irmão Márcio pelo carinho e sensibilidade. Tenho grande orgulho de seu caráter e de seu empenho nas atividades que desenvolve. Guardo a esperança de trabalharmos juntos em vários projetos.

A minha irmã Alice, pois me traz alegria e também caminha comigo pelo mundo do ensino. É uma grande satisfação poder compartilhar desafios e satisfações na arte de transmitir conhecimento e valores. Agradeço também as minhas queridas e lindas sobrinhas Letícia e Lívia pelos momentos pueris de muita felicidade.

Ao meu orientador Ricardo Vêncio por toda a ajuda e apoio. Agradeço principalmente pela paciência perante minhas dificuldades, pelas discussões que permitiram o andamento do trabalho e pela grande competência a qual realiza sua pesquisa, será para mim mais que uma referência.

A Patricia Martorelli por todo seu empenho, atenção, dedicação e ajuda que possibilitam minha entrada no programa de Bioinformática.

Ao meu grande amigo, que considero como irmão, Rômulo Franco pelas conversas e troca de ideias, um precioso companheiro nessa caminhada acadêmica.

Aos amigos de trabalho Felipe, Martinez, Cawal, Lívia, Silva, Gabi, Vicente, Torresmo, Marjorie e Khan pela ajuda, companheirismo e momentos de descontração. Agradeço também à professora Tie Koide pelas sugestões e contribuições que foram essenciais e de grande ajuda.

Aos alunos que passaram pela Oficina de Forró Universitário.

Aos amigos da Seção de Atividades Culturais da USP – Ribeirão Preto em especial ao Lelo, Dilson, Camila e Carlos.

Aos amigos da Cia. Minaz, em especial a bassonaria.

Aos amigos do TUSP Ribeirão Preto.

Aos amigos que conheci nas atividades culturais de Ribeirão Preto.

Aos membros da banca pela perspectiva crítica, sugestões e contribuições.

Agradeço às agências de fomento CAPES e FAPESP pelo suporte financeiro e recursos necessários para a realização deste trabalho.

Por fim, aos que contribuíram de forma direta e indireta no desenvolvimento deste trabalho.

“Somewhere, something incredible is waiting to be known.”

(Carl Sagan)

Resumo

FONSECA, Marcos Abraão de Souza. **Identificação *in silico* de ncRNAs no organismo modelo *Halobacterium salinarum* NRC-1**. 2016. 131 pág. Tese (Doutorado em Ciências) – Faculdade de Filosofia, Ciências e Letras - Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2016.

A regulação da expressão gênica ocorre como um fenômeno essencial nos processos celulares em resposta a dinamicidade mútua estabelecida entre um organismo e seu meio. Além dos elementos reguladores já conhecidos, como fatores de transcrição ou modificações pós-transcricionais, observa-se um crescente interesse no papel de regulação desempenhado por moléculas de RNA não codificadores (ncRNA), que podem atuar em vários níveis de processamento da informação biológica. Organismos modelos oferecem uma forma conveniente de pesquisa e diferentes grupos buscam direcionar seus estudos para um entendimento mais amplo no que se refere aos mecanismos celulares presentes nesses organismos. Apesar da existência de alguns elementos conhecidos para o organismo modelo *Halobacterium salinarum,* acreditamos que nem todos seus elementos de ncRNAs foram identificados. Nesse contexto, desenvolvemos uma análise *in silico* para a identificação de novos ncRNAs em *H. salinarum* NRC-1 e aplicamos metodologias para a predição de possíveis interações RNA-Proteína*.* Com base em uma pespectiva de integração de dados e diferentes metodologias existentes, modelos de Aprendizado de Máquina (AM) foram criados e utilizados para a definição de regiões candidatas a ncRNAs. De acordo com os resultados, 42 novos ncRNAs puderam ser identificados e possibilitaram completar o catálogo de genes ncRNAs de *H. salinarum* NRC-1 e aumentar o universo conhecido destes em 82%. A análise dos resultados obtidos por outras abordagens disponíveis para a identificação de ncRNAs corroboram com alguns dos candidatos sugeridos neste trabalho. Adicionalmente, foram aplicados e avaliados métodos, também baseados em AM, para a identificação de candidatos à interação com a proteína de interesse LSm, presente no organismo em estudo, no intuito de incluir uma possível caracterização funcional de ncRNAs. Os resultados alcançados na aplicação metodologias para a predição de interações RNA-Proteína não foram suficientes para a criação de um modelo com predições de alto grau de acurácia porém, contribuem como estudos preliminares e discussões para o desenvolvimento de outras estratégias.

Palavras-chave: Aprendizado de Máquina, Interação RNA-Proteína, RNAs não-codificadores, *Halobactrium salinarum.*

Abstract

FONSECA, Marcos Abraão de Souza. **Identificação *in silico* de ncRNAs no organismo modelo *Halobacterium salinarum* NRC-1**. 2016. 131 pages. Tese (Doutorado em Ciências) – Faculdade de Filosofia, Ciências e Letras - Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2016.

The gene expression regulation occurs on different cell levels in response to dynamics established between an organism and its environment. In addition to the regulatory elements already known, for instance, transcription factors or post-translation modifications, there is growing interests in the regulatory role played by non-coding RNA molecules (ncRNA) whose functions can be performed on different level of biological information processing. Model organisms allow a convenient way to work on laboratory and different research groups aiming to guide their studies for a mutual and wide understanding of the cellular mechanisms present on these organisms. Although some ncRNAs elements have been found in *Halobacterium salinarum* model organism we believe that not enough is knowing about these genomic regions. In these context, an *in silico* analysis for ncRNAs identification and RNA-protein prediction approach were applied to *H. salinarum* NRC-1*.* Considering a data integration perspective and some available methodologies, several machine learning models was built and used to designate candidate ncRNAs genome regions. According to achieve results, 42 new ncRNAs could be identified, increasing 82% the total of known ncRNAs in *H. salinarum* NRC-1. Combing analysis with other available tools, it had been observed that some suggested candidates also was found with different methodologies and thus, it highlights the proposed results. Additionally, we developed and analyzed methods, also machine learning based, to predict ncRNAs candidates to interact with LSm protein, present on the interested model organism aiming a basic ncRNA characterization. The achieved results in this part was not satisfactory since the applied models were not substantially accurate predictions. However, we believe that these preliminary results can contribute with some discussions to new different approaches.

Keywords: Machine Learning. Non-coding RNAs. *Halobacterium salinarum*. RNA-Protein Interaction.

Lista de figuras

Lista de tabelas

Sumário

1 Introdução

O avanço do conhecimento biológico tem sido amplamente guiado pelo uso intensivo de métodos computacionais para a organização e análises das informações. Nesse contexto, pesquisas interdisciplinares na área de Bioinformática tornaram-se fundamentais para a criação de modelos mais abrangentes e capazes de lidar com dados em larga escala. Logo, estudos que envolvem não somente a caracterização de elementos celulares de forma individual e sim, a partir de uma rede de interações mais complexa, em um sistema celular integrado, têm se tornado cada vez mais factíveis (Karr *et al.,* 2012) (Bonneau *et al.,* 2007) (Brooks *et al.*, 2014) (Hogeweg, 2011).

Organismos modelo são fundamentalmente utilizados nos três domínios da vida para a descoberta e entendimento de mecanismos biológicos. Em um esforço conjunto e comparativo, oriundo de diversos grupos de pesquisa, existe a expectativa de que tais mecanismos possam ainda ser generalizados para outros organismos. Apesar da grande variedade de formas de vida, a junção desses estudos sobre um pequeno subconjunto de organismos contribui para um entendimento mais amplo dos processos celulares, que são fundamentais para manter a vida (Müller & Grossniklaus, 2010) (Ankeny & Leonelli, 2011) (Leonelli & Ankeny, 2013).

Pesquisas baseadas nesses tipos de organismos são orientadas de acordo com vários interesses incluindo, econômicos, agriculturais, saúde e ambientais. A fácil manipulação em estudos experimentais também pode ser uma grande vantagem de alguns organismos modelo, sendo estes geneticamente modificáveis, com curto ciclo de vida e simples de cultivar em laboratório (Hedges, 2002) (Müller & Grossniklaus, 2010). Em procariotos, *Escherichia coli* é o modelo clássico de biologia molecular. Outros organismos como *Bacillus subtilis,* amplamente utilizado em biotecnologia e alguns agentes causadores de doenças como: *Mycobacterium tuberculosis* (tuberculose), *Mycoplasma pneumoniae* (pneumonia) e *Vibrio cholerae* (cólera) também são exemplos de organismos modelo neste domínio (Hedges, 2002). Em eucariotos,

Drosophila melanogaster contribui há mais de um século em diversos estudos sobre hereditariedade e desenvolvimento genético. Pesquisas baseadas em *Saccharomyces cerevisiae* elucidaram diversos processos envolvidos no ciclo celular em eucariotos, inclusive o controle da divisão celular, *Caenorhabditis elegans* possui as vantagens de manipulação genética e ciclo de vida semelhante a de um micro-organismo com as seguintes características: um sistema completo de órgãos, comportamento sexual, social e de aprendizado e ainda a particularidade de ser possível traçar a linhagem de cada uma das suas quase 1000 células. *Mus musculus* é um dos organismos modelo mais próximos ao homem e compartilha estratégias de desenvolvimento e doenças como: hipertensão, diabetes, câncer, osteoporose glaucoma e outros, o que possibilita a compreensão de mecanismos para a busca de medicamentos para o tratamento de tais doenças (Müller & Grossniklaus, 2010).

Antes mesmo da descoberta e formalização do terceiro domínio da vida, organismos do domínio Archaea tornaram-se fundamentais para o entendimento da evolução da vida na Terra e ainda, uma vez que alguns desses vivem em ambientes extremos, forneceram subsídios para a comunidade de astrobiologia expandirem seus horizontes na busca de vida extraterrestre (Cavicchioli, 2011). Dentre os exemplos de organismos modelo em estudo presente neste domínio archaea, pode-se incluir *Methanococcus jannaschii, Sulfolobus solfataricus, Pyrococcus furiosus, Haloferax volcanii,* entre outros, com diversos trabalhos na literatura que buscam compreender habilidades não usuais, como sobreviver em condições extremas ou capacidade de gerar metano sob baixas concentrações de oxigênio (Farkas *et al.,* 2013) (Cavicchioli, 2011).

O organismo modelo *Halobacterium salinarum* NRC-1, também membro do domínio Archaea, está incluído neste contexto, diversas caracterizações e análises como perturbações ambientais (Baliga *et al.,* 2004)(Kaur *et al.,* 2006)(Whitehead *et al.,* 2006)(Schmid *et al.,* 2007), caracterização da estrutura do transcritoma (Koide *et al.,* 2009a), interação entre fatores de transcrição e DNA (Facciotti *et al.,* 2007), caracterização proteômica (Van *et al.,* 2008), têm sido realizadas de forma

a contribuir para seu entendimento como um todo (Baliga *et al.,* 2004) (Koide *et al.,* 2009a). Apesar das contribuições e do significante avanço nos estudos prévios relacionados ao organismo *H. salinarum* NRC-1, pouco se sabe sobre todas as moléculas de RNAs não-codificadores (ncRNAs) presentes em seu genoma. Mesmo o modelo computacional para a predição de mudanças em genes reguladores de transcrição (*Environment And Gene Regulatory Influence Network – EGRIN*) (Bonneau *et al.,* 2007) (Books *et al.,* 2014) não inclui informações sobre o papel desempenhado por ncRNAs na rede de regulação.

Sabe-se que ncRNAs estão envolvidos em um amplo conjunto de processos biológicos e atuam em diferentes níveis de processamento que incluem, regulação da transcrição, replicação, modificação e processamento de RNAs, estabilidade e tradução de mRNAs e ainda na degradação de proteínas (Storz, 2002). Devido sua importância, muitos trabalhos têm sido desenvolvidos com o objetivo de identificar e caracterizar essa classe de moléculas de forma a tornar possível a compreensão dos seus diversos mecanismos de regulação (Mattick, 2009).

Abordagens computacionais desenvolvidas para a identificação de ncRNAs procuram considerar propriedades inerentes de tais moléculas, tais como: conservação de sequência e estrutura (Lu *et al.,* 2011) (Washietl *et al.,* 2005), tamanho da sequência, expressão dos transcritos (Lagemberger *et al.,* 2010) (Leung *et al.,* 2013), motivos funcionais conhecidos (Gautheret & Lambert, 2001) (Chang *et al.,* 2013), entre outros. Infelizmente, apesar da existência de múltiplas metodologias para a busca de ncRNAs, é inviável confiar somente em tais ferramentas disponíveis, como estratégia para determinar possíveis regiões candidatas, pois não existe uma abordagem capaz de generalizar e englobar todas as particularidades presentes em moléculas de ncRNAs.

Outra descoberta relativamente recente refere-se à interação de moléculas de ncRNA com proteínas para desempenhar funções regulatórias em nível pós-transcricional (Straub *et al.,* 2009) (Fischer *et al.,* 2011) (Stortz *et al.,* 2011). Proteínas da família Sm estão presentes nos três domínios da vida e são elementos chave na rede de regulação. Assim,

identificar seus parceiros de interação torna-se um desafio promissor para a descoberta e caracterização dos papeis funcionais exercidos por moléculas de ncRNA.

1.1 *Halobacterium salinarum*

Carl Woese (Woese & Fox, 1977) ao revisitar o problema de classificação taxonômica aproveitou as técnicas de sequenciamento de ácidos nucléicos emergentes e propôs uma nova perspectiva sobre a evolução da vida. A escolha por pequenas subunidades de RNAs ribossômicos (*small subunit ribosomal RNA - SSU rRNA*), como uma assinatura molecular foi visionária, uma vez que estas apresentam conservação estrutural e de sequência. Diferentes trechos da molécula de rRNA possuem variações nas taxas de substituição de bases, o que permite uma análise filogenética mais precisa (Allers & Mevarech, 2005) (Cavicchioli, 2011). As implicações desses resultados conduziram para três distintas divisões taxonômicas e foram posteriormente formalizadas como os três domínios da vida, Bacteria, Archaea e Eucarya.

Particularidades sobressalentes dos microorganismos do domínio Archaea compartilham características dos outros dois domínios da vida: por um lado, mecanismos de processamento das informações genéticas são semelhantes aos encontrados em eucariotos (Hickey *et al.,* 2002) (Albers & Meyer, 2011), e por outro, a simplicidade genômica, metabolismo e organização celular são semelhantes a bactérias (Cavicchioli, 2011) (Bell & Jackson, 1998). Outras características não usuais também determinam a subdivisão do domínio Archaea como a capacidade de metanogenesis e a não evidência de qualquer membro patógeno para animais ou plantas (Cavicchioli, 2011).

Halobacterium salinarum é um organismo modelo unicelular presente no domínio Archaea. Diversos trabalhos têm evidenciado sua importância na compreensão de diferentes mecanismos celulares, com estudos envolvendo diferentes perturbações ambientais (Baliga *et al.,* 2004) (Kaur *et al.,* 2006) (Whitehead *et al.,* 2006) (Schmid *et al.,* 2007),

caracterização da estrutura do transcritoma (Koide *et al.,* 2009a), interação entre fatores de transcrição e DNA (Facciotti et al, 2007), caracterização proteômica (Van *et al.,* 2008) entre outros. *H. salinarum* possui um genoma pequeno de ~2,6Mbp, com ~2400 genes, em uma organização genômica compacta, com poucas regiões intergênicas (Ng *et al.,* 2000). Existem diversas motivações para o desenvolvimento de pesquisas com esse organismo as quais incluem, por um lado, a relativa facilidade de cultivo e manipulação em laboratório com ciclo de vida curto (1 a 2 dias de *doubling time* Müller & DasSarma, 2004) e por outro, como citado anteriormente, contribuições de diversos trabalhos com modelos preditivos e quantitativos para estudos em Biologia Sistêmica (Koide *et al.,* 2009a), como por exemplo, um modelo para predizer a influência dos genes na rede de regulação (*Environment and gene-regulatory influence network - EGRIN*) (Bonneau *et al.,* 2007) (Brooks *et al.,* 2014).

H. salinarum pode ser encontrado naturalmente em ambientes com elevada concentrações de sal (~4,5 M), aproximadamente dez vezes superior à concentração de sal do mar, como salinas e lagos onde observa-se a cor púrpura. Possuem vesículas de gás que permite às células flutuarem próximo à superfície da água, que favorecem a sobrevivencia em ambientes com condições de pouco oxigênio. A cor púrpura é devido aos pigmentos na proteína de membrana bacteriorodopsina (Cavicchioli, 2011). Estudos dessa proteína também indicam um interesse por aplicações biotecnológicas com seu uso em dispositivos optoeletrônicos (Oren, 2010) (Walczak *et al.,* 2011).

1.2 RNAs não-codificadores

A descoberta e caracterização funcional de RNAs não codificadores (ncRNAs) é fundamental para a compreensão dos mecanismos de regulação da expressão gênica. Desde o seu achado, moléculas de ncRNAs têm sido observadas como elementos chave em uma grande variedade de processos que incluem, regulação da transcrição, replicação, processamento e modificação de RNA, estabilidade e tradução de RNAs

mensageiros (mRNA) e até mesmo na degradação de proteínas (Storz, 2002) (Babski *et al.,* 2014). Devido a sua importância é possível observar um crescimento de trabalhos na literatura que envolve pesquisas que procuram caracterizar e identificar essas moléculas regulatórias (Mattick, 2009).

Presentes nos três domínios da vida, diversas classes de ncRNAs têm sido descritas nos últimos anos que se diferenciam em tamanho da sequência, especificidades de interação com outras moléculas, organismo, entre outros. Em eucariotos, miRNAs (microRNAs) são pequenas moléculas de RNA com aproximadamente 20 – 25 nucleotídeos. Muitos são evolutivamente conservados e derivam de regiões intergênicas do genoma. Acredita-se que as funções desempenhadas por miRNAs são principalmente como moduladores de tradução e estabilidade de mRNA, com interações em regiões 3' UTR, atuam também como papel chave em modificação epigenética de cromatina (Mattick, 2003).

Os snoRNAs (*small nucleolar RNAs*), que são classificados principalmente em duas famílias: “C/D box” e “H/ACA” snoRNAs, exercem modificações em rRNAs e tRNAs (Eddy, 2001) (Babski *et al.,* 2014). SnoRNAs também têm sido descritos em Archaeas, pois apresentam correlação quanto ao elevado número de moléculas encontradas em conjunto com temperatura de crescimento ótima para o organismo (Soppa *et al.,* 2009) (Straub *et al.,* 2009).

Em bactérias, pequenas moléculas de ncRNA (*small RNAs – sRNA*) são requeridas para a regulação gênica e atuam funcionalmente de maneira diversa. Em *E. coli*, por exemplo, podem afetar a tradução por obstrução da ligação do ribossomo ou ainda promover a tradução ao impedir a formação de uma estrutura inibidora do mRNA (Massé *et al.,* 2003) (Storz, 2002). É estimado que o genoma de *E. coli* codifique aproximadamente 200 – 300 sRNAs, o que corresponde à cerca de 5% dos genes presentes no organismo (Soppa *et al.,* 2009).

Outra classe de ncRNAs, presente em eucariotos, refere-se a longas cadeias de RNAs não codificadores lnRNA (*long non-conding RNA*) que podem ser definidos, de forma geral, como ncRNAs maiores que 200

nucleotídeos e tipicamente expressos de uma maneira estágio-específico durante o desenvolvimento da célula. São altamente conservados e a maioria possui pequenas ORFs (*open reading frame*). Assim como em genes codificadores, muitos lnRNAs aparentam ser transcritos pela RNA polimerase II e possuem estruturas típicas de pre-mRNA incluindo 5' Cap e cauda poli A+ (Lv *et al.,* 2013).

Alguns ncRNAs, denominados snRNA (*small nuclear RNAs*), são associados a proteínas para a formação de complexos de ribonucleo-protéicos (*Ribonucleo-proteins*) e podem ser encontrados como componentes da maquinaria de *splicing* (Eddy, 2001).

Outra classe de moléculas de ncRNA, presente nos três domínios da vida, é a dos RNAs associados ao início de transcrição (*Trascription start site associated RNAs - TSSaRNAs*) que se localizam próximos a regiões de início de transcrição. Pouco é conhecido sobre as funções desempenhadas por tais moléculas porém, podem estar associadas a regulação epigenética ou ainda como mecanismo regulatório para a prevenção do processo de início de transcrição (Zaramela *et al.,* 2014). Na Figura 1 são ilustrados alguns exemplos de ncRNAs e o contexto geral no qual atuam em diferentes etapas de processamento e regulação da informação genética.

Em Archaea apesar da existência de alguns trabalhos que buscam a identificação e caracterização funcional de moléculas de ncRNAs, sabe-se pouco sobre tais elementos e uma vez que estes têm sido descritos como elementos chave em processos de regulação, o entendimento dos mecanismos celulares torna-se incompleto ou mal compreendidos por não considerar a existência dessas moléculas. Em *H. volcanii* foram identificados 39 sRNAs envolvidos na regulação da expressão gênica atuando em conjunto com a proteína LSm (Straub *et al.,* 2009) além da identificação de outros 150 possíveis ncRNAs (Soppa *et al.,* 2009). Sabe-se também da presença de snoRNAs (*small nucleolar RNAs*) preditos em espécies de *Pyrococcus* e experimentalmente identificados em *Sulfolobus acidocaldarius* (Dennis & Omer, 2005) (Babski *et al.,* 2014). Recentemente foi descrita a existência de transcritos sobrepondo a região senso (*sense*

overlapping transcripts - sotRNAs) de transposases sendo estes diferencialmente expressos sob diversas perturbações ambientais cujos sinais de expressão variam de forma relativa a suas transposases cognatas (Gomes-Filho *et al.,* 2015).

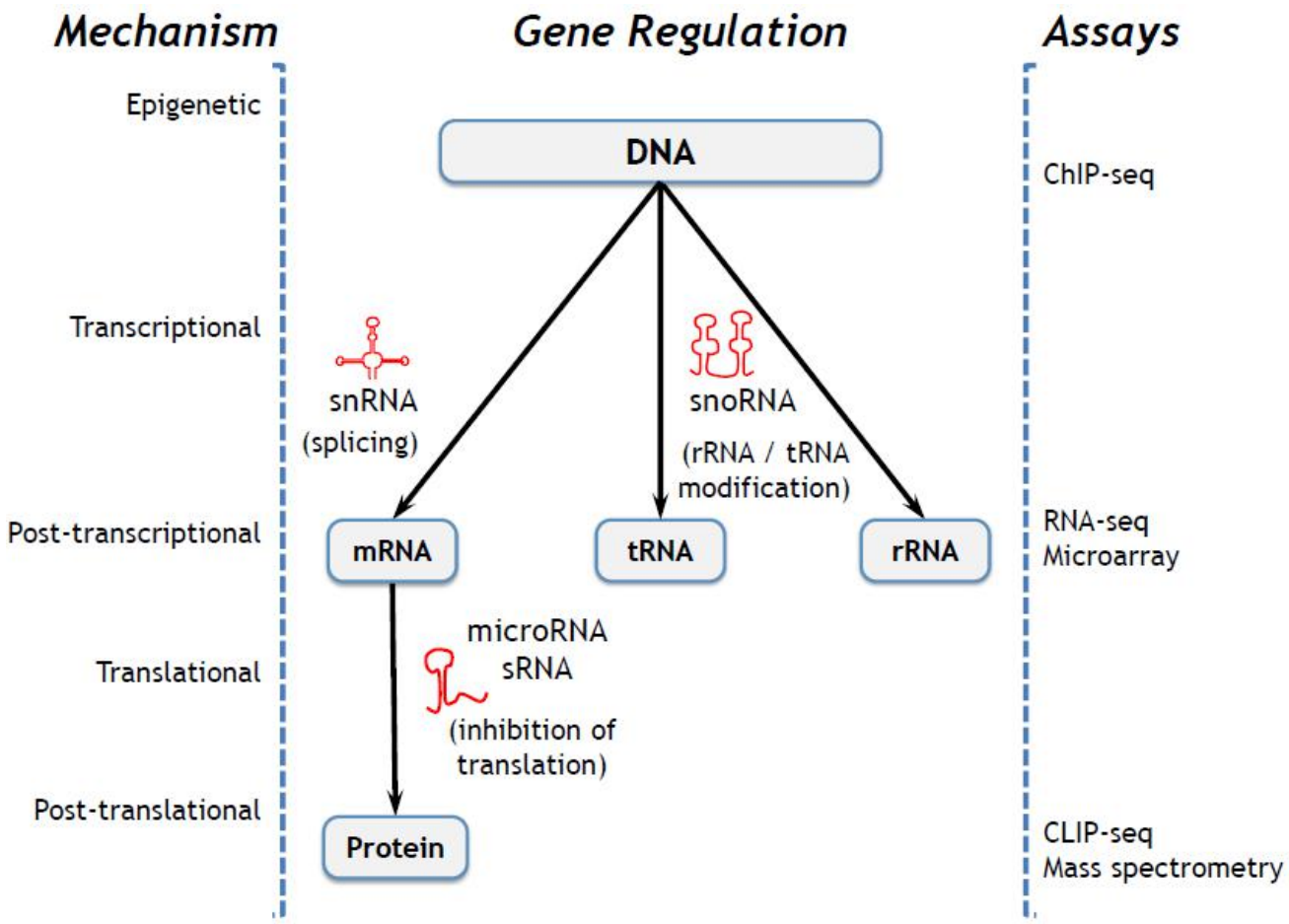


Figura 1 – Visão geral dos diversos níveis de regulação gênica com a atuação de moléculas de ncRNAs em nível pós-transcrição (*post-transcriptional*). Diversos mecanismos de regulação (*Mechanism*) estão presentes na rede de regulação assim como técnicas experimentais (*Assays*) que possibilitam o estudo desses elementos (adaptado de Matsui *et al.,* 2013).

Junto às classes de moléculas de ncRNAs, diversos motivos funcionais têm sido identificados e associados a importantes papéis em processos biológicos. As atividades biológicas desempenhadas por essas regiões são alcançadas pela combinação de estruturas secundárias específicas assim como o padrão da sequência primária. Análises de

motivos funcionais em sequencias de RNA podem fornecer informações úteis sobre os mecanismos regulatórios de tais moléculas (Chang *et al.,* 2013).

Outros trabalhos recentes indicam ainda um crescente interesse na identificação e compreensão dos papéis regulatórios exercidos por pequenas moléculas de RNA não-codificadores (*small RNAs - sRNA*), (Massé *et al.,* 2003) (Straub *et al.,* 2009) (Fisher *et al.,* 2010) que se ligam a proteínas para que este complexo exerça suas funções (Vogel & Luisi, 2011) (Beggs, 2005). Um aspecto importante nessa perspectiva de estudo é que com a análise dessas proteínas, torna-se possível também a identificação e caracterização dos ncRNAs parceiros de interação.

Alguns dos desafios no contexto de ncRNAs compreendem a descoberta, classificação e caracterização desses tipos de moléculas. Neste trabalho voltamos nossos esforços para a identificação de novos ncRNAs presentes em *H. salinarum* NRC-1 e a predição de interações entre esses RNAs e a proteína LSm.

1.3 Interação RNA-Proteína

Devido a uma quantidade considerável de condições sob as quais um organismo está sujeito em seu meio, como estresse causado pela escassez de nutrientes ou oxigênio, mecanismos de regulação surgiram evolutivamente em resposta a essas condições como forma de propiciar um controle sobre o nível de expressão aos quais os produtos funcionais são gerados durante o crescimento das células (Lewin, 2004) (Trun & Trempy, 2003). O processo de regulação envolve diferentes níveis de informações e é alcançado por meio de elementos interconectados que atuam em multi-etapas de processamento, como por exemplo, nos processos de transcrição, pós-transcrição e tradução (Trun & Trempy, 2003). Trabalhos recentes têm focado nas especificidades das funções exercidas por moléculas de RNAs não codificadores (ncRNAs) em uma regulação pós-transcricional (Straub *et al.,* 2009) (Fischer *et al.,* 2010) (Stortz *et al.,* 2011). Em Straub et al., 2009, por exemplo, foram

identificadas 39 pequenas moléculas de RNA (small RNAs – sRNAs) envolvidos na regulação da expressão gênica a partir de sua interação com a proteína *Like-Sm* (LSm). Em Fisher *et al.,* 2010 é confirmada a atuação da proteína LSm em diferentes processos celulares, sendo ainda incluído no trabalho a identificação de diversos tipos de interações com RNAs e outras proteínas. Proteínas da família Sm estão presentes nos três domínios da vida, possuem o domínio Sm conservado estruturalmente e são elementos chave na rede de regulação. Em Bactéria a proteína é denominada Hfq e em Eucariotos e Archeas Sm ou *Like-Sm* (*LSm*). Uma vez que diversos ncRNAs exercem suas funções em conjunto com a proteína, identificar seus parceiros de interação torna-se um desafio promissor para a descoberta e caracterização dos papeis funcionais exercidos por tais moléculas.

Embora métodos experimentais baseados em técnicas de imunoprecipitação, como Rip-chip, PAR-CLIP, HITS-CLIP, sejam capazes de identificar as interações que ocorrem entre RNAs e proteínas de maneira confiável, sua realização requer gastos com pessoas, recursos e equipamentos, além de consumir demasiado tempo (König *et al.,* 2012). O uso de métodos computacionais como ferramenta de apoio para descoberta de novas informações torna-se uma alternativa interessante neste contexto. Abordagens baseadas em técnicas de Aprendizado de Máquina (AM) têm sido aplicadas (Pancaldi e Bähler, 2011) (Muppirala *et al.,* 2011) uma vez que possuem a habilidade de construir modelos de representação a partir de dados de treinamento (Russell e Norving, 2010). Estas aborgens também têm sido aplicadas recentemente na predição de sítios de ligação em proteínas (*Binding sites prediction*) (Terribilini *et al.,* 2006) (Liu *et al.,* 2010), proteínas ligantes a RNA (*RNA-Binding proteins*) (Han *et al.,* 2003) e na predição de interações entre RNA e proteínas (*RNA-Protein prediction*) (Pancaldi e Bähler, 2011) (Muppirala *et al.,* 2011). No entanto, são específicas para um determinado contexto ao qual os dados do organismo modelo *H. salinarum* NRC-1 considerado não se adequa (Pancaldi e Bähler, 2011) ou ainda, como será apresentado posteriormente, os resultados das predições não evidenciam uma robustez

para a identificação dos elementos. Dessa forma, também é apresentado neste trabalho como objetivo secundário o estudo e aplicação de estratégias para a predição de interações entre RNA e proteína presentes no organismo modelo *H. salinarum* NCR-1 em estudo.

1.4 Abordagens computacionais para identificação de ncRNAs

Métodos computacionais desenvolvidos para a identificação de ncRNAs buscam considerar propriedades típicas dessas moléculas, as quais incluem conservação da sequência, estabilidade da conformação estrutural (Lu *et al.,* 2011) (Washietl *et al.,* 2005) tamanho da sequência, informações de expressão dos transcritos (Langenberger *et al.,* 2010) (Leung *et al.,* 2013), motivos (*motifs*) funcionais conhecidos (Gautheret *&* Lambert, 2001) (Chang *et al.,* 2013), entre outras. O uso de abordagens computacionais como auxílio para a identificação e caracterização de moléculas de ncRNAs têm se apresentado como uma alternativa interessante por fornecer subsídios significativos à validação experimental de potenciais candidatos.

Como mencionado, abordagens computacionais tomam diferentes tipos de dados para a predição de novos elementos e essas informações características originam-se de tecnologias e metodologias até então disponíveis. A plataforma RNAspace (Cros *et al.,* 2011), por exemplo, provê uma ferramenta integrada, e de fácil uso, para a busca e anotação de ncRNAs baseadas em algumas das características mencionadas como, similaridade de sequência e estrutura. Outras abordagens exploram dados experimentais mais específicos como sequenciamento de bibliotecas de pequenas moléculas de RNA (*small RNA-Seq - sRNA-Seq*). Uma vez que que o sequenciamento é realizado considerando apenas RNAs pequenos, de tamanhos de até 200 pares de bases, por exemplo, é esperado que características particulares dessa classe de moléculas estejam presentes nos *reads* sequenciados. Ambas abordagens Dario (Fasold *et al.,* 2011) e CoRAL (Leung *et al.,* 2013) utilizam informações de dados de *sRNA-Seq.*

Essas informações referem-se aos *reads* mapeados ao longo do genoma. Na abordagem Dario é utilizada as informações sobre o agrupamento (*cluster*) dos *reads* mapeados. Os autores esperam que as propriedades estruturais de moléculas de ncRNAs definam um perfil para o mapeamento e que a partir das propriedades extraídas do agrupamento, seja possível a criação do modelo de AM para que o mesmo seja utilizado na predição de novos elementos. A abordagem CoRAL, utiliza informações sobre a distribuição do tamanho dos RNAs processados, abundância de transcritos anti-senso, distribuição das posições 5' e 3' de cada *read*, composição de nucleotídeos e energia livre mínima (*minimum free energy - MFE*) predita com a ferramenta RNAfold (Hofacker *et al.,* 1994).

