UNIVERSIDADE DE SÃO PAULO

FACULDADE DE FILOSOFIA, CIÊNCIAS E LETRAS DE RIBEIRÃO PRETO

DEPARTAMENTO DE BIOLOGIA

Identificação e caracterização *in silico* de pequenas proteínas na archaea *Halobacterium salinarum*

André Bordinassi Medina

Monografia apresentada ao Departamento de Biologia da Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da Universidade de São Paulo, como parte das exigências para a obtenção do título de Bacharel em Ciências Biológicas.

Ribeirão Preto

2015

UNIVERSIDADE DE SÃO PAULO

FACULDADE DE FILOSOFIA, CIÊNCIAS E LETRAS DE RIBEIRÃO PRETO

DEPARTAMENTO DE BIOLOGIA

Identificação e caracterização *in silico* de pequenas proteínas na archaea *Halobacterium salinarum*

André Bordinassi Medina

Monografia apresentada ao Departamento de Biologia da Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da Universidade de São Paulo, como parte das exigências para a obtenção do título de Bacharel em Ciências Biológicas.

Orientadora: Profa. Dra. Tie Koide

Co-Orientadora: Dra. Lívia Soares Zaramela

Ribeirão Preto

2015

Agradecimentos

Agradeço a todas as pessoas e instituições que contribuíram para a realização desse trabalho, seja por apoio financeiro, científico ou emocional. Meu muito obrigado à vocês:

Profa. Dra. Tie Koide pela oportunidade dada ao me aceitar no laboratório e pela excelente orientação, conselhos, profissionalismo, dedicação e por ter me guiado em minhas decisões e me ajudado a trilhar um caminho tão importante para a minha futura carreira científica.

Dra. Lívia Soares Zaramela pela orientação, ensinamentos, paciência e por ter me aceitado como co-orientando mesmo em um momento difícil de seu doutorado. Agradeço por me ensinar a planejar meus experimentos, me socorrer nas horas de sufoco e me mostrar como ser mais independente.

Dr. Gilvan Pessoa Furtado por meu auxiliar nos primeiros passos deste trabalho e por toda a contribuição com os experimentos responsáveis pela existência deste projeto. Muito obrigado pela orientação e auxilio, mesmo à distância.

Silvia Helena Epifânio por todo o suporte técnico, preparação de soluções, géis, meios de cultura e todo o material necessário para meus experimentos. Muito obrigado pelos bons momentos e conversas.

Colegas do Laboratório de Biologia Sistêmica de Microrganismos: Felipe ten Caten, José Vicente Gomes Filho e Diego Martinez Salvanha (além dos já citados anteriormente), pelos bons momentos de conversa, apoio e suporte científico, dicas e conselhos que me fizeram ter uma visão diferente sobre ciência.

Professores Dr. Ricardo N. Z. Vêncio, Dr. Rafael Rocha Silva, Dr. Marcelo Damário Gomes, Dr. Roberto do Nascimento Silva e Dra. Ângela K. Cruz e suas respectivas equipes pelo suporte acadêmico e científico e/ou por cederem o espaço e equipamentos de seus laboratórios para a realização de experimentos pra este trabalho.

Aos meus amigos do curso de Ciências Biológicas, em especial à turma 47, com quem passei junto a maior parte da minha graduação. Obrigado pelos momentos de descontração, festas e suporte emocional.

Sina Laubenstein, por todo apoio emocional e por sempre me motivar e acreditar na minha capacidade. Obrigado pelos ótimos momentos juntos e por me alegrar nas horas difíceis.

Meus pais, Roberto Medina e Maisa Bordinassi Medina por permitirem que eu chegasse até aqui, me apoiando e suportando em todas as minhas decisões. Amo vocês!

A toda minha família, por me proporcionarem momento de alegria e por me apoiarem. Vocês são muito importante para mim.

À FAPESP pela bolsa de Iniciação científica fornecida e por todo o suporte financeiro.

Ao CNPq, CAPES e FAEPA pelo apoio financeiro dado ao projeto.

A todos mais uma vez, muito obrigado!

SUMÁRIO

LISTA DE SIGLAS E ABREVIAÇÕES 7
RESUMO 8
1. INTRODUÇÃO 9
1.1. Archaea: o terceiro domínio da vida 9
1.2. Maquinaria genética em Archaeas 13
1.3. *Halobacterium salinarum* como modelo de estudo de Archaea 14
1.4. Anotação de genes e dificuldades na identificação de ORFs 18
1.5. smORFs e pequenas proteínas 20
2. OBJETIVOS 23
2.1. Objetivos Gerais 23
2.2. Objetivos Específicos 23
3. MATERIAL E MÉTODOS 24
3.1. Análises *in silico* de proteínas 24
3.1.1. Obtenção dos dados por espectrometria de massa 24
3.1.2. Identificação de pequenas proteínas 25
3.1.3. Caracterização *in silico* de pequenas proteínas 26
3.2. Construções moleculares 27
3.2.1. Cultivo de *H. salinarum* NRC-1 27
3.2.2. Oligonucleotídeos 28
3.2.3. Marcação cromossômica 29
3.2.4. Padronização do *Chromosomal tagging* para o gene *lsm* 29
3.2.5. Construção do gene controle marcado com FLAG 29
3.2.6. Transformação 31
3.2.7. Recombinação cromossômica (*crossover*) 31
3.2.8. Construção de vetores para a expressão do gene controle (*lsm*) 32
3.2.9. Lise celular e quantificação de proteínas 34
3.2.10. Imunoprecipitação 34
3.2.11. Western blot e Dot blot 35
4. RESULTADOS E DISCUSSÃO 37
4.1. Resultados e Discussão – Parte 1: Identificação e caracterização de pequenas proteínas 37
4.1.1. Análise geral dos dados de espectrometria de massa 37
4.4.2. Seleção de possíveis smORFs intergênicas e antisenso 43

4.1.3. Análise funcional *in silico* das pequenas proteínas 45
4.2. Resultados e discussão – Parte 2: Construções moleculares 52
4.2.1 Padronização do *chromosomal tagging* com o gene controle lsm 52
4.2.2 Detecção da proteína marcada com FLAG (*chromosomal tagging*) 54
4.2.3 Construções de vetores para a expressão do gene controle lsm 57
5. CONSIDERAÇÕES FINAIS..... 61
6. CONCLUSÃO..... 62
7. REFERÊNCIAS BIBLIOGRÁFICAS 64
8. MATERIAL COMPLEMENTAR..... 72

LISTA DE SIGLAS E ABREVIAÇÕES

aa: aminoácidos

CDS: Coding sequence (Sequência codificante)

CM: Complete Media (meio completo)

dRNAseq: differential RNAseq

D.O.600$_{nm}$:Densidade Ótica

GC%: porcentagem de nucleotídeos Guanina e Citosina

GGB: Gaggle Genome Browser

kDa: kilo Dalton

LC –MS: Liquid Chromatography - Mass spectrometry (Cromatografia líquida seguida de espectrometria de massa)

MM: Massa molecular

ORF: Open Reading Frame

ORFans: Orphan ORFs (ORFs pouco conservadas)

smORF: small ORFs (Pequenas ORFs)

TBS: Tris-buffered saline

TSS: Trasncription Start Site (Sítio de início de transcrição)

UTR: Untranslated region (Região não traduzida)

RESUMO

A utilização da biologia sistêmica como abordagem que integra análise de dados em escala genômica com ferramentas de bioinformática permitiu, por exemplo, a descoberta de diversas regiões genômicas codificadoras de pequenas proteínas e peptídeos (smORFs). Estas moléculas eram até então negligenciadas ou em alguns casos, chamadas erroneamente de RNAs não codificantes (ncRNAs). Tais descobertas levaram a diversos estudos sobre as funções biológicas destas proteínas, que possuem importantes papéis regulatórios em todos os domínios da vida. Assim, o presente estudo visa a busca por novas pequenas proteínas na haloarquea *Halobacterium salinarum*, um importante modelo no estudo sistêmico de organismos halófilos. Para tal, foram utilizadas ferramentas de bioinformática para análise de smORFs codificantes de proteínas pequenas (MM<10kDa) identificadas a partir de peptídeos obtidos por LC-MS. Esta abordagem permitiu a identificação de 5 possíveis proteínas codificadas por smORFs intergênicas ou antisenso, e a identificação de regiões conservadas e motivos estruturais em comum que podem auxiliar a sugerir funções para estas moléculas. Como consequência destas análises, foi possível sugerir melhorias na anotação do genoma deste organismo. Além disso, foi realizada a padronização de técnicas de biologia molecular, como a marcação cromossômica e a expressão por vetores a fim de validar a expressão de proteínas com marcadores moleculares adicionados na região C-terminal, permitindo a identificação destas por Western blot.

1. INTRODUÇÃO

O desenvolvimento tecnológico permitiu que conceitos gerados a partir da ciência reducionista pudessem ser analisados no contexto de sistemas integrados, ou seja, do organismo como um todo. Desde então, grandes avanços como, por exemplo, o sequenciamento do genoma humano, vem impulsionando cientistas a terem uma nova visão da biologia, chamada de biologia sistêmica (Hood, 2003), com ênfase em estudos que procuram entender como os diversos componentes dos sistemas biológicos interagem e são regulados (Aderem, 2005).

Os estudos em biologia sistêmica envolvem a formulação de hipóteses que integram dados oriundos de tecnologias em larga escala como transcriptômica, proteômica e metabolômica (Levesque & Benfey, 2004) com outras áreas da ciência, como a bioinformática e estatística (Hood, 2003). Estas hipóteses são testadas em organismos vivos selvagens ou mutantes, possibilitando o monitoramento de seus elementos moleculares como genes, proteínas e vias metabólicas. Os resultados obtidos são analisados novamente com o auxílio de ferramentas de bioinformática e estatística, gerando um estudo cíclico e integrado (Ideker, 2001).

1.1. Archaea: o terceiro domínio da vida

Organismos modelo unicelulares dos três domínios da vida têm sido utilizados para entender fenômenos biológicos complexos, utilizando tanto abordagens reducionistas quanto abordagens sistêmicas. A simplicidade dessa organização unicelular fez com que *Escherichia coli* e *Saccharomyces cerevisiae* tenham sido consagrados como organismos modelos para bactérias e eucariotos, respectivamente (Davis, 2004). Organismos do domínio Archaea apresentam também incrível complexidade em sua maquinaria molecular, fazendo deles um atrativo modelo para o estudo metabólico e de regulação

celular (Kletzin, 2007; Jenney, 2007). Apesar disso, ainda tem sido pouco estudados em comparação aos outros domínios.

A descoberta do grupo Archaea como um novo domínio da vida se deu por estudos realizados na década de 1970 visando uma abordagem evolutiva dos organismos unicelulares, na qual pesquisadores utilizaram sequenciamento de marcadores genéticos como o RNA ribossômico para definir relações filogenéticas (Woese *et al*., 1990); Cavicchioli, 2011). A notável diferença encontrada no conteúdo genético desses organismos, que eram até então chamados de metanogênicos, estimulou os pesquisadores a intensificar os estudos para esse grupo, o que aumentou a descoberta de novas características e argumentos para a criação de três divisões taxonômicas, formalmente proposta por Woese na década de 1990, como sendo os domínios "Archaea", "Bacteria" e "Eucarya" (Figura 1) (Woese et al., 1990); Cavicchioli 2011).

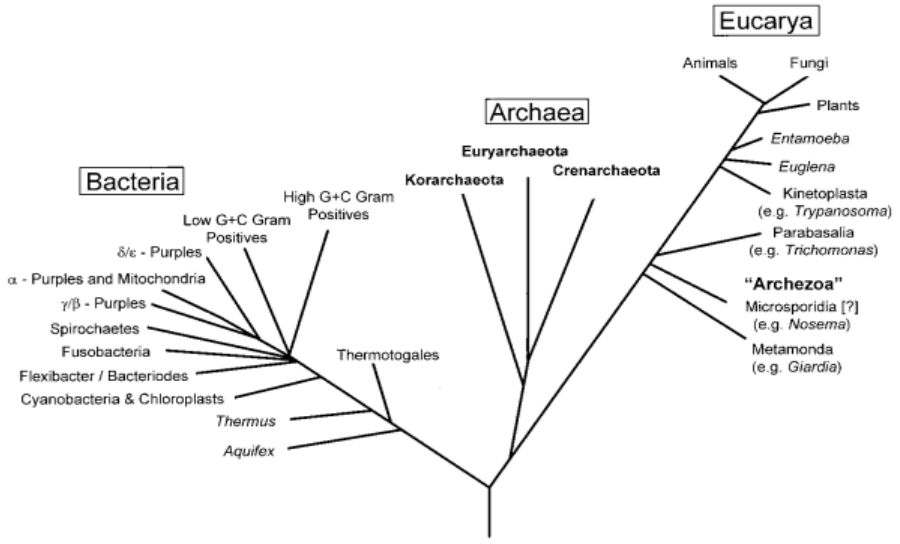

Figura 1. Representação esquemática da árvore evolutiva, construída a partir de análises de rRNA, mostrando a posição filogenética dos três grandes grupos Bacteria, Archaea e Eucarya (modificado de Brown & Doolittle, 1997).

Dentre as novas características descobertas para o grupo Archaea, o que mais chama a atenção são as semelhanças compartilhadas com os domínios Bacteria e Eucarya.

Podemos destacar a maior similaridade com Bacteria na organização celular, como a ausência de núcleo e organelas, genoma compacto em cromossomo circular e proteínas relacionadas a vias metabólicas. A maior semelhança com Eucarya está relacionada à maquinaria de replicação, transcrição e tradução (Kletzin, 2007).

Algumas características tornam o grupo Archaea exclusivo, como a presença de membrana plasmática com lipídios compostos por ligação do tipo glicerol-éter ao invés de glicerol-éster; ausência de peptidoglicanas na parede celular e presença de organismos metanogênicos (Tabela 1) (Kletzin, 2007; Cavicchioli, 2011).

Tabela 1: Principais características que diferem organismos dos três domínios da vida.

Característica	Bactéria	Archaea	Eucarioto
Ligação entre lipídios da membrana celular	Éster	Éter	Éster
Fosfatos dos lipídios da membrana celular	Glicerol-3-fosfato	Glicerol-1-fosfato	Glicerol-3-fosfato
Metabolismo	Bacteriano	Similar ao bacteriano	Eucariótico
Aparato da transcrição	Bacteriano	Similar ao eucariótico	Eucariótico
Fatores de elongação da tradução	Bacteriano	Similar ao eucariótico	Eucariótico
Núcleo	Não	Não	Sim
Organelas	Não	Não	Sim
Metanogênese	Não	Sim	Não
Patógenos	Sim	Não	Sim

(Modificado de Cavicchioli, 2011)

Outra característica importante deste grupo é a presença de vários organismos extremófilos. Estes são tradicionalmente classificados em metanogênicos, termoacidófilos e halófilos. Essa denominação se dá por sua capacidade de sobreviver em habitats anaeróbicos; ou habitats em condições extremas de temperatura, como águas termais e lagos vulcânico; extremas condições de pH, como em lagos ácidos ou alcalinos; ou de extremas condições de osmolaridade, como em salinas e o mar morto (Kletzin, 2007; Cavicchioli, 2011).

Os organismos extremófilos para osmolaridade são denominados halófilos e necessitam de habitats hipersalinos com concentrações de sais que podem chegar a 5,2M.

Para isso, necessitam de adaptações que permitam manter concentrações de moléculas citoplasmáticas isosmótica ao ambiente, seja por acumulação de moléculas orgânicas ou por acúmulo de K^+ ou Na^+ (Kletzin, 2007). Esta segunda opção exige uma adaptação por parte das proteínas, que possuem um predomínio de aminoácidos ácidos na região C-terminal (DasSarma *et al*., 2006; DasSarma *et al*., 2013). Esta propriedade proporciona a estas um ponto isoelétrico ácido variando de 3,5 a 5,5 (Figura 2), fazendo com que sejam dependentes das altas concentrações de sal para o adequado funcionamento (Kletzin, 2007). Essa especialidade permite que os organismos halófilos também sejam utilizados na biotecnologia, como para biorremediação de ambientes hipersalinos (Goo *et al*., 2003).

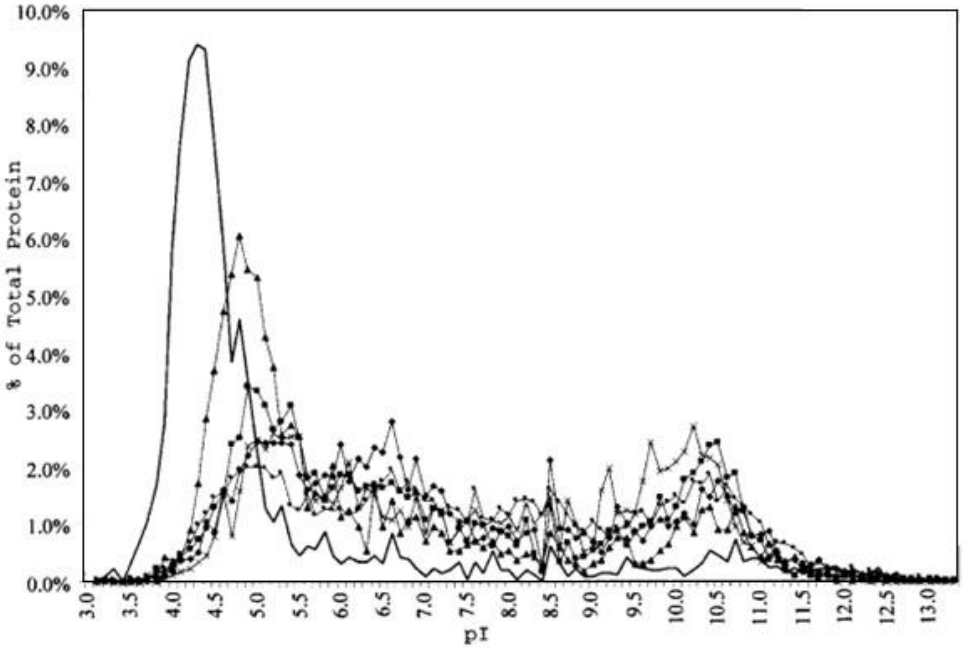

Figura 2: Porcentagem de proteínas em relação ao ponto isoelétrico para o proteoma de 6 organismos: *Halobacterium salinarum* NRC-1 (linha contínua sem marcação), *Methanobacteriuum thermoautotrophicum* (linha com triângulos), *Methanococcus jannaschii* (linhas com x), *Escherichia coli* (linhas com diamantes), *Bacillus subtilis* (linha com quadrados), *Saccharomyces cerevisae* (linhas com círculos) (Kennedy et al. 2001).

1.2. Maquinaria genética em Archaeas

Além das características descritas para Archaea anteriormente, é importante ressaltar aspectos do funcionamento da sua maquinaria genética, que possui grande semelhança com Eukarya e em alguns aspectos com Bacteria (Brown & Doolittle, 1997; Ng et al., 2000; Allers & Moshe, 2005).

A transcrição em Archaea envolve a presença de uma única RNA polimerase, semelhante à RNA polimerase II de Eukarya, composta por subunidades codificadas por 12 genes (Ng et al., 2000, Soppa et al., 2005). Além disso, é conhecido na região promotora a presença de sequências como TATA box e BRE (*B recognition elements*), que funcionam como sítios de ligação para fatores de transcrição TBP (*TATA box binding protein*) e TFB (*Transcription Factor B*) responsáveis pelo recrutamento da RNA polimerase (Wan et al. 2004). Surpreendentemente *H. salinarum*, *Haloferax volcanii* e outros organismos halófilos possuem diversas cópias dos fatores de transcrição TBP e TFB, sugerindo o uso alternativo de pares TFB-TBP na regulação transcricional (Ng et al., 2000; Soppa et al., 2005) Estes sítios de reconhecimento, quando presentes, estão geralmente localizados cerca de 30 nucleotídeos *upstream* ao TSS (*Transcription Start Site*) (Soppa et al., 2005).

Porém, assim como em Bacteria, organismos do domínio Archaea possuem RNAs policistrônicos e não possuem cap na extreminadade 5' e nem cauda poli-A na extreminadade 3' terminal, ambas encontradas em Eukarya (Sartorius-Neef & Pfeifer , 2004; Srinivasan et al. 2006).

Já o processo de tradução em Archaea é um mosaico de características associadas aos outros dois domínios. Ng et al. (2000), encontraram para a Archaea *Halobacterium salinarum* NRC-1 um operon do rRNA semelhante à Bacteria em sua organização. Análises bioquímicas e genômicas sugerem homologia entre alguns dos fatores de início de

tradução entre Archaea e Eukarya (Srinivasan et al. 2006), porém assim como em Bateria, os mRNAs possuem sequências Shine Dalgarno, que em Archaea estão relacionadas à sequência GGAGGUCA, onde o ribossomo se liga para a realização da tradução, geralmente localizado 3-10 nucleotídeos *upstream* ao códon de início (Sartorius-Neef & Pfeifer). Entretanto, em alguns organismos esta sequência não é encontrada com frequência, como por exemplo, *Pyrobaculum aerophilum* e *Halobacterium salinarum*, em que a maioria dos genes não possuem um padrão regulatório conhecido (Srinivasan et al. 2006).

O início da tradução em Archaea ocorre frequentemente no códon AUG (metionina), que é semelhante estruturalmente à metionina de Eukarya. Já em Bacteria o códon AUG é representado pela isoforma formil-metionina (RajBhandary, 2000). Porém em ambos os procariotos, códons de início alternativos GUG, UUG, CUG, AUU, AUC e AUA também são utilizados com menor frequência (Srinivasan et al. 2006, NCBI, tabela 11, http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=tgencodes#SG11).

Um dos modelos biológicos para o estudo da maquinaria genética em Archaea é a halófila *H. salinarum.* Mais informações sobre este organismo serão descritos no tópico a seguir.

1.3. *Halobacterium salinarum* como modelo de estudo de Archaea

Dentre os organismos pertencentes ao domínio Archaea, *H. salinarum* destaca-se como importante modelo de estudo no grupo dos halófilos. Esta é adaptada a extremas concentrações de salinidade (até 5,2 M), concentração que supera em 10 vezes a da água do mar (Ng *et al*., 2000). Este organismo foi descoberto e isolado a partir de peixe salgado em 1922 por Harrison & Kennedy (Leigh *et al*., 2011), apresenta formato de bastonete com

flagelos (Figura 3a) e pode ser encontrado em salinas, lagos de água salgada ou até mesmo, alojado em cristais de sal, como apontado por estudos que identificaram *H. salinarum* e outros halófilos vivos dentro de rochas salinas (halites) datados em mais de 97.000 anos (Figura 3c) (Mormile *et al*., 2003). Além do mais, são anaeróbicos facultativos e apresentam vesículas de gás (Figura 3b), que os permitem flutuar em ambientes aquáticos e em locais onde a disponibilidade de luz é ótima para o seu crescimento (Kletzin, 2007, DasSarma *et al*., 2013). Possuem coloração que varia de rosa a laranja, devido à presença de pigmentos como as bacterioruberinas (carotenóides) e proteínas de membrana fotossensíveis como a bacteriorodopsina (Kletzin, 2007, Jenney, 2007). Devido ao ambiente em que vivem, sofrem grande exposição à luz solar, porém possuem tolerância a altos índices de radiação ultravioleta, devido à sua eficiente maquinaria de reparo de DNA e de outros danos causados no citoplasma. Esta característica impulsionou estudos a utiliza-la como modelo para o melhor entendimento dessas propriedades de reparo celular (McCready, 1996; McCready & Marcello, 2003; Baliga *et al*. 2004).

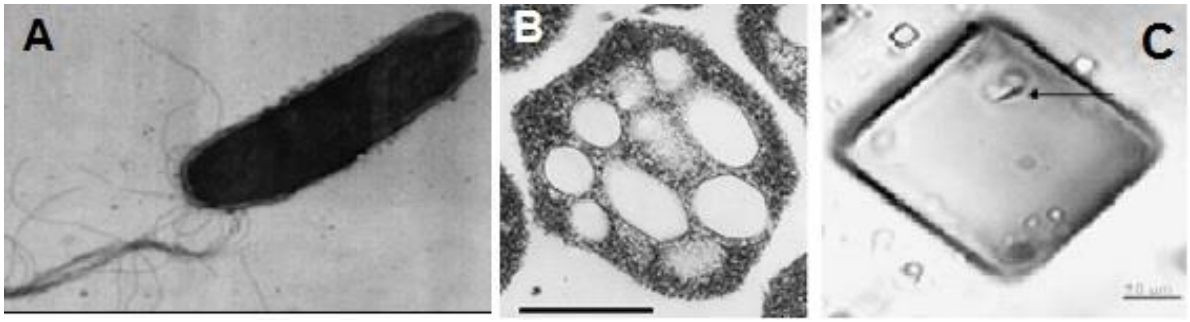

Figura 3. Microscopia eletrônica de *H. salinarum.* Foto em magnificação de 13.500 x, na qual é possível observar a forma bastonete e a presença de flagelos (A). Retirada de (https://www.biochem.mpg.de/522218/Org_Hasal). Corte transversal (B) evidenciando as vesículas de gás (barra preta – 325nm), (modificado de DasSarma *et al*., 2013). *H. salinarum* (seta) (C) encontrada em cristal de sal datado em 97.000 anos. A barra preta inferior à direita representa um tamanho de 10µm (Modificado de Mormile *et al*., 2003).

H. salinarum tem sido o modelo mais utilizado dentro do grupo das Haloarqueas, para diversos outros estudos na área da biologia sistêmica e genômica funcional, em

pesquisas relacionadas ao metabolismo, fisiologia, regulação gênica e interações moleculares (DasSarma *et al*. 2006; Jenney, 2007; Leigh *et al*., 2011; DasSarma *et al*., 2013).

O genoma deste organismo foi sequenciado no ano 2000, apresentando 2.571.010 pares de base, ricos em sequências GC (68%), distribuídos em um grande cromossomo circular (2.014.239 pb) e dois mega plasmídeos, pNRC100 (191.346 pb) e pNRC200 (365.425 pb) (Ng et al., 2000), com 2.629 genes anotados de acordo a última reanotação realizada pelo *National Center for Biotechnology Information* (NCBI) em 2 de Agosto de 2015. Destes, 2540 são considerados CDS (*Coding sequences* - Sequências codificantes de proteínas) e surpreendentemente 40 genes que codificam proteínas com funções essenciais para a célula como DNA polimerase, citocromo oxidase e fatores de transcrição, estão localizados nos mega plasmídeos (Ng et al., 2000). Uma comparação realizada entre as duas principais linhagens de *H. salinarum*, NRC-1 (utilizada neste estudo) e R1, mostrou que a segunda possui 4 plasmídeos, onde há 210 kb que não estão presentes em NRC-1, porém ambas são muito semelhantes em relação à composição do DNA cromossômico e possuem grande quantidade de sequências repetidas entre o cromossomo e os plasmídeos. Esta característica está relacionada ao grande número de inserções de sequências de DNA (91elementos de transposição) o que gera instabilidade ao genoma (Ng, et al., 2000; Brugger, et al. 2002; Pfeiffer et al. 2008). Comparações realizadas entre sequências de nucleotídeos e elementos de transposição concluíram que ambas as linhagens são originadas do mesmo organismo isolado em 1922 e desde então têm divergido em laboratórios (Pfeiffer *et al*. 2008).

Desde o sequenciamento do seu genoma, uma grande quantidade de dados genéticos e moleculares tem sido gerados por meio de análises sistêmicas para *H. salinarum* a partir de tecnologias *high throughput* para análise do transcritoma (microarray,

tiling microarrays, RNAseq), proteoma (espectrometria de massa, géis 2D), e predições *in silico* utilizando ferramentas de bioinformática (Levesque & Benfey, 2004; DasSarmas *et al*., 2006), em diversas condições ambientais e *backgrounds* genéticos (Kaur et al, 2006, Kaur et al, 2010, Koide et al, 2009, Bonneau et al, 2007). Este grande conjunto de dados permitiu a formulação de um modelo global de regulação gênica a partir de dados de transcritoma focado em genes que codificam proteínas (Bonneau et al, 2007, Brooks et al, 2014).

Além disso, diversos estudos de proteômica em *H. salinarum* tem sido realizados, desde estudos *in silico* para predição de estruturas terciárias de proteínas não caracterizadas (Bonneau et al, 2004), como abordagens experimentais utilizando géis bidimensionais e espectrometria de massa para auxiliar na identificação de proteínas citoplasmáticas e proteínas de membrana (Tebbe et al. 2005; Klein et al. 2005; Klein et al.2007). Outro trabalho importante realizado com dados de proteômica foi a criação do banco de dados Peptide Atlas (PA), administrado pelo Institute for Systems Biology, em Seattle, EUA (Van et al., 2008). O objetivo inicial deste projeto foi o mapeamento do proteoma de diversos organismos, (abrangendo diferentes tipos de células e tecidos), incluindo *H. salinarum*, para a qual foi feita a integração de dados de uma variedade de experimentos de espectrometria de massa utilizando peptídeos identificados com um alto índice de confiança. Para o organismo em questão, foram compilados dados de 497 corridas de espectrometria de massa para diversos experimentos diferentes, incluindo o fracionamento de proteínas, enriquecimento por imunoprecipitação e análises quantitativas da variação do proteoma em diferentes condições ambientais, totalizando 16.163 peptídeos distintos. (Van et al., 2008). Estes dados representam a expressão de 63% das proteínas preditas para *H. salinarum* de acordo com a primeira anotação do genoma. Segundo Van et al. (2008), a incapacidade de detecção de peptídeos componentes das demais proteínas pode estar

relacionada principalmente à baixa abundância destas na célula.

Outros elementos regulatórios importantes como RNAs não codificantes foram identificados em *H. salinarum* pela análise de tiling microarrays (Koide et al, 2009) e RNA seq (Zaramela & Vêncio et al, 2014, Gomes-Filho et al, 2015). A existência de uma grande quantidade de pequenos transcritos expresssos e diferencialmente regulados aponta para um papel funcional destas moléculas; a presença de possíveis códons de início e parada levanta também a possibilidade de que muitos desses transcritos, classificados como não-codificantes possam de fato produzir pequenas proteínas.

A grande quantidade de dados produzidos para *H. salinarum* é um fator que facilita o estudo molecular deste organismo, assim como auxilia na correta anotação e identificação de novos genes, permitindo uma visão mais ampla e completa sobre as capacidades funcionais deste organismo (Warren, et al. 2010).

1.4. Anotação de genes e dificuldades na identificação ORFs

Experimentos de sequenciamento e validação da expressão de proteínas são extremamente importantes e contribuem significativamente para anotação mais precisa de genes codificantes de um organismo. Quando um genoma é sequenciado, trabalhos de anotação de genes são necessários para compreender a informação e entender a ligação entre a fisiologia e os genes identificados. Essas anotações são realizadas em sua maioria através de predições de ORFs (*Open Reading Frames*) por softwares que utilizam modelos matemáticos, como modelos de Markov ocultos (HMMs, *Hidden Markov Models*), utilizado por diversas ferramentas de anotações de genomas em busca por padrões que auxiliem no reconhecimento de possíveis sequências funcionais de nucleotídeos (Delcher *et al*. 2007, Kelley et al. 2015).

Muitos softwares podem apresentar precisões elevadas, com alta percentagem de

acertos na identificação de ORFs, (até 98% para alguns organismos) (Delcher *et al*. 2007), porém essa precisão pode variar de acordo com os limiares estabelecidos para a seleção de ORFS menores, implicando a geração de muitos falsos positivos e negativos e consequentemente, uma anotação pobre para essas sequências pequenas (Warren et al., 2010). Além do problema em relação ao tamanho, softwares de anotação enfrentam dificuldades em genomas extremamente compactos devido à grande proximidade entre os genes, que muitas vezes se sobrepõem em *frames* diferentes. Um exemplo da problemática em anotação de genes é a diferença encontrada entre as linhagens NRC-1 e R1 de *H. salinarum*, quanto aos números, posições e tamanho de genes anotados, mesmo estas possuindo genomas praticamente idênticos. Essas variações acontecem principalmente na posição do códon de início, evidenciando erros cometidos pelas ferramentas computacionais (Pfeiffer *et al*. 2008).

Outra etapa importante na anotação de um gene é a identificação de CDS (*coding sequences* – sequências codificantes), que podem ser geradas experimentalmente, por sequenciamento e/ou validação da expressão, ou através de softwares de predições. Estes softwares fazem comparações e buscas por homologias com genes ortólogos de outros organismos. Em geral, comparações com identidade superior a 50% podem ser usadas para predições funcionais e evolutivas entre os genes e proteínas. Para proteínas com identidade menor de 50%, análises em diferentes bancos de dados são necessárias (Vasconcelos & Almeida, 2012). As proteínas identificadas através de CDS são submetidas a um banco de dados, como por exemplo, o EMBL-Bank/GenBank/DDB e após checagem manual são integradas à banco de dados não redundantes como UniProtKB/Swiss-Prot. A grande maioria das proteínas anotadas em banco de dados é identificada somente por predições e análises de bioinformática e somente cerca de 5% são obtidas experimentalmente (informação disponível na página http://www.uniprot.org/help/sequence_origin). Este é um

fato que diminui a precisão das anotações e caracterizações *in silico*, devido a incerteza em relação a existência da proteína. Muitas vezes comparações geram alinhamentos somente com proteínas hipotéticas e dessa forma não é possível estabelecer legitimidade e função. Portanto, o aumento do número de experimentos de proteômica pode contribuir com a identificação de outras proteínas ainda não conhecidas.

Outro problema enfrentado em relação à anotação de genes é a identificação de smORFs (small ORFs – pequenas ORFs menores de 300 nucleotídeos), pois geralmente são estabelecidos limiares para a seleção de ORFs maiores que possuem menor chance de gerarem resultados falso positivos. Este fato é compreensível, pois se os softwares fossem ajustados para encontrar ORFs de qualquer tamanho, haveria um resultado de milhares ou milhões destas dependendo do organismo em questão. Isso diminui o índice de acerto e consequentemente da qualidade dos dados gerados. (Samayoa, 2011; Storz et al, 2014).

1.5. smORFs e pequenas proteínas

Ao longo dos anos, as pequenas proteínas, definidas como menores que 50 ou até mesmo 100 aminoácidos, foram mal compreendidas ou ignoradas (Hobbs *et al*., 2011; Storz *et al*., 2014), devido aos desafios em idenficá-las em misturas complexas de proteínas. Características como pequena massa, tamanho e baixa abundância na célula (Ma *et al*., 2014) fazem com que geralmente sejam perdidas quando se utilizam técnicas tradicionais de proteômica (Klein *et al*. 2007). Além disso, pequenas proteínas são difíceis de serem identificadas e anotadas, devido às limitações e imprecisões estatísticas de softwares para análise de pequenos peptídeos (Samayoa *et al*., 2011) e de smORFs que muitas vezes não usam AUG como códon iniciador (Ma *et al*., 2014; Stortz et al, 2014).

Um exemplo da complexidade das smORFs codificadoras de pequenas proteínas, é a diversidade de posições em que estas podem ser encontradas no genoma, podendo ser

originárias de posições em frames alternativos como *upstream* à região codificante; *upstream* à região codificante porém com sobreposição; dentro da região codificante em frame de leitura alternativo; truncada em uma região codificante; extendida à região codificante (Ingolia, 2014) (figura 4).

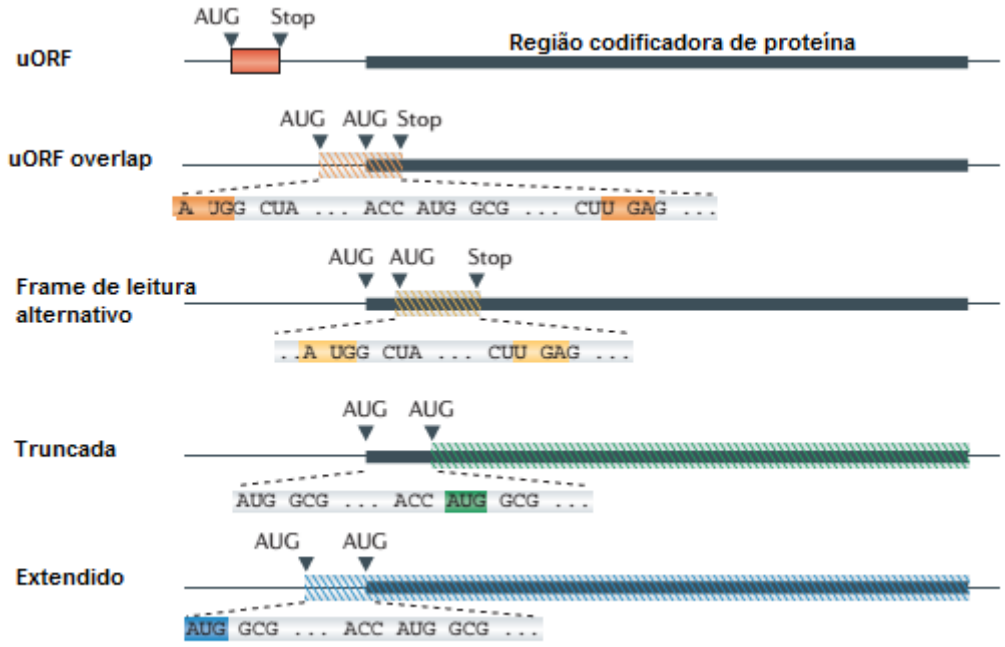

Figura 4: Esquema dos diferentes tipos de ORFs alternativas. uORF: *upstream* ao códon de início de um gene (dentro da região 5'UTR); uORF overlap: *upstream* ao códon de início com sobreposição à região codificante do gene; Frame de leitura alternativo: dentro da região codificante de um gene, porém em *frame* diferente; Truncada: truncada em uma região codificante, com o mesmo códon de parada mas com o códon de início *downstream* ao início do gene; Extendida: dentro da região codificante, porém com tamanho extendido, com o mesmo códon de parada mas com códon de início *uptream* ao início do gene. Figura modificada de Ingolia, (2014).

Devido à variedade de smORFs que podem codificar pequenas proteínas, a maioria destas são de difícil identificação e geralmente são encontradas por acaso, devido à mutações em regiões não anotadas do cromossomo ou por inspeção de transcritos, genes ou regiões promotoras (Storz et al. 2014). Porém, avanços na sensibilidade de técnicas de espectrometria de massa (Andrews & Rothnagel, 2014; Ma *et al*., 2014), *tiling-arrays* e melhor integração de bioinformática, genômica e proteômica tem contribuído para o crescente descobrimento do número de genes codificadores de proteínas menores que 50

aminoácidos (Hobbs et al., 2011; Andrews & Rothnagel, 2014). Os avanços dessas técnicas também contribuíram para a identificação de pequenas proteínas e peptídeos que antes pensava-se ser ncRNAs (Andrews & Rothnagel, 2014). Por exemplo Washietl et al., (2011) identificaram em *Escherichia coli*, um ncRNA que de fato codificava uma pequena proteína.