O mecanismo de busca da abordagem denominada smyRNA (Salari *et al.,* 2009) é baseado em certos trechos da sequência primária (*motifs*) que são importantes para determinar a estrutura da molécula. Esses trechos possuem uma distribuição diferenciada em relação as demais regiões do genoma e a frequência de suas ocorrências nos dados treinamento são consideradas no modelo. Para averiguar a habilidade preditiva da abordagem foi aplicada uma estratégia que consiste em treinar o modelo com o ncRNAs conhecidos ao longo do genoma e em seguida embaralhar as bases de todo o genoma deixando os trechos utilizados no treinamento intactos. Em seguida, considerando o genoma modificado, as predições foram então realizadas e avaliadas. Em outros resultados, os autores tomam como dados de treinamento ncRNAs conhecidos em *E. coli* e aplicam em outros sete organismos. De acordo com os resultados o método proposto foi capaz de identificar ao menos 69% dos ncRNAs conhecidos em *Salmonella enterica* (domínio *Bacteria*) e 90% em *Cyanophora paradoxa cyanelle* (domínio *Eukaryota*).

Bao *et al.,* 2012 pesquisara sobre o aprimoramento da abordagem anterior (smyRNA) por incluir informações sobre estrutura secundária e algumas considerações sobre o conteúdo GC da sequencia além dos motivos (*motifs*) presentes na sequencia primária. Na comparação realizada com smyRNA, o método proposto, denominado ncRNAscout, identificou cerca de 88% dos ncRNAs conhecidos contra 73%. Ambos os

métodos foram aplicados a 4 diferentes genomas. Uma vez que ncRNAscout utiliza basicamente o mesmo conceito de smyRNA, os parâmetros adicionais sobre estrutura secundária e conteúdo GC tornam a primeira mais acurada.

Num trabalho recente baseado em Aprendizado de Máquina (AM), Lertampaiporn *et al.,* 2014 usam como representação dos exemplos de treinamento *features* baseadas em estrutura, propriedades da sequência, trincas particionadas de estruturas secundárias, informações sobre a robustez estrutural e pareamento de nucleotídeos, totalizando 369 atributos. Os autores aplicam diversos classificadores e a técnica baseada em combinação de árvores de decisão (*Ramdom Forest*) alcançou melhores resultados na relação entre a taxa de falsos positivos e sensitividade. Após a realização de um procedimento de seleção de atributos, 20 *features* resultantes foram destacadas como mais informativas para os modelos de AM. Quase todas as categorias de atributos, ou seja, atributos sobre estrutura secundária, propriedades das sequências, motivos (*motifs*) e robustez estrutural estiveram presentes nesse sub-grupo. Ao que parece, informações sobre as trincas geradas a partir de trechos da representação sobre estrutura secundária não se mostram informativas para o problema. Os autores também definem um novo tipo de atributo gerado a partir de uma regressão logística. O modelo logístico foi baseado em 5 atributos, considerados significantes, os quais envolvem similaridade de sequencia e robustez estrutural.

Diversas avaliações foram feitas como forma de evidenciar a capacidade de predição do método proposto. Inicialmente os autores aplicaram a abordagem em dados de *E. coli* e obtiveram resultados de sensitividade e especificidade melhores que outro método baseado em Redes Neurais. Os autores também compararam o método proposto com as abordagens smyRNA e ncRNAscout e obtiveram uma percentagem maior de elementos identificados corretamente nos quatro casos testados. A metodologia foi utilizada em outra avaliação que consistiu em analisar a performance das predições perante todo o genoma através de janelas genômicas. Dessa forma, a partir de vários trechos de tamanho fixo, e

com sobreposições de nucleotídeos, todo o genoma é percorrido e aplicado ao modelo. De acordo com os resultados apresentados ao menos 78% dos ncRNAs conhecidos puderam ser identificados.

Outro trabalho recente (Panwar *et al.,* 2014) propôs explorar variações na composição de di, tri, tetra e penta-nucleotídeos como conjunto de atributos para técnicas de AM para a identificação de ncRNAs. Em uma etapa posterior, aplicam uma metodologia que representa a estrutura secundária predita em grafos e extração de propriedades para a classificação dos ncRNAs identificados. Os resultados apresentados indicam que apesar de utilizarem uma representação simples, a metodologia foi capaz de obter melhores resultados que outra abordagem que também é baseada na composição de nucleotídeos e ainda inclui outras informações sobre estrutura e propriedades da sequência.

Ao visar a integração de diversos tipos de informações, Lu *et al.,* 2011 aplicam abordagens baseadas em AM para a identificação de novas moléculas de RNAs não codificadores em *C. elegans.* Com resultados que indicam alta acurácia do modelo ao integrar diversas características de dados, os autores conseguem identificar diversos elementos novos e validam com métodos experimentais alguns desses elementos. A abordagem foi denominada incRNA (*integrated ncRNA finder*) e de acordo com o método proposto, para a criação do modelo de AM inicialmente foi realizado um alinhamento com o genoma de *Caenorhabditis briggsae*, organismo evolutivamente próximo, como forma de obter regiões com informações já conhecidas. Essas regiões definem as classes (rótulos de cada exemplo) que foram consideradas no modelo, sendo estas: regiões codificadoras (*Coding DNA Sequences - CDS*), regiões não traduzidas (*Untranslated Regions - UTR*), ncRNAs conhecidos e regiões intergênicas. Dessa forma, ao considerar as regiões conservadas oriundas do alinhamento genoma a genoma, trechos com informações conhecidas em *C. briggsae* foram transferidos para *C. elegans*. Essas regiões foram então usadas como dados de treinamento na criação do modelo de AM. Para a predição da estrutura secundária dessas regiões, cada trecho foi subdividido em tamanho de no máximo 150 bases.

Cada trecho do genoma anotado foi então representado por um conjunto de nove atributos, que incluem: dados de expressão gênica, propriedades da sequência primária e informações sobre estrutura predita. Para os dados de expressão gênica foram consideradas bibliotecas de pequenos RNAs (small RNA-seq), poli A+ RNA-seq, microarranjos de RNA total e microarranjos poli A+. Como propriedades da sequência foram consideradas: conteúdo GC da região, conservação do DNA e proteína. Por fim, como informações da estrutura são consideradas a estabilidade e conservação da estrutura predita. De acordo com os resultados, o classificador *Random Forest* obteve a melhor acurácia na validação cruzada em comparação a outros classificadores disponíveis na ferramenta WEKA (Hall *et al.,* 2009). Também foi considerado um conjunto de teste independente para a escolha do modelo. Apesar do desbalanceamento das classes em relação ao número de exemplos em que cada uma possui, grande parte de seus respectivos exemplos foram classificados corretamente e dessa forma, elementos da classe ncRNA foram bem separados das demais classes. Dados de regiões sem informações foram então aplicados sobre o modelo e 7237 elementos foram preditos como candidatos a ncRNAs. Para validação das predições, diversas estratégias foram adotadas como: abordagens experimentais, medidas de conservação, predição de sítios de ligação de polimerase e fatores de transcrição, além do uso de dados independentes.

A partir dos trabalhos da literatura podemos conhecer algumas das informações utilizadas na construção de métodos computacionais para o problema de identificação de ncRNAs. Ao que parece, as abordagens têm sido desenvolvidas e aplicadas em diferentes tipos de organismos, presentes nos três domínios da vida, não se restringindo portanto a informações mais específicas no que refere a propriedades particulares de um ou outro organismo. Mesmo seguindo características, à princípio, gerais, os resultados alcançados indicam que o comportamento dos algoritmos são robustos para a aplicação em diversos organismos. Apesar da diversidade de ncRNAs presentes na célula, muitas características apresentam-se comuns aos diferentes tipos de ncRNAs. Dessa forma,

informações condizentes a estrutura da molécula, por exemplo, tem se apresentado relevante e permeia as considerações de diversas abordagens.

Constatamos então que métodos baseados em AM têm sido utilizados como metodologias para a identificação de moléculas de ncRNAs e apresenta resultados promissores. Dessa forma, buscamos explorar tal perspectiva para a adequação e aplicação de estratégias existentes para a identificação de trechos candidatos a pertencerem à classe ncRNA no genoma de *H. salinarum* NRC-1.

1.5 Abordagens computacionais para predição de interações RNA-Proteína

Com o surgimento de experimentos em larga escala para análises de proteínas ligantes à RNAs, houve também um aprimoramento no conhecimento sobre as informações relativas aos padrões de interações entre moléculas. Dados experimentais puderam ser utilizados na construção de modelos computacionais visando a tentativa de contornar o alto custo e tempo gasto em tais experimentos laboratoriais. No entanto, existem ainda poucos trabalhos na literatura que propõem abordagens computacionais para a identificação de interações RNA-Proteína, uma vez que pouco se sabe sobre os mecanismos de interação e por ser um tema recente de pesquisa (Muppirala *et al.,* 2013).

Dentre as abordagens computacionais para a identificação dos parceiros de interação RNA-proteína, observamos em alguns trabalhos propostos o uso de metodologias baseadas em Aprendizado de Máquina (AM). Pancaldi & Bähler, 2011 aplicam os classificadores Random Forest e SVM para a predição dos possíveis parceiros de interação. Os exemplos considerados no processo de treinamento são oriundo de dados experimentais de imunoprecipitação realizado especificamente para a identificação de interações entre mRNAs e proteínas presentes em levedura (*Saccharomyces cerevisiae*). Os autores reúnem mais de 100 características para uso como atributos de AM, que buscam descrever

diversas propriedades relativas aos exemplos de treinamento. De acordo com os resultados apresentados o classificador Random Forest obteve resultados ligeiramente melhores que o classificador SVM, com uma acurácia de 70% comparado a 69%. A principal limitação da abordagem proposta está na obtenção de todas as informações consideradas como atributo de AM, onde nem todas podem estar disponíveis para aplicação em outros organismos.

Muppirala *et al.,* 2011 sugerem que somente informações relativas a sequência primária são suficientes para atingir resultados próximos ao da abordagem de Pancaldi e Bähler, 2011. Os autores consideram como atributos de AM apenas informações sobre a composição de aminoácidos e ribonucleotídeos e ao aplicarem sobre o mesmo conjunto de dados, utilizados em Pancaldi e Bähler, obtiveram resultados de acurácia muito próximos. Buscamos explorar as considerações dessa abordagem em estudo, denominada RPISeq, reproduzindo a metodologia e analisando algumas variações. Dessa forma, mais detalhes da abordagem serão apresentados em outra seção.

Bellucci *et al.,* 2011 utilizaram informações físico-químicas como: informações sobre pontes de hidrogênio, interações van der Waals e estrutura secundária para o cálculo do perfil de interação entre os pares RNA-Proteína. Os autores apresentam um modelo físico constituído de uma medida denominada *discriminative power* (DP) para determinar a propensão de interação. Dados de complexos RNA-proteína foram extraídos do Banco de Dados de Proteínas (*Protein Data Bank - PDB*) (Berman *et al.,* 2000) que inclui diversos tipos de proteínas (ribossômicos, proteínas de transporte, RNA polimerase, sintetases, ligases, entre outras) assim como diversas classes de RNA (rRNAs, snoRNAs, tRNAs) utilizados como exemplos de treinamento. O PDB é um repositório de informações sobre estruturas tridimensionais de moléculas biológicas. Diversos tipos de informações podem ser obtidos através do *website* (http://www.rcsb.org) onde até a presente data estão disponíveis 113130 estruturas. A abordagem denominada catRAPID obteve uma acurácia de 89% nos resultados da predição de um conjunto de dados independente.

Em um trabalho recente, Cheng *et al.,* 2015 discutem um problema comum apresentado em outros trabalhos desenvolvidos para a predição de interações RNA-Proteína. Têm-se assumido que exemplos negativos de pares RNA-Proteína são aqueles cujos pares positivos originais foram trocados de forma aleatória por outro elemento, ou seja, a partir de um embaralharamento dos exemplos positivos. Os autores mostram o uso de um classificador SVM que considera apenas dados positivos e exemplos não rotulados (exemplos sem a definição de positivo ou negativo). Como parte dos dados de treinamento os autores utilizam exemplos do PDB (Berman *et al.,* 2000) e da base NPInter (Wu *et al.,* 2006), que consiste de exemplos de pares de interação obtidos experimentalmente em diversos organismos. De acordo com os resultados, a abordagem proposta foi capaz de identificar a maioria dos exemplos, com uma acurácia média de 91% e para exemplos não rotulados verificou-se que alguns exemplos validados experimentalmente, e disponíveis na base de dados NPInter, foram preditos corretamente.

Outro trabalho recente descreve a abordagem denominada RPI-Pred (Suresh *et al.,* 2015) para a predição de pares de interação RNA-Proteína. Os autores apresentam um método de AM baseado em SVM. O classificador é construído à partir de exemplos de treinamento extraídos das bases de dados Nucleic Acid Database - NDB (Berman *et al.,* 1992) e Protein-RNA Interface Database - PRIDB (Lewis *et al.,* 2011). O PRIDB é um base de dados que oferece um acesso simples as estruturas do PDB referentes à complexos RNA-Proteína, permitindo a obtenção de diversos conjuntos de exemplos. Além de informações sobre a sequencia primária, são considerados como atributos de AM a representação de estruturas 3D de proteínas e informações de estrutura secundária de RNA como: Stem, Hairpin, Loop, Bulges e Internal loop. No total, 132 features com frequência dos aminoácidos, ribonucleotídeos e informações sobre estrutura são utilizados na representação dos exemplos. Para proteínas ou RNAs sem estruturas definidas foram realizadas predições. Os resultados alcançados indicam uma melhor acurácia quando comparado aos resultados de Muppirala *et al.,* 2011 para os exemplos positivos

disponíveis na base de dados NPInter (Wu *et al.,* 2006). Dentre outras considerações é observado que a abordagem é amplamente influenciada pelas informações sobre estrutura das moléculas e predições ruins podem prejudicar os resultados da predição.

Como observado em alguns trabalhos da literatura, desenvolvidos para a predição de interações RNA-Proteína, poucas informações sobre os mecanismos de interação têm sido utilizadas. Usualmente, cada par RNA-proteína é representado em atributos de AM com informações relativas as propriedades individuais e gerais das moléculas. Dentre essas informações, propriedades físico-químicas, estruturais ou da sequência primária têm sido utilizadas. Mesmo abordagens com informações baseadas apenas na sequência primária têm mostrado bons resultados na predição dos pares de interação.

2 Objetivos

Algumas regiões ao longo do genoma de *H. salinarum* NRC-1 foram sugeridas e identificadas como trechos pertencentes à classe de moléculas de ncRNAs (Koide *et al.,* 2009b) (Zaramela *et al.,* 2014) (Gomes-Filho *et al.,* 2015) porém, acreditamos na existência de muitas outras moléculas, uma vez que ao observarmos dados de expressão dos transcritos constatamos que diversas regiões expressas permanecem sem anotações disponíveis. Esta é a hipótese científica original testada na presente Tese. O objetivo principal da Tese é o de adaptar e aplicar diferentes metodologias para a predição de novas moléculas de RNAs não-codificadores possivelmente presentes no organismo modelo *Halobacterium salinarum* NRC-1 através de uma análise *in silico.* Além de contribuir na identificação, é objetivo secundário da Tese a caracterização básica dessa importante classe de elementos reguladores, por meio de predição e organização de informações sobre interação com a proteína chaperona LSm.

3 Materiais e métodos

3.1 Aprendizado de Máquina

O conceito Aprendizado de Máquina (AM) possui várias definições. Uma dessas definições é apresentada em Michell, 1997 como sendo: “A capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência” (Faceli *et al.,* 2011). O princípio geral seguido por algoritmos que utilizam o conceito de AM é ilustrado na Figura 2.

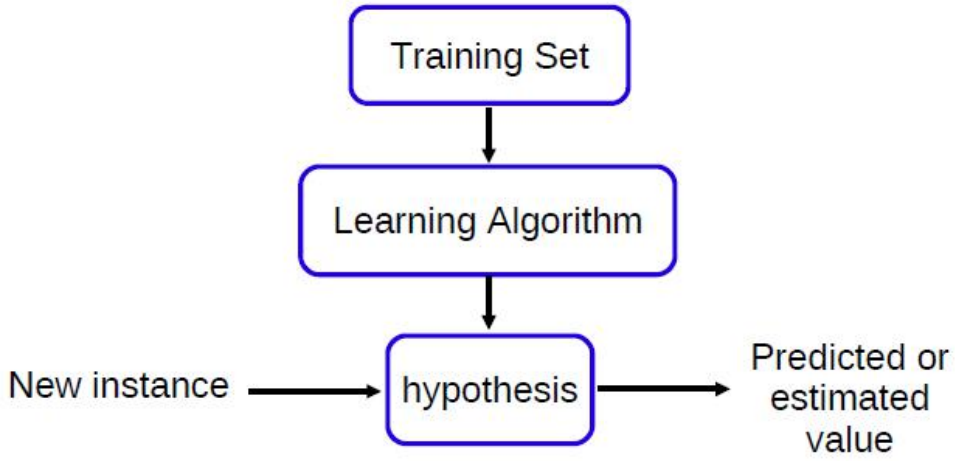


Figura 2 – Cenário típico de algoritmos baseados em AM (Modificado de Mitchell, 1997).

Com base em um conjunto de dados de treinamento (*Traning Set*), também chamado de conjunto de exemplos, um algoritmo de aprendizado de máquina (*Learnig Algorithm*) é aplicado e interpretando o conjunto de atributos (também chamado de características ou *features*) que descrevem os dados de treinamento, uma hipótese (*Hypothesis*) sobre os mesmos é determinada. Isso significa que o padrão encontrado no conjunto de exemplos fornecido ao algoritmo é representado por um modelo matemático e por meio deste, é possível realizar uma inferência com um novo dado (*New instance*) tendo como resultado um valor predito ou estimado (*Predicted or estimated value*) como saída do algoritmo.

Em um aprendizado supervisionado (*Supervised Learning*) os mesmos princípios descritos anteriormente são válidos e para cada exemplo do conjunto de treinamento um rótulo é associado. Esse rótulo determina a classe ao qual cada exemplo pertence, por exemplo, um dos atributos especifica que determinado exemplo é da classe "codificador" ou da classe "não-codificador". Formalmente, a tarefa de um aprendizado supervisionado é:

> Dado um conjunto de treinamento com N exemplos de com suas respectivas classes:
> $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N),$
> onde cada y_i foi gerado por uma função conhecida $y = f(x),$
> Descubra uma função h que se aproxime da verdadeira função f.

A função h é uma hipótese, o aprendizado se refere a uma busca dentro do espaço de possíveis hipóteses por aquela que melhor represente os dados de treinamento e generalize bem o conhecimento para a predição de novos elementos (Russel e Norvig, 2010).

Por outro lado, em um aprendizado não supervisionado (*Unsupervised Learning*) os exemplos não possuem um rótulo explícito e são comumente tratados como dados para técnicas de agrupamento (*clustering*). O objetivo dessas técnicas é encontrar uma estrutura de grupos que compartilham alguma característica ou propriedade relevante para o domínio do problema em estudo (Faceli *et al.,* 2011). Por exemplo, pode ser desenvolvido o conceito de "dias com bom tráfego" e "dias com mau tráfego" por meio desse aprendizado mesmo sem ser especificado exemplos para cada um desses dias (Russel e Norvig, 2010).

Outra consideração é quanto à saída apresentada pelas técnicas de aprendizado, quando o resultado da predição é uma determinada classe, por exemplo, "codificador" ou "não codificador" o problema é característico de classificação. Por outro lado, quando a saída da predição é um valor numérico o problema é característico de regressão.

Na aprendizagem indutiva, toda técnica de AM procura por uma

hipótese, no espaço de hipóteses possíveis, que melhor se ajuste aos dados de treinamento e que seja capaz de descrever em uma forma generalizada as relações entre os exemplos. Para isso, cada algoritmo utiliza uma preferência ou viés (*bias*) para forma de representação e uma preferência para a forma de busca. O viés de representação descreve a hipótese induzida e pode restringir o conjunto de hipóteses. Como exemplo de um viés de representação, árvores de decisão utilizam uma estrutura em árvore em que cada nó interno é representado por uma pergunta referente ao valor do atributo e cada nó externo está associado a uma classe. O viés de busca indica a forma como o algoritmo busca a hipótese. Por exemplo, também considerando indução em árvores de decisão, o algoritmo ID3, tem preferência de busca por árvores com poucos nós (Faceli *et al.,* 2011).

Nas próximas seções, apresentaremos brevemente uma introdução a alguns dos classificadores utilizados durante as atividades desenvolvidas, baseados principalmente em (Faceli *et al.,* 2011), (Russel e Norvig, 2010), (Hastie *et al.,* 2009) e (Bishop, 2006) (Mitchell, 1997).

3.1.1 Árvores de decisão

Árvores de decisão é um método simples e muito utilizado em aprendizado de máquina. Formalmente, uma árvore de decisão é representada por um grafo acíclico direcionado em que cada nó pode ser um nó raiz, que indica o início da árvore, nós de divisão, com dois ou mais sucessores, ou um nó folha, que indica o rótulo de saída (Figura 3). Condições são formadas, considerando os nós de divisão, envolvendo os valores do domínio de um atributo em particular e operadores condicionais (por exemplo, =, >, etc.). Um nó folha possui um valor presente no domínio das classes. No Exemplo da Figura 3, o atributo "*outlook*" possui três sucessores, na condição em que o valor de "*outlook*" é igual a "*Sunny*" outro nó é então avaliado com valores do atributo "*humidity*". Os domínios de rótulos que a saída da árvore assume nesse exemplo correspondem a "*Yes*" ou "*No*", indicando a decisão de jogar ou

não uma partida de tênis (Faceli *et al.,* 2011), (Russel e Norvig, 2010).

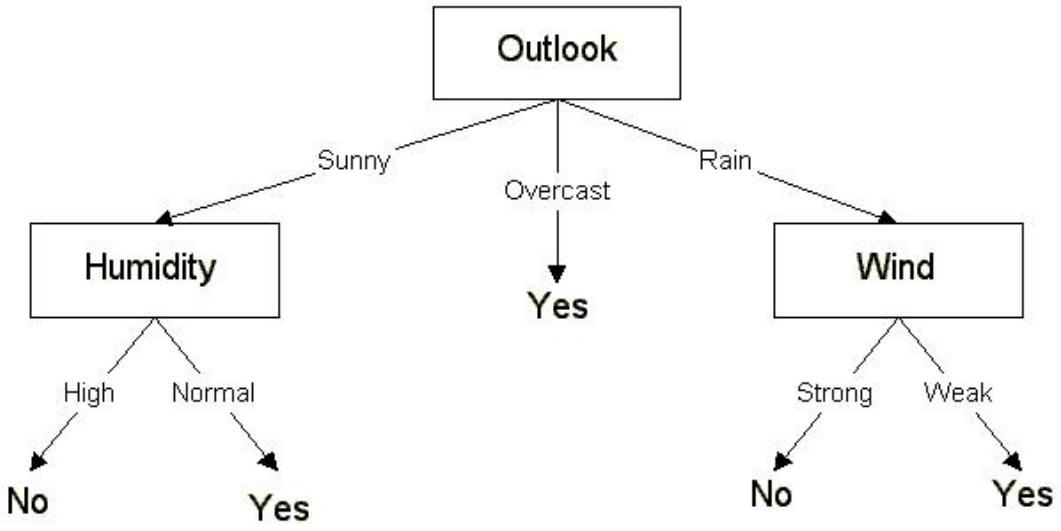


Figura 3 – Exemplo de árvore de decisão para decidir se deve ou não jogar uma partida de tênis (Extraído de Russel e Norvig, 2010).

O processo de indução em uma árvore de decisão consiste em construir, a partir de um conjunto de dados de treinamento, a estrutura da árvore de forma que esta seja consistente com o padrão dos dados e que seja menor possível. Para isso, é considerado um grau de importância para cada um dos atributos (no exemplo, *outlook, humidity, wind*) e várias medidas podem ser utilizadas como: escolha aleatória, atributos com mais ou menos valores ou ainda baseadas no grau de impureza como Entropia (Equação 1), Gini e Erro de Classificação (Faceli *et al.,* 2011), (Russel e Norvig, 2010). Em Teoria da Informação, a medida de entropia, mais especificamente Entropia de Shannon, pode ser definida como:

$$Entropy(S) = -\sum_{j=1}^{k} p(C_j, S) \times \log_2(p(C_j, S)) \quad (1)$$

onde: $p(C_j, S)$ é a frequência relativa da classe j no conjunto S.

k é número total de classes.

Dessa forma, ao considerar o grau de impureza obtido com o particionamento gerado a partir dos valores presentes cada atributo o ganho de informação no conjunto S refere-se a redução esperada na entropia quando se sabe o valor do atributo A (Equação 2).

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (2)$$

onde: *Values(A)* refere-se ao conjunto de todos os valores possíveis do atributo A.

Sv é um subconjunto de *S*, no qual o atributo *A* possui valor *v*.

Após a escolha dos atributos e construção da estrutura, um processo de poda pode ser aplicado com o objetivo de reduzir a influência de ruídos presentes nos dados e aumentar o poder de generalização do algoritmo. Esse procedimento pode ser realizado durante o processo de indução ou ainda ao final sobre a árvore final gerada, ao considerar o desempenho da árvore para a classificação dos exemplos. Dentre as vantagens dessa técnica, é possível interpretar a hipótese gerada a partir das regras de decisão obtidas. O processo de indução possui um baixo custo computacional e ainda é possível averiguar quais atributos são importantes no problema. No que se refere as desvantagens, sabe-se que pequenas variações nos dados de testes podem produzir árvores com diferentes desempenhos, o que torna a técnica instável. Para atributos com valores desconhecidos, é necessário um tratamento especial do algoritmo, uma vez que é necessário a definição dos valores para definir por qual ramo seguir (Faceli *et al.,* 2011), (Russel e Norvig, 2010).

3.1.2 Random forest

Random forest utiliza um conceito interessante presente na técnica denominada *bagging ou bootstrap aggregation,* que consiste em reduzir a o viés de variância da hipótese gerada. Para isso, os dados de treinamentos são selecionados por amostragem com reposição e vários classificadores são gerados. Espera-se que a variabilidade aleatória dos classificadores seja reduzida com esse procedimento e uma superfcie de decisão mais complexa pode ser gerada (Faceli *et al.,* 2011) (Hastie *et al.,* 2009).

A técnica Random forest consiste em produzir uma coleção de árvores de decisão correlatas e então ponderar as saídas por um sistema de voto. No algoritmo da Figura 4 são apresentados os principais procedimentos para a construção de modelo baseado na técnica Random forest. No algoritmo, *B* árvores de decisão são criadas a partir os dados de bootstrap gerado. Um subconjunto de *m* variáveis é escolhido de forma aleatória e ao final, a combinação é realizada com um sistema simples de votos, em que a classe predita pela maioria dos *B* classificadores é então escolhida como o resultado da classificação.

Algoritmo: Random forest para classificação

---

1. Para b = 1 to B:

   (a) crie um bootstrap Z* de tamanho N a partir dos dados de treinamento

   (b) construa uma árvore T_b para os dados de bootstrap; por recusividade, repita as seguintes etapas para cada nó folha da árvore, até que o nó de tamanho mínimo n_{min} seja alcançado.

      (i) selecione *m* variáveis, de forma aleatória, das *p* variáveis.
      (ii) obtenha a melhor variável entre as *m*
      (iii) divida o nó em dois nós filhos.

2. Retorne a combinação de árvores $\{T_b\}^B$

Como resultado da classificação, a classe predita pelo *b*-ésimo classificador $C_b(x)$.
Então, $C_n(x) = maior\ voto\ \{C_b(x)\}^B$

Figura 4 – Algoritmo Random forest (Adaptado de Hastie *et al.,* 2009).

3.1.3 *Naive* Bayes

A abordagem *Naive* Bayes utiliza o princípio de que, a partir de uma probabilidade a *priori* e a verossimilhança de um novo dado é possível calcular a probabilidade a *posteriori* de um determinado evento (teorema de Bayes). Nessa técnica é assumido que os valores dos atributos de um exemplo são independentes entre si, dado o valor da saída. A simples Equação 3, denominada teorema de Bayes, constitui modernas técnicas de Inteligência Artificial para uma inferência probabilística (Russel &

Norvig, 2010).

$$p(C_k|X)=\frac{p(X|C_k)P(C_k)}{P(X)} \tag{3}$$

Em termos de modelo probabilístico Bayesiano, a Expressão 3 pode ser desenvolvida para assumir independência condicional entre cada uma das variáveis consideradas e ser obtido um modelo probabilístico com a Expressão 4.

$$\begin{aligned} p(C_k|x_1,\ldots,x_n) &\propto p(C_k,x_1,\ldots,x_n) \\ &\propto p(C_k)\,p(x_1|C_k)\,p(x_2|C_k)\,p(x_3|C_k)\cdots \\ &\propto p(C_k)\prod_{i=1}^{n} p(x_i|C_k) \end{aligned} \tag{4}$$

Por fim, é possível construir um classificador Naive Bayes que utiliza o modelo probabilístico da Equação 4 em conjunto com a regra de decisão da Equação 5.

$$\hat{y}=\underset{k\in\{1,\ldots,K\}}{argmax}\; p(C_k)\prod_{i=1}^{n} p(x_i|C_k) \tag{5}$$

A expressão acima indica a classe C_k com maior probabilidade de estar associada ao conjunto de atributos x_n.

3.1.4 Redes Bayesianas

A técnica de Redes Bayesianas também utiliza o teorema de Bayes apresentado anteriormente porém, assumem o conceito de independência condicional entre as variáveis. De maneira geral, esse conceito contribui para os casos em que existe uma relação estatística entre duas variáveis

quando uma terceira variável é conhecida (Faceli *et al.,* 2011).

Redes Bayesianas são representadas como um grafo acíclico direcionado cujas arestas indicam a dependência entre as variáveis e cada nó representa os atributos considerados. Para cada nó são associados alguns parâmetros numéricos que se refere as probabilidades condicionais entre as variáveis. No exemplo da Figura 5 o atributo “Alarm” tem probabilidade condiciona aos atributos “Burglary” e “Earthquake”.

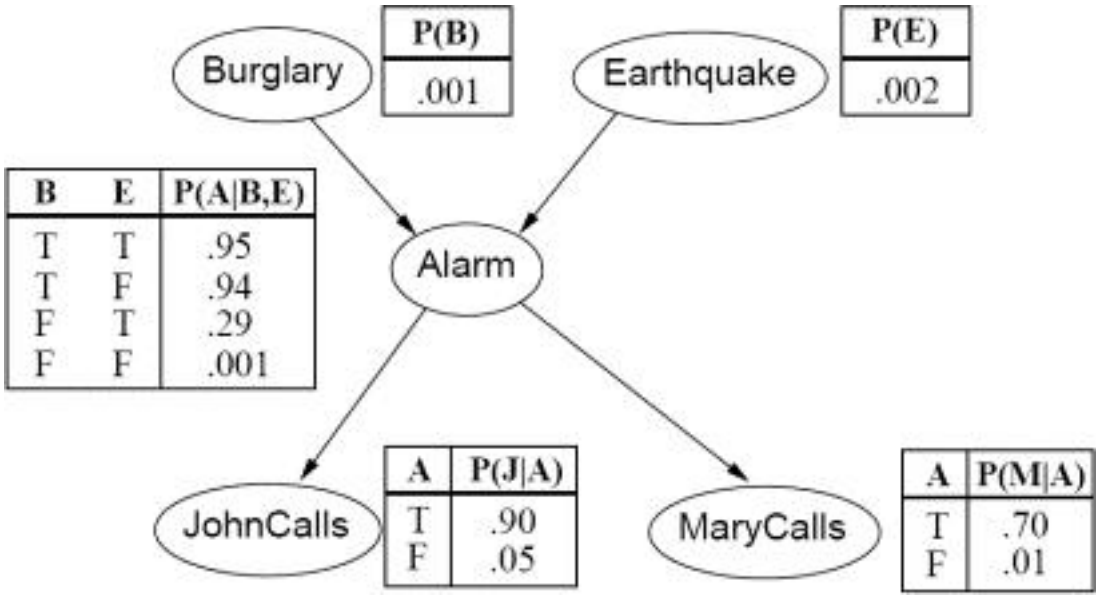


Figura 5 – Um exemplo típico de Redes Bayesianas para designar a probabilidade de um roubo em uma casa com alarme. Na estrutura é ainda considerado o disparo do alarme por terremoto e dois vizinhos, John e Mary, que prometeram ligar quando ouvissem o disparo do alarme (extraído de Russell & Norvig 2009).

O método para a construção da rede consiste em satisfazer, de forma iterativa, a propriedade local *Markov Blanket* que, de forma geral, verifica se uma variável alvo é condicionalmente independente de seus nós não descendentes dado seus atributos pais. Dessa forma, os nós avaliados referem-se aos nós pais do atributo alvo, seus nós filhos e todos os outros possíveis pais de cada um dos nós filhos (Faceli *et al.,* 2011).

3.1.5 Máquinas de vetores de suporte

Máquinas de vetores de suporte (*Suport Vector Machines - SVMs*) possui algumas propriedades que a tornam uma técnica interessante de se aplicar em diferentes problemas. Dentre essas, SVMs constroem fronteiras de decisão de maneira a definir um modelo bem generalizado. Utilizam a estratégia de mapear os dados de treinamento de seu espaço original para um novo espaço de maior dimensão como forma de separar os dados com froteiras de decisão mais simples nesse outro espaço de maior dimensão. Possui a perspicácia de utilizar os exemplos mais importantes para a construção das fronteiras de decisão e são resistentes a um super ajuste (*overfitting*) sobre dados de treinamento (Russel & Norvig, 2010).