As funções das pequenas proteínas são ainda muito inexploradas (Cheng et al., 2011), porém diversos trabalhos recentes já tem descrito a importância e função destas em todos os domínios da vida (Samayoa *et al.*, 2011). Peptídeos ativos, codificados por smORFs têm sido identificados por possuir papéis regulatórios em células eucarióticas (Andrews & Rothnagel, 2014). Por exemplo, 22 smORFs foram descobertas em *Saccharomyces cerevisiae* codificando peptídeos necessários para o crescimento em altas temperaturas, transporte, metabolismo intermediário, segregação de cromossomos, estabilidade do genoma, dentre outras funções (Kastenmayer *et al*., 2006). Além destas, pequenas proteínas que fazem parte de componentes da fotossíntese em plantas também foram descobertas (Shi & Schröder, 2002).

Em *Bacillus subtilis*, foi encontrado um pequeno gene (ComX) que codifica um peptídeo de 10 aminoácidos na sua forma ativa, atuando como feromônio (Hobbs et al., 2011). Além disso, outros trabalhos também citam funções biológicas de pequenos peptídeos encontrados em eucariotos e bactérias, atuando como nucleases, proteínas de membrana (Cheng et al., 2011), antibióticos, reguladores de transcrição e de proteínas maiores, e participantes do metabolismo energético e sinalização celular (Hobbs et al., 2011).

Pequenas proteínas também foram encontradas em Archaea, porém ainda há poucos trabalhos a respeito destas. Por exemplo, Humbard et al. (2010), identificaram 2 pequenas proteínas relacionadas à atividade de ubiquitinase em *Haloferax volcanii*. Em outro estudo,

Tarasov et al. (2008), relataram a função de uma pequena proteína de apenas 60 aminoácidos, codificada na região intergênica entre os genes *brp* e *bop* e que estava relacionada à regulação deste último. Mais recentemente Prasse et al. (2015) descreveram três pequenas proteínas codificadas por *small putative* RNA (spRNA) em *Methanosarcina mazei*.

Além disso, Klein et al., realizaram em 2007 um trabalho com enfoque nas proteínas de baixo peso molecular (< 20 kDa) em *H. salinarum*, onde de 1105 proteínas preditas teoricamente, somente 380 foram identificadas, sendo que destas, pouco sucesso foi obtido para as realmente pequenas, menores que 5 kDa. Isso mostra que lacunas ainda permanecem, sugerindo a necessidade de trabalhos em Archaea onde o enfoque seja a identificação dessas pequenas proteínas.

Os exemplos acima mostram em diferentes espécies, as diversas funções das pequenas proteínas e peptídeos e seus importantes papéis biológicos na célula. Assim, o aprofundamento dos estudos para identificação da existência e função de novas proteínas se mostra essencial, podendo trazer contribuições significativas tanto no estudo básico, quanto no aplicado.

2. OBJETIVOS

2.1 Objetivos gerais

Como parte de um estudo na área da biologia sistêmica, este trabalho visa a identificação de pequenas ORFs com potencial para expressão de pequenas proteínas de baixo peso molecular (até 100 resíduos de aminoácidos), incluindo proteínas expressas a partir de ncRNAs identificados em *Halobacterium salinarum* NRC-1, utilizando dados de identificação de peptídeos por espectrometria de massa e análises de bioinformática.

2.2 Objetivos específicos

-Identificação e caracterização *in silico* de pequenas proteínas

-Padronização de técnicas utilizadas para a validação da expressão de pequenas proteínas

3. MATERIAL E MÉTODOS

3.1. Análises *in silico* de proteínas

Neste tópico será descrita a metodologia utilizada para a identificação e caracterização *in silico* de pequenas proteínas em *H. salinarum* NRC-1**.**

3.1.1. Obtenção dos dados por espectrometria de massa

Para o estudo das pequenas proteínas, um projeto conduzido em 2014 no Laboratório de Biologia Sistêmica de Microrganismos (LaBiSisMi) pelo Dr. Gilvan Pessoa Furtado (Processo Fapesp 2013/23712-6), realizou experimentos de LC-MS (cromatografia líquida seguida por espectrometria de massas), onde as proteínas de *H. salinarum* NRC-1 foram extraídas e selecionadas através de um sistema de filtração Vivaspin 20 (GE Healthcare) com filtros de separação de 10KDa, a fim de enriquecer as amostras de proteínas pequenas. Essas amostras foram dessalinizadas, liofilizadas e enviadas para o *Mass Spectrometry and Proteomics Resource Laboratory*, uma *facility* para serviços de identificação e sequenciamento de peptídeos em amostras complexas da Universidade de Harvard, em Boston, EUA. Essa *facility* possui espectrômetros do tipo Orbitrap e o grupo possui experiência com identificação de SEPs (*small ORFs Encoded Polypeptides*) (Slavoff et al. 2013). Foram realizadas duas corridas e estas foram analisadas no software MaxQuant e Proteome Discoverer (Thermo Fisher), utilizando o software percolator como pacote estatístico. Ao todo, foram identificados 2441peptídeos com alto *score* (q-

valor limiar < 0,01) em duas corridas de LC-MS, sendo que destes, 2339 se alinharam em uma única posição no genoma, 2239 com o cromossomo, 16 com o pNRC100 e 84 com o pNRC200. Além disso, um *script* foi implementado em linguagem R para detectar os peptídeos que se alinham somente em posições sem genes anotados, ou seja, intergênicos ou antisenso. Esta análise resultou na identificação de 159 peptídeos no cromossomo, 7 no pNRC200 e 3 no pNRC100.

3.1.2. Identificação de pequenas proteínas

A partir dos dados obtidos no estudo anterior, um total de 169 peptídeos localizados em posições sem genes anotados foram visualizados através do software GGB (Gaggle Genome Browser, Bare et al., 2010), que permite a integração e visualização de dados oriundos de diversos tipos de experimentos produzidos em larga escala, com o intuito de identificar as possíveis pequenas proteínas codificadas por smORFs menores de 300 nucleotídeos. Foram consideradas as seguintes características: (i) a presença de códon de início e de parada no mesmo frame do peptídeo detectado, (ii) presença de sinal de transcrição por RNAseq e Tilling array (Koide *et al*., 2009; Zaramela et al. 2014) e (iii) presença de TSS (*Transcription Start Site*) identificados por experimentos de dRNAseq (Zaramela et al. 2014; ten-Caten et al., em preparação). A obtenção desses últimos dados é feita através de uma análise seletiva de transcritos primários, que são aqueles que apresentam extremidades 5' trifosfato (5' PPP). Estes são enriquecidos devido ao tratamento com a enzima TEX (Terminator 59 Phosphate-Dependent Exonuclease), que degrada RNAs processados (extremidades 5' monofosfato). A comparação entre amostras tratadas e não tratadas com a enzima TEX permite a identificação dos inícios de cada transcrito (Sharma et al, 2010).

Além disso, foram feitas análises dos peptídeos localizados na mesma posição de

genes anotados, com o foco em genes hipotéticos menores que 300 nucleotídeos; e peptídeos localizados em regiões gênicas em frames diferentes ao do gene anotado.

3.1.3. Caracterização *in silico* de pequenas proteínas

Para a caracterização *in silico* de pequenas proteínas encontradas em regiões sem genes anotados foram utilizadas diversas ferramentas disponíveis *online* e banco de dados públicos.

Busca por similaridade de sequências e domínios conservados foram feitas através da ferramenta BLAST (*Basic Local Alignment Search Tool*), disponível online no NCBI (*National Center for Biotechnology Information*). Estas análises foram realizadas através de 2 tipos de BLAST. O BLASTp que faz buscas a partir de sequências de aminoácidos no banco de dados de proteínas; BLASTx: que realiza buscas a partir de sequências de nucleotídeos traduzidas nos seis *frames* de leitura no banco de dados de proteínas (Altschul et al. 1990). Além disso, estas ferramentas fazem buscas no CDD (*Conserved Domain Database*), para a identificação de domínios conservados (Marchler-Bauer et al. 2015).

Também foram realizadas análises utilizando a ferramenta BLAST do banco de dados Uniprot (*Universal Protein Resource*), que é um banco de dados não redundante de sequências de proteínas e anotações funcionais. O Uniprot foi formado a partir da junção do *Swiss-Prot knowledge Database*, TrEMBL e PIR (The UniProt Consortium, 2008).

Para análises de conservação de domínios funcionais e estruturais foram utilizados os softwares Phyre^2 (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index) e MyHits (http://myhits.isb-sib.ch/cgi-bin/motif_scan). O Phyre^2 é a segunda versão do software desenvolvido por Kelley et al. (2015) e utiliza sequências de aminoácidos para escanear banco de dados em busca por homologia. O alinhamento múltiplo de sequências é utilizado para predição de estruturas secundárias, tanto a sequência alinhada quanto a estrutura

secundária predita são combinadas e convertidas em um *Hidden Markov Model* (HMM), que identifica padrões de mutação em sequências de aminoácidos a fim de estabelecer uma impressão digital evolutiva da sequência. Esse resultado é escaneado contra um banco de dados de HMM de proteínas já conhecidas, gerando alinhamentos. Aqueles com maiores escores são utilizados para a construção da estrutura da proteína de interesse. Já o Myhits é um banco de dados de domínios proteicos e motivos estruturais, cuja ferramenta "Motif scan" procura por homologia de sequências, a fim de identificar motivos estruturais conservados que permitem estabelecer relações de estrutura e função de proteínas não conhecidas (Pagni et al., 2004).

Além dos softwares descritos acima, foram usadas ferramentas *online* disponíveis no ExPASy (Bioinformatics Resource Portal) para calcular o ponto isoelétrico e massa molecular teórica das proteínas estudas (http://web.expasy.org/compute_pi/) e no *endmemo* para calcular a porcentagem de nucleotídeos GC nas sequências das smORFs (http://www.endmemo.com/bio/gc.php).

3.2. Construções moleculares

Nesta etapa do estudo, foram aplicadas técnicas de biologia molecular para a construção de linhagens recombinantes de *H. salinarum* NRC-1, para a padronização das metodologias de validação da expressão de proteínas, o *Chromosomal tagging* (marcação cromossomal) e expressão de genes através de vetores.

3.2.1 Cultivo de *H. salinarum* NRC-1

As linhagens de *Halobacterium salinarum* NRC-1 e NRC-1 Δ*ura3* (Knockout para o gene ura3) e demais linhagens recombinantes, foram cultivadas em meio completo (CM) sólido ou líquido, composto por NaCl 250 g/l, $MgSO_4.7H_2O$ 20 g/l, KCl 2 g/l, citrado de

sódio 3 g/l, peptona 10 g/l e água destilada. Dependendo da condição, o meio foi suplementado com uracila 50 μg/mL, mevinolina 20 μg/ml ou *5-fluoroorotic acid* (5-FOA) 300mg/ml. Para o meio sólido, houve adição de 15g/l de ágar. Estas cresceram em laboratório em estufa a 37°C e sob condições de alta luminosidade, por aproximadamente 7 a 10 dias em meio sólido para a formação de colônias visíveis ou 3 dias em meio líquido, sob agitação de 125 a 250 rpm, até chegar à fase exponencial crescimento ($D.O.600_{nm}$ = 0,5).

3.2.2. Oligonucleotídeos

Os oligonucleotídeos utilizados neste trabalho foram sintetizados pelas empresas Invitrogen ou IDT e estão listados na Tabela 2.

Tabela 2: Oligonucleotídeos utilizados neste estudo.

Nomes	Sequências	Aplicação
F1	5’ – ATCTGAATTCGGCGACGGCCCCGTGGTGAT – 3'	Chromosomal tagging
R1	5’ – TTACTTGTCGTCATCGTCTTTGTAGTCTGGTTTGATGGTGACG – 3'	Chromosomal tagging
F2	5’ - GACGATGA CGACAAGTAACTGGCGCAGGAACCCCGA – 3'	Chromosomal tagging
R2	5’ – AGATAAGCTTTGGGCGACGATGCCCGCCGA – 3'	Chromosomal tagging
F1'	5’ – TTGGGTTCGACCTCGACGTG – 3'	Chromosomal tagging
FLAG rev	5’ – CTTGTCGTCATCGTCTTTGTAGTC – 3'	Chromosomal tagging
5' Pfer Pac I	5' – GCGCTTAATTAAGGCCGGCAGCACCTG – 3'	Construção de vetores
F24	5' – CGCCAGGGTTTTCCCAGTCACGAC – 3'	Construção de vetores
R24	5' – AGCGGATAACAATTTCACACAG – 3'	Construçãode vetores
3xFLAG_pRMG_PstI_HindIII_F	5' - GGACTACAAAGACCATGACGGTGATTATAAAGATCAT GACATCGATTACAAGGATGACGATGACAAGTGAA – 3'	Construção de vetores
3xFLAG_pRMG_PstI_HindIII_R	5' - AGCTTTCACTTGTCATCGTCATCCTTGTAATCGATGTCATGA TCTTTATAATCACCGTCATGGTCTTTGTAGTCCTGCA – 3'	Construção de vetores
cmyc-PstI-F:	5' – ATGAGACTGCAGATGATCCCCGGGTTAATT – 3'	Construção de vetores
cmyc-HindIII-R	5' – GCGCGCAAGCTTTCACTAGTGATTGATTAA – 3'	Construção de vetores

Lsm-compromotor-EcoR1-F	5' – TATAGAATTCGGACGGCGGGTGGCG – 3'	Construção de vetores
Lsm-BamH1-R	5'- GGATCCTGGTTTGATGGTGACG – 3'	Construção de vetores
Lsm-sempromotor-EcoR1-F	5' - ATGCGAATTCATGGATGCCACCACCG – 3'	Construção de vetores

3.2.3. Marcação cromossômica

Foi usado o marcador (*epitope-tag)* do tipo FLAG, composto por uma sequência de oito aminoácidos (DYKDDDDK), para avaliar a expressão gênica através da técnica da marcação cromossômica (*chromosomal tagging*). Tal procedimento já foi utilizado em *H. salinarum* por Wilbanks (2012) e colaboradores, porém com o peptídeo marcador HA (*Human influenza hemagglutinin*).

3.2.4. Padronização do *Chromosomal tagging* para o gene *lsm*

Com o intuito de padronizar o método utilizado para uma futura validação da expressão de pequenas proteínas, um controle positivo foi utilizado. Para isso, foi escolhido o gene *lsm* (VNG_RS05825, antiga VNG1496G) de *H. salinarum*, pertencente à grande família das proteínas do tipo Hfq (Bacteria), Sm e Sm-like (LSm)(Eucariotos e Arqueias) (Fischer et al., 2010). A proteína LSm está relacionada ao metabolismo de RNA e é considerada um componente chave da rede global pós-transcricional (Wilusz & Wilusz, 2005; Vogel & Luisi, 2011). A massa molecular desta proteína em *H. salinarum* é de aproximadamente 7KDa, codificada por uma ORF de 210 pb. Além do tamanho pequeno, a taxa de expressão do gene também é considerada baixa (Fischer et al., 2010), o que faz da LSm um controle semelhante às pequenas proteínas a serem estudadas.

A inserção do marcador FLAG no DNA cromossômico foi feita através da recombinação homóloga, sem o uso de enzimas de restrição, como sugerida por Horton e colaboradores em 1989.

3.2.5. Construção do gene controle marcado com FLAG

A primeira etapa da técnica de recombinação consistiu na geração de um fragmento de ~1000pb por PCR overlap (Heckman & Pease, 2007). Para isso, foram amplificados dois fragmentos de ~500pb, um *upstream*, primers F1, R1 e outro *downstream* ao códon de parada do gene lsm, primers F2, R2, com a adição da sequência codificadora do FLAG (através dos primers F2 e R1) em uma das extremidades de cada fragmento, de forma que ambos os fragmentos amplificados possuíssem uma região complementar entre eles (Figura 5). Assim sendo, os fragmentos puderam se unir através de um PCR *overlap* (Horton *et al*., 1989; Heckman & Pease, 2007), no formato 500pb::FLAG::500pb (Figura 5). A adição do FLAG implicou na troca de posição do stop códon nativo do gene lsm, que foi substituído pela sequência de nucleotídeos do FLAG com um stop códon no final. O fragmento final foi inserido por digestão enzimática no vetor pHsal-S (Figura 5), um vetor suicída, construído para esta finalidade (Silva-Rocha et al., 2015), que possui origem de replicação bacteriana, mas não para arqueia, além do gene *ura3* e os genes de resistência à mevinolina e ampicilina.

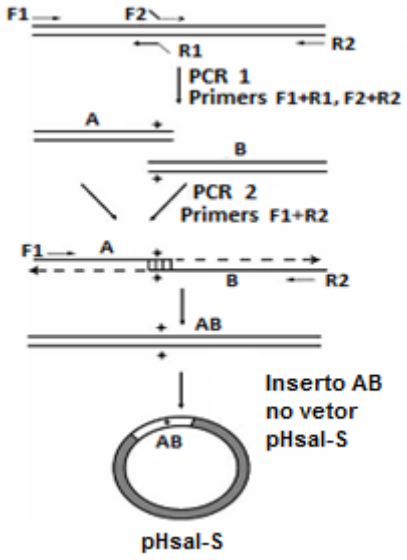

Figura 5. Reação de PCR *overlap* para a construção do fragmento 500pb::FLAG::500pb, inserido no vetor. Na reação de PCR 1, os *primer* F1 + R1 amplificam 500pb *upstream* ao gene LSM (fragmento A) e os *primers* F2 + R2 amplificam os 500pb *downstream* ao gene LSM (fragmento B). Nota-se que a sequência

overlap (correspondente ao FLAG) acompanha os *primers* F2 e R1. Na reação de PCR 2, ambos fragmentos de 500pb (A e B) são unidos pela região *overlap* e amplificados pelos *primers* F1 e R2, produzindo um fragmento de ~ 1000pb (AB), que é inserido ao vetor pHsal-S por digestão enzimática. (Esquema modificado de Heckman & Pease, 2007).

3.2.6. Transformação

Antes da transformação em *H. salinarum*, o vetor pHsal-S contendo o fragmento com *o FLAG* foi transformado em *E. coli* DH5α para replicação e posterior extração do DNA plasmidial, que foi sequenciado, confirmando a construção.

Os plasmídeos foram inseridos em *H. salinarum* NRC-1 *Δura3*, seguindo o protocolo de transformação química sugerido por Dyall-Smith, (2009), no qual utiliza EDTA para a remoção da parede celular, produzindo esferoplastos. Em seguida, os transformantes foram cultivados em meio CM líquido até atingirem a fase exponencial de crescimento, densidade óptica de 0,5 em 600nm**.**

3.2.7. Recombinação cromossômica (*crossover*)

Após o crescimento, os transformantes foram plaqueados em meio CM sólido com adição de mevinolina, agente seletor para arqueias, no qual somente os indivíduos transformantes que possuem o plasmídeo sobrevivem, devido ao gene de resistência mev^R presente neste. Uma segunda seleção foi feita, visando a expulsão do plasmídeo. Nesta, as arqueias foram plaqueadas em meio sólido com adição de *5-fluoroorotic acid* (5-FOA), agente seletor negativo para a presença do gene *pyrF* (presente exclusivamente no plasmídeo), de forma que a ausência deste é necessária para a sobrevivência, indicando a perda do plasmídeo nas colônias sobreviventes e consequentemente a recombinação cromossômica concluída (Figura 6). Para checar se a recombinação foi bem sucedida, as colônias que cresceram foram analisadas, amplificando-se o fragmento upstream ao gene *lsm*, com um novo *primer* F1' que se liga a uma região do DNA genômico anterior aos

500pb, eliminando a possibilidade de amplificação do fragmento contido no vetor (caso esse não tivesse sido eliminado) e R', *primer* que se liga ao FLAG. Além disso, esta construção foi confirmada por sequenciamento.

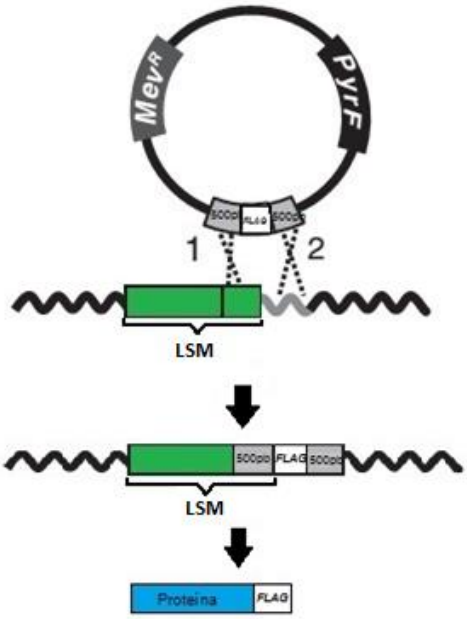

Figura 6. Marcação cromossômica: Esquema simplificado representando um plasmídeo (linha continua preta) com um inserto portador do marcador FLAG (branco) e dois fragmentos de 500pb (quadrados cinza), no qual um é homólogo a região terminal do gene *lsm* (verde) e outro é homólogo a uma região do DNA cromossômico (linha ondulada cinza). Após a recombinação, o inserto substitui a região homóloga no cromossomo e o DNA cromossômico correspondente passa a fazer parte do plasmídeo. O gene, agora ligado ao marcador FLAG codificará uma proteína (azul) também marcada. Imagem modificada de Wilbanks *et al.* (2012).

3.2.8. Construção de vetores para a expressão do gene controle (*lsm*)

Além da marcação cromossômica com o peptídeo FLAG, outros dois tipos de construções foram idealizadas, visando-se a avaliação da expressão do gene *lsm* a partir dos vetores plasmidiais pHsal-C e pHsal-E (Silva-Rocha et al., 2015), ambos com origem de replicação para bactéria e arqueia e genes de resistência à Carbenicilina e Mevinolina. Porém o primeiro vetor não possui promotor e o segundo possui um promotor modificado da ferredoxina (VNG_RS08265), adequado para superexpressão de genes (Figura 7).

Ao todo, quatro tipos de vetores foram panejados a fim de se testar em *H.*

salinarum NRC-1 diferentes condições de expressão com diferentes marcadores (*epitope tags*), 3xFLAG e Cmyc. Os novos vetores construídos foram: pHsal-C-Cmyc-Lsm (com promotor nativo), pHsal-C-3xFLAG-Lsm (com promotor nativo), pHsal-E-Cmyc-LSm e pHsal-E-FLAG-LSm. O marcador 3xFLAG foi produzido sinteticamente por meio de 2 grandes primers complementares, F: 3xFLAG_pRMG_PstI_HindIII_F e R: 3xFLAG_pRMG_PstI_HindIII_R, com sítios de restrição para PstI e HindIII e o marcador Cmyc, amplificado a partir de uma outra construção no vetor pHsal-S pelos primers F: cmyc-PstI-F e R: cmyc-HindIII-R, com sítios de restrição para PstI e HindIII. O gene *lsm* foi amplificado sem e com promotor nativo (~200pb *upstream* ao códon de início do gene), com os *primers* F: Lsm-sempromotor-EcoR1-F e R: Lsm-BamH1-R para o fragmento do gene sem o promotor e os *primers* F: Lsm-compromotor-EcoR1-F e R: Lsm-BamH1-R para o fragmento do gene com promotor. Devido ao acoplamento de um marcador após o gene, o *primer* reverse do lsm exclui o códon de parada do mesmo.

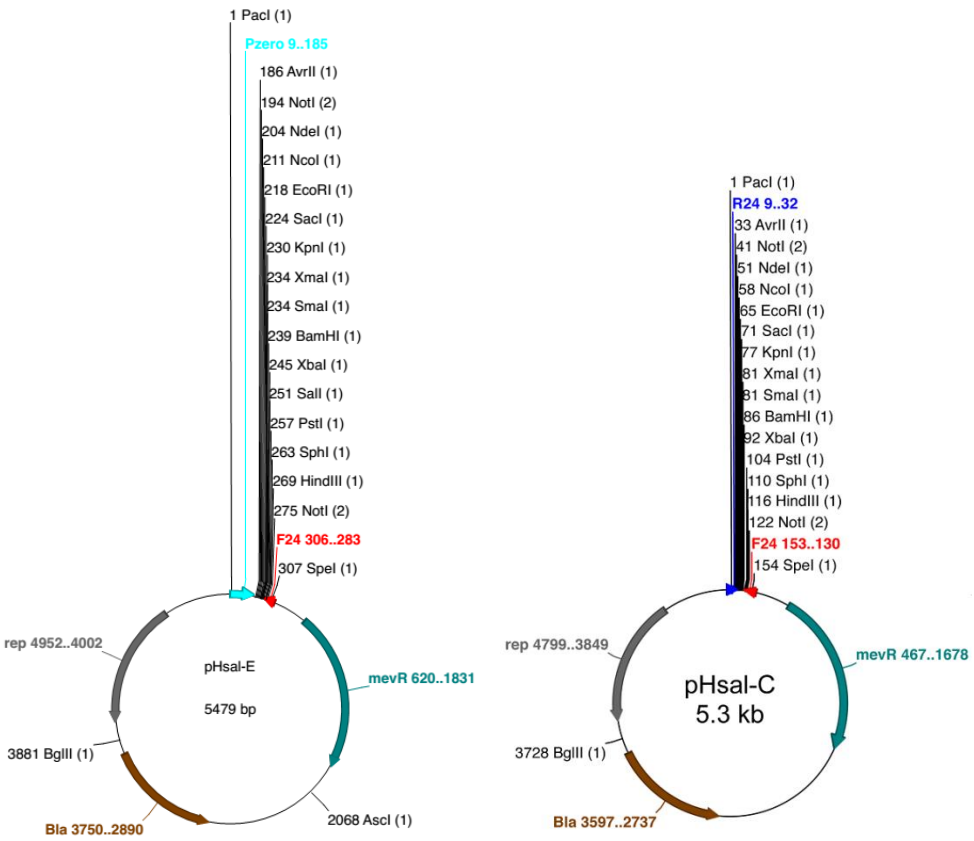

Figura 7. Vetor pHsal-E (esquerda), utilizado juntamente com o vetor pHsal-C (direita) como base para as construções pHsal-C-Cmyc-lsm(com promotor nativo), pHsal-C-3xFLAG-lsm (com promotor nativo), pHsal-E-Cmyc-lsm e pHsal-E-FLAG-lsm. A diferença entre os dois vetores está somente na presença do promotor Pzero (seta azul clara) (criado a partir do promotor da Ferredoxina de *H. salinarum NRC-1*) no vetor pHsal-E. Nota-se o gene de resistência à mevinolina (mevR em verde), gene de resistência à Carbenicilina (Bla, em marrom), origem de replicação (rep em cinza). Imagem modificada de Silva-Rocha et al., (2015). R24: sítio de ligação do oligo *reverse*. F24: sítio de ligação do oligo *forward*.

3.2.9. Lise celular e quantificação de proteínas

Os organismos recombinantes através da inserção de vetores para a expressão foram cultivados em 50mL de CM com adição do antibiótico mevinolina e os recombinantes através da técnica da marcação cromossômica foram cultivados em 50mL de CM com adição de uracila. Após atingir a $D.O._{600nm}$=0.5, as culturas foram transferidas para tubos de polipropileno e centrifugadas (centrífuga Eppendorf 5804 R) a 8.000g por 5 minutos a 4°C. O *pellet* foi lavado com 20mL de solução basal de sal (NaCl 250g/l; $MgSO_4.7H_2O$ 20 g/l; KCl 2g/L; citrato de sódio 3 g/L) e centrifugado novamente a 8000g por 5 min a 4°C.

Para os recombinantes com vetores de expressão, o *pellet* foi ressuspendido e utilizado diretamente para a análise por Western blot. Para a linhagem recombinante por marcação cromossômica, o *pellet* foi resuspendido em 1mL de TBS suplementado com inibidor de protease (Sigma S8830-20TAB) e em seguida transferido para microtubos plásticos de 1,5mL. As células foram lisadas por sonicação (Q125 Qsonica) em potência de 40% com pulsos de 15 segundos com intervalos de 15 segundos por 6 minutos. O produto da lise foi centrifugado a 11.000g por 40 minutos a 4°C, o precipitado foi descartado e o sobrenadante aspirado e transferido para outro microtubo de 1,5 mL. O produto da extração de proteínas foi quantificado pelo teste de Bradford (Bradford, 1976).

3.2.10. Imunoprecipitação

Esta etapa foi realizada somente para os organismos recombinantes cromossomais. Esta etapa consistiu no enriquecimento das amostras com proteínas marcadas por FLAG. Para tal, um volume de 50 µL (por reação) de *Beads* magnéticas acopladas à Anti-mouse IgG (Dynabeads Pan Mouse IgG*:* Invitrogen REF 11041) foi centrifugado a 2.000g por 2 minutos, o supernadante removido e o precipitado lavado com TBS; este procedimento foi repetido 2 vezes. Para o acoplamento do anticorpo com as *beads*, estas foram centrifugadas novamente a 2.000 g por 2 minutos e resuspendidas em 250 µL TBS com adição de 2 µL de anticorpo Anti-FLAG M2 (Sigma: F1804) e incubadas por 3 horas à temperatura ambiente. Em seguida estas foram lavadas 2x com TBS para eliminar os anticorpos não ligados. Após a lavagem, foram centrifugadas e ressuspendidas em 30µL (6µg/µL) do produto da lise celular, para que a proteína marcada com o FLAG se ligasse ao Anti-FLAG acoplado as *beads*. E foram incubadas por 5 horas a 4°C sob agitação rotativa e em seguida lavadas 5x em tampão de lise (TBS). As proteínas foram eluídas em 50 µL de TE + SDS 0,1% a 65°C por 10 minutos e centrifugadas em velocidade máxima por 30 minutos. O sobrenadante foi aspirado e transferido para um microtubo de 1,5ml.

3.2.11. Western blot e Dot blot

Os pellets das culturas recombinantes com vetores e o produto da imunoprecipitação das proteínas dos organismos recombinantes cromossomais foram misturados ao tampão de amostra (glicerol 30% (v/v), SDS 9,2% (p/v), Azul de bromofenol 1% (p/v), β-mercaptoetanol 20% (v/v), tris-HCl 1M pH 7,0 (0,25M)), aquecido a 95°C por 5 minutos e corrido em gel de poliacrilamida 12%, 1,5mm, em tampão de corrida (tris 0,25M, Glicina 1,9M, EDTA 10mM, SDS 35mM). As proteínas foram transferidas para membrana de nitrocelulose de poro 0,22 µm umidecida em tampão de

amostra de sistema semi-seco (Tris 48mM, Glicina 39mM, Metanol 20%, SDS 0,04%, pH 8,5) por western blot semi-seco, a 25V por 30 minutos. A membrana foi bloqueada em leite desnatado (molico) 5% em TBS-Tween (0,1% v/v Tween 20) durante 1 hora em temperatura ambiente e lavada 2 vezes em TBS-Tween para eliminar o excesso de leite. Em seguida esta foi incubada com anticorpo primário Anti-FLAG M2 em diluição de 1:500 em TBS-Tween, overnight a 4°C sob agitação rotativa e lavada 3 vezes com TBS-Tween. Por fim, foi incubada com anticorpo secundário Anti-mouse IgG (Sigma: A4416) conjugado com peroxidase diluído em TBS-Tween na razão 1:1000 por 1 hora. A atividade da peroxidase foi estimulada com solução quemiluminescente de ECL (Amersham™ ECL Western Blotting Detection Reagentes, GE Healthcare) e a revelação foi feita em filme fotográfico (Amersham Hyperfilm ™ ECL GE Healthcare 28906836) e pelo fotodocumentador (ImageQuant™ LAS 4000, GE Healthcare).

Além do western blot, testes de Dot blot foram realizados para verificar a presença da proteína de interesse na amostra de lisado celular. Para isso, 5µL de cada amostra foram aplicados diretamente à membrana de nitrocelulose umidificada previamente com Tampão de transferência de sistema semi-seco. A membrana foi incubada a 4°C overnight para o adequado secamento da amostra aplicada. Uma proteína controle, gentilmente cedida pelo Prof. Dr. Marcelo Damário Gomes, FBXO25 de ~70KDa marcada com 3xFLAG (sequência de 3 FLAGs) foi usada como controle positivo. Os procedimentos seguintes se repetem aos descritos anteriormente para Western blot.

4. RESULTADOS E DISCUSSÃO

A seção Resultados e Discussão será dividida em duas partes para facilitar a organização e compreensão. Na primeira parte serão apresentados os resultados referentes às análises de bioinformática dos dados obtidos por espectrometria de massa, com o intuito de identificar e caracterizar *in silico* as pequenas proteínas anotadas e não anotadas de *H. salinarum*. Na segunda parte será apresentada a padronização das técnicas moleculares de marcação cromossomal e superexpressão através de vetores, que serão utilizadas para a validação da expressão de pequenas proteínas.

4.1. Resultados e Discussão – Parte 1: Identificação e caracterização de pequenas proteínas

O experimento de espectrometria de massa gerou 2441 peptídeos com alto escore e estes foram selecionados quanto ao número de vezes que se alinham ao genoma; quanto à sua localização (em região com gene anotado ou região sem gene anotado, Figura 8); e quanto ao tamanho da possível proteína codificada. Esta seleção foi feita com o intuito de identificar e caracterizar *in silico* pequenas proteínas em *H. salinarum NRC-1*.

4.1.1. Análise geral dos dados de espectrometria de massa

A análise dos 2339 peptídeos que se alinham em uma única posição no genoma resultou na identificação de 169 peptídeos que se alinham em regiões sem genes anotados e 2170 que se alinham em regiões com genes anotados (Figura 8). Destes últimos, alguns se alinham a genes bem conhecidos e que são codificados por smORFs. Alguns exemplos

desses genes podem ser observados na Tabela 3.

Tabela 3: Exemplos de genes conhecidos que apresentam smORFs e possuem peptídeos alinhados identificados por LC-MS.

Gene	Função	Tamanho da smORF	Nº de peptídeos da LC-MS alinhados	Nº peptídeos do Peptide Atlas alinhados
VNG_RS10455, antiga VNG2668G	RNA polimerase subunidade H	226pb	2	4
VNG_RS03360, antiga VNG0860G	RNA polimerase subunidade L	285pb	1	5
VNG_RS06625, antiga VNG1706G	Proteína ribossomal 30S subunidade S14	159pb	1	4
VNG_RS08025 , antiga VNG2076G	Proteína ribossomal 50S subunidade L40	144pb	1	1

A identificação de smORFs de genes conhecidos com peptídeos alinhados é um importante controle positivo para o experimento, mostrando que os dados obtidos por espectrometria de massa corroboram com outros dados já conhecidos. Apesar da identificação de somente 1 ou 2 peptideos por LC-MS, também foram encontrados peptídeos do banco de dados Peptide Atlas (Van et al., 2008), fornecendo evidências experimentais adicionais para identificação dessas smORFs.

Além de peptídeos localizados em regiões de genes conhecidos, outros se alinharam a genes anotados considerados hipotéticos. De acordo com a anotação do NCBI anterior a 02 de Agosto de 2015, (quando parte deste estudo foi realizado), *H. salinarum* NRC-1 possuia 2682 genes anotados, sendo que destes 530 eram smORFs menores de 300 nucleotídeos e destes, 313 eram considerados genes hipotéticos. Tendo em vista este fato, foram identificados ao todo 75 genes hipotéticos menores de 300 nucleotídeos alinhados com 175 peptídeos, representando uma cobertura de 23,96%.

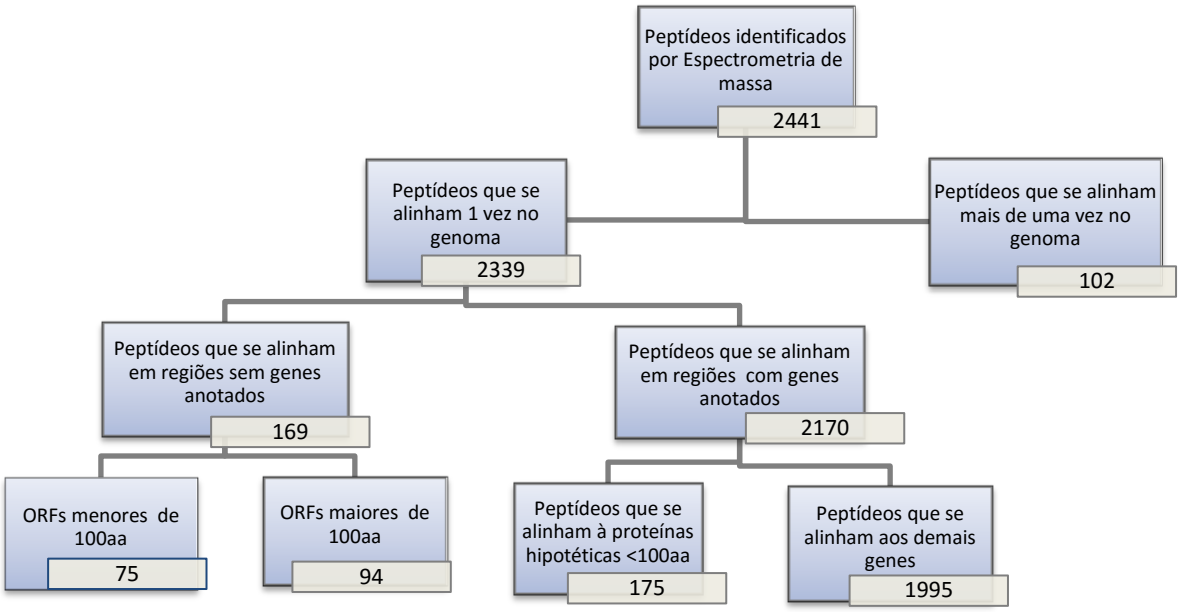

Figura 8: Fluxo das análises para a identificação das pequenas proteínas a partir de peptídeos identificados por espectrometria de massa.

Dos 75 genes hipotéticos pequenos, 39 possuem entre 2 e 7 peptídeos alinhados identificados pelo experimento de LC-MS (Tabela 4) e muitos desses genes ainda possuem alinhamento de peptídeos do banco de dados Peptide Atlas (Van et al., 2008). Esta ocorrência é sugestiva, indicando que esses genes podem ser de fato codificantes.

Tabela 4: Genes hipotéticos < 100aa com peptídeos alinhados

Número de genes	Número de peptídeos alinhados
9	2
15	3
6	4
4	5
4	6
1	7

Além disso, análises visuais no software GGB permitiram a visualização de 33 genes hipotéticos pequenos que possuem alinhamento exclusivamente com peptídeos obtidos pelo experimento de LC-MS (Figura 9 A). Porém, dos 175 peptídeos analisados que se

alinham a genes hipotéticos pequenos, 85 se alinham na mesma posição de peptídeos encontrados no *Peptide Atlas* (Figura 9 B).