Para ilustrar o princípio empregado por SVMs na separação dos dados, considere a Figura 6 a seguir. Ao observar os elementos do espaço bidimensional apresentado, podemos constatar que é possível sugerir diversas outras margens de separação para os dados porém, SVMs buscam otimizar a margem de separação destes maximizando a distância entre alguns exemplos, como é indicado na Figura pela linha em negrito entre as linhas pontilhadas.

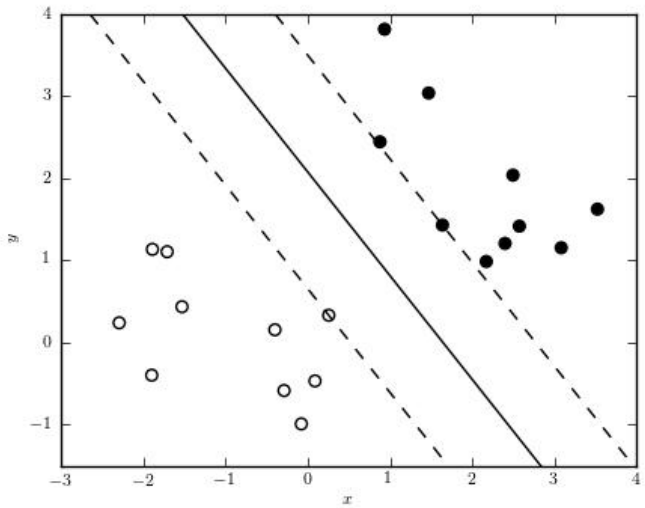


Figura 6 – Exemplo de definição da fronteira de decisão (extraído de Hastie *et al.,* 2009).

Para fronteiras de decisão mais complexas, SVMs possuem a habilidade de aumentar a dimensão do espaço de entradas, usando uma função *kernel* (ø), no intuito de que nesse novo espaço de dimensões a fronteira que separa os exemplos torne-se mais simples. Esse princípio é ilustrado na Figura 7. De maneira geral, a função recebe dois pontos do espaço de entradas original e calcula o produto escalar desses objetos no espaço de dimensões aumentado (Faceli *et al.,* 2011) (Bishop, 2006).

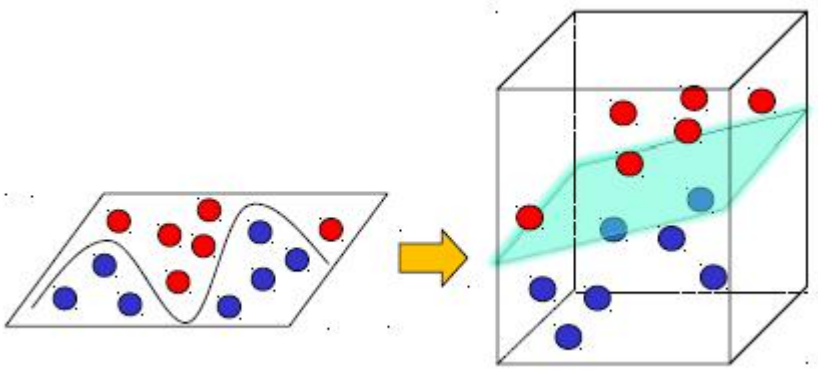


Figura 7 – Exemplo do mapeamento do exemplos em um espaço bidimensional para o espaço tridimensional, realizado por uma função *kernel*.

Diferentes funções kernel podem ser utilizadas para realizar o mapeamento dos atributos em diferentes planos dimensionais. Dentre as utilizadas neste trabalho, incluem Função Linear, (Gaussian) *Radial Basis Function* (RBF) e Polinomial. Cada uma difere na operação realizada sobre os exemplos do espaço de característica e são apresentadas a seguir:

- Kernel Linear: uma função kernel simples que possui uma constante opcional *c.*

$$K(x, y) = x^T y + c$$

- Kernel (Gaussian) *Radial Basis Function* (RBF): *sigma* é um parâmetro ajustável.

$$K(x,y)=\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$$

- Kernel Polinomial: possui três parâmetros ajustáveis, *alfa*, a constante *c* e o grau do polinômio *d.*

$$K(x,y)=(\alpha x^T+c)^d$$

3.2 Medidas de avaliação

Nesta seção, apresentaremos as principais medidas utilizadas para a avaliação dos modelos gerados durante o desenvolvimento do trabalho. São critérios amplamente empregados na literatura e fornecem subsídios para a avaliação do desempenho de cada modelo em relação às predições realizadas. Contudo, uma vez que é necessário um desdobramento sobre o significado biológico dos resultados alcançados, procuramos considerar também essa perspectiva como forma de direcionar as atividades desenvolvidas.

Dentre as estimativas baseadas em erro de predição o método de validação cruzada com *k* partições tem se destacado e amplamente aceito na comunidade de mineração de dados (Refaeilzadeh *et al.,* 2009). Dentre as vantagens do método, ele é capaz de avaliar o grau de generalização dos modelos com uma estimativa acurada.

No método de validação cruzada com *k* partições (*k-fold cross-validation*), os dados de treinamento são divido em *k* subconjuntos de tamanho aproximadamente igual. Então, *k – 1* subconjuntos são utilizados no treinamento de um classificador e o subconjunto restante é tomado como dados de teste. Esse processo é repetido *k* vezes e em cada ciclo um subconjunto de teste diferente é utilizado. O desempenho final do classificador é dado pela média dos desempenhos que foram observados

ao aplicar cada subconjunto de teste. Na Figura 8 é ilustrado esse processo de particionamento, treinamento e teste (Hastie *et al.,* 2009) (Faceli *et al.,* 2011).

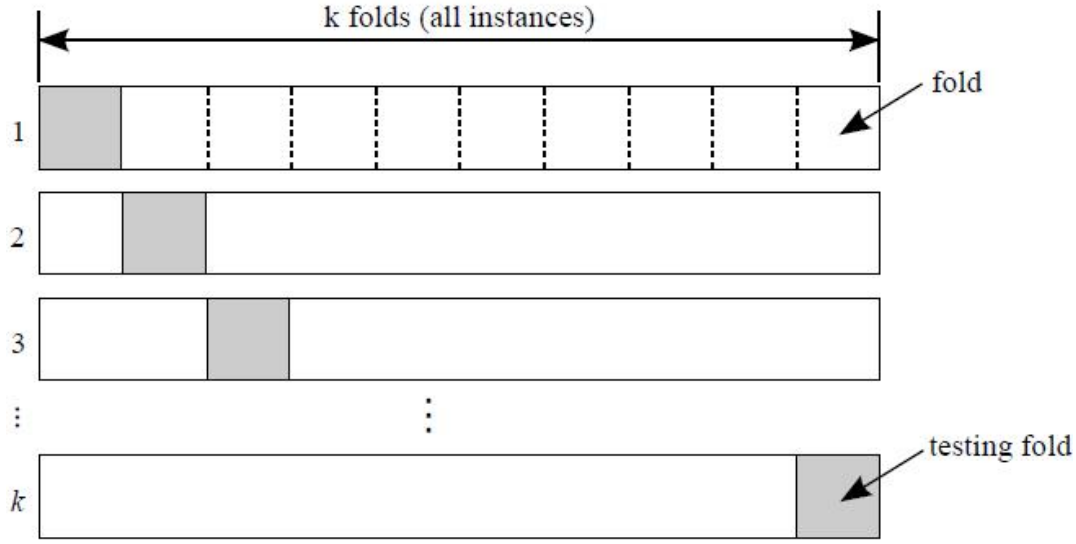


Figura 8 – Exemplo de particionamento em validação cruzada com k-fold adaptado de (adaptado de Borovicka, 2012).

Em um problema de duas classes, uma classe é denotada como positiva (+) e a outra negativa (-) e pode-se obter uma matriz de confusão como ilustrada na Tabela 1, onde:

- TN corresponde ao número de verdadeiros negativos (*True Negative*). Exemplos negativos que foram preditos corretamente como negativos.
- FP corresponde ao número de falsos positivos (*False Positive*). Exemplos que foram preditos como positivos mas pertencem a classe negativa.
- FN corresponde ao número de falsos negativos (*False Negative*). Exemplos preditos como negativos mas pertencem a classe positiva.
- TP corresponde ao número de verdadeiros positivos (*True Positive*). Exemplos positivos que foram preditos corretamente como positivos.

Tabela 1 – Matriz de confusão para uma classificação binária.

		Classe predita	
		+	-
Classe verdadeira	+	TP	FN
	-	FP	TN

É importante destacar que os mesmos princípios podem ser aplicados para problemas que envolvem mais de duas classes. Portanto, a contagem dos erros e acertos é realizada da mesma forma para cada classe ao longo das demais, sempre considerando os exemplos da classe rotulada conhecida em relação a saída predita. Considerando ainda a matriz de confusão descrita anteriormente, é possível obter vários valores estatísticos como:

- Precisão (precision): proporção de exemplos positivos classificados corretamente entre todos os preditos como positivos.

$$precision = \frac{TP}{TP + FP}$$

- Sensibilidade (recall):
  taxa de acerto na classe positiva.

$$recall = \frac{TP}{TP + FN}$$

- Medida-F (F-Measure): Combina precisão e recall em uma única medida de forma a determinar a exatidão e completude do modelo.

$$F - measure = 2\,x\,\frac{precision\,x\,recall}{precision + recall}$$

- Acurácia: mede a proporção de predições verdadeiras dentre todas as predições.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Essas medidas permitem avaliar o grau de generalização e comportamento do modelo de acordo com os resultados das predições. O método estatístico de validação cruzada é um procedimento para estimar o grau de generalização do modelo buscando evitar efeito de viés dos dados (Refaeilzadeh *et al.,* 2009). Isto é, que os resultados sofram algum tipo de tendência que não corresponda as informações presente nos dados.

3.3 Gaggle Genome Browser

A ferramenta *Gaggle Genome Browser* (Bare *et al.* 2010) possibilita a visualização de dados genômicos de uma forma bem simples e oferece vários recursos para manipulação e representação de dados produzidos em larga escala. Informações sobre dados de expressão, anotações existentes para um determinado organismo, dados de proteômica ou conservação (entre outros), podem ser adicionados facilmente e assim oferecer uma análise integrada com diversos níveis de informações.

A representação dos dados pode ser realizada seguindo os recursos disponíveis na ferramenta, com diferentes categorias e formatos, como por exemplo, dados baseados em segmentos, dados posicionais ou informações sobre genes. Cada formato pode ser indicado como caixas, *heatmap*, marcadores, barras verticais, entre outros. Com isso, é possível percorrer o genoma de interesse e navegar por suas regiões de forma a observar o conjunto de informações reunidas no navegador (também chamado de *browser*). Na Figura 9 é apresentada a janela principal do programa *Gaggle Genome Browser* com diversos dados, de forma a ilustrar alguns dos tipos de faixas genômicas (*track*) para representação dos mesmos.

O banco de dados incorporado à ferramenta é baseado em Sqlite e possui algumas características como: interface para banco de dados maiores, habilidade para tratar faixas genômicas que não cabem na memória, facilita a importação de bases de dados e faixas genômicas e possui menor consumo de memória.

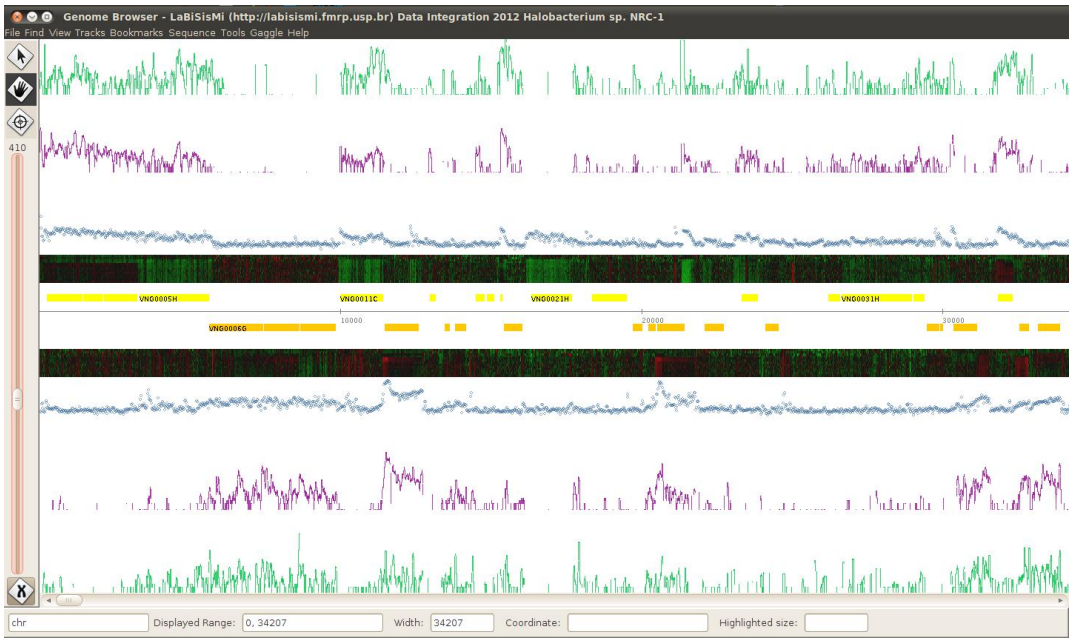


Figura 9 – Janela principal da ferramenta *Gaggle Genome Browser*. Caixas em amarelo indicam genes anotados para a fita *foward* e em laranja para a fita *reverse*. Dados importados no exemplo ilustram alguns dos tipos de faixa genômica (*track*) como recurso de representação fornecido pela ferramenta. Uma faixa genômica do tipo *heatmap* (cores tendo do verde para o vermelho em ambas as fitas), faixas genômicas posicionais (em lilás e verde em ambas as fitas) e faixa genômica do tipo segmentos (em azul claro para ambas as fitas).

3.4 Weka

A ferramenta WEKA (Hall *et al.,* 2009) é um arcabouço de algoritmos de aprendizado de máquina que inclui diversos recursos para pré-processamento de dados, classificação, regressão, agrupamento, regras de associação e visualização. Além de uma interface gráfica é possível fazer o uso de uma API (*Application Programming Interface*) que torna fácil a utilização de todo código fonte da ferramenta em projetos mais específicos. O uso da API WEKA envolve a implementação de algoritmos em linguagem JAVA que utilizando os métodos já desenvolvidos na

ferramenta podem-se construir modelos e aplicar avaliações sobre os mesmos.

Para o uso da ferramenta, tanto por meio da interface gráfica como pela API, é necessária a construção de um arquivo baseado em formato de atributo-relação (*Atribute-Relation file format – ARFF*). Nele são especificados o conjunto de atributos para a representação dos dados e seus respectivos valores.

3.5 Ambiente de pré-processamento

Para a manipulação dos dados na fase de pré-processamento foram desenvolvidos algoritmos na linguagem de programação R (R Development Core Team). Para o uso da API WEKA e implementação dos procedimentos para o acesso aos classificadores foi utilizada a linguagem de programação JAVA. Importamos as bibliotecas disponíveis na ferramenta WEKA através da IDE Net Beans. Todas atividades foram desenvolvidas em ambiente Linux.

3.6 Tecnologias de sequenciamento

O transcritoma é o conjunto completo de todas as moléculas de RNAs, incluindo mRNAs, tRNAs, rRNAs e outros RNAs não-codificadores, presentes na célula em um determinado momento ou condição (Wang *et al.,* 2009). O estudo e compreensão desses elementos reflete parte da dinamicidade da célula, uma vez que diferentes classes de transcritos emergem como informações expressas ao longo do genoma.

Tecnologias têm sido desenvolvidas para a quantificação do nível de expressão relativo aos elementos presentes no transcritoma e incluem abordagens baseadas em hibridização ou sequenciamento (Wang *et al.,* 2009) (Metzeker, 2010). A tecnologia de microarranjos (*microarrays*) utiliza o conceito de hibridização onde coleções de trechos alvo do genoma são utilizados em pequenos *spots* anexados em um chip, esses trechos são então hibridizados e transformados em moléculas de cDNA ou

cRNA em uma transcrição reversa. Essas moléculas possuem nucleotídeos modificados para carregar moléculas fluorecentes, que por sua vez, reagem ao serem excitadas por laser. Esse procedimento permite detectar a quantidade de RNA expresso a partir da intensidade de fluorescência (Hoheisel, 2006). Por fim, imagens são geradas e processadas, envolvendo importantes etapas de processamento computacional como tratamento de ruídos e normalização, para a medição do nível de expressão associado aos transcritos.

Diferente de métodos de microarranjos, abordagens baseadas em sequenciamento determinam de forma direta a sequência de cDNA (Wang *et al.,* 2009). Plataformas de sequenciamento de nova geração em larga escala (*Next-genereation sequencing* - *NGS*) utilizam diferentes *workflows* para preparo e execução dos experimentos. De forma geral, os procedimentos envolvem a construção da biblioteca de moléculas a serem sequenciadas, aplicação da tecnologia de sequenciamento e análise dos trechos sequenciados. A construção da biblioteca requer o planejamento prévio do experimento a ser realizado, onde os RNAs a serem estudados são devidamente isolados e purificados, a aplicação da tecnologia implica em diferente formas e etapas para os procedimentos de sequenciamento envolvendo fragmentação dos transcritos, uso de adaptadores moleculares e geração de pequenas sequencias lidas. Por fim, os pequenos trechos podem ser alinhados ao genoma referência como forma de gerar a expressão dos transcritos (Wang *et al.,* 2009).

Diversas aplicações têm sido empregadas com o uso de NGS, dentre estas: avaliar o nível de expressão dos transcritos; detectar novos transcritos ou isoformas, mapear estruturas do gene com informações precisas de início e fim, análise de *splicing* alternativo, análise variações da sequência (como, por exemplo, identificação de SNPs) (Wang et al., 2009) (Metzker, 2010).

4 Identificação *in silico* de ncRNAs em *Halobacterium salinarum*

Na tentativa de alcançarmos os objetivos propostos nesta Tese, aplicamos uma metodologia baseada na adaptação de abordagens existentes para a identificação de ncRNAs. Dessa forma, obtivemos como resultado secundário "sub-produto" a criação de um *workflow* com os procedimentos desenvolvidos. O *workflow* foi baseado na abordagem *incRNA* (Lu *et al.,* 2011) e será destacado nas próximas subseções com um nível maior de detalhamento.

4.1 Adaptação da metodologia incRNA

Desde a perspectiva inicial de adaptação da metodologia proposta em Lu *et al.,* 2011, que de forma geral consiste em integrar diversas fontes de dados e regiões anotadas para a criação de um modelo de AM, diversas modificações foram desenvolvidas. Essas modificações foram necessárias uma vez que muitas das etapas propostas pelos autores não tinham correspondência com o arranjo de informações disponíveis para o organismo em estudo nesta Tese, *H. salinarum* NRC-1. Por exemplo, os autores utilizam mais de uma fonte de dados oriundas de experimentos baseados em tiling array e sRNA-seq, o que não está disponível para *H. salinarum* NRC-1.

No entanto, o método sugerido no artigo mostrou-se como uma perspectiva interessante de ser adaptada uma vez que explora a diversidade de informações que podem contribuir para uma metodologia mais robusta. Como discutido na seção anterior, diferentes métodos buscam incluir as características pertinentes ao problema tratado, porém não de uma forma mais integrada, com diversos níveis de informação, no que se refere as diferentes etapas do processamento da informação biológica, sendo incluídos. Dessa forma, dirigimos os esforços para ajustar os princípios sugeridos na metodologia aos dados que temos disponíveis para *H. salinarum NRC-1*. A disponibilidade dos dados de expressão em

larga-escala (*tiling array e RNA-seq*) também foi uma motivação extra para a escolha da adaptação.

Inicialmente foram reunidas as informações de conservação, expressão, sequencia primária e propriedades estruturais de regiões ao longo do genoma de *H. salinarum* NRC-1. Posteriormente, diversos algoritmos de AM foram aplicados para a criação e escolha do modelo computacional, gerado a partir dos dados de treinamento obtidos. Por fim, regiões sem anotações ao longo do genoma foram definidas em uma etapa posterior visando predizer potenciais candidatos a moléculas de RNAs não codificadores no organismo modelo *H. salinarum* NRC-1.

De forma geral, a metodologia que adaptamos possui duas etapas principais, a primeira consiste no pré-processamento dos dados para a criação do modelo e a segunda envolve a aplicação de um procedimento de janela deslizante ao longo do genoma de forma a possibilitar a definição e identificação de regiões genômicas com maior probabilidade de transcreverem moléculas não-codificadoras, ou em outras palavras, utilizando o jargão de AM, de pertencerem à classe ncRNA. Os principais procedimentos contidos nessas duas etapas são ilustrados na Figura 9 sendo que, os procedimentos A, B e C pertencem a primeira etapa e os procedimentos D, E e F pertencem a segunda etapa de processamento e análise das informações.

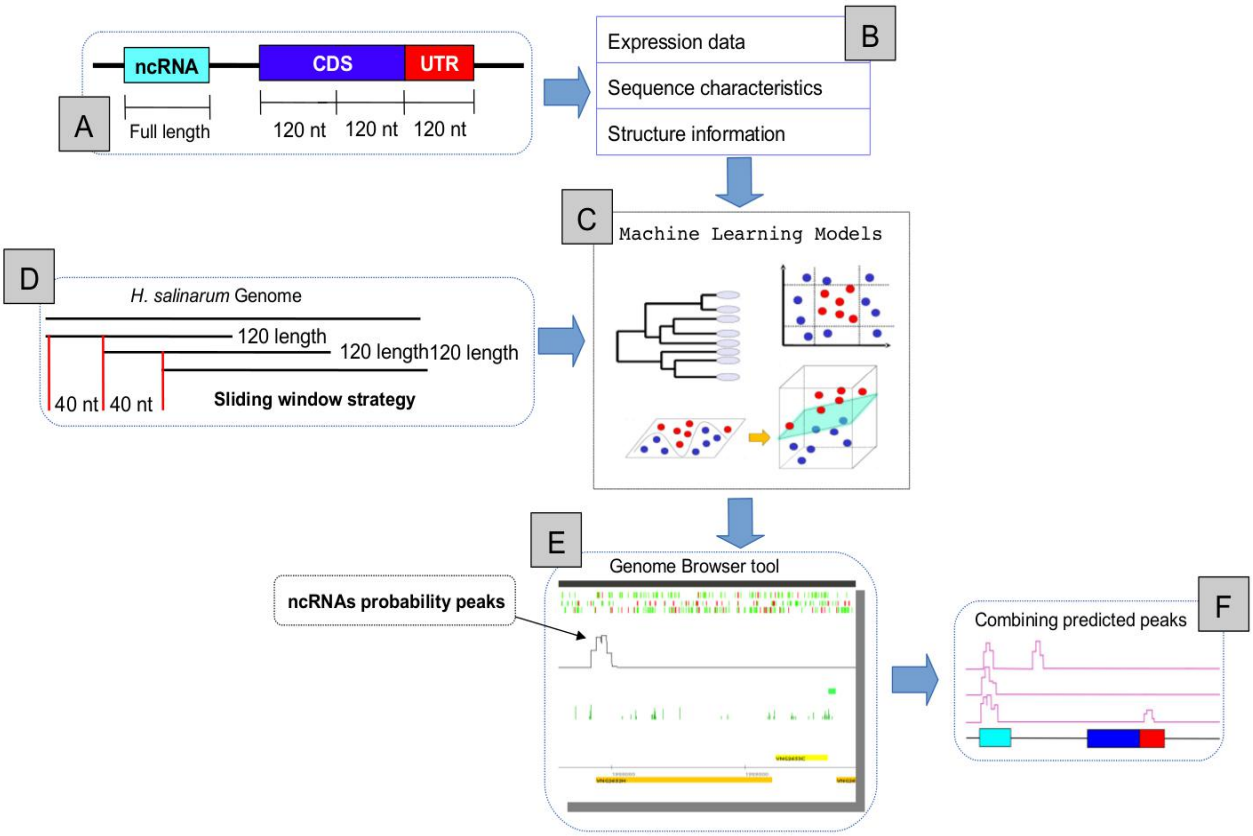


Figura 10 – *Workflow* da abordagem desenvolvida com os principais procedimentos envolvidos na criação do modelo de AM para a identificação de trechos genômicos com probabilidade de transcrever moléculas de ncRNA. Na primeira etapa, são consideradas as anotações existentes para o genoma de *H. salinarum* NRC-1(A) e dados de expressão, estrutura e propriedades da sequência primária para cada região anotada (B). Essas informações são então utilizadas na construção de modelos de AM (C). Na segunda etapa, é aplicado um procedimento de janela deslizante em cada modelo de AM gerado (D) com isso, picos de probabilidades associadas a classe ncRNAs são gerados ao longo do genoma (E). Finalmente, esses picos são combinados e regiões que possuem picos gerados por vários classificadores em conjunto são selecionadas como potenciais candidatos a ncRNAs (F).

Cada procedimento corresponde aos tópicos a seguir e cada um será descrito em uma subseção correspondente:

Procedimento A: Obtenção das anotações disponíveis para o organismo em estudo.

Procedimento B: Obtenção de dados genômicos, de expressão e conservação disponíveis para o organismo em estudo como forma de definir os atributos (*features*) de AM.

Procedimento C: Criação e avaliação de modelos de AM.

Procedimento D: Definições de regiões para aplicação nos modelos de AM gerados a partir do particionamento do genoma com sobreposições.

Procedimento E: Geração da faixa genômica. Dados com valores posicionais ao longo do genoma que podem ser representados no Genome Browser. São considerados os valores de probabilidade associados a cada trecho aplicado anteriormente (procedimento D).

Procedimento F: Combinação dos trechos que possuem picos de probabilidade e que estão presentes em conjunto nos resultados de vários classificadores.

Nas próximas seções, são apresentados cada um dos procedimentos de forma mais detalhada.

4.1.1 Anotações disponíveis para *H. salinarum*

Coletamos regiões genômicas anotadas para *Halobacterium salinarum* NRC-1 e utilizamos cada trecho como dados de treinamento para algoritmos de AM. Dentre as anotações disponíveis, 2635 genes foram obtidos em http://www.microbesonline.org/ (Dehal *et al.*, 2010). Koide *et al.,* (2009b) identificaram através de dados de *tiling array* 61 regiões como candidatos putativos a ncRNAs. Adicionalmente, baseado na integração de vários tipos de dados os autores também identificaram 1377 regiões como 3' e 5' não traduzidas (*Untranslated Regions* - UTR) associadas a diversos genes. Obtivemos outras 41 regiões pertencentes à classe ncRNAs que por sua vez foram preditas utilizando a ferramenta *snocan* (Lowe & Eddy, 1999), a qual busca motivos (*motifs*) C/D box presentes em moléculas da classe de snoRNAs.

4.1.2 Integração de dados e definição de atributos

Dentre as fontes de dados disponíveis para o organismo em estudo, obtivemos dados experimentais de expressão (*Expression data*) oriundos de bibliotecas de pequenas moléculas de RNA (*RNA-seq small RNAs*) e dados de 13 pontos ao longo da curva de crescimento obtidos por técnicas de microarranjos (*Tiling array growth curve*) (Koide *et al.,* 2009b). Na Tabela 2 são listadas todos as categorias de atributos que foram utilizados.

Outra informação importante que nos ajuda a distinguir exemplos pertencentes a trechos codificadores dos não-codificadores é o conjunto de três nucleotídeos que correspondem a um códon de finalização e inicialização. Consideramos as informações da tabela com os códigos genéticos do domínio *Archaea* e calculamos a distância (número de nucleotídeos) entre o início da região de interesse para o códon de início mais próximo, da mesma forma calculamos a distância entre o valor final da região de interesse para o códon de finalização mais próximo. Esse atributo foi denominado como *ORF distance*.

Tabela 2 – Resumo das categorias de atributos utilizados na representação dos dados de treinamento.

Feature group	Name	No. of features	Feature name
Expression data	RNA-seq small RNAs	7	small_ExpMean, small_ExpMedian, small_ExpInterval, small_ExpSD, small_ExpObliq, small_ExpKurt, small_ExpPercentage
	Tilling array (growth curve	13	tiling_01, tiling_02, tiling_03, tiling_04, tiling_05, tiling_06, tiling_07, tiling_08, tiling_09, tiling_10, tiling_11, tiling_12, tiling_13
Sequence characteristics	Conservation	7	cons_ExpMean, cons_ExpMedian, cons_ExpInterval, cons_ExpSD, cons_ExpObliq, cons_ExpKurt, cons_ExpPercentage
	GC content	1	%gc
	ORF Distance	2	dist5Prime, dist3Prime,
	No. of codons	1	CountsStop
Structure information	Minimum free energy (MFE)	8	n_hairpin, n_multiloop, n_interloop, n_bulge, loops, tpaired, tunpaired, MFE
	Structure features		
Total		39	

A medida de conservação de sequência foi calculada com base no método proposto em Marchais *et al.,* 2009. A partir de um alinhamento de cada posição do genoma, utilizando a ferramenta BLAST (Zhang & Madden, 1997), um índice de conservação é gerado e corresponde ao número de

genomas em cada posição cujo peso associado a essa contagem baseia-se na proximidade filogenética do genoma em relação ao genoma de *H. salinarum NRC-1*. Como parte das características da sequência primária (*Sequence characteristics*), também incluímos o conteúdo GC do trecho anotado.

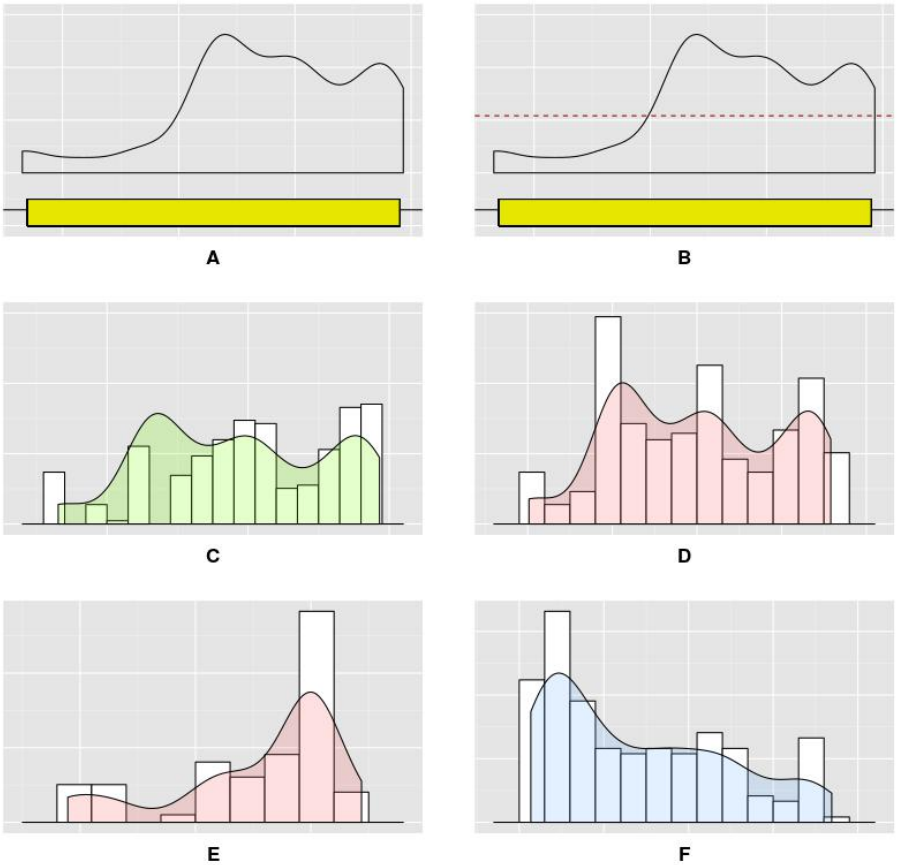


Figura 10 – Ilustração de algumas das medidas de espalhamento e distribuição. (A) representação simples de uma região anotada como gene (caixa amarela) e seu respectivo sinal hipotético de sRNA-seq ao longo do comprimento do gene (linha preta). (B) a mesma representação anterior porém, com uma linha vermelha tracejada indicando o valor médio do sinal de sRNA-Seq. Medida de distribuição com valor de curtose negativa (C), isso indica que o histograma dos dados apresenta um achatamento mais baixo do que uma distribuição normal. Curtose positiva (D), o histograma dos dados apresenta um achatamento mais alta e concentrada que uma distribuição normal. Obliquidade negativa (E) a distribuição concentra-se mais no lado direito. Obliquidade positiva (F) a distribuição concentra-se mais no lado esquerdo.

Finalmente, as informações que se referem a estrutura secundária (*Structure information*) foi incluída com base nos resultados da predição utilizando a ferramenta ContextFold (Zakov *et al.,* 2011). A anotação da estrutura predita foi interpretada e suas subestruturas foram definidas como uma coleção de outros atributos, os quais incluem: número de grampos (*hairpins*), loops internos (*internal-loops*), multi-loops, *budges*, *loops*, número de bases pareadas e não pareadas e energia livre da estrutura.