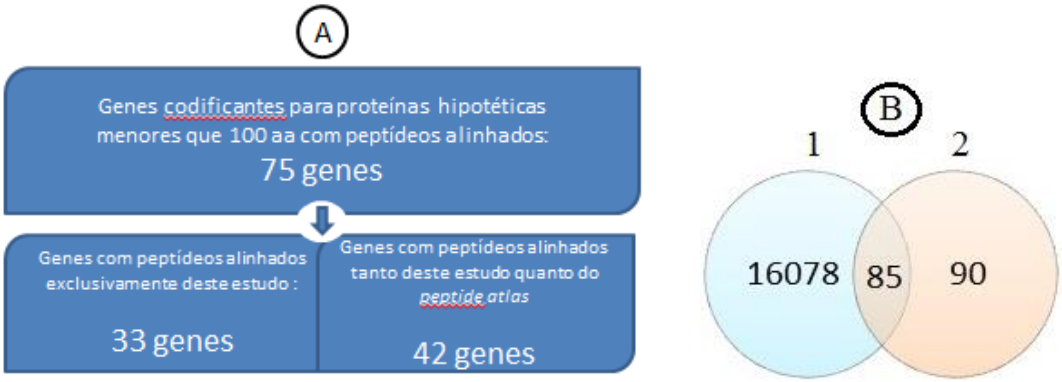

Figura 9: Comparação dos dados deste estudo com os dados do banco Peptide atlas. A: Genes hipotéticos que possuem peptídeos alinhados deste estudo (experimento de LC-MS na facility de Harvard) e do Peptide atlas. B: 1: Número de total de peptídeos do banco Peptide Atlas para *H. salinarum*. 2: Peptídeos deste estudo alinhados a genes hipotéticos <300 nucleotídeos. Na intersecção há 85 peptídeos comuns entre este estudo e o Peptide Atlas que se alinham na mesma posição em genes hipotéticos <300 nucleotídeos.

Devido à reanotação de genes realizada pelo NCBI em agosto de 2015, o número de genes hipotéticos menores de 100 aminoácidos presentes em *H. salinarum* NRC-1 passou de 313 para 227, e dos 75 genes alinhados a peptídeos neste estudo, 14 foram excluídos do banco de dados (Tabela 5), porém a maioria destes possuem diversas evidências de que são expressos como, a presença de sinal de transcrição e presença de *Transcription Start Site* (TSS), obtidos por outros experimentos realizados no Laboratório onde este estudo foi conduzido e evidências de que codificam proteínas, como a presença de peptídeos alinhados e presença de códon de início e de parada (Figura 10). Este fato mostra que anotações automáticas possuem problemas e estudos mais minuciosos precisam ser levados em consideração.

Tabela 5: Genes que foram descontinuados do banco de dados NCBI, porém há evidências de que são expressos e de que são codificantes de proteínas.

Número	Genes	Posição no genoma	Número de peptídeos de LC-MS	Presença de TSS/ nº de experimentos de dRNAseq
1	VNG0532Hd	Reverse [409836, 409991]	1	Sim/ 4
2	VNG1292H	Reverse [966430, 966543]	1	Sim/ 4
3	VNG1376H	Reverse [1026595, 1026759]	1	Sim/ 4
4	VNG1960H	Reverse [1447469, 1447639]	7	Sim/4
5	VNG2129H	Reverse [1565967, 1566074]	1	Sim/4
6	VNG0069H	Forward [1565967, 1566074]	5	Não
7	VNG0287H	Forward [230608, 230883]	1	Sim/3
8	VNG0990H	Forward [756010, 756183]	1	Não
9	VNG1423H	Forward [1060045, 1060191]	4	Não
10	VNG1591H	Forward [1189961, 1190134]	1	Sim/ 1
11	VNG2039Hd*	Forward [1505330, 1505425]	2	Sim/ 4
12	VNG2385H	Forward [1789808, 1789984]	4	Sim/ 2
13	VNG2451H	Forward [1838406, 1838513]	3	Sim/ 4
14	VNG2626H	Forward [1965732, 1965824]	1	Sim /4

* Este gene nunca foi anotado no banco de dados NCBI, porém está presente em outros bancos de dados).

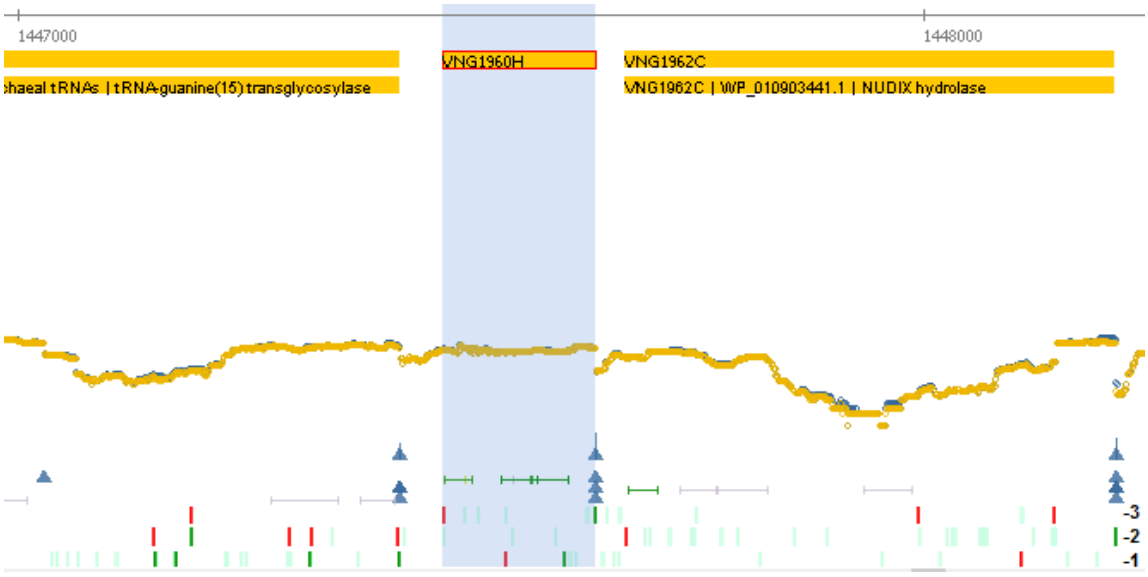

Figura 10: Vizualização no GGB de um exemplo de um gene anotado como proteína hipotética (VNG1960H) que foi descontinuado na recente reanotação realizada pelo NCBI. Porém é possível notar a presença de diversos elementos que o colocam como um gene funcional. Para esta região (hachurado em azul) há 7 peptídeos alinhados (traços horizontais verdes, alguns sobrepostos), presença de um alto sinal de transcrição (linhas laranjas), com início marcado pela presença de TSS (sítio de início de transcrição) (triângulos azuis) produzidos por quatro experimentos de dRNAseq. As barras laranjas representam os genes anotados na fita *reverse* (Superior: anotação antiga. Inferior: anotação nova). As barras verticais verdes representam os códons de início e as barras vermelhas, os códons de parada. Os frames estão indicados no canto direito inferior (-3,-2,-1).

Outro dado interessante para peptídeos alinhados a genes anotados e conhecidos, foi a presença de peptídeos em frames diferentes aos das proteínas codificadas, o que pode ser um indício de tradução alternativa, assim como relatado por Vanderperre *et al*. (2013) em um estudo com células humanas. Uma análise visual da posição de todos os peptídeos no software Gaggle Genome Browser permitiu a identificação de 9 peptídeos em frames diferente, como exemplificado na Figura 11.

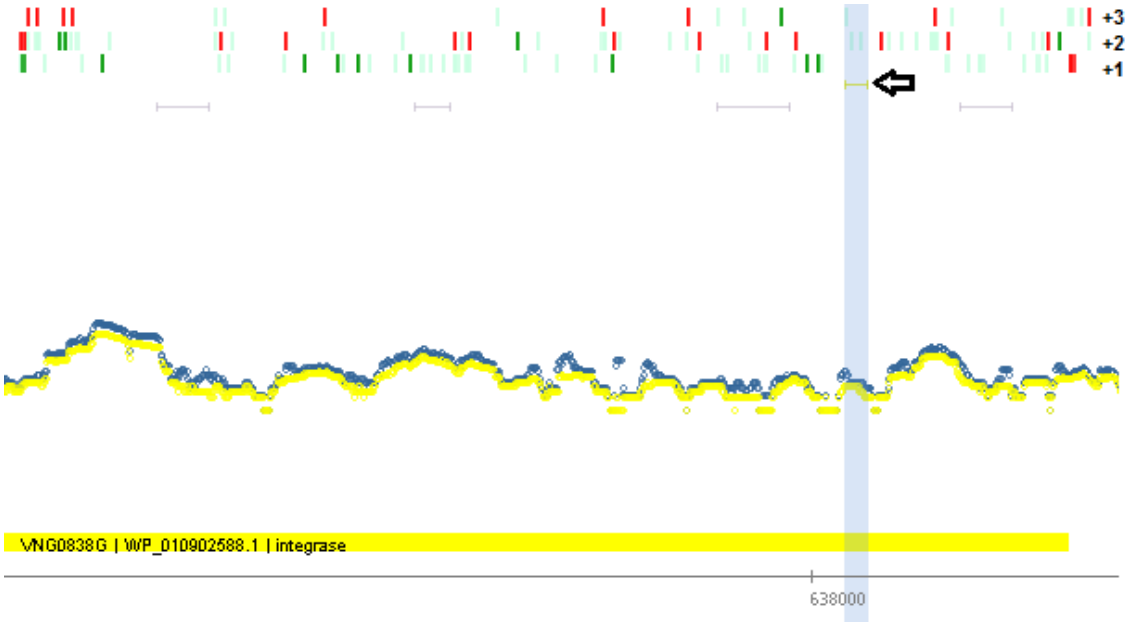

Figura 11. Visualização no GGB de um peptídeo intragênico (traço amarelo indicado pela seta e hachurado em azul), localizado na fita *forward*. Este peptídeo é codificado no *frame* +3, enquanto que o gene VNG0838G (barra amarela) codifica a proteína integrase no *frame* +1. Os traços cinza na posição superior representam peptídeos do banco de dados *Peptide atlas* e assim como o gene, estes estão no frame +1. As linhas amarelas e azuis representam sinais de transcritos obtidos por experimentos de RNAseq, indicando a ocorrência de expressão gênica no local. As pequenas barras em verdes verticais indicam a presença de códon de início para cada *frame*. As barras vermelhas indicam os códons de parada e as barras em verde escuro representam o códon de início canônico AUG.

Dos 9 peptídeos intragênicos identificados, somente 4 possuem ORFs (Tabela 6). As sequências de aminoácidos foram analisadas através do BLASTp no banco de dados NCBI e no banco de dados Uniprot.

Tabela 6: Identificação de peptídeos intragênicos

Número	Posição do peptídeo no cromossomo	Tamanho da ORF	Identidade com outra proteína
1	Forward [90120, 90233]	114	-
2	Forward [219692, 219905]	213	44,6%
3*	Reverse [505852, 506227]	450*	31,4%
4	Forward [637972, 638125]	159	47,5%

*Apesar de o foco do estudo ser com smORFs (<300 nucleotídeos), esta ORF foi considerada para análise por ser uma caso interessante de possível tradução alternativa intragênica.

A similaridade de sequência com proteínas de outros organismos para as ORFs alternativas pode ser um indicativo de que a proteína encontrada é conservada. Porém neste estudo, somente 3 ORFs alternativas apresentaram identidade com outras proteínas e essa identidade foi baixa (<50%).

O fato de essas ORFs possuírem sinais de transcritos e presença de peptídeos devem ser levados em consideração, mas evidências adicionais são necessárias para provar a sua existência, como por exemplo, dados adicionais de espectrometria de massa, dados de *Ribossomal profile*, presença de domínios conservados e validação *in vivo*. A soma de evidências de que uma proteína pode estar sendo expressa contribui significativamente para a qualidade do resultado.

4.1.2. Seleção de possíveis smORFs intergênicas e antisenso

Com o intuito de identificar possíveis pequenas proteínas, 169 peptídeos localizados em regiões sem genes anotados (região intergênica e antisenso) foram analisados quanto a presença de sinais de transcrição evidente e de códon de início e de parada no mesmo frame do peptídeo. O tamanho das smORFs foi delimitado pela presença de um códon de início canônico (metionina) *upstream* ao peptídeo e de um códon de parada *downstream* ao peptídeo. Nos casos onde não havia uma metionina (ATG) presente, foram considerados os códons de início alternativo (TTG, CTG, ATT, ATC, ATA e GTG) localizados mais próximos do peptídeo.

Essa análise resultou na seleção de 75 smORFs menores de 300 nucleotídeos, sendo que 71 foram localizadas no cromossomo, 43 na fita *forward* e 28 na fita *reverse* e 4 foram localizadas no plasmídeo pNRC200, 3 na fita *forward* e 1 na fita *reverse*. Destes, somente 5 possíveis smORFs situadas no cromossomo foram selecionadas por serem as mais prováveis a codificar uma pequena proteína, devido a evidências de expressão do gene, como a presença de sinais de transcritos fortes, presença de TSS e evidências de tradução, como a presença de um ou mais peptídeos identificado por espectrometria de massa, localizados entre um códon de início e códon de parada. Também foram calculados o ponto isoelétrico da possível proteína e a porcentagem de nucleotídeos GC da smORF. Nas tabelas 7 e 8 é possível verificar as características gerais das proteínas selecionadas. Um exemplo de proteína que se encaixa em todos os critérios pode ser observado na figura 12.

Para facilitar a abordagem, foi criada uma identificação para cada peptídeo: GPF1, GPF2, GPF3, GPF4 e GPF5. As sequências de nucleotídeos e aminoácidos de cada smORF podem ser encontradas com mais detalhes nos materiais complementares, na tabela 10.

Tabela 7: Proteínas identificadas de acordo com os critérios de seleção

Identificação da smORF	Peptídeos alinhados	Posição da smORF no cromossomo	Número de aa da proteína	%GC da ORF	PI	MM
GPF1	ADIHAELDAR GADADVSR	[1505330 – 1505425] *Forward* frame 2	31	67.70	4.65	3488.80
GPF2	ALVDTQPGLVR	[882933 – 883085] *Reverse* frame 2	50	66.01	4.12	5329.93
GPF3*	DLPIQQIREMYRQAAR	[1187834 – 1188115] *Reverse* frame 3	93	56.02	4.69	10475.52
GPF4	LTGDSAGGFEVVTTPSPVGR	[1332769 – 1333068] *Forward* frame 1	99	70.66	12.53	11433.17
GPF5	QQQQQQITVSDNSSSKPK#	[752956 – 753141] *Forward* frame 1	61	50.00	4.75	6569.22

PI: Ponto isoelétrico (teórico). MM: Massa molecular (Da) (teórica). * Proteína hipotética anotada no NCBI em agosto de 2015. GPF1 e GPF2 estão anotadas em *H. salinarum* R1, com a mesma sequência, mas com códon de início canônico (Metionina). #Peptídeo identificado duas vezes, em duas corridas de LC-MS.

Tabela 8: Possível sequência de aminoácidos das pequenas proteínas putativas.

smORF	Sequência de aminoácidos
GPF1	MTAWQTLFERGADADVSRADIHAELDARRGE
GPF2	VVLNRLRALVDTQPGLVRECRDCGTTLGEDSDDAT VCPTCGSSEIATYDL
GPF3	LGQTGSDQILLGHAHHEHLRDLPIQQIREMYRQAARRLLGQVSETEEFYRAGIVAIDVPESDPFTGDRAGDEDEIIGTKENT DESQYITKPLS
GPF4	LVASTRLTGDSAGGFEVVTTPSPVGRRCSTTRRTGCPRGRRRRVRRHRPSGCRRRFRRTRRPRSRRRRSRSRRRRASQGLP GRPGCRRPAGTRPAGCRW
GPF5	MGVVGCCGSKQQQQQQITVSDNSSSKPKRICPDCG MENPTEANHCGDCGFTFKSPEDTDDK

As smORFs GPF1 e GPF5 foram as únicas a apresentar mais de um peptídeo alinhado. Na primeira foram identificados dois peptídeos distintos, enquanto que na segunda o mesmo peptídeo foi identificado duas vezes. Este fato nos fornece uma evidência maior de que pequenas proteínas são codificadas por estas smORFs.

Também foram identificados para a maioria das possíveis proteínas pontos isoelétricos ácidos e sequência de nucleotídeos com alta percentagem de GC. Estas características corroboram com o padrão descrito para *H. salinarum,* de alto conteúdo GC e baixo pI.

Um exemplo de proteína que se encaixa em todos os critérios pode ser observado na figura 12.

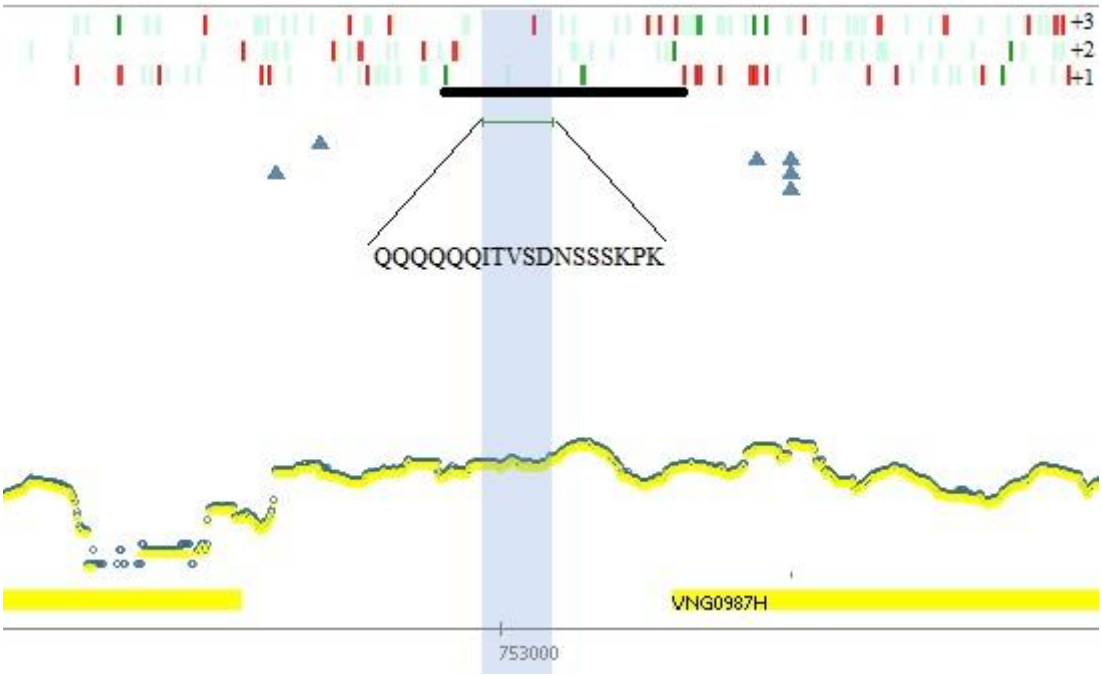

Figura 12. Visualização no GGB da smORF GPF5 do peptídeo QQQQQQITVSDNSSSKPK, (hachurado em azul), localizado na fita *forward* no *frame* +1. Este se encontra em uma região onde há sinais de RNAseq (linhas amarelas e azuis), indicando a ocorrência de expressão gênica. Apesar de se encontrar em uma região intergênica uma parte da sua possível proteína codificada no frame +1 (61aa, barra preta) sobrepõe o início do gene VNG0987H (barra amarela), com início no frame +2. Os triângulos azuis indicam os TSS. As pequenas barras em verdes indicam a presença de códon de início para cada *frame*. As barras vermelhas indicam os códons de parada e as barras em verde escuro representam o códon de início canônico ATG.