Buscamos ainda aprimorar a representação dos valores de conservação e de sRNA-seq e aplicamos algumas medidas baseadas nas seguintes observações: Originalmente as informações de RNA-seq correspondem ao logaritmo base 2 da contagem de *reads* mapeados no genoma de *H. salinarum* NRC-1. Assim, para cada posição do genoma uma contagem é associada e o sinal torna-se ruidoso com vários decaimentos, oscilações e pontos de quebra (Figura 10 – A). Inicialmente, havíamos considerado apenas o valor da média do sinal em cada região anotada. Por exemplo, na Figura 10 – B a linha vermelha tracejada indica o valor médio do sinal (linha preta) que corresponde ao sinal de expressão de determinado gene (caixa amarela). Visando então uma melhor representação, utilizamos outras medidas de espalhamento e distribuição que procuram sumarizar a forma com que os dados se organizam. Essas medidas são: obliquidade (*skweness*), curtose, média, mediana, desvio padrão, intervalo (dado pelo valor máximo da região subtraído pelo valor mínimo) e a porcentagem dos valores de expressão que estão acima do valor médio da região. Vale ressaltar que as mesmas considerações foram aplicadas tanto aos dados de sRNA-seq quanto aos de conservação.

Os princípios sugeridos pelas medidas de espalhamento sugerem uma maneira interessante de aprimorar a representação dos dados e dessa forma optamos por sua aplicação na definição dos atributos utilizados.

4.1.3 Construção e avaliação de modelos de AM

Aplicamos e avaliamos diferentes algoritmos de AM na tentativa de verificar se determinada técnica era suficiente para separar os exemplos com trechos genômicos codificadores dos não-codificadores. Cada algoritmo utilizado está descrito na seção 3.1 deste documento.

Na avaliação consideramos a validação cruzada com 10 partições (*10-fold cross validation*), mas não sobre a totalidade dos dados, como é tradicional, e sim sobre 2/3 deles. Os 1/3 remanescentes são separados logo de início e não considerados como parte do universo total. Deixamos 1/3 dos dados fora da validação cruzada e aplicamos como dados de teste independente para avaliar o comportamento do melhor modelo obtido no resultado com a validação cruzada. Outras medidas descritas na seção 3.2 também foram utilizadas durante o processo de avaliação dos modelos gerados.

Este procedimento encerra a primeira etapa da abordagem que envolve as definições dos dados de treinamento, atributos e criação e avaliação de modelos baseados em AM.

4.1.4 Aplicação da estratégia baseada em janela deslizante

Após o procedimento de criação dos modelos de AM seguimos para a segunda fase da abordagem. Inicialmente, definimos uma estratégia baseada no particionamento de todo o genoma considerando ainda uma certa sobreposição de nucleotídeos em cada trecho. A estratégia é bem simples e possibilita observar como se dá o comportamento dos modelos na medida em que diversos trechos são aplicados e a probabilidade associada ao trecho aplicado é então obtida.

No Capítulo 5 serão apresentados os resultados obtidos em algumas variações tanto dos modelos quanto na maneira de se especificar o tamanho do trecho e o tamanho das sobreposições. É importante destacar que, da mesma forma como é feito com os dados de treinamento, para

cada trecho do genoma particionado todas os atributos discutidos são também calculados, ou seja, informações sobre expressão, conservação, estrutura e propriedades da sequência são também calculados.

Como resultado de saída desse procedimento cada classificador define uma probabilidade, para cada trecho particionado, de pertencer à uma das classes consideradas (CDS, UTR, ncRNAs) . Em seguida, para cada resultado dos classificadores, geramos uma faixa genômica com um valor de probabilidade associado a cada posição ao longo do genoma. Uma vez que cada trecho possui certa sobreposição sobre os trechos vizinhos, no cálculo da probabilidade foi necessário considerar esse aspecto para atribuição dos valores em cada posição ao longo do genoma. E dessa forma, em cada posição é realizada uma média das probabilidades dos trechos que se sobrepõem.

4.1.5 Processamento dos sinais de probabilidade

O processamento dos picos gerados consiste em obter regiões, considerando cada classificador de forma independente, cujos valores são maiores que a média de todos os valores de probabilidade ao longo do genoma para um determinado classificador. Na Figura 11 é ilustrado esse princípio considerando a existência de 3 classificadores C1, C2 e C3. A linha azul tracejada indica o valor médio das probabilidades ao longo do genoma para seus respectivos classificadores. Isso é necessário porque simplesmente adotar como interessante as posições com probabilidade alta/máxima ou apenas diferente de zero, trechos curtos ou muito longos seriam obtidos e dessa forma não seria possível a captura de trechos mais significativos.

Para realizar esse processamento utilizamos um algoritmo que interpreta os valores da faixa genômica verificando posição a posição se ocorre a mudança dos valores em relação a uma referência, que nesse caso seria o valor médio das probabilidades ao longo do genoma, quando ocorrer a mudança de valores para acima do valor médio é porque ocorreu um início do trecho com o pico. Do contrário, quando ocorre a mudança

dos valores para abaixo do valor médio é porque ocorreu um fim do trecho. Essas posições são salvas e assim são definidos os valores de início e fim.

Figura 11 – Ilustração de faixas genômicas e o respectivo valor médio da probabilidade definida por cada classificador.

A partir de cada valor médio, o início e fim de cada trecho contendo os picos são então obtidos como é ilustrado na Figura 12. Na ilustração, as posições de início são representadas por triângulos vermelhos e as posições de fim são apresentadas por triângulos verdes.

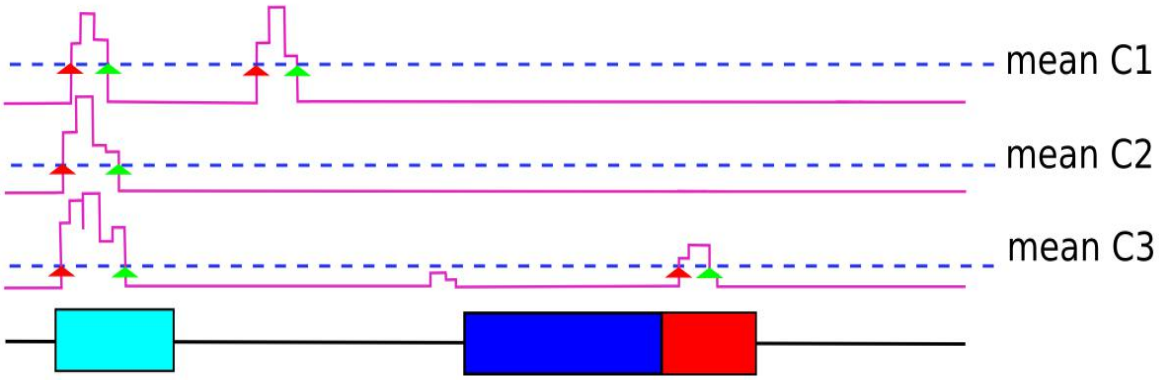


Figura 12 – Definição das posições de início (triângulo vermelho) e fim (triângulo verde) a partir dos picos obtidos em cada uma das faixas genômicas.

Ao final do procedimento de definição das posições de início e fim dos trechos com picos, notamos que em alguns casos, os trechos possuíam tamanhos maiores que 500 pares de base pois, em alguns

trechos o valor da média se mantinha por regiões mais longas. Dessa forma, foi necessário realizar uma filtragem dessas regiões e mantivemos somente as que são menores que 400 pares de base. Definimos esse tamanho com base nos dados de treinamento, uma vez que o tamanho médio dos exemplos de treinamento contidos na classe ncRNAs é 160 e o exemplo com maior tamanho possui até 400 pares de base.

4.1.6 Combinação das regiões preditas

O procedimento de combinação das regiões preditas consiste em verificar quais trechos, obtidos na seleção dos picos preditos, são ditos como ncRNAs por vários classificadores. A combinação consiste num sistema simples de voto, onde regiões são escolhidas com o maior número de classificadores possível, observamos o intervalo de combinação com votos de 5 a 8 classificadores e os resultados serão apresentados posteriormente no Capítulo de resultados. Uma vez selecionado os trechos, é então realizado um procedimento para verificar se já existe anotação para essas regiões selecionadas. Basicamente, ao aplicarmos a estratégia de janela deslizante diversas partições coincidem com trechos já anotados e dessa forma, é provável que alguns picos preditos coincidam com a anotação existente. Esses trechos então são removidos das análises. Da mesma forma, no Capítulo de Resultados são apresentados mais informações sobre este procedimento.

4.2 Aplicação de abordagens disponíveis para a identificação de ncRNAs

Nessa seção, são apresentadas as atividades desenvolvidas para a aplicação de algumas das abordagens apresentadas na seção 1.4, que estão até então disponíveis para a predição de ncRNAs e não foram desenvolvidas nesta Tese, apenas utilizadas diretamente.

4.2.1 Aplicação da abordagem Dario

Buscamos a aplicação de metodologias baseadas em dados de sRNA-seq, uma vez que possuíamos dados disponíveis no grupo de pesquisa, e dentre as disponíveis encontramos a abordagem Dario (Fasold *et al.,* 2011). Verificamos a possibilidade de utilizar o genoma referência de *H. salinarum* NRC-1*,* porém este não está disponível no *website* ferramenta (até a presente data temos os seguintes genomas Human (hg18), Human (hg19), Rhesus monkey (rhemac2), Mouse (mm9), Fruit fly (dm3), Worm (ce6), Zebrafish (danRer6)). Dessa forma, estudamos as considerações do método de predição baseado em AM e desenvolvemos os procedimentos necessários para sua aplicação.

De maneira geral a abordagem consiste em gerar agrupamentos de *reads,* oriundos de dados de sRNA-Seq*,* e utilizar informações desses agrupamentos como atributos para métodos de AM. A primeira etapa de mapeamento dos *reads* e agrupamento foi realizada com a aplicação da ferramenta blockbuster7 (Langenberger *et al.,* 2009). O programa combina blocos de *reads* que são mapeados de acordo com o alinhamento no genoma referência e então gera agrupamentos (*cluster*) desses blocos. Em seguida, com base nos atributos definidos pelos autores, coletamos as informações dos atributos que são baseadas nos agrupamentos (*cluster*) gerados anteriormente, sendo: número de blocos de *reads* dentro do agrupamento, tamanho do agrupamento, número de nucleotídeos cobertos por pelo menos dois blocos, tamanho máximo, mínimo e médio do bloco e distância máxima, mínima e média de nucleotídeos entre dois blocos consecutivos.

Essas informações procuram mapear o comportamento dos *reads* em diferentes regiões genômicas e como isso estabelecer os possíveis padrões presentes em trechos codificadores e não-codificadores. Dessa forma, verificamos dentre os dados de agrupamento quais coicidem com regiões já anotadas, considerando informações sobre as classes CDS, CDS com UTRs conhecidas e ncRNAs conhecidos. Essas anotações se referem aos mesmos dados da seção 4.1.1. Dentre os grupos que batem em

regiões anotadas, 1651 estão em trechos pertencentes a classe CDS, 1333 em trechos de CDS com UTRs conhecidas e 68 batem com regiões de ncRNAs conhecidos. Esses dados foram então utilizados para a criação do modelo de AM. Agrupamentos sem anotações totalizaram 4225 outros trechos, que foram utilizados como dados de teste na tentativa de identificar novos ncRNAs. Os resultados da aplicação dessa metodologia serão discutidos no Capítulo 5.

4.2.2 Aplicação da abordagem smyRNA

Como mencionado, o mecanismo de busca da abordagem denominada smyRNA (Salari *et al.,* 2009) se baseia em certos trechos da sequência primária (*motifs*) que são importantes para determinar a estrutura da molécula. Para a aplicação da abordagem smyRNA também consideramos as anotações disponíveis para *H. salinarum* NRC-1 e que estão descritas na seção 4.1.1. Basicamente as informações necessárias para a aplicação da abordagem consiste apenas de exemplos conhecidos de ncRNAs e do genoma referência. A criação do modelo estabelece uma taxa de verosimilhança entre motivos (*motifs*) gerados a partir dos exemplos de treinamento e a mesma é usada no cálculo de probabilidade de exemplos novos pertencerem a classe ncRNA.

4.2.3 Aplicação da abordagem RNASpace

A plataforma RNASpace fornece uma interface para diversas ferramentas de predição de ncRNAs. A maioria dessas abordagens são baseadas em homologia, com buscas por similaridade de sequência e de estrutura. Por meio da ferramenta é possível aplicar, por exemplo, a ferramenta BLAST (Altschul *et al.* 1990) ao bando de dados RFam (Gardner *et al.,* 2009). A plataforma ainda inclui, no contexto de metodologias baseadas em homologia, as ferramentas: YASS (Noé e Kucherov, 2005) que efetua busca por similaridade em bancos de dados, Infernal (Nawrocki

et al. 2009), Erpin (Gautheret e Lambert 2001) e Darn! (Zytnicki *et al.* 2008) que utilizam informações sobre similaridade de sequência e estrutura secundária para a busca, RNAmmer (Lagesen *et al.* 2007) para a busca de RNAs ribossomais e tRNAscan-SE (Lowe e Eddy, 1997) para a busca de RNAs transportadores. A ferramenta também oferece uma metodologia de busca através de uma análise comparativa de sequencias. Para isso, é possível selecionar algumas espécies para comparação. Primeiro é realizada um alinhamento usando BLASTN ou YASS e em seguida são gerados agrupamentos de regiões conservadas. Por fim é determinado um *score* para a conservação da estrutura secundária usando RNAz (Washietl *et al.* 2005) ou caRNAc (Touzet e Perriquet, 2004). Uma terceira categoria de ferramenta incluída na plataforma motipara a busca de ncRNAs refere-se a uma metodologia *ab initio* denominada AtypicalGC. Foi desenvolvida pelos próprios autores da plataforma e o princípio explorado pela ferramenta considera o viés da composição de nucleotídeos de regiões que pertencem a ncRNAs em comparação ao resto do genoma.

Aplicamos todas as três categorias de busca ao genoma de *H. salinarum* e os resultados serão discutidos no Capítulo 5.

4.2.4 Aplicação da abordagem CoRAL

A abordagem CoRAL foi desenvolvida para a classificação de RNAs em algumas categorias funcionais utilizando dados de sRNA-Seq. Para a representação dos ncRNAs conhecidos e criação dos modelos são consideradas características dos *reads* como: variações no tamanho dos reads, abundância de *reads* na região anti-senso, distribuição das posições 5’ e 3’ de cada *read* e ainda a energia livre míninima (MFE) predita. Os autores sugerem que essas características podems refletir propriedades subjacentes a bibliotecas de sequenciamento de sRNA-Seq, contribuindo na identificação e classificação de diversas classes de ncRNAs. Para a aplicação da abordagem CoRAL consideramos os dados de sRNA-Seq disponíveis para *H. salinarum* NRC-1. Realizamos alguns procedimentos

para o pré-processamento dos dados seguindo as especificações no guia da metodologia para usuários, que está disponível em Ryvkin *et al.,* 2014. Foi necessário incluir as informações dos ncRNAs conhecidos ao arquivo de anotações em formato GFF. Além de seguir as etapas do *workflow*, algumas alterações nos *scripts* foram necessárias para adequar a execução local dos arquivos. Da mesma forma como nas abordagens anteriores, os resultados serão apresentados no Capítulo 5.

4.3 Predição de interação RNA-Proteína

Para a predição de possíveis ncRNAs candidatos a interação com a proteína de interesse LSm, presente no organismo modelo em estudo, aplicamos uma metodologia também baseada em AM que segue o esquema da Figura 13 a seguir.

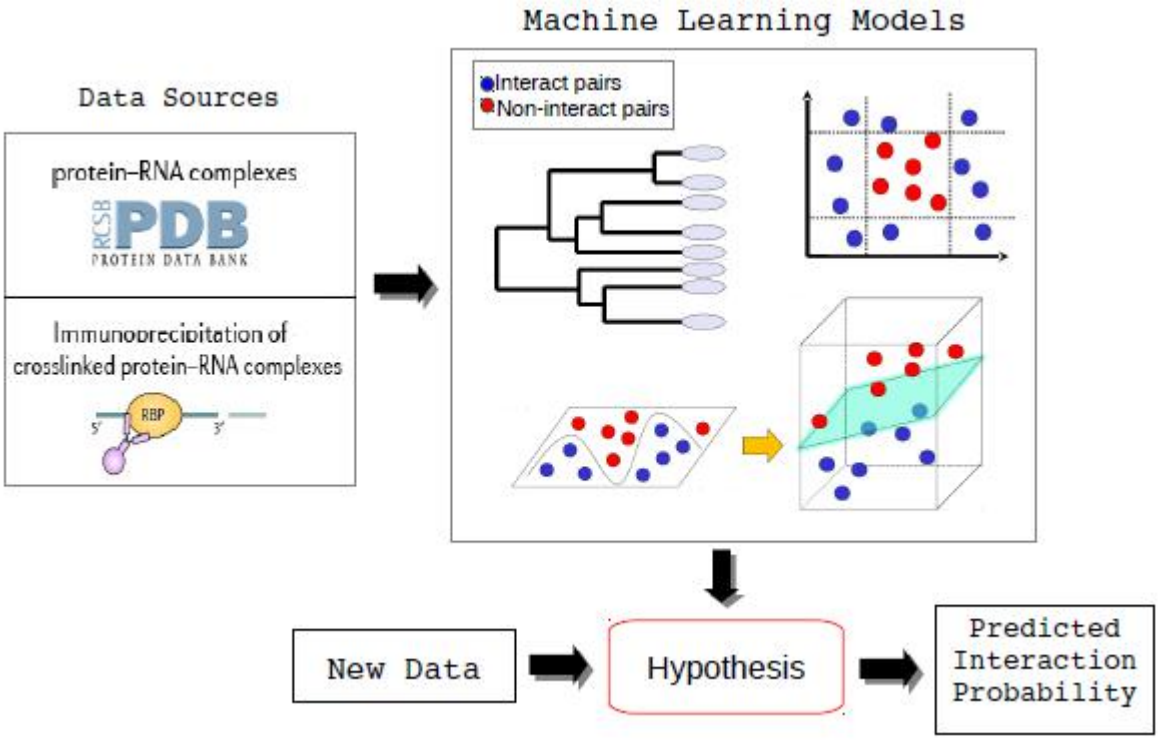


Figura 13 – Principais aspectos da abordagem aplicada na predição de interações RNA-Proteína. A partir de um conjunto de dados de treinamento (*Data Source*) disponíveis em bases de dados como o Protein Data Bank – PDB ou por meio de técnicas de imunopreciptação, modelos de AM são criados (*Machine Leaning Models*) como forma de interpretar e distinguir entre os pares de exemplos positivos (pares que interagem – *interact pairs*) dos pares de exemplos negativos (pares de RNA-proteína que não interagem entre si – *non-interact pairs*) e assim, determinar uma hipótese (Hypothesis), ou fronteira de decisão que separe os exemplos. Considerando essa hipótese, novos elementos (new data) podem ser inferidos sobre o modelo com o objetivo de obter um valor de probabilidade de interação para o mesmo.

A metodologia consiste em explorar diversas estratégias para a criação de diferentes modelos de AM seguindo o esquema da Figura 13. Dentre essas alternativas foi explorada a proposta de representação de atributos apresentada em Muppirala *et al.,* 2011, uma das primeiras abordagens desenvolvidas para a identificação de parceiros de interação que utiliza uma representação simples dos exemplos de treinamento e com bons resultados na identificação dos pares RNA-Proteína. Este grupo forneceu a comunidade a ferramenta computacional denominada *RPIseq*. Inicialmente realizamos a reprodução da abordagem que por sua vez possibilitou verificar a influência de RNAs ribossomais no comportamento do modelo para as predições além de utilizarmos dados de treinamento mais específico, por considerar apenas pares de interação com a proteína de interesse. Cada uma das etapas de processamento será brevemente descrita a seguir em suas respectivas subseções.

4.3.1 Fontes de dados

Duas principais fontes de dados foram utilizadas. A primeira advém dos dados que os próprios autores da abordagem RPI-Seq geraram. A geração dos exemplos consistiu do uso da base de dados Protein-RNA Interface Database - PRIDB (Lewis *et al.,* 2011) que coleta exemplos de parceiros de interação do Bando de Dados de Proteínas (*Protein Data Bank – PDB*) (Berman *et. al.,* 2000). Para a segunda fonte de dados foram considerados exemplos mais específicos no que se refere a proteína em

estudo e a partir de um levantamento bibliográfico, foram reunidos e utilizados os dados apresentados na Tabela 3, a seguir.

Tabela 3 – Exemplos de interação entre as proteínas Hfq/LSm e seus respectivos RNAs.

Organism	True positives	True negatives	Reference (Positives/Negatives)
Escherichia coli	20	154	Olejniczak, 2011; Zhang *et al.,* 2006 / Zhou e Rudd, 2013 – EcoGene
Bacillus subtilis	23	177	Dambach *et al.,* 201 / Karp *et al.,* 2005 – BioCyc Database
Haloferax volcanii	39	58	Straub *et al.,* 2009 / Karp *et al.,* 2005 – BioCyc Database
Listeria monocytogenes	3	85	Christiansen *et al.,* 2006 / Karp *et al.,* 2005 – BioCyc Database
Salmonella typhimurium	128	109	Chao *et al.,* 2012 / Karp *et al.,* 2005 – BioCyc Database

Na Tabela 3, os exemplos positivos referem-se aos RNAs (sRNAs e mRNAs) que foram identificados a partir de abordagens experimentais, como parceiros de interação da proteína Hfq em *E. coli*, *B. subtilis*, *L. monocytogenes* e *S. typhimurium* e também da proteína LSm em *H. volcanii*. Com exceção do organismo *S. Typhimurium* que possuem elementos positivos da categoria de RNA mensageiro (*mRNA*) e pequenos RNAs (*sRNAs*) os demais organismos possuem elementos somente da categoria de pequenas moléculas de RNA (sRNAs). Para os exemplos negativos foram considerados os RNAs presentes no banco de dados BioCyc e EcoGene. Para as análises, assumimos como exemplos negativos os demais RNAs disponíveis para o organismo que não fazem parte dos exemplos positivos. A escolha dos organismos apresentados anteriormente ocorre em função dos dados experimentais disponíveis até a presente data que possuem exemplos de pares que interagem com as proteínas de interesse.

Essa variação na fonte de dados, utilizada no processo de treinamento dos algoritmos de AM, conduz tanto para um aspecto mais

geral na predição de parceiros de interação por considerar diversos tipos de moléculas quanto para a tentativa de uma predição mais específica dos elementos que interagem com as proteínas Hfq/LSm. Dessa forma, com a escolha dos exemplos mais específicos, espera-se que os dados de treinamento sejam mais indicativos para a uma avaliação considerando a predição dos elementos já conhecidos e também por possibilitar o uso de exemplos de treinamento mais próximos dos parceiros a serem identificados.

4.3.2 Adaptação da abordagem *RPIseq*

Para a aplicação da metodologia que visa a predição de possíveis interações entre ncRNAs com a proteína LSm, consideramos o trabalho de Muppirala *et al.,* 2011 e desenvolvemos uma reprodução da abordagem. É um dos primeiros trabalhos propostos para o problema que utiliza apenas informações sobre a sequencia primária dos pares RNA-Proteína, apesar da proposta de representação dos dados de treinamento ser simples a abordagem foi capaz de separar bem diversos exemplos em diferentes organismos e dessa forma, dirigimos nossos estudos preliminares na investigação tanto da forma de representação dos exemplos de treinamentos quanto ao uso de exemplos de parceiros de interação disponíveis. A partir dessas considerações, nesta seção são descritos com mais detalhes os princípios envolvidos no trabalho de Muppirala *et al.,* 2011.

Utilizando pares de RNA-Proteína, extraídos a partir da base de dados PRIDB (Lewis *et al.,* 2011), como dados de treinamento e a partir de uma representação desses dados com informações extraídas apenas de suas respectivas sequências primárias, dois classificadores são aplicados para a criação do modelo de AM, um baseado em um conjunto de árvores de decisão (*Random Forest – RF*) e outro baseado em Máquinas de Vetores de Suporte (*Suport Vector Machines – SVM*). Os classificadores apresentam os resultados de forma independente, cada qual com seu próprio viés indutivo sobre os dados tanto na representação dos mesmos quanto na

forma de busca como descrito na seção 3.1. Vale ressaltar que a abordagem baseada em RF é um tipo específico de combinação de classificadores, por ponderar o resultado de diversas árvores de decisão através de um sistema simples de voto. De acordo com os autores, a escolha dessas técnicas provém de seu amplo uso em problemas relacionados e do êxito em tais aplicações para a obtenção de bons resultados na classificação de novos elementos.

Dois conjuntos de dados, denominados RPI2241 e RPI369, foram gerados para avaliação da performance de ambas as técnicas. Os autores extraíram os exemplos de pares de interação do *Protein data bank - PDB* utilizando a ferramenta *PRIDB* (disponível em http://pridb.gdcb.iastate.edu/index.php) (Lewis *et al.,* 2011), que atua como uma espécie de filtro para estruturas que compreendem RNA e proteína. Dessa forma, os autores obtiveram 2241 pares não redundantes, para o conjunto RPI2241, no qual são incluídos diversos tipos de moléculas de RNA como RNAs mensageiros e outros tipos de RNAs não codificadores (RNAs ribossomais, RNAs transportadores, micro-RNAs, entre outros). Para o conjunto RPI369 foram excluídos os pares envolvendo RNAs ou proteínas ribossomais resultando em 369 pares. Os exemplos negativos foram gerados a partir de um embaralhamento aleatório desses pares positivos onde os exemplos com mais de 30% de identidade na sequência primária ou que estão presentes no conjunto positivo são descartados. Essa geração de exemplos negativos possibilita que os padrões nos exemplos positivos que contribuem na interação sejam desfeitos.

Sobre esses conjuntos de dados foi aplicada uma representação essencialmente baseada na frequência de nucleotídeos e aminoácidos presentes em cada par RNA-proteína. Esse tipo de representação foi anteriormente aplicado à predição de interação entre proteína-proteína (*Protein-protein interactions*) descrito em Shen *et al.,* 2007, e também na predição de proteínas ligantes a RNA (*RNA-binding proteins*) e dessa forma inspiraram os autores a optarem por essa mesma codificação de dados. Na Figura 14 é ilustrado o esquema de representação.

Cada par RNA-proteína é representado como um vetor de 599

atributos, no qual 343 são usados para codificar a sequência da proteína e 256 para codificar a sequência de RNA (Na Figura 14 é ilustrada apenas a representação da proteína). No método proposto por Shen *et al.,* 2007, os 20 aminoácidos são classificados em 7 grupos de acordo com as informações de dipolo e volume das cadeias: {*A, G, V*}, {*I, L, F, P*}, {*Y, M, T, S*}, {*H, N, Q, W*}, {*R, K*}, {*D, E*}, {*C*}.

A partir de uma janela deslizante de 3 aminoácidos a sequência da proteína é percorrida e a contagem da frequência de cada trinca é armazenada no vetor de características *F* na posição correspondente a trinca no vetor *V,* uma vez que cada posição da trinca representa um dos subconjuntos, o tamanho total do vetor *V* é de 7x7x7 = 343. Por exemplo, a primeira posição do vetor *V* (ilustrada na cor rosa) representa uma trinca em que cada um dos 3 elementos corresponde ao primeiro dos 7 subconjuntos, ou seja, {*A, G, V*}. Durante a leitura da sequência primária, em cada ocorrência de uma das variações possíveis dos aminoácidos presentes no subconjunto é feito o incremento da contagem armazenada na primeira posição do vetor *F* que por sua vez, se refere aos elementos da primeira posição do vetor *V*.

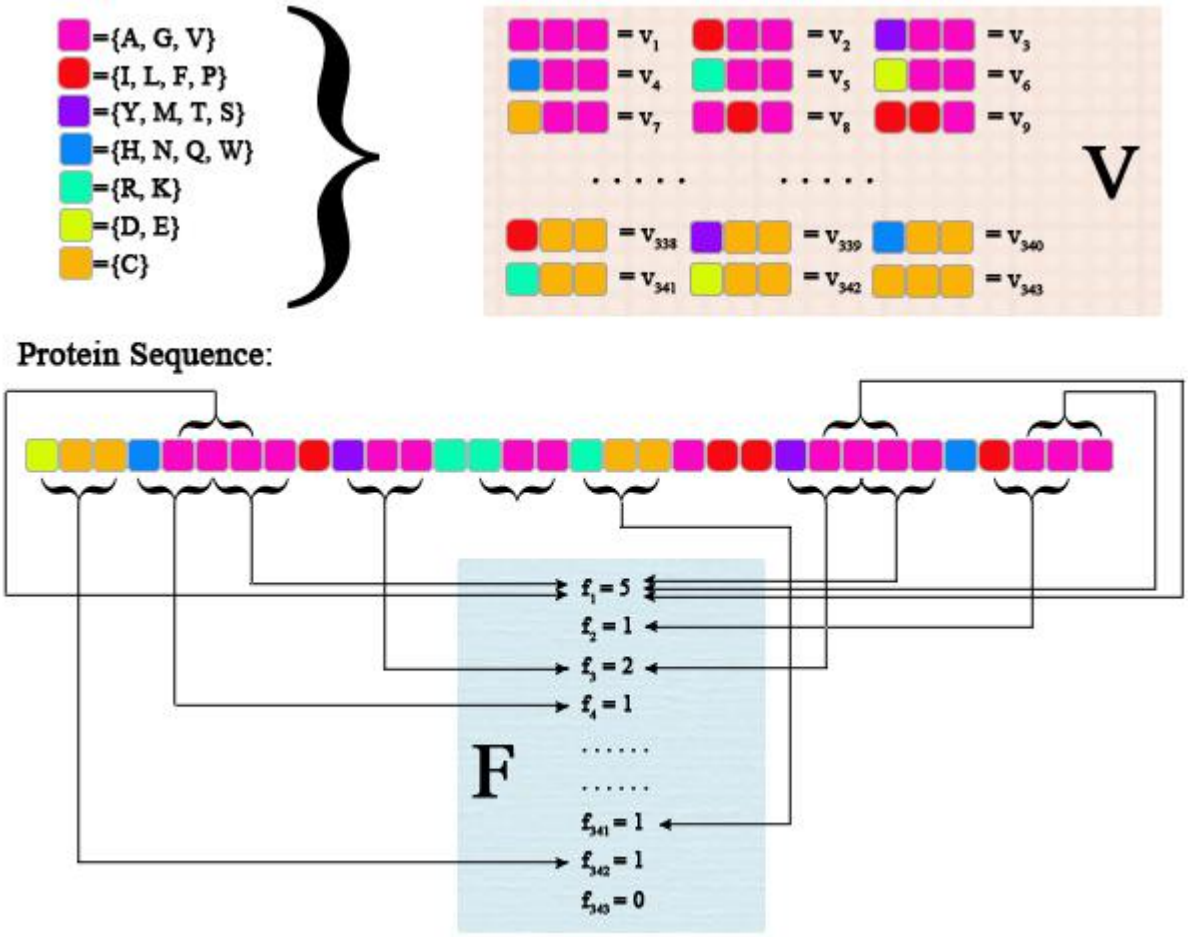


Figura 14 – Esquema de representação do conjunto de dados baseado em frequência de cada aminoácido. As cores correspondem os subcojuntos de aminoácidos. O vetor V corresponde a todas as possíveis combinações de trincas de aminoácidos geradas considerando o subconjunto. F é a contagem de todas as ocorrências das trincas em um determinada sequencia de proteína (protein sequence) O mesmo princípio é aplicado a sequência de RNA. (extraído de Shen et al., 2007).

As sequências de RNAs são codificadas da mesma forma, porém não foram subdivididos em subgrupos como os aminoácidos e também existe uma variação do tamanho da janela de leitura, com deslocamento de 4 nucleotídeos por leitura. Isso implica num outro vetor V2 com tamanho de 4x4x4x4 resultando em 256 outras posições. Por fim, cada instância é então codificada na forma de um vetor de 599 posições que provém da

junção de ambas as representações (par proteína-RNA) mencionadas.

5 Resultados

Neste Capítulo são apresentados os resultados alcançados no processo de identificação de novos ncRNAs em *H. salinarum NRC-1* e na aplicação metodologias para a predição dos parceiros de interação com a proteína LSm presente no organismo modelo em estudo. Separamos os resultados em dois subcapítulos, cada qual tratando os temas abordados nesta Tese.

5.1 Identificação de ncRNAs

Como parte dos resultados obtidos nos procedimentos desenvolvidos para a tentativa de descoberta de novos RNAs não codificadores (ncRNAs) em *H. salinarum* NRC-1, exploramos diversas variações para a adaptação da metodologia *incRNA* (Lu *et al.,* 2011) até chegarmos ao *workflow* final, apresentado na seção 4.1 (Figura 9), e que possibilitou atingir o objetivo proposto de descobrir novos ncRNAs em *H. salinarum* NRC-1. Incluímos nessas variações: diferentes maneiras de definir as regiões anotadas para treinamento do modelo de AM, análises com remoção de atributos, diferentes estratégias para definição e aplicação dos dados na etapa de predição e finalmente o estudo e interpretação dos resultados.

Como mencionado, essas variações foram necessárias uma vez que as informações disponíveis para o organismo modelo não correspondem a todas os procedimentos da metodologia incRNA. Dessa forma, ao invés de obter informações de anotações de um organismo próximo utilizamos as anotações até então disponíveis do próprio organismo em estudo. Inicialmente usamos essas anotações seguindo a definição original, ou seja, respeitando as informações apresentadas para o início e fim de cada trecho. Em seguida, particionamos os trechos anotados e utilizamos de outra forma as anotações. Com isso, analisamos a influência dessas anotações na criação do modelo e no processo de inferência.

Este capítulo está organizado da seguinte maneira, na primeira

seção é apresentado a primeira proposta de adequação da abordagem incRNA, onde integramos diversas fontes de dados e avaliamos algumas propostas para inferência de regiões sem anotações. Em seguida, redefinimos os modelos experimentando de forma distinta os exemplos de treinamento e propomos outra estratégia para a identificação de regiões a partir de um procedimento predição com sobreposição de trechos ao longo do genoma. Os resultados das análises são apresentados na seção 5.1.4 com a descrição dos procedimentos desenvolvidos para a filtragem e seleção das regiões.

5.1.1 Integração de dados e uso de regiões anotadas

No que se refere as modificações realizadas na manipulação das regiões anotadas, inicialmente utilizamos as informações de posição inicial e final de acordo com as definições apresentadas originalmente em cada trecho. Nessa etapa, utilizamos apenas os dados de Koide *et al.,* 2009b como exemplos de treinamento da classe denominada como “ncRNA” e deixamos os dados do *Genome Browser* da Universidade Santa Cruz Califórnia (UCSC) para realizamos um teste como conjunto de dados independente. Na Figura 15 é apresentada a distribuição dos exemplos em suas respectivas classes de anotação.

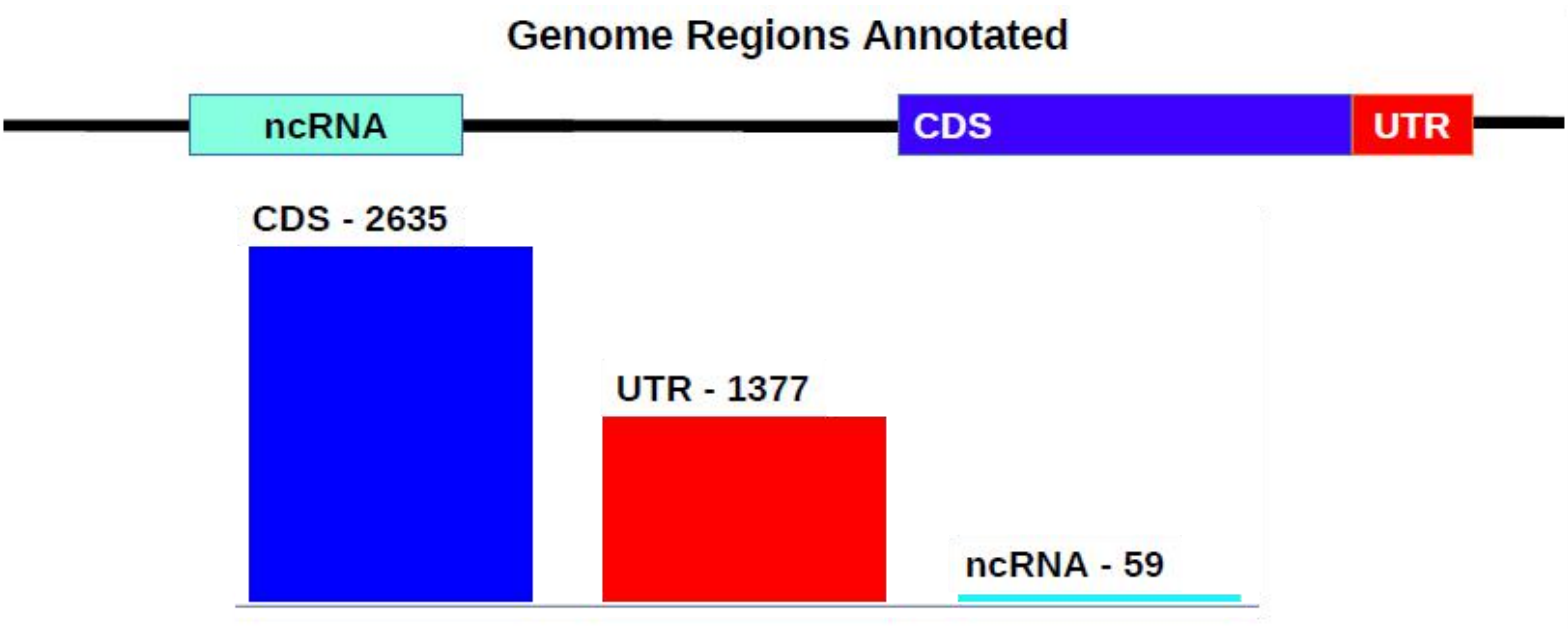


Figura 15 – Regiões do genoma com o número de exemplos utilizadas na criação do modelo de AM.

Avaliamos o modelo utilizando validação cruzada (*10-fold-cross-validation*). Nas Tabelas 4 e 5 são apresentados os resultados para os classificadores Random Forest e J48 uma vez que ambos obtiveram melhores resultados quando comparados a outras técnicas disponíveis na ferramenta WEKA. Apesar de o classificador Random Forest ter como resultado uma ligeira melhor acurácia para a predição de todas as classes, 93,1 % dos elementos classificados corretamente contra 92,7 %, o classificador J48 conseguiu separar melhor os exemplos da classe "ncRNA" confundindo-se em apenas quatro exemplos, sendo 3 ncRNAs preditos como UTR e 1 exemplo predito como CDS (Tabela 5).

Tabela 4 – Resultados para a avaliação cruzada considerando o classificador Random Forest (RF).

		Predict as:		
		CDS	**UTR**	**ncRNA**
Current class	**CDS**	2554	81	0
	UTR	186	1190	1
	ncRNAs	5	8	46

Tabela 5 – Resultados para avaliação cruzada considerando o classificador J48.

		Predict as:		
		CDS	**UTR**	**ncRNA**
Current class	**CDS**	2522	113	0
	UTR	173	1198	6
	ncRNAs	1	3	55

Analisando os possíveis erros das predições, foi constatado que o exemplo da classe “ncRNA” predito como CDS nos resultados obtidos pelo classificador J48 na verdade não estava incorreto, de acordo com as atualizações dos dados, um dos elementos que em Koide *et al.,* 2009b foi indicado como ncRNA na verdade refere-se a um gene (Figura 16) e dessa forma, o modelo de AM gerado foi capaz de corrigir as informações com base no padrão presente nos próprios dados de treinamento.

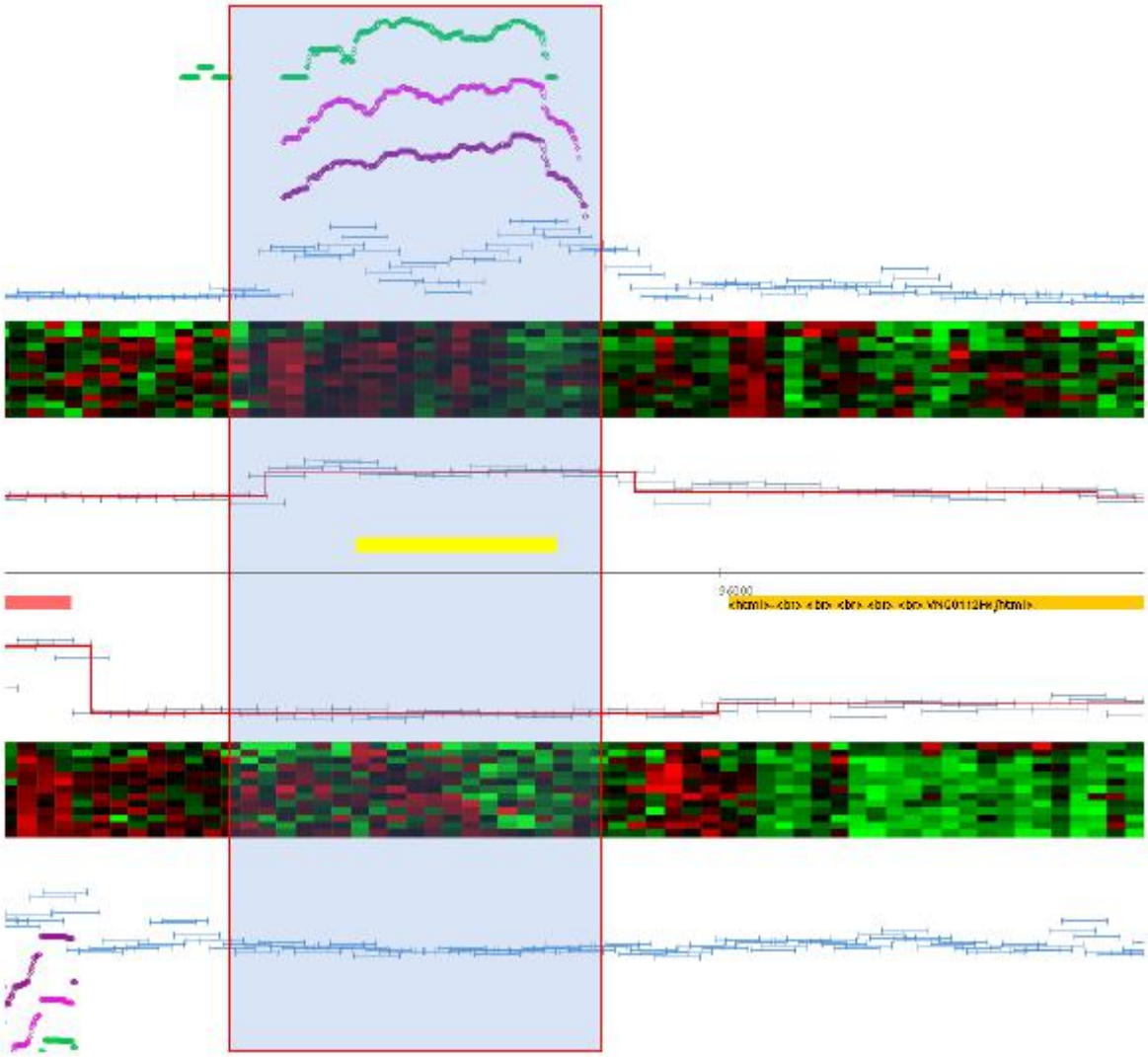


Figura 16 – Região selecionada em azul indicando o trecho que coincide com um gene codificante (em amarelo).

Esses resultados indicaram que a adaptação da metodologia pode representar e separar bem os exemplos das classes associadas a cada região anotada para o organismo em estudo. Elementos da classe ncRNA foram separados das demais e poucos erros foram cometidos na predição. Nenhum elemento da classe CDS foi predito como ncRNA. Em outros casos, nem todos elementos da classe UTR e ncRNAs foram corretamente classificados uma vez que são regiões mais difíceis de separar devido as próprias características de ambos os trechos genômicos. Em certa medida,

uma região UTR é, de fato, não-codificadora apesar de tecnicamente não ser um ncRNA.

Em uma segunda avaliação foi considerado um conjunto de teste independente, com exemplos obtidos através do *Genome Browser* da Universidade da Califórnia Santa Cruz (UCSC). De acordo com as anotações disponíveis no *website* do navegador, 41 elementos foram preditos pela ferramenta snoscan (Lowe & Eddy, 1999) como possíveis ncRNAs. A ferramenta utiliza outros procedimentos para identificação de trechos pertencentes a essa classe de ncRNAs ao longo do genoma. Dentre as características considerada na abordagem, é incluído informações de motivos conhecidos para as famílias “C/D Box” de snoRNA. Após a remoção de alguns elementos redundantes, 38 exemplos foram utilizados como teste.

Ao aplicarmos os dados de teste aos modelos descritos anteriormente obtivemos que 45% dos exemplos foram preditos corretamente utilizando o classificador J48 contra 29% utilizando Random Forest. Ambos classificadores confundiram os demais exemplos com a classe UTR. Novamente, nenhum exemplo dito como ncRNA foi predito como pertencente a classe CDS, mostrando que as características representadas nos dados são capazes de distinguir bem cada subconjunto. UTR são, de fato, não-codificadoras e portanto a confusão de qualquer modelo matemático em separar ncRNA e UTR rigidamente não é indício grave de falha. Os números na tabela, por mais que tenham indicado mais erros do que acertos, se considerarmos estritamente as classes, não indicam um resultado ruim dada essa dificuldade de separação UTR e ncRNA. O importante é que jamais uma CDS seja classificada como ncRNA e vice-versa. A classificação aqui é, portanto, considerada suficientemente bem sucedida.

Finalmente, procuramos definir uma metodologia para a aplicação dos trechos sem anotações visando a busca de novos trechos candidatos a ncRNAs. Dessa forma, realizamos o seguinte procedimento. Trechos do genoma já anotados e também que não possuem sinal de expressão em bibliotecas de sRNA-Seq foram descartados como é ilustrado na Figura 17.

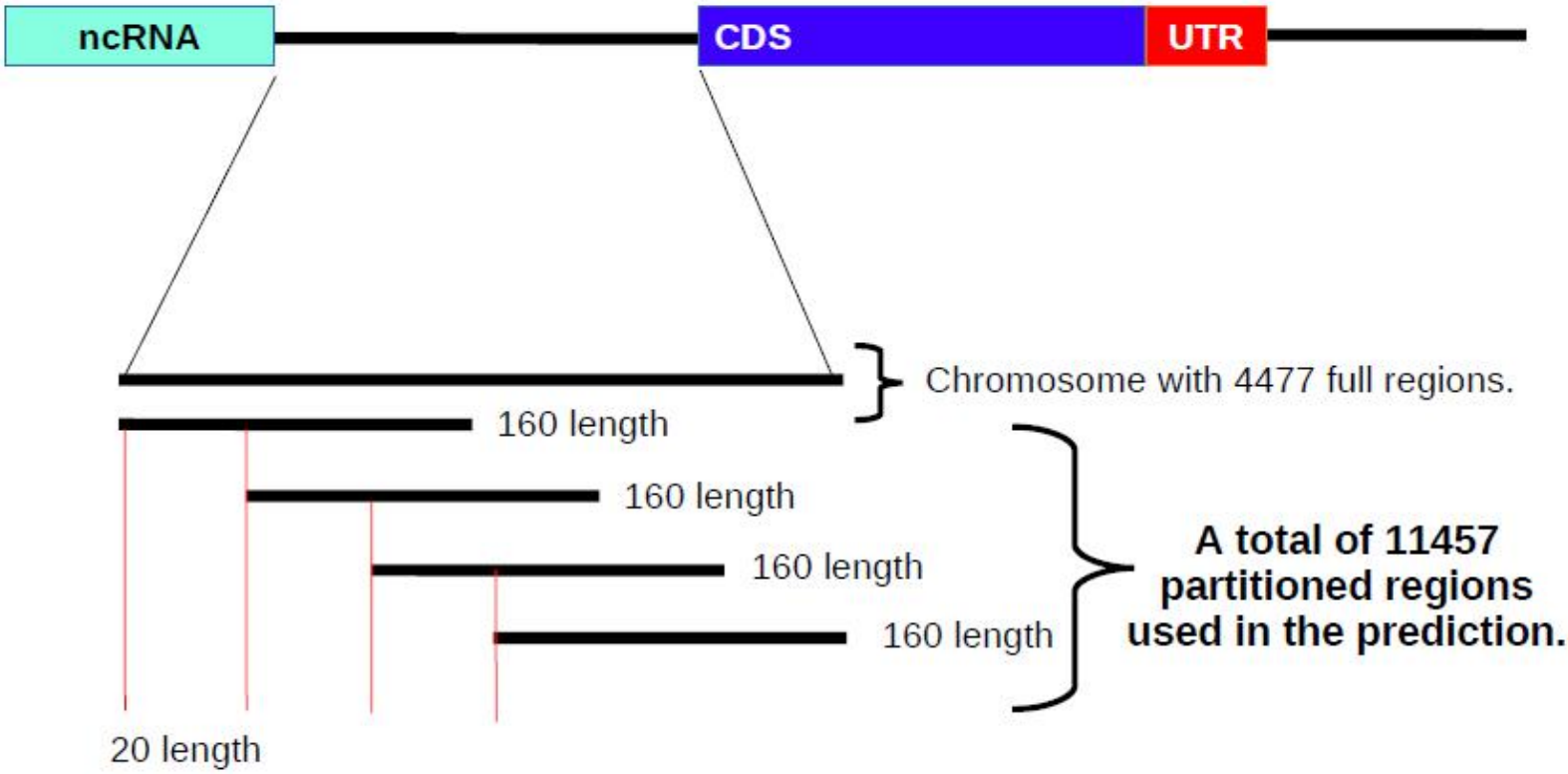
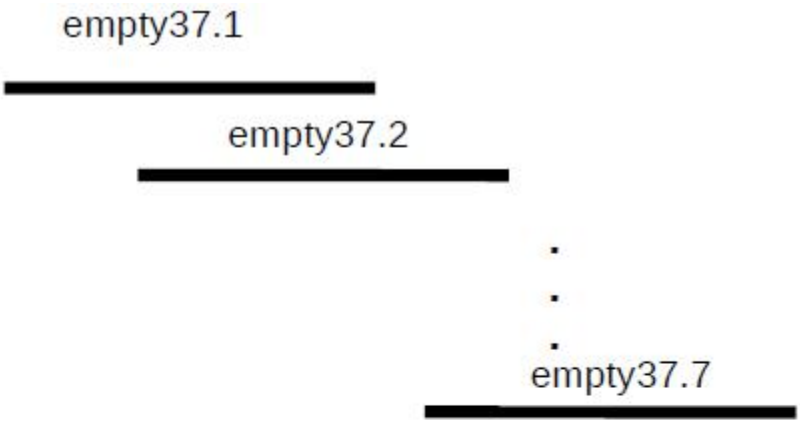


Figura 17 – Procedimento para a definição dos trechos sem anotações à serem preditos.

No total foram obtidos 4477 trechos para o cromossomo e uma vez que esses trechos possuem tamanhos diversos foi aplicado um particionamento de todas as regiões em que cada parte possui tamanho máximo de até 160 pares de base. Esse valor foi baseado no tamanho médio dos ncRNAs conhecidos para *H. salinarum* NRC-1 (Koide *et al.,* 2009b). Por serem regiões sem informação prévia, o início do transcrito, assim como o fim, pode estar em qualquer parte do trecho e dessa forma, foi aplicada um deslocamento de 20 pares de base ao longo do particionamento no intuito de explorar diferentes posições de início e fim (Figura 17). No total 11457 regiões foram utilizadas na predição.



Name	Probability	Class
chr_empty37.1	0.8	ncRNA
chr_empty37.2	0.9	ncRNA
chr_empty37.3	1	ncRNA
chr_empty37.4	0.9	ncRNA
chr_empty37.5	0.9	ncRNA
chr_empty37.6	0.9	ncRNA
chr_empty37.7	0.9	ncRNA

Figura 18 – Probabilidade associada a cada trecho de pertencer a classe ncRNA.

Como resultado da predição, cada trecho possui um valor de probabilidade associado ao elemento de pertencer à determinada classe (Figura 18). Uma vez que o número de elementos a serem preditos é muito grande, muitos desses elementos foram classificados como pertencentes à classe ncRNA e alguns critérios foram considerados para selecionar os candidatos. Somente elementos com probabilidade igual a 1 e flanqueados por outros dois elementos com probabilidade igual a 1 foram escolhidos para a lista de possíveis ncRNAs. Após esse procedimento, dentre os 11457 elementos utilizados na predição, restaram ainda 5328 elementos que foram preditos como possíveis ncRNAs.

Essa quantidade não corresponde com o esperado para um organismos de genoma pequeno como *H. salinarum*. Assim como ocorre em outros organismos, como por exemplo em *H. volcanii* e *E. coli*, cerca de apenas algumas centenas de ncRNAs estão presentes ao longo do

genoma (Soppa *et al.,* 2009). O número elevado de candidatos deve-se principalmente ao procedimento utilizado na definição das regiões a serem inferidas no modelo. Dentre as razões, cada trecho possui várias sobreposições e as definições tornaram os exemplos muito distintos dos exemplos de treinamento. Dessa forma, o modelo tendeu a ser otimista para os possíveis candidatos, gerando como saída muitos elementos putativos e certamente muitos falsos positivos.

Buscamos uma alternativa para a definição das regiões a serem preditas explorando informações existentes sobre os dados de expressão, ao invés de particionar todo o genoma, obtemos somente os trechos que possuem sinais de expressão e definimos o início e fim do trecho com base no início e fim das contagens de *reads* alinhados. Esse procedimento está ilustrado na Figura 19.

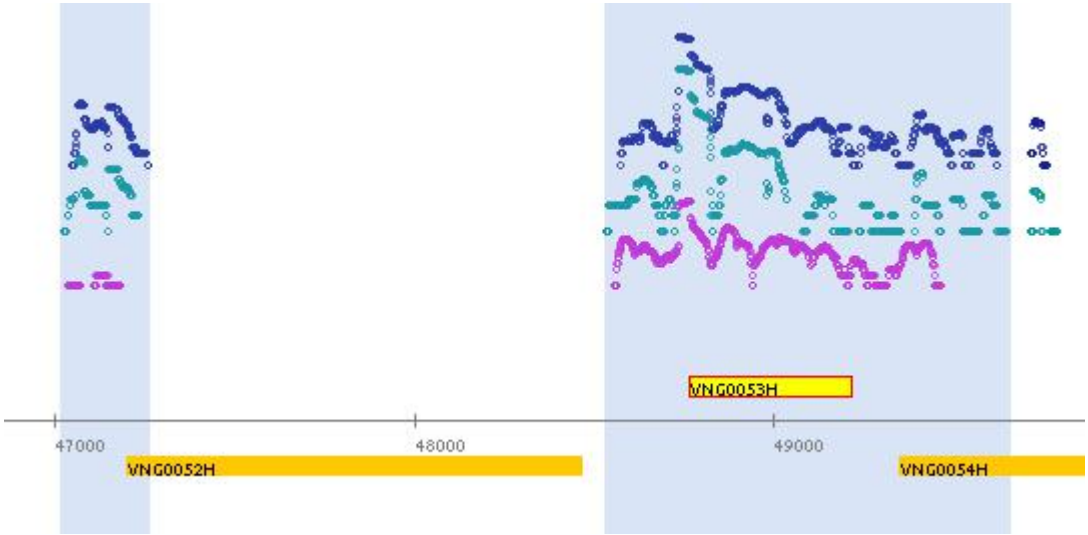


Figura 19 – Definição dos trechos a serem utilizados no processo de inferência com base nos sinas de expressão. Trecho em destaque indica o início e fim de cada região.

Na tentativa de adequar melhor os dados de treinamento para essa nova configuração nos dados de inferência, decidimos associar as regiões UTRs aos genes uma vez que ao aplicarmos trechos diversos do genoma, com sinais de expressão, é improvável que somente regiões UTRs isoladas sejam expressas sem os genes associados. Então treinamos o modelo com

essa nova configuração e denominamos a classe UTR como CDS/UTR.

Mesmo com essa nova configuração não obtivemos sucesso no processo de inferência e dessa forma, vários elementos ainda foram ditos como ncRNAs, tornando inviável posteriores análises.

5.1.2 Redefinição dos modelos de AM

Uma vez que ao utilizarmos informações sobre as anotações considerando início e fim como definido originalmente dificultou a etapa de inferência no modelo, buscamos outras alternativas para a criação do modelo e, posteriormente, definição de novos trechos a serem preditos. Verificamos uma variação nos dados de treinamento que consiste em particionar os exemplos para que o tamanho desses trechos não ocasionasse algum tipo de viés no modelo, uma vez que os trechos de ncRNAs são menores e todos os atributos levam em consideração a região definida. No particionamento consideramos um tamanho fixo de 120 nucleotídeos (Lertampaiporn *et al.,* 2014) para realizar as subdivisões de cada exemplo da classe CDS e CDS/UTR e mantemos como na forma original os trechos pertencentes à classe ncRNA. Com essa modificação, a distribuição dos exemplos foi alterada como é ilustrado na Figura 20.

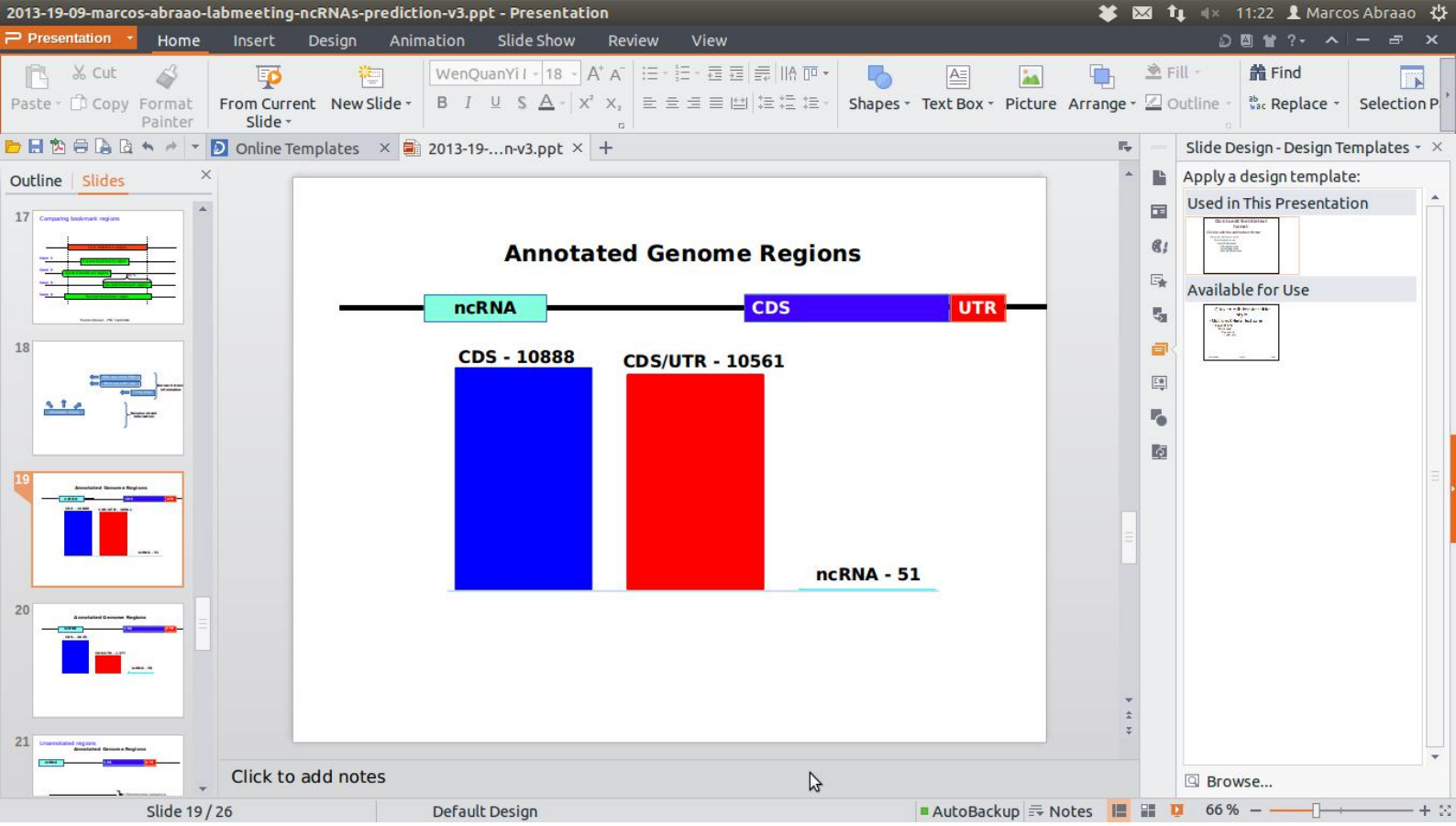


Figura 20 – Distribuição dos exemplos e suas respectivas anotações genômicas. Os valores indicam o número de exemplos gerados com particionamento das regiões que pertencem às classes CDS e CDS/UTR, exemplos da classe ncRNAs foram filtrados e alguns que não possuíam sinal de expressão foram removidos.

É necessário notar que alguns exemplos da classe ncRNAs foram removidos por não possuírem sinais de expressão nos dados coletados. Ainda, como discutido anteriormente, nos resultados preliminares que obtivemos durante a criação dos primeiros modelos de AM, identificamos um erro de anotação em um dos exemplos de ncRNAs e atualizamos todas essas informações.

A partir dessas definições, calculamos os valores de cada atributo, seguindo as especificações apresentada na seção 4.1.2, e analisamos o comportamento dos classificadores em uma validação cruzada. Variamos os modelos de forma a avaliar a influência dos dados de treinamento considerando três configurações. A primeira refere-se ao uso dos dados de treinamento com as classes CDS e CDS com regiões UTRs associadas. Isso significa que particionamos os dados de treinamento tanto para os exemplos com anotações CDS e CDS/UTR (Figura 20).

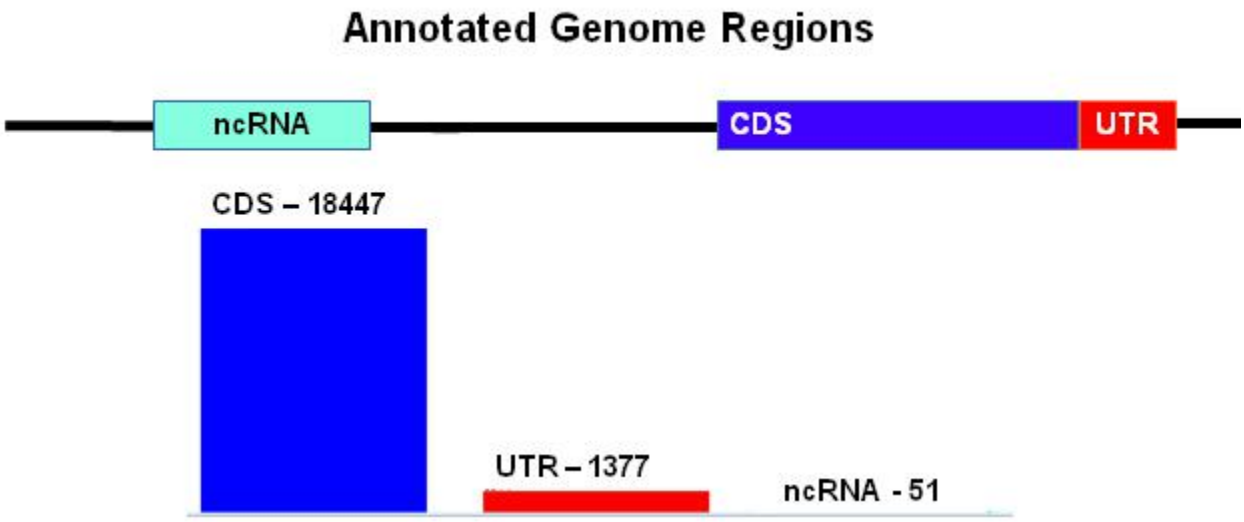


Figura 21 – Distribuição dos exemplos com suas respectivas anotações genômicas. Os valores indicam o número de exemplos gerados com particionamento das regiões que pertencem às classes CDS e UTR.

Nas outras duas configurações particionamos somente os exemplos da classe CDS e mantivemos os trechos pertencentes a classe UTR como na definição original. Variamos nessas configurações os exemplos da classe ncRNA e em um dos modelos consideramos os exemplos filtrados da classe ncRNA (Figura 21) e na segunda incluímos todas as anotações disponíveis, ou seja, tanto anotações de Koide *et al.,* 2009b quanto os dados do *Genome Browser* da Universidade Santa Cruz Califórnia (UCSC) (Figura 22). Os resultados dessas análises para uma avaliação baseada em validação cruzada estão apresentados nas Tabelas 6, 7 e 8.

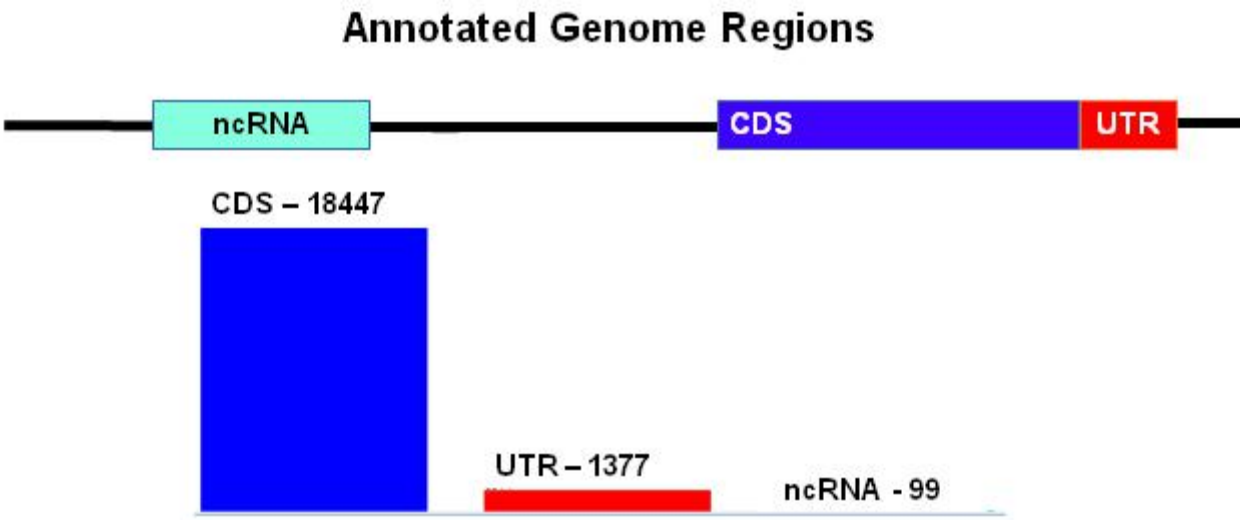


Figura 22 – Distribuição dos exemplos com suas respectivas anotações genômicas. Os valores indicam o número de exemplos gerados com particionamento das regiões que pertencem às classes CDS. Exemplos da classe UTR foram mantidos como na anotação original. Incluímos nessa variação todos os exemplos disponíveis para a classe ncRNA (Koide *et al,*2009b, snoRNAs).