4.1.3. Análise funcional *in silico* das pequenas proteínas

As cinco smORFs localizadas em regiões sem genes anotados selecionadas pelas análises de bioinformática foram caracterizadas *in silico* a partir de ferramentas disponíveis online, com o intuito de fazer buscas por similaridade com outras espécies, identificação de domínios conservados, e predição estrutural a partir das sequências de nucleotídeos e aminoácidos. Somente os resultados com altos índices de confiança foram selecionados e podem ser visualizados na tabela 9.

Tabela 9: Caracterização *in silico* das 5 pequenas proteínas identificadas.

smORF	Software utilizado	% Identidade	E-value ou índice de confiança	Organismo
GPF1	BLASTp/BLASTx	100% de identidade com proteína não caracterizada (OE_3859B1F) de 31aa	2e-12	*H. salinarum* R1
GPF2	BLASTp/BLASTx	98% de identidade com uma proteína CPxCG relacionada a dedos de zinco.	7e-26	*H. salinarum* R1
		47% de identidade com uma proteína hipotética de 48aa	9e-06	*Halanaeroarchaeum sulfurireducens*
	Phyre2	39% de identidade com Ribonucleotídeo redutase anaeróbica	95%	-
		31% de identidade com *Metal-binding protein* relacionada à família HypA	93%	-
GPF3*	BLASTp/BLASTx	78% de identidade com uma transposase (IS4-*like*) de 345aa	4e-33	*Halorhabdus tiamatea* SARL4B
GPF4	Uniprot	50% de identidade com proteína não caracterizada rica em sequências de arginina de 241aa	1,9e-07	*Frankia* sp. BMG5.23
	Myhits	Motivo estrutural: sequências ricas em arginina	7,6e-17	-
GPF5	BLASTp/BLASTx	52% de identidade com uma proteína hipotética de 77aa	3e-08	*Halorhabdus tiamatea*
	BLASTp/BLASTx	Domínio conservado: *zinc ribbon* (folhas de zinco)	2,56e-04	-
	BLASTp/BLASTx	Dominio conservado: *Double zinc ribbon,*	9,18e-04	-
	Phyre2	35% de identidade com super família de proteínas ribossomais que se ligam a íons zinco	95%	-
	Myhits	Motivo estrutural de proteínas ribossomais 50S	0,052#	-
	Myhits	Motivo estrutural para proteínas ribosomais 30S	0,35#	-

* Proteína hipotética anotada no NCBI em agosto de 2015. #Apesar de um e-value alto, o software Myhits considera a identificação desses motivos estruturais como confiáveis.

A busca por similaridade para a smORF GPF1 encontrou uma proteína hipotética (OE_3859B1F) de 31aa em *H. salinarum* R1 com 100% de identidade, indicando que pode ter ocorrido um erro de anotação. Além disso, a smORF GPF1 possui dois peptídeos

diferentes alinhados e TSSs identificados em quatro experimento de dRNAseq, sugerindo que esta região esta sendo transcrita e provavelmente, codificando uma proteína.

A smORF GPF2 possui alta identidade (98%) com uma proteína CPxCG relacionada a dedos de zinco em *H. salinarum* R1, sugerindo mais uma vez um erro de anotação. Além disso, essa pequena proteína possui identidade de 50% com outra proteína CPxCG relacionada a dedos de zinco (OE_6112F) de 48aa em *Halobacterium salinarum* R1 localizada em um dos plasmídeos. A existência dessas sequências repetidas entre o cromossomo e o plasmídeo pode ser uma evidência da ação de elementos de transposição, que são abundantes em *H. salinarum (*Ng, et al., 2000).

A análise feita no software Phyre[2] identificou domínios de ribonucleotídeo redutase anaeróbica com identidade de 39% e índice de confiança de 95% e *metal-binding protein* relacionada à família HypA, com identidade de 31% e índice de confiança de 93% (Figura 23, material complementar) A família HypA, que em *E. coli* foi definida como uma metaloproteína, possui capacidade de ligação com os metais zinco e níquel (Atanassova & Zamble, 2005) e também possui capacidade de maturação de hidrogenases (Hube, *et al*. 2002). Além disso, foi descrito por Watanabe *et al*. (2009) a propriedade de chaperona de hidrogenases de níquel/ferro na Archaea *Thermococcus kodakaraensis* KOD1, onde HypA é responsável pela inserção do átomo de níquel na enzima hidrogenase. As proteínas do tipo HypA podem apresentar diversidade functional (Watanabe *et al*. 2009), permitindo a associação com os outros domínios encontrados, como o de ribonucleotídeo redutase anaeróbica, que é uma enzima catalizadora de deoxirribonucleotídeo a partir de ribonucleotídeos em organismo anaeróbicos facultativos. Em *E. coli*, essas enzima possui sítio de ligação de zinco (Luttringer *et al*., 2009).

Além do mais, a região da smORF GPF2 possui TSS identificado em quatro experimentos de dRNAseq e possui sinais de transcritos diferencialmente expressos ao

longo da curva de crescimento, indicando que este gene é de fato expresso em *H. salinarum* NRC-1.

Assim como apontado na tabela acima, a análise da smORF GPF3 mostrou que esta foi anotada recentemente como proteína hipotética (VNG_RS06180) em *H. salinarum* NRC-1 e possui identidade de 93% com uma proteína hipotética de *H. salinarum* R1 (OE_RS06170) de 120aa, indicando que são praticamente a mesma proteína, porém em R1 foi anotada em uma ORF com códon de início *upstream*. Além disso, as análises de BLASTp sugerem que esta proteína é uma transposase do tipo IS4, devido à alta identidade >70% e baixo e-value (4e-33) identificado para diversas proteínas de outros organismos.

Um fato curioso observado no software GGB, foi a presença de uma transposase ISH3 de tamanho maior, localizada downstream à smORF, porém em frame diferente (Figura 13), indicando que ambas podem ter sido uma única transposase, mas que durante o processo evolutivo foram separadas.

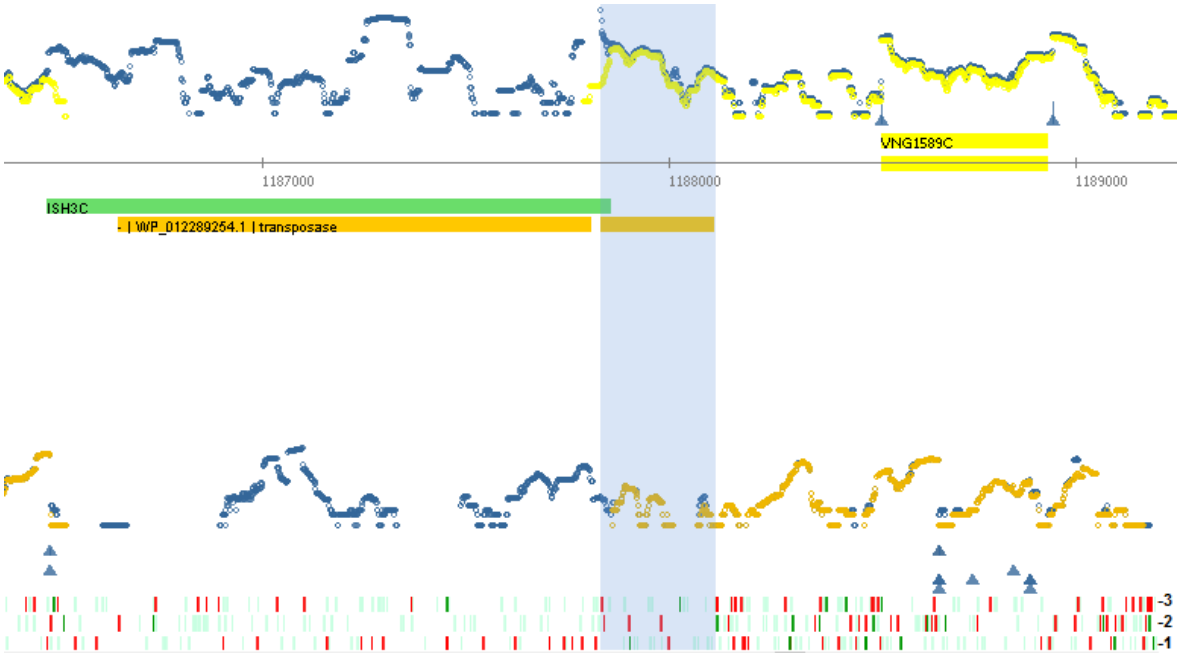

Figura 13: Visualização no GGB da smORF GPF3, (barra laranja, hachurado em azul), localizada na fita *reverse* no *frame* -3. *Downstream* a GPF3 é possível observar a presença de um gene anotado como

transposase (barra laranja) localizado no *frame* -2. A barra em verde indica a região identificada como transposase através do banco de dados ISfinder. Sinais de RNAseq (linhas amarelas e azuis), indicando a ocorrência de expressão gênica. Os triângulos azuis indicam os TSS. As pequenas barras em verdes indicam a presença de códon de início para cada *frame*. As barras vermelhas indicam os códons de parada e as barras em verde escuro representam o códon de início canônico ATG.

A smORF GPF4 apresenta similiaridade com uma proteína rica em aminoácidos arginina da bactéria *Frankia* sp. BMG5.23 e a forte evidência da presença de motivos estruturais ricos em argininas sugere uma função relacionada a presença destes aminoácidos para esta proteína. Motivos estruturais ricos em arginina podem estar relacionados à proteínas que se ligam a diversos tipos de RNA, participando do processamento destes (Bayer et a. 2005).

A smORF GPF5 apresentou similaridade com uma proteína hipotética de outro halófilo, além de alguns domínios conservados. Um deles é pertencente a uma família de proteínas não caracterizadas UPF0547, mas que possuem um motivo *zinc ribbon* (fitas de zinco). O outro domínio é uma família de *Double zinc ribbon*, que possuem um par de motivos *zinc ribbon* (Figura 24, material complementar). Além disso, o software Phyre[2] identificou com índice de confiança de 95% e identidade de 35% que essa proteína pode possuir uma região relacionada a uma super família de proteínas ribossomais que se ligam a íons zinco, além de índice de confiança >90% para outras proteínas ribossomais 40S e 50S (Figura 25, material complementar). Também foi identificado pelo software MyHits motivos estruturais de proteínas ribossomais 50S (Figura 26, material complementar).

A relação destes domínios com dedos de zinco sugerem possíveis funções para essa pequena proteína. Estes são pequenos domínios proteicos no qual o zinco auxilia na estabilização da molécula. Os dedos de zinco possuem uma variedade de estruturas e funções podendo apresentar papeis na replicação, transcrição e tradução; metabolismo; divisão celular, nos quais se ligam à moléculas como DNA, proteínas e substratos lipídicos

(Krishna et al. 2003). Dentre os subtipos de dedos de zinco, o *zinc ribbon* é um dos mais abundantes e foi o mais encontrado nas análises funcionais realizadas neste estudo. Esse motivo estrutural está presente em proteínas da maquinaria de tradução e transcrição, fatores de transcrição, RNAs polimerases, topoisomerases e proteínas ribossomais (Krishna et al., 2003). Este fato é um indicativo quanto à possível função da proteína codificada pela smORF GPF5, pois foram encontrados domínios e estruturas relacionadas à proteínas ribossomais, indicando que uma das funções desempenhadas por essa proteína pode estar relacionada à subunidades ribossômicas ou algum papel regulador em proteínas ribossômicas. Além do mais, *zinc ribbons* são compostos estruturalmente por β –hairpin e três ou mais β-fitas (Krishna et al., 2003). A análise estrutural da smORF revelou a provável formação de quatro β-fitas na proteína (Figura 14), portanto esta é mais uma evidência que suporta a presença de motivos zinc ribbons.

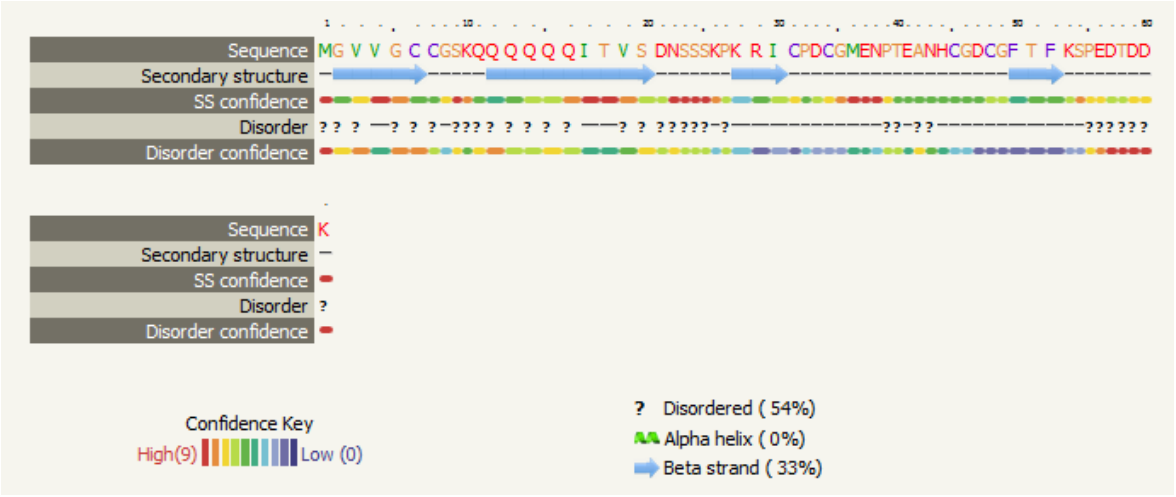

Figura 14: Software Phyre[2]: Predição de estrutura secundária a partir de sequências de aminoácidos da smORF GPF5. É possível observar a presença de quatro β-fitas (seta em azul), presente em 33% da possível proteína.

Um fato curioso encontrado foi que ambos os genes localizados *upstream* (VNG0986H) e *downstream* (VNG0987H) à smORF GPF5 são hipotéticos, porém possuem identidade com proteínas de outras arqueias com estruturas relacionadas aos dedos de zincos. Isso pode ser uma evidência de uma família de proteínas nesta região do

genoma, porém estas não possuem similaridade entre as sequências. Outro fato importante foi a identificação de dois peptídeos iguais para esta smORF, identificados nas duas corridas de LC-MS.

Análises no GGB, mostraram que a smORF GPF5 possui transcritos diferencialmente expressos ao longo da curva de crescimento de *H. salinarum* e há TSS em uma região próxima encontrado em um experimento de RNAseq, fortalecendo a evidência de que esta smORF esta sendo transcrita.

Em geral, somente um ou dois peptídeos foram encontrados para cada smORF, o que de certa forma é esperado, já que a região codificante é pequena. Esta característica também foi encontrada para as pequenas proteínas conhecidas, citadas na tabela 3. Porém, existe muito debate em relação ao quão confiável é a identificação de uma proteína baseada em somente um peptídeo ("one-hit wonder"). Alguns dos peptídeos identificados posicionados em regiões sem genes anotados não apresentam ORFs e alguns destes foram localizados até mesmo entre dois códons de parada. Estes fatos geram dúvidas na identificação destes peptídeos. Portanto, apesar de este estudo ter identificado smORFs possivelmente codificantes, futuros experimentos de validação *in vivo* devem ser realizados para a comprovação da existência da proteína.

4.2 - Resultados e discussão – Parte 2: Construções moleculares

Nesta parte serão apresentados os resultados referentes às construções moleculares realizadas a fim da padronização dos métodos *chromossomal tagging* e expressão através de vetores para o gene *lsm*, que foi escolhido devido às suas características de pequenas proteína, como baixo peso molecular e baixa abundância na célula. Estes métodos serão utilizados para validar a expressão de pequenas proteínas.

4.2.1 Padronização do *chromosomal tagging* com o gene controle lsm

A construção do fragmento referente ao segmento 500pb::FLAG::500pb, descrito na seção 3.2.5, com o gene lsm, resultou em uma sequência de 1111pb. Esta sequência é resultante do fragmento *upstream* ao códon de parada localizado após o FLAG com 569pb e do fragmento *downstream* ao códon de parada após o FLAG com 539pb, com uma região de 18 nucleotídeos de sobreposição entre ambos, necessária para a união dos fragmentos por PCR overlap. O esquema a seguir ilustra tal construção (Figura 15).

GGCGACGGCCCCGTGGTGATGCGCGGCGGCACGAACCACCCCGTGGTGGTGTCCGCGCTTCGT
GACGCCGCAGCCGACGCCGGGGCGTCGCTGCAGGTCGCCGCCCACCGGATCTCGACTGGCACG
GACGCCGACGCGTTCGGGACGGCGGGTGGCGGGGTGCCGACCGCCGACCTGGGGATCCCGAA
CCGGTATATGCATACGCCGGCGGAGGTCGTTGATCTGGCGGACCTCGCTGCGGGGGCGGACGT
GCTCGCGGCGTTCGCCGCCCACGCCGGGGACCGGGACTCGTTTGGGGTCTCGGTCTGAGCGTC
GGGCGGGAGCGGTCCGGTACGCTTATGGATGCCACCACCGGCGGTAGGTACATGAGCGGCCG
ACCACTGGATGTGCTGGAGGAGTCACTCGAAGAAACCGTCACCGTCCGCCTGAAGGACGGCG
ACGAGTTCACCGGCGTGCTGACGGGCTACGACCAGCACATGAACGTCGTCATTGAGGGCGAAG
ACACAACGATTATCCGTGGCGATAACGTCGTCACCATCAAACCAGACTACAAA**GACGATGAC**
GACAAGTAACTGGCGCAGGAACCCCGAGCCAGGGGAAGAAGAACACCACGACGCACACGAA
GTGCCGACGCTGCGGTGAGAAGTCCTACCACACGAAGAAGAAGGTCTGCAGTTCCTGCGGCTT
CGGTGCTTCCGCCAAGCGCCGGGATTACGAGTGGCAGGGCAAGACCGGCGACAACTAAGCAG
TTCTTTCTCCGCGTTTCGTTCCGGTCCGGACAGCCACTGGTCCGTGATCGCGCCCGGTTCGCCG
AGTCGTCCGTCTCTACCCCCGGCCATGGGTGCTGAACACACCCCTCACTGCAGTGTTGTGCATA
CTTCACGGCTATCGGGGCGTGTATTCGACACGGTTGCGGAAGGTTTACCCTATGGTGTCCGCAA
GCGGGAGGTATGTCAGGCGGCCCCGCGGACGGACCCATGAGTGACCATCCAACCGAGAAATG
TGGGGTCGTCGGTGCCTCACTGTCGGCCCGTGACGCCGCGCTTCCGACGTACTACGCGCTGTAC
GCCCTCCAGCACCGCGGCCAGGAGTCGGCGGGCATCGTCGCCCA

Figura 15: Representação da sequência do fragmento 500pb::FLAG::500pb , com 1111pb, na qual a sequência em azul representa o gene *lsm* e a sequência em vermelho representa o FLAG. A região em negrito e sublinhada é referente à sobreposição entre os dois fragmentos.

Após a inserção do plasmídeo pHsal-S contendo o inserto 500pb::FLAG::500pb no citoplasma da arqueia, esta foi plaqueada em meio CM com mevinolina e em seguida em meio CM com 5-FOA, como descrito na seção 3.2.7. Após ocorrida a recombinação seguida pela expulsão do plasmídeo da célula, 10 colônias foram selecionadas para a confirmação por PCR, utilizando-se os *primers* descritos na seção 3.2.2. Como um dos *primers* se liga ao genoma do organismo em uma região externa ao fragmento *upstream* e o outro na sequência correspondente ao FLAG, o produto da amplificação por PCR gerou um fragmento de 673pb (Figura 16). Ao todo, 7 colônias foram diagnosticadas positivas

para a presença do gene *lsm* marcado com o FLAG no cromossomo da arqueia, totalizando 70% de recombinação (Figura 16). Além da confirmação da recombinação por PCR, a sequência esperada foi confirmada por sequenciamento.