Tabela 6 – Resultados da aplicação de uma validação cruzada (10 fold crossvalidation) com os dados da Figura 20. Valores da medida de AUC em cada classe para cada um dos classificadores.

Model 02 - Training CDS and CDS_UTR bins, filtered ncRNAs.			
Algorithm	Prediction performance AUC		
	CDS	UTR	ncRNA
Naive Bayes	0.65	0.64	0.95
Bayes Net	0.70	0.70	0.93
Decision Tree	0.62	0.62	0.69
Rules Based	0.69	0.69	0.64
Random Forest	0.73	0.73	0.83
Logistic Regression	0.69	0.69	0.83
SVM with Linear Kernel	0.58	0.58	0.78
SVM with 2nd degree Poly Kernel	0.64	0.64	0.92
SVM with RBF Kernel	0.68	0.68	0.92

Tabela 7 – Resultados da aplicação de uma validação cruzada (10 fold crossvalidation) com os dados da Figura 21. Valores da medida de AUC em cada classe para cada um dos classificadores.

Model 03 - Training CDS bins, UTR, filtered ncRNAs.			
Algorithm	Prediction performance AUC		
	CDS	UTR	ncRNA
Naive Bayes	0.96	0.94	0.89
Bayes Net	0.98	0.97	0.95
Decision Tree	0.90	0.88	0.40
Rules Based	0.90	0.90	0.62
Random Forest	0.99	0.99	0.91
Logistic Regression	0.90	0.90	0.86
SVM with Linear Kernel	0.82	0.82	0.83
SVM with 2nd degree Poly Kernel	0.88	0.88	0.85
SVM with RBF Kernel	0.97	0.96	0.83

Tabela 8 – Resultados da aplicação de uma validação cruzada (10 fold crossvalidation) com os dados da Figura 22. Valores da medida de AUC em cada classe para cada um dos classificadores.

Model 04 - Training CDS bins, UTR, all ncRNA examples.			
Algorithm	Prediction performance AUC		
	CDS	UTR	ncRNA
Naive Bayes	0.96	0.94	0.85
Bayes Net	0.97	0.98	0.95
Decision Tree	0.92	0.90	0.59
Rules Based	0.89	0.90	0.61
Random Forest	0.99	0.99	0.94
Logistic Regression	0.89	0.89	0.78
SVM with Linear Kernel	0.79	0.79	0.72
SVM with 2nd degree Poly Kernel	0.90	0.91	0.77
SVM with RBF Kernel	0.96	0.97	0.88

De acordo com os resultados das Tabelas 6, 7 e 8 as diferentes configurações propostas nos dados de treinamento provocam variações nas performances de cada algoritmo. De maneira geral, dados de treinamento com exemplos da classe ncRNAs filtrados, ou seja, dados de treinamento com a remoção de: ncRNAs sem sinal de expressão, tRNAs e rRNAs, porém mantidos snoRNAs com sinal de expressão, apresentaram melhores resultados na medida de AUC para a classe ncRNA. Também notamos que usando informações de UTR em conjunto com seus respectivos trechos CDS melhora a performance relativo a quando

utilizados trechos com anotações de UTR como uma classe separada.

Tabela 9 – Resultados da aplicação de uma validação cruzada (10 fold crossvalidation) com os dados da Figura 15. Valores da medida de AUC em cada classe para cada um dos classificadores.

Model 01 - Training annotated regions with original length.			
Algorithm	Prediction performance AUC		
	CDS	UTR	ncRNA
Naive Bayes	0.97	0.96	0.86
Bayes Net	0.98	0.97	0.94
Decision Tree	0.98	0.96	0.73
Rules Based	0.99	0.98	0.86
Random Forest	0.99	0.99	0.97
Logistic Regression	0.98	0.97	0.86
SVM with Linear Kernel	0.96	0.94	0.79
SVM with 2nd degree Poly Kernel	0.97	0.96	0.86
SVM with RBF Kernel	0.98	0.98	0.87

Verificamos ainda que ao utilizarmos as anotações sem o particionamento dos trechos o modelo conseguiu separar melhor os exemplos e atingiu uma medida de AUC superior em 8 dos 9 classificadores utilizados (Tabelas 6 e 9). Buscamos então avaliar o comportamento dos modelos na identificação de novos trechos candidatos a pertencerem à classe ncRNA a partir de um procedimento de janela deslizante ao longo do genoma, como apresentado na seção 4.1.4. Os resultados estão descritos no capítulo a seguir.

5.1.3 Geração da faixa genômica

Inicialmente, aplicamos o procedimento de janela deslizante ao cromossomo na fita *forward* para avaliar o comportamento dos modelos até então gerados. No total 50354 trechos foram utilizados nessa primeira análise. Esses trechos referem-se ao particionamento de toda a fita *forward* do cromossomo com tamanho de 120 nucleotídeos e descolcamento de 40 pares de bases entre um trecho e outro, dessa forma, 80 bases são sobrepostas entre dois trechos consecutivos. Esse procedimento equivale a ilustração da Figura 18 no que se refere ao

deslocamento e sobreposições porém, com a alteração no tamanho do trecho e no deslocamento. A escolhas do tamanho do trecho e quantidade de bases no deslocamento citadas anteriormente foram baseadas nas mesmas considerações apresentadas em (Lertampaiporn *et al.,* 2014).

Cada um dos 50354 trechos foi aplicado aos modelos e a probabilidade associada a cada uma das classes foi então inferida. Para a geração da faixa genômica consideramos a probabilidade obtida para a classe ncRNA e normalizamos os valores de acordo com as sobreposições. Ou seja, se por exemplo, três trechos sobrepõem uma determinada posição do genoma, calculamos a média da probabilidade dessa posição com base no valor associado a cada um dos três trechos. Ao final, cada posição possui apenas um valor de probabilidade e esses valores podem então ser visualizados como um sinal ao longo do genoma, compondo assim a faixa genômica. Cada Figura a seguir representa a faixa genômica somente para os três melhores classificadores, baseado na media de AUC, de acordo com cada uma das configurações. Por exemplo, para o modelo que utiliza os exemplos sem o particionamento (Modelo 01), os classificadores com melhor performance foram Redes Bayesianas (Bayes Net), Random Forest e SVM com *kernel* RBF cujos valores de AUC são 0,94, 0,97 e 0,87, respectivamente.

Na Figura 23, as faixas genômicas indicam a tendência de determinado trecho pertencer a classe ncRNA na medida em que o sinal possui picos de probabilidade superiores a outros trechos da mesma faixa. Para facilitar a visualização dos picos a área das curvas foram preenchidas. Com isso, notamos que o sinal gerado com o modelo sem particionamento (Modelo 01) possui picos em regiões pertencentes a classe CDS. Deste modo, vários trechos levam a prováveis falsos positivos. Para a definição dos trechos de maior probabilidade utilizados um procedimento que determina o início e fim do pico avaliando os valores ao longo do sinal e identifica a variação dos valores em relação à média de todos os valores da faixa genômica. Por exemplo, para o sinal da abordagem Random Forest (sinal em verde) os picos obtidos com o procedimento estão apresentados na Figura 24. Esse procedimento também foi descrito na

seção 4.1.5.

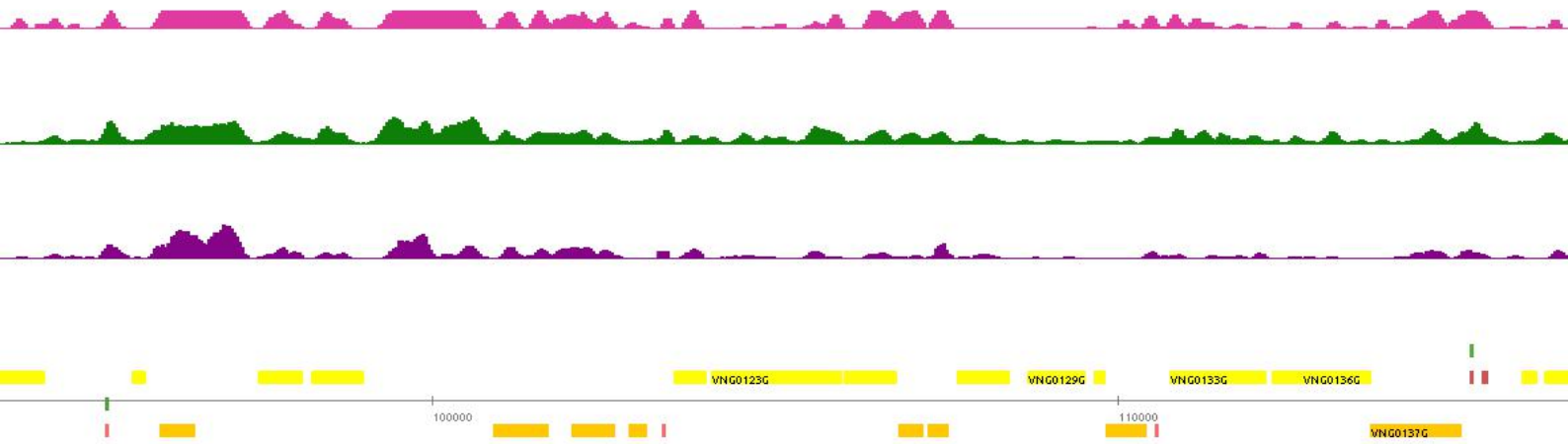


Figura 23 – Genome browser com a representação das faixas genômicas obtidas com o Modelo 01 (sem particionamento dos exemplos de treinamento). Em lilás os valores obtidos com a abordagem baseada em Redes Bayesianas (Bayes Net), em verde os valores da abordagem Random Forest e em roxo os valores da abordagem SVM com kernel RBF. Caixas em amarelo indicam os genes anotados da fita *forward* e em laranja as anotações dos genes da fita *reverse*.

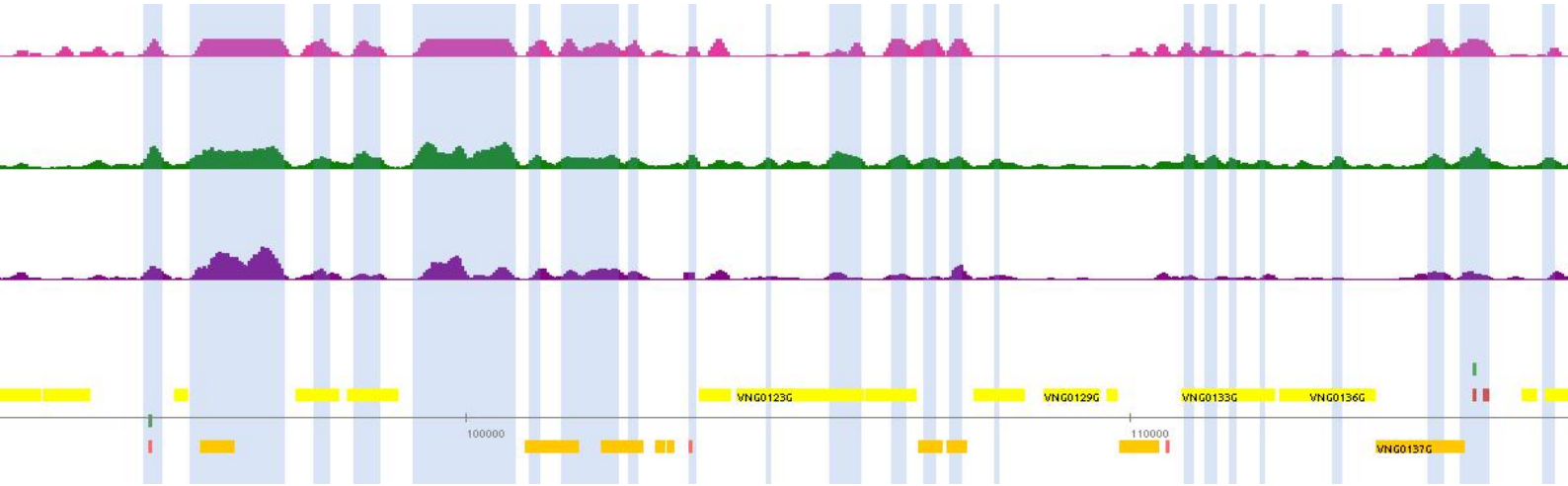


Figura 24 – Genome browser com a representação dos trechos que foram determinados a partir dos picos de probabilidade obtidos com o classificador Random Forest no Modelo 01 (sinal em cor verde). Na imagem os trechos identificados estão destacados por faixas verticais em azul claro.

A partir dos trechos determinados com o procedimento descrito anteriormente, verificamos quais destes coincidem com regiões já anotadas. Os resultados dessa avaliação estão na Tabela 10. Como pode ser observado, certa de 45% dos trechos obtidos pelo classificador baseado em Redes Bayesianas possuem anotações, dessas 44% são falsos positivos. Os trechos gerados pelos dois outros classificadores possuem cerca de 40% e 30% de regiões coincidindo com anotações. Outro problema com os resultados desse modelo é a quantidade total de trechos obtidos. Por exemplo, a abordagem Random Forest gerou para o cromossomo na fita *forward* 2232 trechos com tendência de serem ncRNAs e de acordo com a avaliação preliminar, favorece a inclusão de muitos falsos positivos.

Tabela 10 – Comparação dos resultados obtidos com 3 melhores classificadores do modelo 01 e anotações existentes.

Model 01 - Top 3 classifiers sliding window results								
	Total	**%**	**CDS**	**%**	**UTR**	**%**	**ncRNAs**	**%**
Bayes Net	1838	44.8	742	40.4	75	4	7	0.4
Random Forest	2232	41	825	37	84	3.8	8	0.3
SVM RBF	1781	31	495	28	50	2.8	8	0.4

Avaliamos outros dois modelos apresentados anteriormente e os resultados foram mais promissores e importante para o andamento das análises. Ao utilizarmos dados de treinamento com particionamento, ocorreu uma redução no número dos picos gerados e também uma redução no número de falsos positivos. Os modelos considerados incluem em primeiro caso aqueles que foram gerados a partir do particionamento nos exemplos da classe CDS e com os exemplos UTR mantidos de acordo

com a definição de início e fim original (Modelo 03) e no segundo caso aqueles cujas anotações UTRs estão associadas aos seus respectivos CDS, e tanto essa classe CDS/UTR quando a classe CDS foram também particionadas (Modelo 02). Os resultados da aplicação do procedimento de janela deslizante e o processamento dos picos estão apresentados nas Tabelas 11 e 12.

Tabela 11 – Comparação dos resultados obtidos com 3 melhores classificadores do modelo 02 e anotações existentes.

Model 02 - Top 3 classifiers								
	Total	**%**	**CDS**	**%**	**UTR**	**%**	**ncRNAs**	**%**
Bayes Net	318	31.3	27	8.5	68	21.3	4	1.2
Naive Bayes	583	27.4	58	9.9	98	16.8	4	0.7
SVM RBF	539	20.9	41	7.6	66	12.2	6	1.1

Tabela 12 – Comparação dos resultados obtidos com 3 melhores classificadores do modelo 03 e anotações existentes.

Model 03 - Top 3 classifiers								
	Total	**%**	**CDS**	**%**	**UTR**	**%**	**ncRNAs**	**%**
Bayes Net	929	12	48	5	61	7	3	0.3
Random Forest	1472	27	198	13	190	13	8	0.6
Naive Bayes	539	22.2	57	10.6	60	11	3	0.6

Como pode ser observado, tanto o número de trechos gerados a partir dos picos quanto o número de anotações existentes para esses trechos diminuíram. Por exemplo, no Modelo 01 foram obtidos 1838 trechos com a abordagem Bayes Net e nos Modelo 02 e 03 para a mesma abordagem os números foram reduzidos para 318 e 929, respectivamente. Já no que se refere as anotações que coincidem com os trechos, o número de trechos falsos positivos que antes nessa mesma abordagem estava em cerca 44% passou para 29.8% (Modelo 02) e 12% (Modelo 03). Vale ressaltar ainda que os 742 trechos que coincidem com CDS, obtidos pela abordagem Bayes Net no Modelo 01 (sem particionamento), equivale a

73% dos CDS existentes na fita *foward* do cromossomo. Nos outros dois modelos o valor é de 2.6% e 4.7% para os Modelos 02 e 03. Na Figura 25 são ilustradas as regiões que foram determinadas pelo classificador Bayes Net, do Modelo 02, com alta probabilidade de pertecerem às classes ncRNAs. Ao contrastarmos com o Modelo 01 (Figura 24) observa-se menos ruídos no sinal, com picos mais destacados e fáceis de identificar até mesmo visualmente ao longo do genoma.

Analisamos ainda o comportamento dos sinais gerados com um modelo treinado com regiões sem particionamento, porém com a remoção de alguns atributos de estrutura secundária, sendo estes: o número de nucleotídeos não-pareados (*unparied*) e número de alças (*loops*) e peso molecular (*mw*) como atributo de propriedades da sequência. Observamos que essa remoção não alterou de forma significativa os resultados da predição e por apresentar uma ligeira piora na classificação, decidimos manter todos os atributos já discutidos. Da mesma forma, o modelo gerado com todos o ncRNAs anotados (Figura 22 e Tabela 08) não apresentou melhorias significativas e com isso buscamos explorar os resultados dos demais modelos gerados.

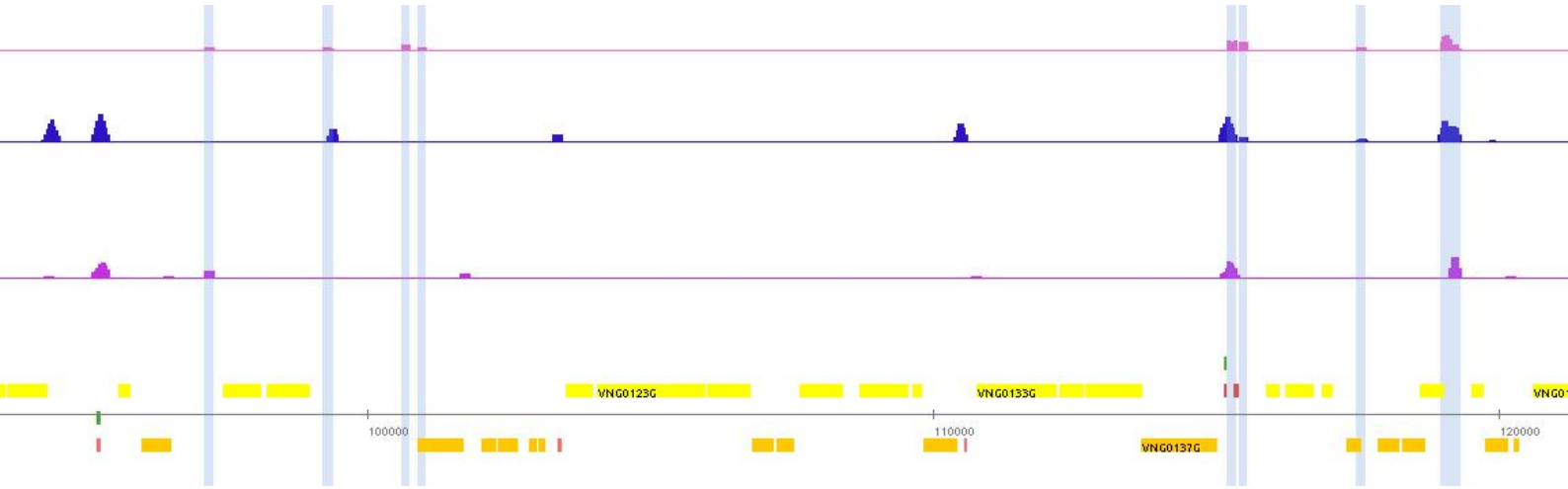


Figura 25 – Genome browser com a representação em destaque dos trechos que foram determinados a partir dos picos de probabilidade obtidos com o classificador Bayes Net.

A partir de uma avaliação considerando o aspecto global das variações de modelos propostas, concluímos que apesar de alguns modelos apresentarem bons resultados na validação cruzada e serem capazes de separar bem os exemplos de treinamentos em suas devidas classes, ao conciliarmos com uma estratégia que possibilite a identificação de novas regiões com probabilidade de serem ncRNAs, esses modelos não apresentaram um comportamento razoavelmente significativo e que torna-se factível uma distinção mais clara de possíveis regiões candidatas. Decidimos então considerar os modelos que apresentaram resultados inferiores na validação cruzada mas que foram capazes de indicar, de forma significativa, regiões de interesse. Ao que indica, essas regiões são mais coerentes e podem de abranger resultados mais plausíveis.

Por fim, os resultados do Modelo 02 foram melhores na medida AUC para a classe ncRNA em 5 classificadores quando comparado aos resultados do Modelo 03 (Tabelas 06 e 07). Com a aplicação do procedimento de janela deslizante, o Modelo 02 apresentou resultados ligeiramente melhores que o Modelo 03 por incluir menos falsos positivos (poucos picos de probabilidade associados a trechos codificadores). Optamos então pela escolha do Modelo 02 para as demais análises.

5.1.4 Análise das faixas genômicas

Após os resultados preliminares discutidos anteriormente, analisamos com mais detalhes as faixas genômicas geradas e buscamos interpretar os picos de probabilidade para todo o cromossomo e plasmídeos em ambas as fitas e não somente na fita *forward* do cromossomo, até então descrito. Com isso, o número total inicial de regiões genômicas com picos de probabilidade foi obtido. Na Tabela 13 são resumidos os resultados obtidos por cada classificador utilizado no *workflow* de processamento.

Tabela 13 – Número total de regiões com picos de probabilidade para a classe ncRNA gerados por cada técnica de AM.

	Classifier Name	**Total regions peaks**
01	Bayes Net	754
02	Decision Tree	644
03	Logistic Regression	2275
04	Naive Bayes	1535
05	Random Forest	5034
06	Rules Based	4053
07	SVM poly 2nd	3421
08	SVM linear	5132
09	SVM RBF	1291

Cada região foi inicialmente analisada por seu tamanho e regiões maiores que 400 nucleotídeos foram descartadas. Como mencionado, esse procedimento foi necessário uma vez que para alguns casos, o algoritmo usado para estabelecer as posições de início e fim de cada região, definiu alguns trechos longos que não condiz com um tamanho esperado para ncRNAs. Em seguida, iniciamos a etapa de combinação dos classificadores, que consiste em verificar dentre todas as regiões, geradas de forma independente por cada técnica de AM, quais ocorrem na mesma região. Dessa forma, optamos por selecionar somente aquelas que foram ditas como ncRNAs por mais de um classificador. Considerando um limiar (*threshold*) que vai de 5 a 9 classificadores, obtivemos os resultados que estão apresentados na Tabela 14.

Tabela 14 – Combinação das regiões preditas com diferentes limiares. Os valores estão separados por cromossomo e fita. Consideramos nas análises posteriores os trechos dos valores que estão em negrito.

	Chromosome		**pNRC200**		**pNRC100**	
Threshold number of classifiers	Forward	reverse	Forward	Reverse	Forward	reverse
=9	10	08	0	01	01	03

≥8	**91**	**86**	03	10	07	10
≥7	209	165	**16**	**21**	**16**	**20**
≥6	375	292	33	46	31	23
≥5	722	584	99	115	50	32

De acordo com os resultados, na medida em que foram estabelecidos limiares mais estringentes e exigido que todos os classificadores coincidam em suas predições, o número de trechos que respeitam o critério se tornou bem pequeno, por exemplo, apenas 10 trechos no cromossomo fita *forward* foram preditos por 9 classificadores. Ao diminuirmos essa restrição para pelo menos 5 classificadores, o número elementos para o mesmo cromossomo e fita se eleva para 722. Decidimos então, considerar a combinação de pelo menos 8 classificadores para a predição das regiões presentes no cromossomo e pelo menos 7 nas regiões presentes nos plasmídeos pNRC200 e pNRC100.

Tabela 15 – Resultados da verificação de anotações e ruídos associadas aos trechos selecionados. Combinação dos trechos obtidos pelos classificadores considerando o cromossomo e plasmídeos. Na tabela são incluídos: trechos que coincidiram com anotações já existentes nos dados de treinamento (True positive), trechos pertencentes aos tRNAs e rRNAs e trechos pertencentes a regiões CDS (False positives).

Combining classifiers - Removed candidates with annotations						
Classes	**Chromosome**		**pNRC200**		**pNRC100**	
	Forward	reverse	Forward	reverse	Forward	Reverse
True positive	4	4	2	2	2	2
tRNA/rRNA	23	16	-	-	-	-
False positive	7	6	4	3	5	8
Total	34	26	6	5	7	10

Com base nos trechos obtidos na combinação de classificadores, analisamos então quais desses trechos coincidem com alguma anotação existente. Dessa forma, simplesmente checamos se a região possui a

anotação de alguma das classes consideradas e consideramos como falsos positivos os exemplos que coincidem com anotações da classe “CDS”. Na Tabela 15 são resumidas as informações obtidas nesse procedimento de verificação e como pode ser observado, das 91 regiões preditas como ncRNAs para o cromossomo fita *forward* somente 7 foram considerados falsos positivos, 4 possuem anotação como ncRNAs, utilizados no treinamento, e 23 tRNAs e rRNAs que não fizeram parte do treinamento também puderam ser identificados. Com a remoção desses trechos, obtemos os resultados da Tabela 16.

Notamos que em média 49% dos candidatos selecionados estão em regiões intergênicas ao longo do genoma, o restante está próximo de regiões CDS e podem corresponder a regiões UTR, ou ainda como será discutido, podem pertencer a outra classe de ncRNAs (TSSaRNA). Essas duas classes, ncRNAs e UTR, por possuírem características de trechos não codificadores em suas definições tornam a distinção bem mais complexa para o modelo. No entanto, conforme discutido, dado o comportamento da metodologia de estretégia deslizante utilizada em que no aspecto global das predições, as regiões candidatas possuem poucos ou nenhum trecho falso positivo dentre os que restam sem anotações, acreditamos que bons candidatos podem ser obtidos a partir desses resultados preliminares.

Tabela 16 – Resultados da verificação de anotações e ruídos associados aos trechos selecionados.

Combining classifiers - overall results							
	Chromosome		**pNRC200**		**pNRC100**		**Total**
	Forward	reverse	Forward	reverse	Forward	reverse	
Total selected	91	86	16	21	16	20	250
Removed	34	26	6	5	7	10	88
Candidates	**57**	**60**	**10**	**16**	**9**	**10**	**162**
Intergenic	**29%**	**33%**	**67%**	**44%**	**78%**	**46%**	**49%**

Outro resultado observado na verificação dos trechos que coincidem com anotações é que parte dos exemplos ditos como ncRNAs são regiões que transcrevem RNAs da classe de RNAs transportadores (tRNAs). Apesar

de ser uma classe de moléculas com características bem próprias, e a princípio regiões de fácil distinção no genoma, é interessante observar que esses exemplos não fizeram parte dos dados de treinamento, uma vez que estamos interessados em identificar outras classes de ncRNAs, e mesmo assim foram identificados nas predições com a estratégia de janela deslizante.

Como resultado final dessas análises preliminares, obtivemos um total de 162 regiões sem anotações e que são candidatas a pertencerem a classe ncRNAs. Essas regiões estão distribuídas ao longo do cromossomo e dos plasmídeos como é ilustrado na Tabela 16. A partir desses resultados, procuramos então analisar os trechos como forma de buscar melhores evidencias e os resultados serão descritos nas próximas seções.

5.1.5 Resultados com a aplicação de algumas abordagens disponíveis para a identificação de ncRNAs

Nesta seção, são apresentados os resultados obtidos com a aplicação das abordagens descritas na seção 4.2. Para cada predição, verificamos se o trecho sugerido corresponde à alguma anotação existente para o organismo. Reunimos os resultados das comparações e estes estão apresentados na Tabela 17. Na Tabela são indicados o número total de trechos preditos,

Tabela 17 – Resultados da verificação de anotações e ruídos associados aos trechos selecionados.

Approach	Total	FP	TP	Matches
Coral	5365	2089	75	2563
Dario	84	24	36	50
RNASpace: YASS	41	0	39	39
RNASpace: Blast	27	0	26	26
RNASpace: ERPIN	76	2	48	50
RNASpace: RNAmmer	3	0	3	3
RNASpace: INFERNAL	23	9	1	10
RNASpace: tRNAscan-SE	47	0	47	47
RNASpace: Darn	182	44	42	86

RNASpace: RNAz	216	2	32	34
RNASpace: AtypicalGC	16	14	0	14

Observando os resultados podemos constatar que a abordagem Coral gerou muitos trechos falsos positivos. O sinal de sRNA-seq considerado pode não ter contribuído para a representação das anotações e posteriormente para a predição de novos trechos. Dessa forma, tornou-se inviável considerar os resultados das predições para essa abordagem.

Alguns dos *clusters* gerados pela abordagem Dario coincidiram tanto com trechos UTRs quanto CDS anotados por isso o número total de trechos que batem em anotações (*Matches*) é menor que o número de trechos falso-positivos (FP) e verdadeiro-positivos (TP). Observamos que uma vez que o agrupamento de *reads* mapeia ambas as classes, algumas das predições que foram ditas como falso-positivas correspondem a trechos de UTR associadas a suas respectivas regiões CDS, o que aparentemente ocasionou uma tendência do *cluster* ser predito como ncRNAs. Cerca de 35% dos trechos preditos como candidatos a ncRNAs pela abordagem Dario estão em regiões intergências e outros 30% podem estar associados a regiões UTR.

As abordagens baseadas em busca por similaridade YASS e Blast basicamente encontraram trechos pertencentes a tRNAs e rRNAs, bem como a ferramenta tRNAscan-SE, que é mais específica para essa classe de ncRNAs. A maioria dos tRNAs e rRNAs anotados também foram identificados pelas abordagens Darn e ERPIN.

Como parte ainda dos resultados da abordagem Darn, 5 snoRNAs-CDbox foram sugeridos sobrepondo trechos pertencentes aos genes VNG1529G, VNG1726G, VNG0318G, VNG1585Cm e VNG1988G. Esses trechos não foram confirmados com a abordagem Snoscan (Lowe & Eddy, 1999) uma vez que não coincidiram com dados disponíveis no *Genome Browser* da Universidade Santa Cruz Califórnia (UCSC). Verificamos ainda que 44 outros trechos preditos sobrepõem anotações CDS.

Dentre as outras predições da ferramenta ERPIN, 2 trechos ditos como *Small nucleolar RNA* (snRNAs) estão sobrepostos aos genes VNG1654G e VNG2176H da mesma forma, esses trechos não puderam ser

confirmados por outras abordagens.

RNAz obteve resultados mais interessantes por incluir apenas 2 trechos coincidindo com CDS anotados. Verificamos que 22 outros elementos preditos correspondem a UTR já descritos os verdadeiros positivos restantes correspondem a tRNAs anotados. Dessa forma, o comportamento do algoritmo tendeu a ser menos ruidoso em relação aos trechos sugeridos pela abordagem.

As ferramentas INFERNAL, Rammer e AtypicalGC não obtiveram muitos trechos como resultados da predição. Rammer apenas coincidiu com anotações de rRNA e INFERNAL com a RNaseP anotada como VNGs01, os outros 9 trechos preditos pela abordagem INFERNAL e os 14 obtidos com AtypicalGC, sobrepõem regiões CDS e foram considerados falsos positivos.

Os resultados da abordagem smyRNA foram difíceis de avaliar. O sinal gerado ao longo do genoma apresentou muitos ruídos, com valores imprecisos no que se refere a uma informação mais clara e que torne possível a distinção dos trechos de regiões não-codificadoras.

De maneira geral, cerca de 90% das abordagens identificaram trechos pertencentes a tRNAs e rRNAs, uma vez que se referem a uma classe de ncRNA bem específica e com propriedades conhecidas, como estrutura e funções. Observando o comportamento dos resultados preliminares obtidos, buscamos então considerar os demais trechos na forma de combinação dos resultados. Dessa forma, verificamos se as regiões sugeridas pelos classificadores através das faixas genômicas também são indicadas por essas outras abordagens aplicadas ao genoma do organismo de interesse. Os resultados dessas análises estão descritos na seção a seguir.

5.1.6 ncRNAs candidatos identificados

Com o objetivo de incluir outras evidências que favoreçam um maior grau de confiabilidade dos 162 trechos sem anotações obtidos anteriormente, analisamos cada uma das regiões candidatas com uma

inspeção visual através da ferramenta *Gaggle Genome Browser*. Além das informações sobre o perfil de expressão ao longo da curva de crescimento e sinal de expressão com dados de RNA-seq, que foram utilizadas para a representação dos trechos como atributos de Aprendizado de Máquina, consideramos os dados de enriquecimento de *reads* alinhados a coordenadas de início do *read*. Que basicamente consiste na identificação da coordenada de início do *read* mais frequente próximo ao códon de início de um CDS (Zaramela *et al.,* 2014).

Na Figura 26 é ilustrado um dos exemplos avaliados. É possível observar que em quase todas as faixas genômicas incluídas, a probabilidade da região em destaque ser da classe ncRNA é acentuada por picos, como discutido na presente metodologia. O mesmo acontece com a região pertencente ao tRNA à esquerda. Existe uma variação na expressão do trecho ao longo da fase de crescimento (*Growth phase*) do organismo como é indicado nos dados de *tiling-array* e ainda*,* ocorre um enriquecimento de *reads* no início do trecho ao que indicam as barras verticais em verde. Na informação de tiling-array para a condição referência (linha em azul) também é possível observar uma elevação no sinal de expressão para o trecho em destaque.

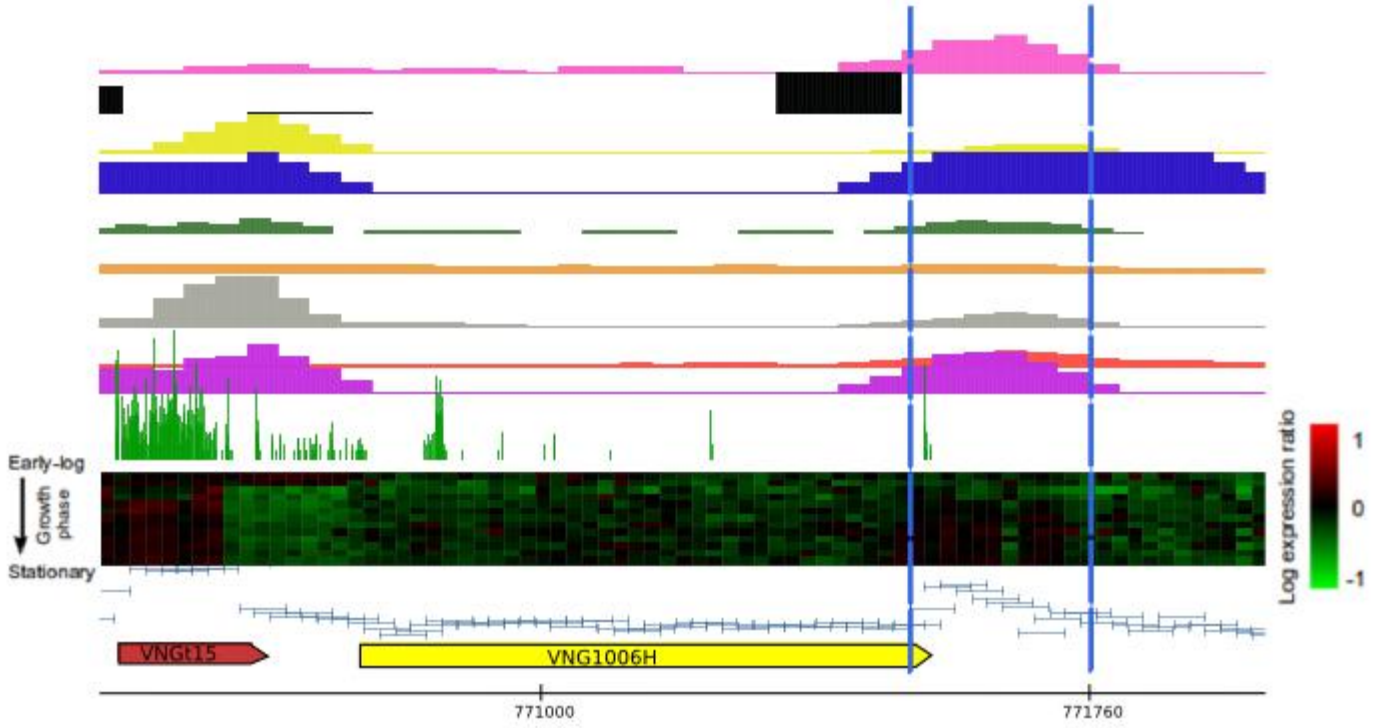


Figura 26 – Exemplo de um trecho candidato a ncRNA. A caixa em amarelo representa o trecho de um gene anotado na fita foward e em vermelho um tRNA. Linhas em azul pontilhadas representam o trecho estimado para a região do ncRNA. As coordenadas do genoma estão indicadas no eixo horizontal. O perfil de expressão ao longo da curva de crescimento é indicado por um heatmap, colorido de acordo com os valores da expressão de cada ponto relativo a condição referência de *H. salinarum*. Linhas horizontais em azul representam o sinal de tiling-array para a condição referência. Informações sobre o enriquecimento de reads estão representadas como faixas verticais em verde. Cada linha superior a informação sobre enriquecimento refere-se as faixas genômicas geradas por cada um dos 9 classificadores.

Aplicamos as mesmas observações para as demais regiões e incluímos os trechos candidatos selecionados na Tabela 16. A partir de uma inspeção visual baseada dos dados de expressão durante a curva de crescimento, descartamos trechos que podem estar associados a regiões UTR, cuja expressão do trecho cognato ao CDS se comporta de maneira semelhante a expressão do gene. Apesar da dificuldade de se definir as informações de início e fim do trecho, estabelecemos tais valores ponderando as informações de expressão na curva de crescimento e

enriquecimento de *reads.*

Como mencionado, optamos por descartar trechos mais difíceis de definir uma vez que suas coordenadas estão próximas às regiões UTR porém, ao compararmos as 162 regiões com dados de RNAs associados a Inicio de Transcrição (*Transcription Start Site Associated RNAs - TSSaRNAs*) disponíveis em Zaramela *et al.,* 2014, satisfatoriamente constatamos que 40 trechos coincidem com essa classe de ncRNAs. Essa evidência corrobora com os resultados da metodologia aplicada por favorecer outros indícios de potenciais candidatos através de uma metodologia distinta. Vale ressaltar ainda que, dentre esses 40 candidatos, 3 deles (TSSaRNA-VNG1213C, TSSaRNA-VNG0101G e TSSaRNA-VNG2293G) são citados como exemplos descritos em Zaramela *et al.,* (2014) e um deles (TSSaRNA-VNG1213C) foi avaliado experimentalmente no trabalho citado. De acordo com os resultados, tornou-se claro que o comportamento dinâmico do TSSaRNA-VNG1213C em relação ao seu gene cognato é semelhante ao longo da curva, porém o nível de expressão é 16 vezes maior que a do gene. Incluímos na Figura 27 o trecho referente ao TSSaRNA. É possível verificar a existência dos picos de probabilidade definidos pelos classificadores em suas respectivas faixas genômicas bem como o sinal de enriquecimento de *reads.* Note que o trecho obtido com a combinação de classificadores sobrepõe a região codificadora, o que torna difícil a avaliação por inspeção visual. Com a evidência obtida através dos resultados de outra metodologia, concluímos que os candidatos selecionados possuem subsídios para serem verdadeiros trechos pertencentes a classe ncRNA.

A Tabela 18 resume o principal produto da presente Tese, a lista dos novos ncRNAs encontrados e foram incluídos apenas os 42 novos candidatos selecionados, os 40 trechos que coincidem com os dados de Zaramela *et al.,* 2014 foram removidos, apesar de serem verdadeiros ncRNAs. Como pode ser observado na Tabela 18, nem todos os candidatos possuem variações de expressão na curva de crescimento porém, todos possuem sinal de enriquecimento de *reads,* o que indica que fatores de transcrição estão associados a região. Verificamos ainda, quais trechos

coincidem com resultados das abordagens aplicadas na seção anterior e 20 dos exemplos foram encontrados por pelo menos uma abordagem. Na Tabela 18, estão destacados em negrito 19 desses exemplos, sendo que 1 deles, que também foi identificado com a abordagem RNAz, se refere a um TSSaRNA e dessa forma, foi retirado da lista de candidatos.

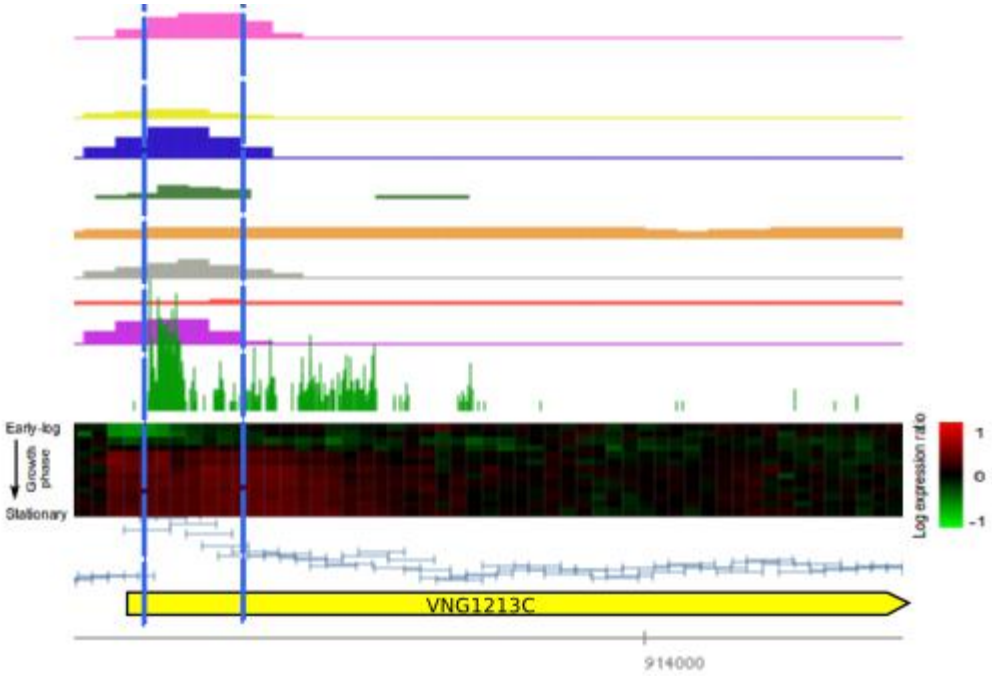


Figura 27 – Trecho obtido com a aplicação da metodologia adaptada que coincide com o TSSaRNA-VNG1213C, validado experimentalmente e apresentado em Zaramela *et al.,* 2014.

A maior parte dos 20 trechos coincidem com as abordagens RNAz e Dario. Essas ferramentas não fornecem nenhuma informação adicional sobre anotação porém, o exemplo ncRNAc01_p05, que foi também obtido com a abordagem ERPIN, pode se referir a um *small nucleolar RNA* (snRNA).

Tabela 18 – Lista de trechos candidatos à ncRNAs. Na tabela são incluídos o cromossomo (*Chromossome*), as posições de início (*Start*) e fim (*End*), Nome (*Name*), fita (*Strand*) e se no trecho existe variações na expressão ao longo da curva de crescimento (*Expr.*). Exemplos em negrito também foram identificados por pelo menos uma das abordagens aplicadas (ver texto).

Chromosome	Start	End	Name	Strand	Expr.

chr	**54801**	**54960**	**ncRNAc01_p05**	**forward**	**no**
chr	65881	66120	ncRNAc02_p06	forward	yes
chr	119121	119320	ncRNAc03_p08	forward	no
chr	**223281**	**223384**	**ncRNAc04_p11**	**forward**	**no**
chr	281761	281840	ncRNAc05_p15	forward	no
chr	**464481**	**464520**	**ncRNAc06_p17**	**forward**	**yes**
chr	**568041**	**568120**	**ncRNAc07_p20**	**forward**	**yes**
chr	**590801**	**590847**	**ncRNAc08_p23**	**forward**	**no**
chr	725792	725920	ncRNAc09_p25	forward	no
chr	**749241**	**749400**	**ncRNAc10_p28**	**forward**	**no**
chr	**768841**	**768880**	**ncRNAc11_p29**	**forward**	**yes**
chr	**771472**	**771760**	**ncRNAc12_p32**	**forward**	**yes**
chr	**990561**	**990840**	**ncRNAc13_p46**	**forward**	**yes**
chr	**1060201**	**1060320**	**ncRNAc14_p48**	**forward**	**no**
chr	**1186001**	**1186160**	**ncRNAc15_p53**	**forward**	**no**
chr	12681	12760	ncRNAc16_p01	reverse	no
chr	**53761**	**53800**	**ncRNAc17_p03**	**reverse**	**no**
chr	54361	54480	ncRNAc18_p04	reverse	no
chr	**153321**	**153440**	**ncRNAc19_p11**	**reverse**	**no**
chr	296961	297240	ncRNAc20_p13	reverse	no
chr	305201	305320	ncRNAc21_p14	reverse	no
chr	634161	634240	ncRNAc22_p22	reverse	no
chr	883041	883160	ncRNAc23_p32	reverse	yes
chr	**1002681**	**1002840**	**ncRNAc24_p35**	**reverse**	**no**
chr	1224361	1224560	ncRNAc25_p44	reverse	yes
chr	**1279521**	**1279640**	**ncRNAc26_p48**	**reverse**	**no**
chr	**1789641**	**1789720**	**ncRNAc27_p76**	**reverse**	**no**
chr	1902361	1902440	ncRNAc28_p79	reverse	no
chr	**1987801**	**1987960**	**ncRNAc29_p85**	**reverse**	**yes**
pNRC100	143801	143960	ncRNAc30_p12	forward	yes
pNRC100	112761	113200	ncRNAc31_p01	reverse	no
pNRC100	115681	115920	ncRNAc32_p05	reverse	no
pNRC100	116841	117040	ncRNAc33_p09	reverse	yes
pNRC100	133641	134000	ncRNAc34_p16	reverse	no
pNRC200	129161	129240	ncRNAc35_p02	forward	no
pNRC200	133161	133320	ncRNAc36_p03	forward	yes
pNRC200	205361	205440	ncRNAc37_p05	forward	no
pNRC200	223321	223520	ncRNAc38_p07	forward	yes
pNRC200	274321	274360	ncRNAc39_p12	forward	yes
pNRC200	155881	156160	ncRNAc40_p04	reverse	no
pNRC200	244401	244560	ncRNAc41_p10	reverse	yes
pNRC200	262561	262600	ncRNAc42_p13	reverse	yes

Os resultados da avaliação com outras metodologias, por proporem de forma independente e a partir de outros tipos de abordagem alguns dos mesmos trechos como candidatos a ncRNAs, também contribuem como indícios para os resultados da seleção de potenciais ncRNAs realizada com a metodologia baseada em faixas genômicas.

Um ponto fraco observado nos resultados da metodologia aplicada refere-se a uma certa sensibilidade ao padrão da fita. Mesmo com dados de expressão com fita específica, outras propriedades consideradas, como

por exemplo a medida de conservação utilizada ou propriedades da sequência primária, podem ter dificultado a definição de probabilidade dos trechos. Essa característica foi apresentada principalmente em regiões pertencentes a tRNAs, em que sinais de ambas faixas genômicas (*forward* e *reverse*) sugeriram a presença de um ncRNA. Apesar desse comportamento indesejado, acreditamos que o viés da fita não destitui os resultados com os candidatos selecionados.

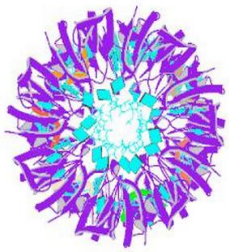
5.2 Predição de interação RNA-Proteína

Nesta seção são incluídos os resultados obtidos na busca do objetivo secundário deste Tese, que consiste na aplicação de metodologias para a predição de ncRNAs candidatos a interação com a proteína LSm, presente no organismo em estudo *H. salinarum* NRC-1. Essa perspectiva é importante uma vez que ao identificarmos um trecho com alta probabilidade de interação com a proteína maiores serão os indícios desse trecho transcrever para um ncRNA, contribuindo assim no conjunto de evidências para o objetivo principal deste Tese.

Como descrito anteriormente, a abordagem denominada *RNA-Protein Interaction Prediction (RPISeq)* disponível em http://pridb.gdcb.iastate.edu/RPISeq/index.html utiliza dois algoritmos de forma independente para a predição, um deles baseado em um conjunto de árvores de decisão (*Random Forest – RF*) e outro baseado em Máquinas de Vetores de Suporte (*Suport Vector Machine – SVM*). O modelo é gerado por cada algoritmo utilizando um conjunto de exemplos derivados do banco de dados de proteínas *Protein Data Base PDB* (Berman *et al.,* 2000). Diversas classes de moléculas de RNA (como por exemplo, RNAs ribossomais, RNAs transportadores, RNAs mensageiros, etc) e proteínas são utilizadas como exemplos positivos de interação. Apesar da grande variedade de proteínas influenciando na diversidade do modelo de AM, dois complexos com proteínas da família Sm e LSm puderam ser encontrados nos dados de treinamento através de uma busca por similaridade disponível no site da ferramenta. Nas Figuras 28 e 29 são

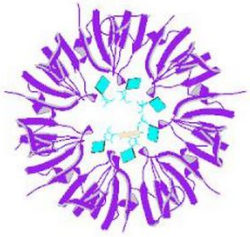
ilustrados tais complexos cujos organismos são *Pyrococcus abyssi* e *Archaeoglobus fulgidus,* respectivamente.

O website da abordagem permite que seja submetido para o cálculo de probabilidade de interação: pares únicos de RNA e proteína, múltiplas sequências de RNAs e uma sequência de proteína ou ainda múltiplas sequências de proteínas e uma sequência de RNA. Para o caso de múltiplas sequências é considerado um arquivo de no máximo 100 sequências em formato FASTA.



1m8v: 14 protein chains, 7 RNA chains
Complex Name: Structure of Pyrococcus abyssii Sm Protein in Complex with a Uridine Heptamer
Resolution: 2.6 Å
Solving Method: X-RAY DIFFRACTION

Figura 28 – Estrutura da proteína Sm de *Pyrococcus abyssii* PDB ID 1M8V.



1i5l: 14 protein chains, 2 RNA chains
Complex Name: CRYSTAL STRUCTURE OF AN SM-LIKE PROTEIN (AF-SM1) FROM ARCHAEOGLOBUS FULGIDUS COMPLEXED WITH SHORT POLY-U RNA
Resolution: 2.75 Å
Solving Method: X-RAY DIFFRACTION

Figura 29 – Estrutura da proteína Sm -Like de *Archaeoglobus fulgidus* PDB ID 1I5L.

Aplicamos os dados reunidos na seção 4.3.1 (Tabela 3) ao site da abordagem RPISeq como o objetivo de verificar como se comporta a abordagem na predição de parceiros de interação das proteínas Hfq e LSm. Os resultados obtidos seguem o esquema da Figura 30.

Organism				
Positive instances				Number
Negative instances				Number
Selection criteria	TN	FP	ACC	**F-measure**
	FN	TP	Precision	Recall

Figura 30 – Esquema de apresentação dos resultados. Organismo ao qual os dados pertencem, número de exemplos positivos e negativos, critérios de seleção para a interpretação das probabilidades obtidas pelo classificador Random Forest (RF) e Suport Vector Machine (SVM) e valores estatísticos considerando a matriz de confusão (*confunsion matrix*), acurácia (*accuracy*), precisão (*precision*), medida-F (*F-measure*) e *recall.*

Utilizamos três critérios para interpretar as probabilidades obtidas por cada classificador. Na primeira opção é considerado como uma classificação positiva de interação os elementos que obtiveram probabilidade maior ou igual a 0,6 em ambos classificadores (RF e SVM >= 0,6), ou seja, se os dois classificadores concordam que determinado elemento tem pelo menos 60% de chance de interagir então ele é positivo. Na segunda e terceira opção é considerado o mesmo valor de probabilidade, porém de forma independente para cada classificador. Assim, considerando apenas para os resultados de um dos classificadores, se a probabilidade de um determinado elemento for maior ou igual a 0,6 então ele é considerado positivo. Como exemplo, considere os dados correspondentes a predição dos elementos pertencentes ao organismo *Bacillus subitilis* com critério de seleção a probabilidade maior ou igual a 0,6 para ambos classificadores simultaneamente, os reseultados são: TN = 128, FN = 20, FP = 49 e TP = 3 para a matriz de confusão e ainda, acurácia = 0,65, precisão = 0,06, medida-F = 0,08 e recall = 0,13 (Figura 31).

Escherichia coli				
Positive instances				22
Negative instances				152
RF and SVM >= 0.6	75	77	0.47	**0.14**
	14	8	0.09	0.36
RF >= 0.6	80	72	0.54	**0.27**
	7	15	0.17	0.68
SVM >= 0.6	124	28	0.75	**0.27**
	14	8	0.22	0.36

Bacillus subtilis				
Positive instances				23
Negative instances				177
RF and SVM >= 0.6	128	49	0.65	**0.08**
	20	3	0.06	0.13
RF >= 0.6	134	43	0.71	**0.22**
	15	8	0.16	0.34
SVM >= 0.6	147	30	0.75	**0.1**
	20	3	0.1	0.13

Haloferax volcanii				
Positive instances				39
Negative instances				58
RF and SVM >= 0.6	54	4	0.56	**0**
	39	0	0	0
RF >= 0.6	54	4	0.6	**0.17**
	35	4	0.5	0.1
SVM >= 0.6	54	4	0.56	**0**
	39	0	0	0

Salmonella typhimurium				
Positive instances				128
Negative instances				109
RF and SVM >= 0.6	36	73	0.45	0.52
	58	70	0.49	0.55
RF >= 0.6	47	62	0.66	**0.74**
	17	111	0.64	0.87
SVM >= 0.6	78	31	0.64	**0.63**
	54	74	0.7	0.58

Listeria monocytogenes				
Positive instances				3
Negative instances				85
RF and SVM >= 0.6	70	15	0.79	**0**
	3	0	0	0
RF >= 0.6	73	12	0.84	**0.12**
	2	1	0.07	0.33
SVM >= 0.6	70	15	0.81	**0.2**
	1	2	0.12	0.66

***Escherichia coli* – sRNAs**				
Positive instances				22
Negative instances				40
RF and SVM >= 0.6	16	24	0.39	**0.3**
	14	8	0.25	0.36
RF >= 0.6	17	23	0.51	**0.5**
	7	15	0.39	0.68
SVM >= 0.6	29	11	0.6	**0.39**
	14	8	0.42	0.36

Figura 31 – Resultados da classificação para dados de interação RNA-proteína conhecidos utilizando o website da abordagem RPISeq.

De acordo com os resultados da Figura 31, a medida de precisão para a identificação de exemplos positivos é baixa em todos os organismos testados, isso indica que ao considerar esses resultados em uma validação experimental, muitos dos exemplos ditos como positivos de interação na realidade não seriam, ocasionando em um desperdício de recursos por incluir exemplos falsos positivos na validação. A medida de acurácia não provê valores mais significativos devido ao desbalanceamento dos exemplos em cada conjunto de dados. Dessa forma ao acertar muitos exemplos de uma determinada classe, que por sua vez possui mais exemplos do que a outra, a medida de acurácia será alta mesmo com muitos exemplos sendo preditos de forma errada para a outra classe. Por exemplo, observando os resultados para os dados de *L. monocytogenes* é possível constatar uma acurácia alta, com valor de 0,79, quando considerado ambos os classificadores (RF e SVM) com probabilidade maior ou igual a 0,6, porém nenhum exemplo da classe positiva foi predito corretamente.

Na Figura 31 também é apresentada uma variação na predição para os dados de *E. Coli*, essa variação se refere à subdivisão dos exemplos negativos por descartar outros tipos de RNAs presentes no organismo, sendo considerado apenas moléculas de pequenos RNAs (*small RNAs – sRNAs*). De acordo com as informações do banco de dados EcoGene, estão presentes 62 sequencias de sRNAs em *E. Coli*, das quais 22 interagem com a proteína Hfq. Ainda de acordo com os resultados, a abordagem baseada em SVM tendeu a ter uma melhor precisão em comparação a abordagem baseada em RF. Esta última por sua vez, tendeu a ter uma melhor medida de recall indicando que de todos os exemplos positivos, grande parte deles puderam ser identificados.

Apesar de alguns elementos serem identificados com sucesso, ao observar as predições como um todo é possível constatar que a confiabilidade dos resultados é baixa por não evidenciar robustez nas predições.

5.2.1 Reprodução da abordagem *RPISeq*

Aplicamos os classificadores *Random Forest* (*RF*) e *Suport Vector Machine* (*SVM*) nos conjuntos de dados RPI2241 e RPI369 e obtivemos os resultados apresentados na Tabela 19 utilizando uma avaliação cruzada (*10-fold cross-validation*). Seguimos as mesmas considerações apresentadas no trabalho original (Muppirala *et al.,* 2011). Os resultados do trabalho original são apresentados na Tabela 20. Como podem ser observados, os resultados da classificação com a reprodução da abordagem desenvolvida estão bem próximos dos resultados originais. Apesar de todas as considerações, quanto aos parâmetros dos classificadores e versão da ferramenta WEKA, serem as mesmas na implementação ocorreram pequenas variações nos resultados. Alguns fatores podem contribuir nessa variação, como por exemplo, o uso de variáveis aleatórias na construção do modelo ou especificidades para a execução da ferramenta WEKA, contudo essas variações não comprometem as análises uma vez que tanto os modelos gerados quanto

a aplicação dos conjuntos de testes nesses modelos passam pelos mesmos procedimentos implementados. O principal objetivo com a reprodução da abordagem *RPISeq* foi explorar o tipo representação proposta em Muppirala *et al.,* 2011 e verificar se ao incluirmos diferentes perspectivas para a construção do modelo de aprendizado de máquina a identificação dos parceiros de interação das proteínas Hfq/Lsm é mais robusta para as sequências pertencentes aos organismos analisados.

Uma dessas possíveis perspectivas consistiu em analisar a influência de cada conjunto de dados (RPI2241 e RPI369) na predição dos elementos, ou seja, se ao ser retirado os pares com RNAs ribossomais obtêm-se algum ganho na classificação. Dessa forma, para a criação do modelo foram usados os mesmos classificadores (RF e SVM), porém com a modificação no conjunto de dados de treinamento.

Tabela 19 – Resultados obtidos usando a implementação própria da abordagem RPISeq.

Dataset	**Classifier**	**Accuracy %**	**Precision**	**Recall**	**F-measure**
RPI2241	Random Forest	89.7	0.91	0.88	0.89
RPI2241	SVM	88.8	0.88	0.9	0.89
RPI369	Random Forest	77.3	0.77	0.73	0.76
RPI369	SVM	76.7	0.75	0.8	0.77

Tabela 20 – Resultados apresentados em Muppirala *et al.,* 2011.

Dataset	**Classifier**	**Accuracy %**	**Precision**	**Recall**	**F-measure**
RPI2241	Random Forest	89.6	0.89	0.90	0.90
RPI2241	SVM	87.1	0.87	0.88	0.87
RPI369	Random Forest	76.2	0.75	0.78	0.77
RPI369	SVM	72.8	0.73	0.73	0.73

Ao analisarmos essas modificações nos dados de treinamento, observamos que a exclusão dos pares com RNAs ribosomais prejudicou a predição dos exemplos positivos em todos os organismos (Figura 32). Como mencionado anteriormente, apesar da grande parte dos exemplos positivos serem pequenas moléculas de RNAs (*sRNA*), o modelo ao considerar dados ribossomais consegue distinguir melhor esses pares. A partir dos valores da medida-F, verificamos que o classificador baseado

em árvores de decisão usando como dados de treinamento o conjunto RPI2241 obteve o melhor desempenho em relação aos outros modelos por conseguir identificar a maior parte dos exemplos positivos e ainda, por incluir poucos exemplos falsos positivos. Adicionalmente, um resultado interessante para esse modelo advém do seu melhor desempenho na classificação dos elementos pertencentes ao organismo *Haloferax volcanii* que por sua vez é o organismo mais próximo evolutivamente do organismo de interesse *Halobacterium salinarum NRC-1*. Como pode ser obersarvado na Figura 32, os valores correspondentes a predição dos elementos pertencentes ao organismo *Haloferax volcanii* considerando o classificador *Random Forest* (RF) com dados de treinamento RPI2241 são: TN = 58, FN = 14, FP = 0 e TP = 25 para a matriz de confusão e ainda, acurácia = 0,86, precisão = 1, medida-F = 0,78 e *recall* = 0,64. Além desses resultados, em um aspecto geral na classificação, muitos elementos ainda não puderam ser preditos corretamente diminuindo assim a confiabilidade desse modelo.

Listeria monocytogenes				
Positive instances				3
Negative instances				85
RF- RPI2241	85	0	99	**0.8**
	1	2	1	0.67
SVM – RPI2241	85	0	96.6	**0**
	3	0	0	0
RF – RPI369	0	85	0	**0**
	3	0	0	0
SVM – RPI369	85	0	0.96	**0**
	3	0	0	0

***Escherichia coli* – sRNAs**				
Positive instances				22
Negative instances				40
RF- RPI2241	19	21	48	**0.41**
	11	11	0.34	0.5
SVM – RPI2241	22	18	50	**0.37**
	13	9	0.33	0.41
RF – RPI369	18	22	40.3	**0.27**
	15	7	0.24	0.32
SVM – RPI369	21	19	39	**0.14**
	19	3	0.14	0.14

Haloferax volcanii				
Positive instances				39
Negative instances				58
RF- RPI2241	58	0	86	**0.78**
	14	25	1	0.64
SVM – RPI2241	58	0	63.9	**0.19**
	35	4	1	0.1
RF – RPI369	58	0	63	**0.14**
	36	3	1	0.07
SVM – RPI369	58	0	62	**0.1**
	37	2	1	0.05

Salmonella typhimurium				
Positive instances				128
Negative instances				109
RF- RPI2241	109	0	83	**0.68**
	41	87	1	0.81
SVM – RPI2241	0	109	26.6	**0.42**
	65	63	0.37	0.49
RF – RPI369	0	109	13	**0.24**
	96	32	0.22	0.25
SVM – RPI369	109	0	63	**0.48**
	87	41	1	0.32

Bacillus subtilis				
Positive instances				23
Negative instances				177
RF- RPI2241	177	0	94	**0.65**
	12	11	1	0.49
SVM – RPI2241	177	0	89	**0.08**
	22	1	1	0.04
RF – RPI369	177	0	90	**0.23**
	20	3	1	0.13
SVM – RPI369	177	0	89.5	**0.16**
	21	2	1	0.08

Figura 32 – Resultados da classificação para dados de interação RNA-proteína conhecidos utilizando a reprodução da abordagem *RPISeq*.

5.2.2 Proposta de representação baseada em propriedade físico-química e estrutural da sequência primária.

Buscamos incluir outra alternativa para a representação dos dados utilizados na abordagem *RPISeq* (RPI2241 e RPI369) e analisamos como essa representação distinta pode contribuir na predição dos parceiros de interação das proteínas Hfq/Lsm. Desenvolvemos e aplicamos algoritmos para a extração de propriedades oriundas da sequência primária da proteína e do RNA. Para a extração de propriedades físico-químicas da

proteína foi usado parte de uma abordagem desenvolvida por Lobley e colaboradores (Lobley *et al.,* 2011). Essas propriedades fazem parte de um conjunto de blocos de informações que são consideradas para o tratamento do problema de predição de funções em proteínas. A abordagem completa inclui várias outras etapas de processamento e por estar disponível uma versão *desktop* (ou *stand-alone*), pode-se explorar somente a etapa de extração de características da proteína. Dentre essas características é possível obter a composição de cada aminoácido, hidrofobicidade, carga, ponto isoelétrico, superfície da área, volume dos resíduos, entre outras. Na Figura 33 são listadas todas as que foram utilizadas nas análises.

Para a extração de algumas propriedades da sequência primária do RNA inicialmente foi usado o *website* http://www.basic.northwestern.edu/biotools/OligoCalc.html e posteriormente foi desenvolvido um algoritmo para o cálculo das características mais relevantes como peso molecular, tamanho da sequência, conteúdo GC e temperatura de *melting.* Também é considerada como característica de representação dos dados as informações sobre estrutura secundária predita do RNA. A predição foi realizada utilizando o aplicativo *RNAFold* (Hofacker *et al.,* 1994), uma abordagem amplamente utilizada na literatura e disponível também em versão *stand-alone,* o que facilita sua execução e manipulação. A estrutura predita pode ser representada por uma anotação em pontos e parênteses. As informações quanto ao número de grampos (*hairpins*), *loops* internos, multi-*loops*, *budges*, *loops*, número de bases pareadas e não pareadas e energia livre da estrutura são obtidas a partir de uma análise dos símbolos e pelo resultado do cálculo de energia livre da estrutura predita. Essas informações correspondem as características que foram usadas para representar cada sequência de RNA e são apresentadas na Figura 34, complementando as demais informações mencionadas.

Feature Group	Name
Amino acids	Percent residue composition
Sequence Features	Sequence Length
	Molecular weight
	Average hydrophobicity
	Charge
	Molar extinction coefficient
	Iso electric point
	Aliphatic index
	Residue volume
	Surface area
	Hydrophobicity
	number of atoms
	number of negative amino acids
	number of positive amino acids
	carbon atoms
	Oxygen atoms
	Nitrogen atoms
	Hydrogen atoms
	Sulfur atoms

Figura 33 – Características extraídas da sequência da proteína.

Feature Group	Name
Sequence Features	Sequence Length
	Molecular weight
	% GC content
	TM (basic)
Predicted Structure information	number of hairpin
	number of multi-loop
	number of internal loop
	number of bulge
	loops
	total na paired
	total na unpaired
	minimum free energy

Figura 34 – Características extraídas da sequência do RNA

Os modelos baseados em características físco-químicas e estruturais (*Physico-Chemical and Structural Features - PCSF*) foram inicialmente avaliados com validação cruzada (*10-fold cross validation*) da mesma forma como realizado no trabalho original. De acordo com os resultados somente a abordagem baseada em árvores de decisão (*Random Forest*) obteve resultados próximos a abordagem *RPIseq,* a abordagem baseada em *Suport Vector Machine (SVM)* não obteve um comportamento interessante sobre esse tipo de representação (Tabelas 21 e 22). Considerando ainda os modelos cujos atributos para representação dos dados são baseados em características físico-químicas e estruturais foram aplicados como conjunto de teste os exemplos com pares relacionados as proteínas Hfq/LSm, apresentado anteriormente. De acordo com os resultados, essa forma de representação não contribui com melhorias significativas para a identificação dos elementos que interagem com as proteínas Hfq/LSm quando comparado com a abordagem com representação *RPISeq.* Na maioria dos casos as técnicas de AM tenderam a classificar todos os elementos como positivos gerando assim muitos

falsos positivos. É importante incluir que avaliamos alguns dos atributos utilizados na representação da proteína, e listados na Figura 33, e verificamos que as informações referentes a contagem do número de átomos não contribuem em um representação significativa da informação.