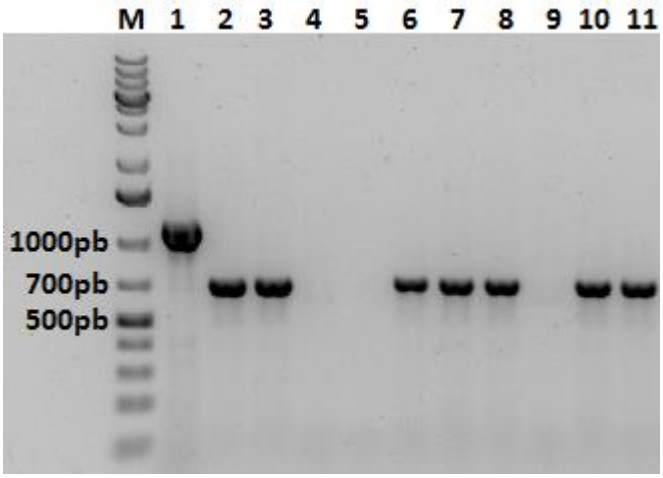

Figura 16. Eletroforese de DNA em gel de agarose 1% para o diagnóstico das colônias positivas para a recombinação cromossômica. A reação de PCR foi realizada com um *primer* que se liga no genoma do organismo, em uma região externa ao fragmento *usptream* ao gene *lsm* (*Forward*) e o outro *primer* que se liga na sequência correspondente ao FLAG (*Reverse*). O produto de amplificação resultante corresponde a um fragmento de 673pb. Para a avaliação, 10 colônias foram selecionadas aleatoriamente e o produto da reação de PCR aplicado nas canaletas 2-11. A presença de bandas indica que a recombinação aconteceu (canaletas 2,3,6,7,8,9,10,11), totalizando 70% de colônias positivas. 1 = Amostra não relacionada ao projeto. M = Marcador de DNA: Gene Ruler™ 1Kb plus DNA Ladder 0.5µg/µl (Fermentas).

4.2.2 Detecção da proteína marcada com FLAG (*chromosomal tagging*)

Após o cultivo das colônias recombinantes, o extrato proteico destas foi imunoprecipitado e analisado por Western blot para a detecção da proteína de interesse marcada com o peptídeo FLAG. Outra proteína, FBXO25, marcada com 3xFLAG foi usada como controle positivo.

A revelação do western blot resultou em bandas no tamanho de ~25kDa e ~50kDa para as quatro amostras testadas (Figura 17). A proteína controle de peso molecular 70kDa, apesar de ter aparecido, formou um arraste na região superior da membrana. Porém nenhuma banda foi revelada na região dos 7kDa, peso molecular da proteína expressa pelo

gene lsm (Figura 17).

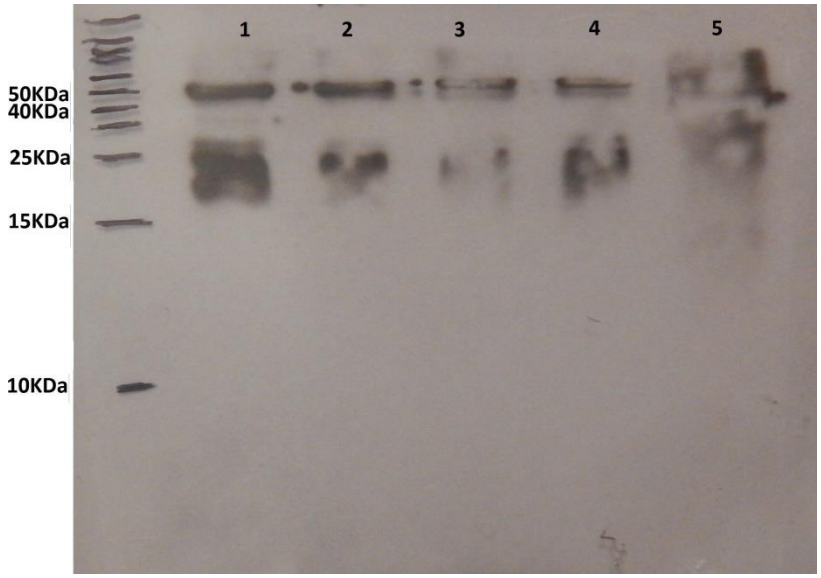

Figura 17. Revelação de Western blot: extrato celular de culturas recombinantes para o gene LSM marcado com FLAG. As amostras de 1 a 4 foram imunoprecipitadas e destas somente a 1 e 2 são de recombinantes do gene lsm. As amostras 3 e 4 não são relacionadas ao projeto. Amostra 5 corresponde ao controle positivo (70KDa). Nenhuma amostra gerou bandas na região dos 7kDa, tamanho da proteína de interesse. Marcador de proteínas: Spectra multicolor (Fermentas: #SM 1841).

Devido ao não aparecimento da banda correspondente ao tamanho da proteína de interesse, um teste de Dotblot foi realizado para detectar a presença da proteína marcada na amostra. Para isso, amostras de quatro culturas da linhagem recombinantes de *H. salinarum* NRC-1 foram analisadas, junto com um controle negativo (purificado proteico do mesmo organismo, porém sem marcação com FLAG) e um controle positivo (mesma proteína utilizada como controle positivo no Western blot, Figura 17). Mais uma vez o teste foi negativo para todas as amostras, exceto para o controle positivo (Figura 18).

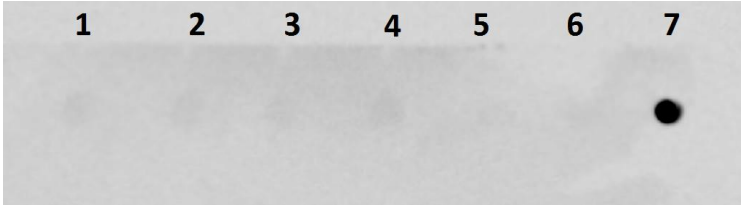

Figura 18: Dotblot para checar a presença da proteína marcada com FLAG expressa pelos organismos

recombinantes. Foram aplicadas diretamente em membrana de nitrocelulose 4 amostras de culturas recombinadas (1-4), 1 controle negativo (6) e um controle positivo (7). A amostra 5 não faz parte do projeto. A revelação mostrou que somente o controle positivo emitiu luz, evidenciando a presença do FLAG. Imagem obtida pelo fotodocumentador ImageQuant™ LAS 4000 (GE Healthcare).

Para a primeira análise, acredita-se que as bandas de 25KDa e 50KDa sejam resultantes das cadeias leves e pesadas (Sigma, Catalogo Número: A2220) do anti-FLAG M2 desnaturado durante o processo de imunoprecipitação, já que este anticorpo foi usado nesta técnica, atuando como molécula intermediária entre a ligação das proteínas marcadas com o FLAG e o anticorpo Anti-mouse IgG acoplado às *Beads*. Além disso, a etapa de eluição da amostra proteica durante a imunoprecipitação pode ser um dos fatores responsáveis pela desnaturação do anticorpo, devido a alta temperatura do processo (65°C por 10 minutos). Outro problema do western blot foi a proteína controle usada, pois devido ao grande tamanho ela não correu propriamente em um gel concentrado para proteínas menores, dessa forma fica difícil a comparação com proteínas de tamanho molecular reduzido.

Já a ausência da proteína marcada com FLAG em ambas as análises pode estar relacionada a alguns fatores como, a provável baixa taxa de expressão do gene lsm, que em *Haloferax volcanii*, haloarchaea próxima filogeneticamente de *H. salinarum*, foi constatada uma expressão de somente 4000 proteínas LSm por célula, contrastando com mais de 50.000 proteínas Hfq (equivalente a LSm para bactéria) expressas por *E. coli* (Fischer *et al*., 2010), implicando em um baixo número destas para a detecção por Western blot. Além disso, a fase de crescimento em que o organismo se encontrava durante a extração proteica ($D.O.600_{nm}$ = 0.5) pode não coincidir com a fase de maior produção desta proteína pela célula. O tamanho pequeno da proteína também pode ser um fator na dificuldade de detecção, já que esta poderia atravessar a membrana de nitrocelulose e ou se soltar facilmente durante as lavagens com TBS-Tween.

Além dos argumentos citados acima, a técnica de validação utilizada, o *chromosomal tagging*, pode ser um fator limitante em alguns casos, pois como *H. salinarum* possui um genoma muito compacto, qualquer modificação pode alterar a função de outros genes, pois é comum observar estes numa fita anti-senso a outro gene (Figura 19) ou sobrepondo-se em *frames* diferentes ou até mesmo na mesma *frame* (Fischer *et al*., 2010).

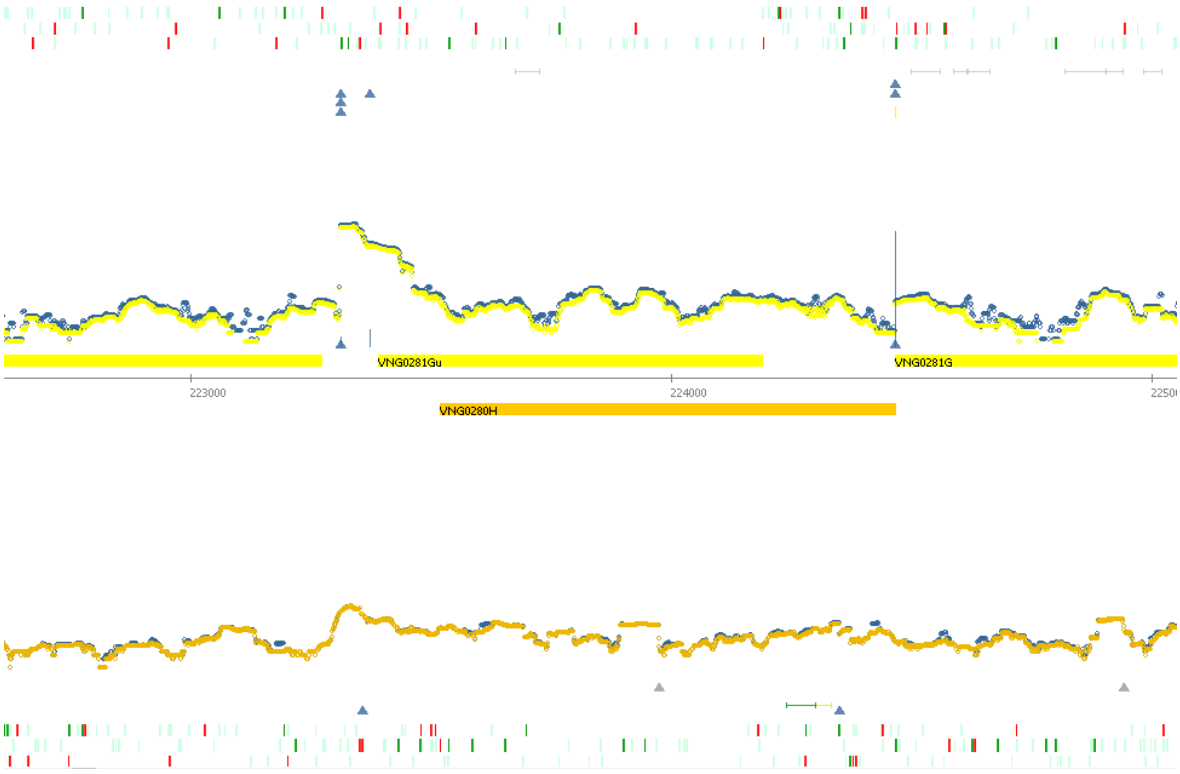

Figura 19. Visualização no Gaggle Genome Browser: gene na fita *forward* (barra amarela) codificante da proteína Luciferase (VNG0281Gu) antisenso a um gene hipotético (VNG0280H) na fita *reverse* (barra laranja). A validação de uma proteína antisenso por chromossomal tagging é problemática, pois a inserção de um marcador para um gene afetaria o outro na fita oposta. As pequenas barras verticais em verdes indicam a presença de códon de início para cada *frame*. As barras vermelhas indicam os códons de parada e as barras em verde escuro representam o códon de início canônico AUG. As linhas amarelas (superior) e laranjas (inferior) são dados de diversos experimentos de RNAseq.

Um dos maiores problemas dos peptídeos identificados nas análises de bioinformática é o fato de muitos se localizarem em sobreposição com outros genes tanto na mesma fita quanto na fita anti-sense. Portanto, uso da expressão de proteínas através de

vetores seria uma alternativa para o *cromossomal tagging* para esses casos.

4.2.3 Construções de vetores para a expressão do gene controle lsm

Com a finalidade de amenizar os danos causados em outros genes pela técnica da recombinação cromossômica, assim como o problema da baixa expressão do gene de interesse, a construção de vetores foi uma alternativa encontrada para suprimir esses fatores, pois suas características permitem tanto a superexpressão do gene lsm a partir de um promotor bastante ativo, como o da ferredoxina, assim como a expressão do gene com seu promotor nativo (localizado ~200pb *upstream* ao códon de início de tradução do gene), visando uma análise mais próxima aos níveis de expressão fisiológicos. A escolha dos marcadores (tags) foi feita pensando na maior eficiência de reconhecimento pelo anticorpo e na contribuição para o problema de tamanho enfrentado em relação às pequenas proteínas. Portanto dois marcadores foram utilizados, o 3xFLAG que além de ser melhor em relação ao reconhecimento pelo anticorpo, devido ao maior número de FLAGs, é 3 vezes mais pesado que o FLAG utilizado para o *cromossomal tagging*, possuindo ~2.5KDa. O outro marcador é o Cmyc, de tamanho maior e peso molecular de cerca de 19KDa. Um esquema das quatro construções pode ser observado na Figura 20.

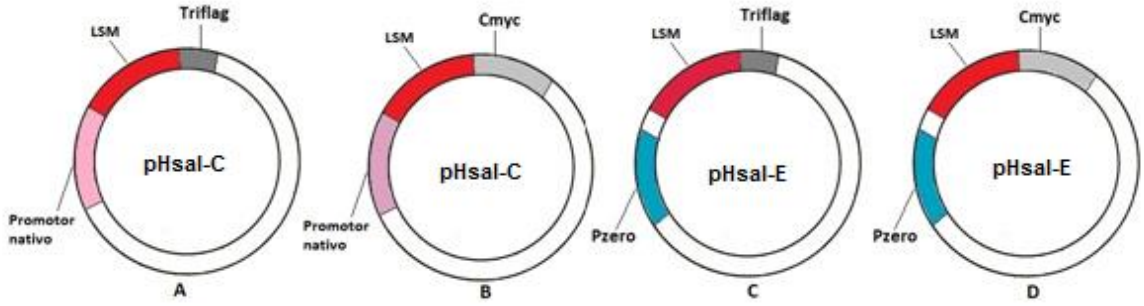

Figura 20. Esquema das quatro construções nos vetores pHsal-C e pHsal-E. A: Vetor pHsal-C com o inserto Promotor nativo + LSm + 3xFLAG. B: Vetor pHsal-C com o inserto Promotor nativo + LSM + Cmyc. C: Vetor pHsal-E com o promotor Pzero e o inserto LSM + 3xFLAG. D: Vetor pHsal-E com o promotor Pzero e o inserto LSM + Cmyc. Pzero é o nome dado ao promotor da ferredoxina modificado e já presente no vetor

previamente às contruções.

A inserção de cada fragmento (gene lsm com a região promotora, gene lsm sem a região promotora, 3xFLAG, Cmyc) nos vetores foi feita separadamente através de digestão enzimática e ligação do inserto no vetor. Cada inserção foi confirmada separadamente através de reações de PCR e digestão enzimática. Das quatro construções realizadas, somente uma não funcionou, a pHsal-E+Lsm+Cmyc, pois não foi possível inserir o gene lsm neste vetor. O resultado final das outras 3 contruções pode ser observado na Figura 19, que mostra as bandas do produto de PCR amplificado através dos oligos M13 do vetor. O tamanho das bandas é correspondente ao tamanho dos insertos + tamanho do sítio de recombinação do vetor + tamanho do promotor da ferredoxina (no caso do vetor pHsal-E): pHsal-C+LSm+3xFLAG = 599pb; pHsal-C+LSm+Cmyc = 1085pb; pHsal-E+LSm+3xFLAG=556pb (Figura 21).

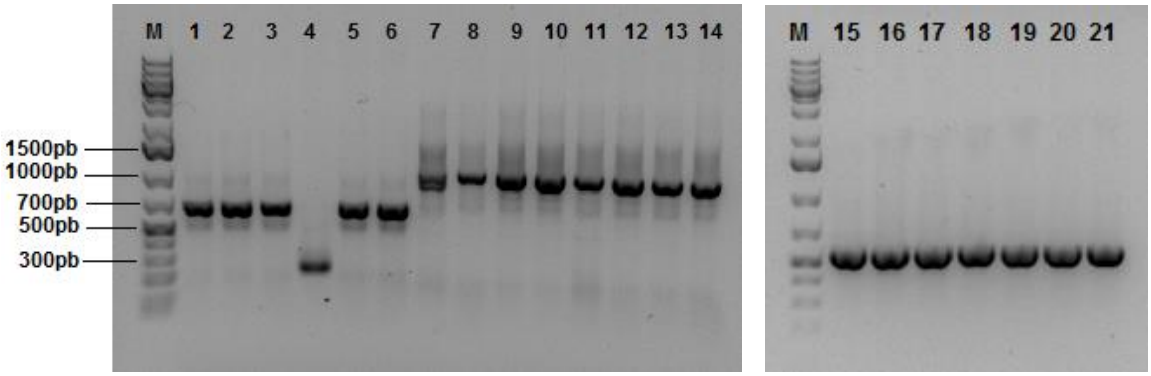

Figura 21: Eletroforese em gel de agarose 1%: As bandas correspondem aos produtos da reação de PCR de colônia para a confirmação da presença dos vetores construídos. 1-6:pHsal-C+LSm+3xFlag, fragmento de tamanho 599pb. 7-14: pHsal-C+LSm+Cmyc, fragmento de tamanho: 1085pb. 15-21: pHsal-E+LSm+3xFlag, fragmento de tamanho: 556pb. Somente a colônia número 4 não possui o vetor correto. M = Marcador de DNA: Gene Ruler™ 1Kb plus DNA Ladder 0.5µg/µl (Fermentas).

A comparação entre os vetores em todas as etapas das construções pode ser observada na

Figura 22.

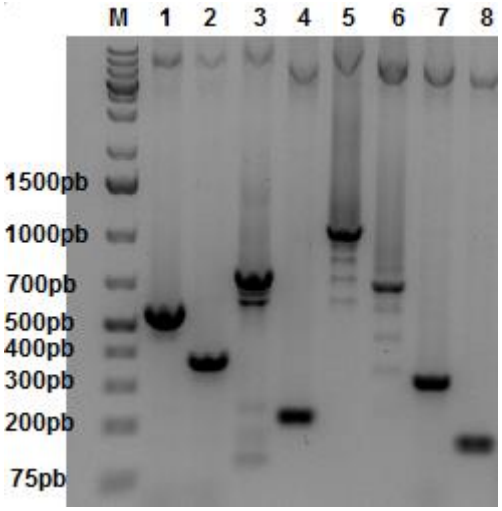

Figura 22. Eletroforese em gel de agarose 1%. Comparação dos vetores de todas as etapas das construções. Amplificação com os oligos M13 do vetor e oligo do promotor da ferredoxina (para o vetor pHsal-E). Vetor pHsal-C (vazio) (8): tamanho de 145pb. Vetor pHsal-C+Cmyc (6): tamanho de 694pb. Vetor pHsal-C+3xFLAG (4): tamanho de 208pb. Vetor pHsal-C+LSm+Cmyc (5): tamanho de 1085. Vetor pHsal-C+LSm+3xFLAG (3): tamanho de 599pb. Vetor pHsal-E (vazio) (7): tamanho de 277pb. Vetor pHsal-E+3xFLAG (2): tamanho de 385pb. Vetor pHsal-E+LSm+3xFLAG (1): 556. M = Marcador de DNA: Gene Ruler™ 1Kb plus DNA Ladder 0.5µg/µl (Fermentas).