Tabela 21 - Resultados obtidos em uma avaliação *10-fold cross-validation* com representação baseada em PCS.

Dataset	Classifier	Accuracy %	Precision	Recall	F-measure
RPI2241	Random Forest	89.4	0.9	0.88	0.89
RPI2241	SVM	79.2	0.79	0.76	0.78
RPI369	Random Forest	74.3	0.74	0.7	0.73
RPI369	SVM	63.6	0.63	0.65	0.64

Tabela 22 – Resultados apresentados em Muppirala *et al.,* 2011.

Dataset	Classifier	Accuracy %	Precision	Recall	F-measure
RPI2241	Random Forest	89.6	0.89	0.90	0.90
RPI2241	SVM	87.1	0.87	0.88	0.87
RPI369	Random Forest	76.2	0.75	0.78	0.77
RPI369	SVM	72.8	0.73	0.73	0.73

5.2.3 Criação de modelos de AM utilizando dados de treinamento específicos

Outra avaliação realizada consistiu em verificar se utilizando dados de treinamento mais específicos, ou seja, com pares de interação somente considerando a proteína de interesse, podem melhorar a predição dos ncRNAs presentes nos organismos e se assim, as representações utilizadas estão sendo suficientes para a identificação dos parceiros de interação das proteínas Hfq/Lsm. Para isso, foi utilizado como dados de treinamento o conjunto de exemplos presentes em *E. coli* e os demais organismos utilizados como exemplos de teste. Os resultados são apresentados na Figura 34.

Haloferax volcanii				
Positive instances				39
Negative instances				58
RF – RPISeq	58	0	68	**0.34**
	31	8	1	0.2
SVM – RPISeq	58	0	82.5	**0.72**
	17	22	1	0.56
RF – PCS	58	0	61.8	**0.09**
	37	2	1	0.05
SVM – PCS	58	0	61	**0.05**
	38	1	1	0.26

Salmonella typhimurium				
Positive instances				128
Negative instances				109
RF – RPISeq	109	0	59.1	**0.39**
	97	31	1	0.24
SVM – RPISeq	109	0	78.4	**0.75**
	5	77	1	0.6
RF – PCS	109	0	58.6	**0.38**
	98	30	1	0.23
SVM – PCS	109	0	65	**0.52**
	83	45	1	0.35

Listeria monocytogenes				
Positive instances				3
Negative instances				85
RF – RPISeq	85	0	96.6	**0**
	3	0	0	0
SVM – RPISeq	85	0	98.9	**0.8**
	1	2	1	0.67
RF – PCS	85	0	98.9	**0.8**
	1	2	1	0.67
SVM – PCS	0	85	0	**0**
	3	0	0	0

Bacillus subtilis				
Positive instances				23
Negative instances				177
RF – RPISeq	177	0	91	**0.36**
	18	5	1	0.2
SVM – RPISeq	177	0	93	**0.56**
	14	9	1	0.4
RF – PCS	0	177	7.5	**0.14**
	8	15	0.08	0.65
SVM – PCS	177	0	89	**0.08**
	22	1	1	0.04

Figura 35 – Resultados da classificação utilizando como conjunto de treinamento dados de *E. coli.*

Os resultados obtidos indicam que com dados de treinamento mais específicos a abordagem baseada em SVM com representação *RPISeq* obteve os melhores resultados na identificação dos parceiros de interação das proteínas Hfq/LSm em todos os organismos. Contudo, nem todos os elementos presentes em cada organismo puderam ser identificados, isso indica que a representação *RPISeq* não foi suficiente para descrever as propriedades de interação para as moléculas analisadas.

As análises realizadas para a predição de parceiros de interação RNA-proteína indicaram que é necessário o uso de abordagens mais sofisticadas para a predição mais confiável dos elementos de interesse, tanto por uma representação mais específica em relação às propriedades de interação, isto é, pelo uso de fatores que contribuem no sistema biológico do organismo para a interação dos elementos, quanto por uma estratégia de combinação que busque considerar aspectos envolvidos nos princípios de interação entre moléculas.

6 Conclusões

O objetivo principal deste trabalho foi o de adaptar e aplicar diferentes metodologias para a análise e identificação *in silico* de novas moléculas de RNAs não-codificadores presentes no organismo modelo *Halobacterium salinarum* NRC-1. Ainda, o objetivo segundário foi o de aplicar metodologias para a predição de pares de interação RNA-Proteína como forma de incluir uma caracterização básica de ncRNAs como possíveis parceiros de interação da proteína LSm, presente no organismo modelo em estudo.

Para atingir o objetivo principal, conciliamos as considerações de uma abordagem existente e adequamos dados disponíveis do organismo em estudo para a construção da metodologia. Como parte das atividades desenvolvidas, reunimos informações de conservação, expressão e propriedades estruturais de regiões ao longo do genoma para a criação de modelos de Aprendizado de Máquina (AM). Em seguida, aplicamos sob esses modelos uma estratégia para identificar trechos com tendência de transcrever a classe de moléculas ncRNAs. Diversas adaptações foram necessárias para alcançarmos resultados mais significativos e exploramos: diferentes formas de definir os trechos de treinamento, representação dos dados representes em cada trecho e estratégias para a definição dos trechos a serem inferidos nos modelos. Selecionamos alguns trechos com base na combinação de diversos preditores e obtivemos como resultado final o estabelecimento de uma lista de 42 ncRNAs desconhecidos em *H. salinarum NRC-1*, aumentando em cerca de 82% (51 + 42) o repertório de ncRNAs candidatos.

Uma das conclusões inespedadas durante o desenvolvimeto do trabalho foi que o modelo de AM não-ótimo segundo critérios clássicos de performance, como validação cruzada e acurácia, foi o que produziu resultados mais coerentes do ponto de vista global e com mais significado , visualizados em faixas genomicas. O modelo ótimo, seria o escolhido naturalmente numa procedimento padrão, mas quando colocado no contexto da inferência de novos elementos a partir da estratégia de janela

deslizante gera faixas com muitos falsos positivos, ou seja, probabilidades altas da classe ncRNA em regiões CDS. Porém ao utilizarmos outros modelos observamos que, mesmo não separando tão bem esses dados, apresentaram resultados mais coerentes no que se refere ao comportamento das predições ao longo do genoma. Ao visualizarmos as faixas genômicas geradas a partir das probabilidades, notamos que picos de probabilidade se distinguem melhor e possuem poucos trechos com probabilidades altas de ncRNAs sob regiões codificadoras. Essa foi uma das características que contribuiram de forma significativa para a interpretação dos resultados alcançados com a estratégia pois, verificamos que apesar de inferir 50354 trechos ao longo de todo o genoma da fita *forward* do cromossomo, por exemplo, poucos falsos positivos foram gerados, o sinal da faixa não se mostrava ruidoso e os trechos com picos de probabilidade se mantiveram bem definidos e esparsos. Essas características propiciaram uma seleção final, a partir da combinação dos classificadores, com poucos trechos candidatos.

De forma a incluir outras evidências que reforcem as sugestões por análises *in silico* de candidatos, aplicamos e adaptamos outras abordagens disponíveis para a identificação de ncRNAs. Verificamos que mesmo se tratando de metodologias distintas, 45% dos candidatos finais sugeridos são também sugeridos por outra abordagem. A princípio tivemos dificuldades em avaliar trechos próximos a regiões CDS, uma vez que existe uma maior complexidade em separar UTRs por também possuírem características de trechos não codificadores porém, satisfatoriamente averiguamos que 25% dos 162 trechos iniciais sugeridos correspondem a classe de TSSaRNAs identificados em outro trabalho do grupo (Zaramela *et al.,* 2014).

Adaptamos metodologias considerando os dados de um organismo específico porém, a partir de algumas modificações no *workflow* acreditamos na possibilidade de gerar uma ferramenta que automatize todos os processos e que esta possa ser aplicada futuramente em outros organismos, uma vez que dados característicos e informações de anotações ao longo do genoma estejam disponíveis para uso. Nesse

quesito, o recente trabalho pode ser expandido de forma a averiguar sua robustez perante outras abordagens propostas, utilizando como referência outros organismos modelos cujos dados de ncRNAs sejam conhecidos e também análises devidas visando a comparação com os resultados de outras abordagens.

Dentre as limitações presentes, observamos que para alguns casos ocorreu uma certa sensibilidade com as propriedades da fita e também estiveram presentes nesses casos picos na fita anti-senso. Essas circunstâncias foram mais evidentes em trechos pertencentes a moléculas de tRNAs.

Vale ressaltar ainda que, apesar dos esforços realizados para minimizar a ocorrência de candidatos falsos positivos é possível que alguns destes estejam presentes nos 162 trechos iniciais sugeridos e acreditamos ser mais improvável a presença dos mesmos na lista final uma vez que buscamos incluir evidências a partir de outros dados expementais e ainda utilizamos os resultados de predições oriundas de outras ferramentas. As regiões obtidas com a abordagem podem servir como candidatas a validação experimental por oferecerem subsídios pautados em diversos tipos de informações relevantes em sua identificação.

Na perspectiva de aplicação de métodos para a identificação de parceiros de interação RNA-Proteína, foram desenvolvidas e avaliadas diversas estratégias para a criação de modelos de AM. Inicialmente foi realizada a reprodução da abordagem *RPISeq* (Muppirala *et al.,* 2011) que por sua vez possibilitou verificar a influência de RNAs ribossomais no comportamento do modelo para as predições. De acordo com os resultados, utilizando a representação *RPISeq* para a classificação dos elementos que interagem com as proteínas Hfq/LSm, presente no organismo modelo em estudo, a remoção dos dados com elementos ribossomais prejudica a predição.

Concluímos que o classificador baseado em um conjunto de árvores de decisão (*Random Forest – RF*) foi o que obteve os melhores resultados durante as análises com a reprodução da abordagem *RPISeq*, sendo capaz

de identificar grande parte dos elementos presentes em *H. volcanii* sem incluir exemplos falsos positivos na predição. Apesar desses resultados, a robustez na classificação para os demais exemplos presentes em outros organismos ainda foi pequena.

Outras perspectivas de criação do modelo de AM se deram pela variação na forma de representação dos atributos com a inclusão de características físico-químicas e estruturais (*Physico-Chemical and Structural – PCS*) das sequências primárias. Da mesma forma como na abordagem *RPISeq* esta foi uma alternativa simples de representação por considerar informações extraídas somente da sequência primária porém, não foram suficientes para atingir as especificidades dos exemplos testados e os resultados foram inferiores ao da representação utilizada na abordagem *RPISeq*. Apesar dos resultados a motivação que torna essa alternativa de representação interessante é a possibilidade de análise de correlação de atributos de forma mais significativa quanto às informações que podem estar contribuindo no aspecto de interação entre as moléculas (Pancaldi e Bähler, 2011), uma vez que as informações utilizadas na abordagem original não contribuem na interpretação dos mecanismos de interação entre moléculas.

Avaliamos ainda a criação de modelos de AM com dados de treinamento mais específicos, que inclui somente exemplos de RNA com a proteína Hfq presente em *E. coli*, os resultados obtidos não foram suficientes para determinar uma abordagem mais robusta, capaz de separar todos os exemplos testados. Contudo, observamos que dados mais específicos podem contribuir para a criação de modelos mais acurados.

Concluímos então que uma nova perspectiva sobre o problema deve ser desenvolvida buscando atender as especificidades que não foram consideradas. Dentre as alternativas, pode-se analisar o uso de características mais relevantes para o processo de representação dos exemplos com o estudo e inclusão de informações que caracterizam as propriedades de interação em si e não propriedades das moléculas de forma independente (Sauer & Weichenrieder, 2011) (Sobti *et al.,* 2010). No

que se refere aos exemplos de interação, pode ser melhor reduzir o escopo para elementos mais próximos do organismo em estudo, utilizando por exemplo somente dados de *H. volcanii* (Soppa *et al.,* 2009) (Fisher *et al.,* 2010) (Straub *et al.,* 2009) para a criação do modelo.

Referências

Albers, S.-V., & Meyer, B. H. (2011). The archaeal cell envelope. Nature Reviews. Microbiology, 9(6), 414–26.

Allers, T., & Mevarech, M. (2005). Archaeal genetics - the third way. *Nature Reviews. Genetics*, *6*(1), 58–73.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J.. Basic local alignment search tool. J. Mol. Biol. 215:403-410 (1990).

Ankeny, R. a., & Leonelli, S. (2011). What's so special about model organisms? Studies in History and Philosophy of Science Part A, 42(2), 313–323.

Babski, J., Maier, L.-K., Heyer, R., Jaschinski, K., Prasse, D., Jäger, D., … Soppa, J. (2014). Small regulatory RNAs in Archaea. *RNA Biology*, *11*(5), 1–10.

Baliga, N. S., Bjork, S. J., Bonneau, R., Pan, M., Iloanusi, C., Kottemann, M. C. H., … DiRuggiero, J. (2004). Systems level insights into the stress response to UV radiation in the halophilic archaeon Halobacterium NRC-1. Genome Research, 14(6), 1025.

Bao, M., Cervantes Cervantes, M., Zhong, L., & Wang, J. T. L. (2012). Searching for non-coding RNAs in genomic sequences using ncRNAscout. Genomics, Proteomics & Bioinformatics, 10(2), 114–21.

Bare JC, Koide T, Reiss DJ, Tenenbaum D, Baliga NS. (2010). Integration and visualization of systems biology data in context of the genome. BMC Bioinformatics 11: 382.

Beggs, J. D. (2005) Lsm proteins and RNA processing. Biochemical Society transactions, 33(Pt 3), 433-8.

Bell, S. D., & Jackson, S. P. (1998). Transcription and translation in Archaea: a mosaic of eukaryal and bacterial features. Trends in Microbiology, 6(6), 222–8.

Bellucci, M., Agostini, F., Masin, M., & Tartaglia, G. G. (2011). Predicting protein associations with long noncoding RNAs a. *Nature Publishing Group*, *8*(6), 444-445. Nature Publishing Group.

Berman. H. M., Westbrook. J., Feng. Z., Gilliland. G., Bhat. T. N., Weissig, H.,

Shindyalov , I. N., Bourne P. E. (2000). The Protein Data Bank. Nucleic Acids Res, 28:235-42. < http://www.pdb.org/> acessado em janeiro 2012.

Bishop, C. M. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc. Secaucus, NJ, USA, ISBN:0387310738, 2006

Bonneau R, Facciotti MT, Reiss DJ, Schmid AK, Pan M, Kaur A, Thorsson V, Shannon P, Johnson MH, Bare JC, et al. (2007). A predictive model for transcriptional control of physiology in a free living cell. Cell 131: 1354–1365

Borovicka, T., & Jr, M. J. (2012). Selecting representative data sets. Advances in Data Mining Knowledge Discovery and Applications, 418.

Brooks, A. N., Reiss, D. J., Allard, A., Wu, W.-J., Salvanha, D. M., Plaisier, C. L., … Baliga, N. S. (2014). A system-level model for the microbial regulatory genome. *Molecular Systems Biology*, *10*, 740.

Cavicchioli, R. (2011). Archaea — timeline of the third domain. Nature Reviews Microbiology, 9(1), 51–61.

Chang, T.-H., Huang, H.-Y., Hsu, J. B.-K., Weng, S.-L., Horng, J.-T., & Huang, H.-D. (2013). An enhanced computational platform for investigating the roles of regulatory RNA and for identifying functional RNA motifs. BMC bioinformatics, 14 Suppl 2(Suppl 2), S4.

Chao, Y., Papenfort, K., Reinhardt, R., Sharma, C. M., & Vogel, J. (2012). An atlas of Hfq-bound transcripts reveals 3’ UTRs as a genomic reservoir of regulatory small RNAs. The EMBO journal, 31(20), 4005–19.

Cheng, Z., Zhou, S., & Guan, J. (2015). Computationally predicting protein-RNA interactions using only positive and unlabeled examples. *Journal of Bioinformatics and Computational Biology*, *13*(3), 1541005.

Christiansen, J. K., Nielsen, J. S., Ebersbach, T., Valentin-Hansen, P., Søgaard-Andersen, L., & Kallipolitis, B. H. (2006). Identification of small Hfq-binding RNAs in Listeria monocytogenes. RNA (New York, N.Y.), 12(7), 1383–96.

Cros, M., Monte, A. De, & Mariette, J. (2011). RNAspace. org: An integrated environment for the prediction, annotation, and analysis of ncRNA.

RNA, 1947–1956.

Dambach, M., Irnov, I., & Winkler, W. C. (2013). Association of RNAs with Bacillus subtilis Hfq. (A. Driks, Ed.)PLoS ONE, 8(2), e55156.

Dennis P. P., Omer A (2005). Small non-coding RNAs in Archaea. Curr Opin Microbiol; 8:685-94; PMID:16256421;

Dehal, P. S., Joachimiak, M. P., Price, M. N., Bates, J. T., Baumohl, J. K., Chivian, D., … Arkin, A. P. (2010). MicrobesOnline: an integrated portal for comparative and functional genomics. Nucleic acids research, 38(Database issue), D396–400.

Eddy, S. R. (2001). Non-coding RNA Genes and the Modern RNA World. Nature Reviews Genetics, 2(December), 919–929.

Faceli, Katti; Lorena, Ana Carolina; Gama, João de Carvalho, A. C. P. L. F. . Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina. 1. ed. Rio de Janeiro: LTC, v. 1. 394p, 2011.

Facciotti MT, Reiss DJ, Pan M, Kaur A, Vuthoori M, Bonneau R, Shannon P, Srivastava A, Donohoe SM, Hood LE, et al. (2007). General transcription factor specified global gene regulation in archaea. Proc Natl Acad Sci 104: 4630–4635.

Fasold, M., Langenberger, D., Binder, H., Stadler, P. F., & Hoffmann, S. (2011). DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Research, 39(Web Server issue), W112–7.

Farkas, J. a, Picking, J. W., & Santangelo, T. J. (2013). Genetic techniques for the archaea. *Annual Review of Genetics*, *47*, 539–61.

Fischer, S., Benz, J., Späth, B., Maier, L.-K., Straub, J., Granzow, M., Raabe, M., et al. (2010). The archaeal Lsm protein binds to small RNAs. *The Journal of biological chemistry*, *285*(45), 34429-38.

Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S,Wilkinson AC, FinnRD, Griffiths-Jones S, Eddy SR, et al. (2009). Rfam:Updates to the RNA families database. Nucleic Acids Res 37: D136–D140.

Gautheret, D., & Lambert, a. (2001). Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. Journal of molecular biology, 313(5), 1003–11.

Gomes-Filho, J. V., Zaramela, L. S., Italiani, V. C. D. S., Baliga, N. S., Vêncio, R. Z. N., & Koide, T. (2015). Sense overlapping transcripts in IS 1341 - type transposase genes are functional non-coding RNAs in archaea. *RNA Biology*, *12*(5), 490–500.

Han, L. Y. (2004). Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. Rna, 10(3), 355-368. doi:10.1261/rna.5890304

Hastie, T., Tibshirani, R., Friedman, J. Springer-Verlag. 763 pages, 2008

Halbeisen, R. E., Galgano, a, Scherrer, T., & Gerber, a P. (2008). Post-transcriptional gene regulation: from genome-wide studies to principles. *Cellular and molecular life sciences : CMLS*, *65*(5), 798-813.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. (2009). The WEKA Data Mining Software: An Update; SIGKDD Explorations, Vol. 11, 1.

Hedges, S. B. (2002). The origin and evolution of model organisms. *Nature Reviews. Genetics*, *3*(11), 838–49.

Hickey, A. J., Conway de Macario, E., & Macario, A. J. L. (2002). Transcription in the archaea: basal factors, regulation, and stress-gene expression. Critical Reviews in Biochemistry and Molecular Biology, 37(6), 537–99.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., & Tacker, M. (1994). Fast Folding and Comparison of RNA Secondary Structures, 188, 167–188.

Hogeweg, P. (2011). The roots of bioinformatics in theoretical biology. PLoS Computational Biology, 7(3), 1–5.

Hoheisel, J. D. (2006) Microarray technology: beyond transcript profiling and genotype analysis. Nature Rev. Genet. 7, 200–210.

Karp, P. D., Ouzounis, C. a, Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahrén, D., Tsoka, S., et al. (2005). Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic acids research, 33(19), 6083–9.

Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival, B., … Covert, M. W. (2012). A Whole-Cell Computational Model

Predicts Phenotype from Genotype. Cell, 150(2), 389–401.

Kaur A, PanM, Meislin M, FacciottiMT, El-Gewely R, Baliga NS (2006) A systems view of haloarchaeal strategies to withstand stress from transition metals. Genome Res 16: 841–854

Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M., & Zavolan, M. (2011). A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nature methods*, *8*(7), 559-64.

Kittler, J., Hatef, M., Duin, R., and Matas, J. (1998) On Combining Classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, 3.

Koide, T., Pang, W. L., & Baliga, N. S. (2009a). The role of predictive modelling in rationally re-engineering biological systems. Nature Reviews. Microbiology, 7(4), 297–305.

Koide, T., Reiss, D. J., Bare, J. C., Pang, W. L., Facciotti, M. T., Schmid, A. K., … Baliga, N. S. (2009b). Prevalence of transcription promoters within archaeal operons and coding sequences. Molecular systems biology, 5(285), 285.

König, J., Zarnack, K., Luscombe, N. M., & Ule, J. (2012). Protein–RNA interactions: new genomic technologies and perspectives. *Nature Reviews Genetics*, *13*(2), 77-83. Nature Publishing Group.

Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35: 3100–3108.

Langenberger, D., Bermudez-Santana, C. I., Stadler, P. F., & Hoffmann, S. (2010). Identification and classification of small RNAs in transcriptome sequence data. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 80–7.

Leonelli, S., & Ankeny, R. a. (2013). What makes a model organism? Endeavour, 37(4), 209–212.

Lertampaiporn, S., Thammarongtham, C., Nukoolkit, C., Kaewkamnerdpong, B., & Ruengjitchatchawalya, M. (2014). Identification of non-coding RNAs with a new composite feature in the Hybrid Random Forest Ensemble algorithm. Nucleic Acids Research,

42(11), e93.

Leung, Y. Y., Ryvkin, P., Ungar, L. H., Gregory, B. D., & Wang, L.-S. (2013). CoRAL: predicting non-coding RNAs from small RNA-sequencing data. Nucleic acids research, 41(14), e137.

Lewin, B. Genes. Oxford University Press, 2004.

Lewis, B. a, Walia, R. R., Terribilini, M., Ferguson, J., Zheng, C., Honavar, V., & Dobbs, D. (2011). PRIDB: a Protein-RNA Interface Database. Nucleic acids research, 39(Database issue), D277–82.

Liu, Z.-P., Wu, L.-Y., Wang, Y., Zhang, X.-S., & Chen, L. (2010). Prediction of protein-RNA binding sites by a random forest method with combined features. *Bioinformatics (Oxford, England)*, *26*(13), 1616-22.

Lobley, A. E., Nugent, T., Orengo, C. A., Jones, D. T. (2008). FFPred: An Integrated Feature-based Function Prediction Server for Vertebrate Proteomes. Nucleic acids research, 36.

Lowe, T.M. & Eddy, S.E. (1999). A computational screen for methylation guide snoRNAs in yeast, Science 283:1168-71

Lowe TM, Eddy SR. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25: 955–964.

Lu, Z. J., Yip, K. Y., Wang, G., Shou, C., Hillier, L. W., Khurana, E., … Gerstein, M. B. (2011). Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome Research*, *21*(2), 276–85.

Lv, J., Liu, H., Huang, Z., Su, J., He, H., Xiu, Y., … Wu, Q. (2013). Long non-coding RNA identification over mouse brain development by integrative modeling of chromatin and genomic features. Nucleic acids research, 41(22), 10044–61.

Marchais, A., Naville, M., Bohn, C., Bouloc, P., & Gautheret, D. (2009). Single-pass classification of all noncoding sequences in a bacterial genome using phylogenetic profiles. *Genome Research*, 1084–1092.

Massé, E., Majdalani, N., & Gottesman, S. (2003). Regulatory roles for small RNAs in bacteria. Current Opinion in Microbiology, 6(2), 120–124.

Matsui, A., Nguyen, A., Nakaminami, K., & Seki, M. (2013). Arabidopsis Non-Coding RNA Regulation in Abiotic Stress Responses. *International Journal of Molecular Sciences*, *14*(11), 22642–22654.

Mattick, J. S. (2009). The genetic signatures of noncoding RNAs. *PLoS Genetics*, *5*(4), e1000459.

Mattick, J. S. (2003). Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology, *25*(10), 930–9.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews. Genetics*, *11*(1), 31–46.

Mitchell, T. M. Machine Learning, MacGraw-Hill, 1997.

Müller, B., & Grossniklaus, U. (2010). Model organisms - A historical perspective. Journal of Proteomics, 73(11), 2054–2063.

Muppirala, U. K., Honavar, V. G., & Dobbs, D. (2011). Predicting RNA-Protein Interactions Using Only Sequence Information. BMC bioinformatics, 12(1), 489.

Muppirala, U. K., Lewis, B. A., & Dobbs, D. (2013). Computational Tools for Investigating RNA-Protein Interaction Partners, Sci, J. C., Biol, S., *6*(4), 182–187.

Nawrocki EP, Kolbe DL, Eddy SR. (2009). Infernal 1.0: inference of RNA alignments. Bioinformatics 25: 1335–1337.

Ng WV, Kennedy SP, Mahairas GG, Berquist B, Pan M, Shukla HD, Lasky SR, Baliga NS, Thorsson V, Sbrogna J, Swartzell S, Weir D, Hall J, Dahl TA, Welti R, Goo YA, Leithauser B, Keller K, Cruz R, DansonMJet al (2000) Genomesequence of Halobacterium species NRC-1. Proc Natl Acad Sci USA 97: 12176–12181

Noé L, Kucherov G. (2005). YASS: enhancing the sensitivity of DNA similarity search. Nucleic Acids Res 33: W540–W543.

Olejniczak, M. (2011). Despite similar binding to the Hfq protein regulatory RNAs widely differ in their competition performance. *Biochemistry*, *50*(21), 4427-40.

Oren, A. (2010). Industrial and environmental applications of halophilic

microorganisms. Environmental Technology, 31(8-9), 825–34.

Peck, R.F., Dassarma, S., and Krebs, M.P. (2000). Homologous gene knockout in the archaeon *Halobacterium salinarum* with ura3 as a counterselectable marker. Mol. Microbiol. 35: 667–676

Pancaldi, V., & Bähler, J. (2011). In silico characterization and prediction of global protein-mRNA interactions in yeast. Nucleic acids research, 1-11.

Panwar, B., Arora, A., & Raghava, G. P. S. (2014). Prediction and classification of ncRNAs using structural information. BMC Genomics, 15(1), 127.

P. Refaeilzadeh, L. Tang, and H. L. (2009). Cross Validation. (M. T. O. Liu, Ling, Ed.)Encyclopedia of Database Systems (EDBS) (Springer). Springer.

Russel, S. Norvin, P. Artificial Intelligence: A Modern Approach. Third Edition. Prentice-Hall, 2010.

Ryvkin, P., Leung, Y. Y., Ungar, L. H., Gregory, B. D., & Wang, L.-S. (2014). Using machine learning and high-throughput RNA sequencing to classify the precursors of small non-coding RNAs. Methods (San Diego, Calif.), 67(1), 28–35.

Salari, R., Aksay, C., Karakoc, E., Unrau, P. J., Hajirasouliha, I., & Sahinalp, S. C. (2009). smyRNA: A Novel Ab Initio ncRNA Gene Finder. PLoS ONE, 4(5), e5433.

Sauer, E., & Weichenrieder, O. (2011). Structural basis for RNA 3’-end recognition by Hfq. Proceedings of the National Academy of Sciences of the United States of America, 108(32).

Schmid AK, Reiss DJ, Kaur A, Pan M, King N, Van PT, Hohmann L, Martin DB, Baliga NS (2007) The anatomy of microbial cell state transitions in response to oxygen. Genome Res 17: 1399–1413.

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y. and Jiang, H. (2007). Predicting protein-protein interactions based only on sequences information. Proceedings of the National Academy of Sciences of the United States of America, 104(11), 4337–41.

Sobti, M., Cubeddu, L., Haynes, P. a, & Mabbutt, B. C. (2010). Engineered rings of mixed yeast Lsm proteins show differential interactions with

translation factors and U-rich RNA. *Biochemistry*, *49*(11), 2335-45.

Soppa, J., Straub, J., Brenneis, M., Jellen-Ritter, A., Heyer, R., Fischer, S., … Marchfelder, A. (2009). Small RNAs of the halophilic archaeon Haloferax volcanii. Biochemical Society transactions, 37(Pt 1), 133–6.

Storz, G. (2002). An expanding universe of noncoding RNAs. Science (New York, N.Y.), 296(5571), 1260–3.

Storz, G., Vogel, J., & Wassarman, K. M. (2011). Regulation by Small RNAs in Bacteria: Expanding Frontiers. *Molecular Cell*, *43*(6), 880-891.

Straub, J., Brenneis, M., Jellen-Ritter, A., Heyer, R., Soppa, J., & Marchfelder, A. (2009). Small RNAs in haloarchaea: identification, differential expression and biological function. *RNA biology*, *6*(3), 281-92.

Suresh, V., Liu, L., Adjeroh, D., & Zhou, X. (2015). RPI-Pred: predicting ncRNA-protein interaction using sequence and structural information. *Nucleic Acids Research*, *43*(3), 1370–1379.

Touzet H, Perriquet O. (2004). CARNAC: folding families of related RNAs. Nucleic Acids Res 142: W142–W145.

Trun, N. J. e Trempy J. E. Fundamental Bacterial Genetics, Wiley-Blackwell; 1 edition October 20, 2003.

Terribilini, M., Lee, J.-H., Yan, C., Jernigan, R. L., Honavar, V., & Dobbs, D. (2006). Prediction of RNA binding sites in proteins from amino acid sequence. RNA (New York, N.Y.), 12(8), 1450-62.

Ule, J., Jensen, K., Mele, A., & Darnell, R. B. (2005). CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods (San Diego, Calif.)*, *37*(4), 376-86.

Van PT, Schmid AK, King NL, Kaur A, Pan M, Whitehead K, Koide T, Facciotti MT, Goo YA, Deutsch EW, Reiss DJ, Mallick P, Baliga NS. (2008) *Halobacterium salinarum* NRC-1 PeptideAtlas: toward strategies for targeted proteomics and improved proteome coverage. J Proteome Res. 7: 3755-3764.

Vogel, J., & Luisi, B. F. (2011). Hfq and its constellation of RNA. *Nature reviews. Microbiology*, *9*(8), 578-89.

Walczak, K. a., Bergstrom, P. L., & Friedrich, C. R. (2011). Light Sensor Platform Based on the Integration of Bacteriorhodopsin with a Single

Electron Transistor. Active and Passive Electronic Components, 2011, 1–7.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, *10*(1), 57–63.

Washietl, S., Hofacker, I. L., & Stadler, P. F. (2005). Fast and reliable prediction of noncoding RNAs. Proceedings of the National Academy of Sciences of the United States of America, 102(7), 2454–9.

Woese, C. R. & Fox, G. E. (1977). The phylogenetic structure of the procaryotic domain: the primary kingdoms. Proc. Natl Acad. Sci. USA 74, 5088–5090.

Whitehead K, Kish A, Pan M, Kaur A, Reiss DJ, King N, Hohmann L, DiRuggiero J, Baliga NS (2006) An integrated systems approach for understanding cellular responses to gamma radiation. Mol Syst Biol 2: 47

Wilusz, C. J., & Wilusz, J. (2013). Lsm proteins and Hfq Life at the 3 ’ end, (April), 1–10.

Wu T, Wang J, Liu C, Zhang Y, Shi B, et al. (2006) NPInter: the noncoding RNAs and protein related biomacromolecules interaction database. Nucleic Acids Res 34: D150-152. 19.

Zakov, S., Goldberg, Y., Elhadad, M., & Ziv-Ukelson, M. (2011). Rich parameterization improves RNA structure prediction. Journal of Computational Biology : A Journal of Computational Molecular Cell Biology, 18(11), 1525–42.

Zaramela, L. S., Vêncio, R. Z. N., ten-Caten, F., Baliga, N. S., & Koide, T. (2014). Transcription Start Site Associated RNAs (TSSaRNAs) Are Ubiquitous in All Domains of Life. PLoS ONE, 9(9), e107680.

Zhang, J & Madden, TL. (1997) PowerBLAST: a new network BLAST application for interactive or automated sequence analysis and annotation. Genome Res, 7: 649-656.

Zhang, Y., Sun, S., Wu, T., Wang, J., Liu, C., Chen, L., Zhu, X., et al. (2006). Identifying Hfq-binding small RNA targets in Escherichia coli. Biochemical and biophysical research communications, 343(3), 950–5.

Zhou, J., & Rudd, K. E. (2013). EcoGene 3.0. Nucleic acids research, 41(Database issue), D613–24.

Zytnicki M, Gaspin C, Schiex T. (2008). DARN! A weighted constraint solver for RNA motif localization. Constraints 13: 91–109.