Após finalizar as construções, os vetores foram inseridos em *H. salinarum* por transformação química e estas foram cultivadas até atingirem a D.O = 0,5. As proteínas foram extraídas para a realização de análises de Western Blot e Dot blot. Mais uma vez a proteína de interesse, LSm, marcada tanto com 3xFlag quanto com Cmyc não foi identificada em nenhuma das análises.

Como já discutido anteriormente, problemas em relação a baixos níveis de expressão da proteína LSm e à fase de crescimento na qual foram extraídas as proteínas, são fatores limitantes para essas análises. Infelizmente, não houve tempo para a repetição dos experimentos e tentativas de melhorar as análises. Portanto outras tentativas devem ser realizadas, como a extração de proteínas em diferentes pontos da curva de crescimento e o enriquecimento das amostras através de técnicas da Imunoprecipitação.

A padronização destes métodos são importante para que as análises futuras de validação de expressão das pequenas proteínas identificadas neste estudo sejam realizadas.

5. CONSIDERAÇÕES FINAIS

O estudo de pequenas proteínas tem se demonstrado desafiador tanto na validação da expressão, quanto na identificação e estabelecimento de homologia e função.

De um total de 169 smORFs intergênicas ou antisenso com peptídeos alinhados, somente cinco foram selecionadas por possuir características consideráveis para expressão e codificação. Este baixo número pode estar relacionado ao fato de muitas regiões do genoma com potencial codificante serem consideradas pouco conservadas (ORFans – Orphan ORFs, regiões que não apresentam homologias com outros genomas) (Shmuely et al., 2004). Em geral genomas recém anotados apresentam até 30% de ORFs que não possuem similaridade com genes codificantes de outros organismos (Fischer and Eisenberg, 1999). Shmuely et al. (2004), encontraram em *H. salinarum* NRC-1 evidências de expressão em diversas ORFans maiores de 222 nucleotídeos com parálogos no genoma e sem homologia encontrada em banco de dados. Estas ORFans identificadas podem ser interpretadas como regiões que expressam ncRNA ou genes que de fato codificam proteínas, mas não apresentam similaridade com proteínas encontradas em outros organismos.

Ao longo dos anos, com a incorporação de novas sequências aos bancos de dados, sequências similares podem ser encontradas, facilitando reanotações do genoma (Pfeiffer & Oesterhelt, 2015). Portanto, o fato de as smORFs analisadas neste estudo não terem sido anotadas até o momento, pode ser uma interpretação de que houveram erros de anotação ou estes são genes que codificam proteínas pouco conservadas, exclusivas ou quase exclusivas para o organismo de estudo.

Outro fato interessante para este trabalho foi a relação entre transcritos e peptídeo alinhados. Estes transcritos podem ser resultado da chamada transcrição pervasiva, que abrange quase todo o genoma em algum momento da vida de um organismo. Estes transcritos nem sempre estão relacionados a funções na célula, porém muitos podem desempenhar papéis regulatórios (Wade & Grainger, 2014). Neste estudo foi identificada a presença de transcritos para todos os 2441 peptídeos detectados por LC-MS. Esta pode ser uma evidência de que algumas regiões do genoma que expressam RNAs considerados não codificantes, podem ser de fato codificar peptídeos.

As funções de peptídeos e pequenas proteínas na célula são ainda pouco conhecidas, porém, como citado neste trabalho, muitas dessas proteínas possuem funções redundantes com outras proteínas e estão relacionados principalmente à regulações finas do metabolismo e maquinaria genética (Storz et al., 2014). Proteínas com múltiplas isoformas e funções redundantes são conhecidas em *H. salinarum* (Ng et al., 2000; Leigh et al., 2011), e as pequenas proteínas identificadas neste estudo podem ser mais uma destas, resultantes de duplicações de genes ou inserções causadas por elementos de transposição.

A origem das smORFs codificantes de pequenas proteína é ainda incerta, podendo corresponder a novas proteínas com novas funções resultantes de um recente e rápido processo evolutivo ou a partir proteínas antigas cujas funções ainda não foram identificadas (Shmuely et al., 2004; Storz et al., 2014).

Portanto, o foco em estudos de pequenas proteínas ainda tem muito a nos revelar e futuros estudos relacionados à validação da expressão e ensaios fenotípicos *in vivo*, têm grande potencial na descoberta de novas características.

6. CONCLUSÃO

Os problemas referentes a anotação de genomas são evidentes e alguns destes foram demonstrados neste estudo, como a presença de genes hipotéticos com indícios de tradução que foram descontinuados nos bancos de dados e de genes anotados em *H. salinarum* R1 que possuem a mesma ORF em *H. salinarum* NRC-1, porém não estão anotados neste organismo ou estão anotados com códon de início diferentes.

Tendo em vista esta problemática, os resultados obtidos por este estudo contribuíram para o melhoramento na anotação do genoma de *H. salinarum* NRC-1 através da análise de dados obtidos por LC-MS integrados com dados de transcriptoma obtidos por outros estudos no nosso laboratório, contribuindo para o aumento de evidências que provam a existência de 75 genes hipotéticos, os quais possuem peptídeos detectados. Também foram demonstradas evidências para a existência de 14 genes hipotéticos descontinuados no banco de dados NCBI.

Além disso, este trabalho contribuiu para a identificação de 5 possíveis novos genes codificantes de pequenas proteínas localizados em regiões intergênicas ou antisenso, assim como a sugestão de possíveis funções para algumas deste pequenas proteínas.

As proteínas codificadas pela smORF GPF2 podem estar relacionados à funções de chaperonas de hidrogenases ou metalproteínas que se ligam a metais zinco e níquel. Já a smORF GPF3 pode estar relacionada à codificação de transposases. Também foram encontradas evidências na smORF GPF4 relacionadas a motivos estruturais ricos em arginina, cuja função está relacionada à regulação de RNAs. E por último, a smORF GPF5 apresenta evidências que relacionam sua função à proteínas com dedos de zinco, que podem desempenhar diferentes papéis como ligantes de moléculas de ácidos nucleicos ou outras proteínas, assim como possível relação com proteínas ribossomais.

Neste estudo também contribuímos para o aprimoramento de técnicas de biologia molecular importantes para a validação da expressão de pequenas proteínas, que serão

utilizadas em validações futuras das possíveis pequenas proteínas identificadas neste trabalho.

7. REFERÊNCIAS BIBLIOGRÁFICAS

Aderem, A. (2005). Systems Biology: Its Practice and Challenges. **Cell** 121: 511–513.

Allers, T., and Moshe, M. (2005) 'Archaeal Genetics — The Third Way'. **Nat Rev Genet** 6.1. 58-73.

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

Andrews, S. & Rothnagel, J. (2014). Emerging evidence for functional peptides encoded by short open reading frames**. Nature reviews | Genetics**. v.15. 193 -204.

Atanassova A, Zamble DB. (2005) Escherichia coli HypA is a zinc metalloprotein with a weak affinity for nickel. **J Bacteriol**.;187:4689–4697.

Baliga NS, Bjork SJ, Bonneau Ret al. (2004) Systems level insights into the stress response to UV radiation in the halophilic archaeonHalobacteriumNRC-1.**Genome Res**14: 1025 1035.

Bare, J.C., T. Koide, D.J. Reiss, D. Tenenbaum, and N.S. Baliga. (2010). Integration and visualization of systems biology data in context of the genome. **BMC Bioinformatics**. 11:382.

Bayer, T.S., Booth, L. N., Knudsen, S. M. and Ellington, A. D. (2005). Arginine-rich motifs present multiple interfaces for specific binding by RNA. **RNA**, 11:1848–1857.

Bonneau R, Baliga NS, Deutsch EW, Shannon P, Hood, L. (2004): Comprehensive de novo structure prediction in a systems-biology context for the archaea Halobacterium sp. NRC-1. **Genome Biol**, 5(8):R52

Bonneau R, Facciotti M.T., Reiss D.J., Schmid A.K., Pan M., et al. (2007) A predictive model for transcriptional control of physiology in a free living cell. **Cell** 131: 1354–1365.

Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. **Anal Biochem**.;72:248-54.

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*(7), 740.

http://doi.org/10.15252/msb.20145160

Brown, J. & Doolittle. W. (1997). Archaea and the Prokaryote-to-Eukaryote Transition. **Microbiology and Molecular Biology Reviews**. v. 61, 4. 456–502.

Brugger, K. et al.(2002). Mobile elements in archaeal genomes, **FEMS Microbiol**. Lett. 206, 131–141.

Cavicchioli, R. (2011). Archaea - timeline of the third domain**. Nature reviews**, v. 9. 51-61.

Cheng, H.; Chan, W.; Li, Z.; Wang, D.; Liu, S.; Zhou, Y. (2011). Small Open Reading Frames: Current Prediction Techniques and Future Propects. **Current Protein and Peptide Science**. 12, 503-507.

DasSarma, S., Berquist, B., Coker, J., DasSarma, P., Müller, J. (2006) Post-genomics of the model haloarchaeon Halobacterium sp. NRC-1. **Saline Systems**. Disponível em: http://www.salinesystems.org/content/2/1/3.

DasSarma, S.;Karan, R.; DasSarma, P.; Barnes, S.; Ekulona, F.; Smith, B. (2013). An improved genetic system for bioengineering buoyant gas vesicle nanoparticles from Haloarchaea. **BMC Biotechnology**. 13:112.

Davis. R. H. (2004) The age of model organism. **Nature Reviews/ Genetics**. Vol. 5. 69-77.

Delcher, A.L.; Bratke, K.A; Powers, E.C. and Salzberg, S.L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer, **Bioinformatics** 23:6, 673-679.

Dyall-Smith. M. (2009). Protocols for halobacterial genetics © Ver 7.2, Disponível online em: web: http://www.haloarchaea.com.

Fischer, S., Benz, J., Späth, B., Maier, L. K., Straub, J., Granzow, M., … Marchfelder, A. (2010). The archaeal lsm protein binds to small RNAs**. Journal of Biological Chemistry**, 285(45), 34429–34438. http://doi.org/10.1074/jbc.M110.118950

Fischer, D. and Eisenberg, D. (1999). Finding families for genomic ORFans. **Bioinformatics**,15, 759–762.

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

Goo, Y., Yi, E., Baliga, N., Tao, W., Pan, M., Aebersold, R., Goodlett, D., Hood, L. Ng, W. (2003). Proteomic Analysis of an Extreme Halophilic Archaeon, Halobacterium sp.NRC-1. **Molecular & Cellular Proteomics** 2.8.

Heckman L. H. & Pease, L. R. (2007). Gene splicing and mutagenesis by PCR-driven overlap extension. **Nature Protocols** 2, 924 – 932.

Hobbs, E.; Fontaine, F.; Yin, X.; Storz, G. (2011) An expanding universe of small proteins. **Current Opinion in Microbiology**.14: 167–173.

Hood, L. (2003). Systems biology: integrating technology, biology, and computation. **Mechanisms of Ageing and Development** 124, 9-6.

Horton, R. M.; Hunt, H. D.; Ho, S. N.; Pullen, J. K.; Pease, L. R. (1989). Engineering hybrid genes without the use of restriction enzymes: gene splicing by overlap extension. **Gene,** 77. 61-68.

Hube M, Blokesch M, Böck A (2002) Network of Hydrogenase Maturation in Escherichia coli: Role of Accessory Proteins HypA and HybF. **Journal of Bacteriology** 184:3879–3885

Humbard, M.; Miranda, H.; Lim, J.; Krause, D.; Pritz, J. ; Zhou, G.; Chen, S.;Wells, L.; Maupin-Furlow. J. (2010). Ubiquitin-like Small Archaeal Modifier Proteins (SAMPs) in *Haloferax volcanii.* **Nature**: 463(7277): *54–60.*

Ideker, T.; Galitski, T.; Hood, L. (2011). A New approach to Decoding Life: Systems Biology. **Annual Review. Genomics Hum. Genet**. 2:343–72

Ingolia. N. T. (2014). Ribosome profiling: new views of translation, from single codons to genome scale. **Nature Reviews | Genetics**. Vol. 15, 205-213.

Jenney, F.; Tachdjian, S.; Chou, C.; Kelly, R.; Adams, M. Functional Genomics. In: Cavicchioli, R. **Archaea: Molecular and Cellular Biology**. American Society for Microbiology. ASM Press. Washington, DC. 2007. p. 434-462

Kastenmayer, J.; Ni, Li.; Chu, A.; Kitchen, L.; Au, W.; Yang, H.; Carter, C.; Wheeler, D.; Davis, R.; Boeke, J.; Snyder, M.; Basrai, M. (2006). Functional genomics of genes with small open reading frames (sORFs) in *S. cerevisiae*. **Genome Research**. 16: 365-373.

Kaur, A., Pan, M., Meislin, M., Facciotti, M. T., El-Gewely, R., & Baliga, N. S. (2006). A systems view of haloarchaeal strategies to withstand stress from transition metals. **Genome Research**, *16*(7), 841–854. http://doi.org/10.1101/gr.5189606

Kaur, A., Van, P. T., Busch, C. R., Robinson, C. K., Pan, M., Pang, W. L., … Baliga, N. S. (2010). Coordination of frontline defense mechanisms under severe oxidative stress. **Molecular Systems Biology**, *6*, 393. http://doi.org/10.1038/msb.2010.50

Kelley, L.A.; Mezulis, S.; Yates, C. M.; Wass, M. N. & Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. **Nature protocols**, Vol.10,N.6 ,845-858.

Kennedy, S. P., Ng, W. V., Salzberg, S. L., Hood, L., & DasSarma, S. (2001). Understanding the Adaptation of Halobacterium Species NRC-1 to Its Extreme Environment through Computational Analysis of Its Genome Sequence. **Genome Research**, 11(10), 1641–1650.

Klein, C., Garcia-Rizo, C., Bisle, B., Scheffer, B., Zischka, H., Pfeiffer, F., Siedler, F. and Oesterhelt, D. (2005). The membrane proteome of *Halobacterium salinarum*. **Proteomics**, 5: 180–197. doi: 10.1002/pmic.200400943

Klein, C.; Aivaliotis, M.; Olsen, J.; Falb, M.; Besir, H.; Scheffer, B.; Bisle, B.; Tebbe, A.; Konstantinidis, K.; Siedler, F.; Pfeiffer, F.; Mann, M.; Oesterhelt, D. (2007). The Low Molecular Weight Proteome of *Halobacterium salinarum.* **Journal of Proteome Research**, 6. 1510-1518.

Kletzin, A. General Characteristics and important Model Organisms. In: Cavicchioli, R. **Archaea: Molecular and Cellular Biology**. American Society for Microbiology. ASM Press. Washington, DC. (2007). p. 14-92.

Krishna, S. S., Majumdar, I., & Grishin, N. V. (2003). SURVEY AND SUMMARY: Structural classification of zinc fingers. **Nucleic Acids Research**, *31*(2), 532–550.

Koide, T.; Reiss, D.; Bare, J.; Pang, W.; Facciotti, M.; Schmid, A.; Pan, M.; Marzolf, B.; Van, P.; Lo, F.; Pratap, A.; Deutsch, E.; Peterso, A.; Martin, D.; Baliga, N. (2009). Prevalence of transcription promoters within archaeal operons and coding sequences. **Molecular Systems Biology** 5.

Leigh, J. A.; Albers, S.; Atomi, H.; Allers, T. (2011). Model organisms for genetics in the domain Archaea:methanogens, halophiles,Thermococcales and Sulfolobales. **FEMS Microbiology Review**. 35. 577–608.

Levesque, M. P. & Benfey, P. N. (2004). Systems biology. **Current Biology**. V. 14. 5. p. R179–R180.

Ma, J.; Ward, C., Jungreis, I., Slavoff, S., Schwaid, A., Neveu, J., Budnik, B., Kellis, M., Saghatelian, A. 2014. Discovery of Human sORF-Encoded Polypeptides (SEPs) in Cell Lines and Tissue. **Journal of Proteome Research**. 13, 1757−1765.

Luttringer F, Mulliez E, Dublet B, Lemaire D, Fontecave M (2009) The Zn center of the anaerobic

ribonucleotide reductase from *E. coli*. **J Biol Inorg Chem** 14: 923–933

Ma, J., Ward, C. C., Jungreis, I., Slavoff, S. A., Schwaid, A. G., Neveu, J., … Saghatelian, A. (2014). Discovery of Human sORF-Encoded Polypeptides (SEPs) in Cell Lines and Tissue. **Journal of Proteome Research**, *13*(3), 1757–1765. http://doi.org/10.1021/pr401280w

Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Bryant SH. (2015). CDD: NCBI's conserved domain database. **Nucleic Acids Res.** 28;43 (Database issue):D222-2. doi: 10.1093/nar/gku1221.

McCready, S., Marcello, L. (2003).: Repair of UV damage in *Halobacterium salinarum*. **Biochem Soc Trans**; **31**(Pt 3): 694–698.

McCready, S. (1996). The repair of ultraviolet light-induced DNA damage in the halophilic archaebacteria,*Halobacterium cutirubrum*, *Halobacterium halobium* and *Haloferax volcanii*. **Mutat Res**;364:25–32.

Mormile, M. R., Biesen, M. A., Gutierrez, M. C., Ventosa, A., Pavlovich, J. B., Onstott, T. C. and Fredrickson, J. K. (2003), Isolation of *Halobacterium salinarum* retrieved directly from halite brine inclusions. **Environmental Microbiology**, 5: 1094–1102

Ng, W. V.; Kennedy, S. P.; Mahairas, G. G. et al. (2000). Genome sequence of Halobacterium species NRC-1. **Proceedings of the National Academy of Sciences of the United States of America** 97(22): 12176-12181.

Pagni, M., Ioannidis, V., Cerutti, L., Zahn-Zabal, M., Jongeneel, C. V., Hau, J., … Falquet, L. (2007). MyHits: improvements to an interactive resource for analyzing protein sequences. *Nucleic Acids Research*, *35*(Web Server issue), W433–W437. http://doi.org/10.1093/nar/gkm352

Pfeiffer F., Schuster S.C., Broicher A., Falb M., Palm P., Rodewald K., Ruepp A., Soppa J., Tittor J., Oesterhelt D. (2008) Evolution in the laboratory: The genome of *Halobacterium salinarum* strain R1 compared to that of strain NRC-1. **Genomics**.;91:335–346.

Pfeiffer, F., & Oesterhelt, D. (2015). A Manual Curation Strategy to Improve Genome Annotation: Application to a Set of Haloarchael Genomes. **Life**, *5*(2), 1427–1444. http://doi.org/10.3390/life5021427

Prasse, D., Thomsen, J., De Santis, R., Muntel, J., Becher, D., Schmitz, R. (2015). First description of small proteins encoded by spRNAs in Methanosarcina mazei strain Gö1. **Biochimie**. http://dx.doi.org/10.1016/j.biochi.2015.04.007

RajBhandary, U. L. (2000). More surprises in translation: Initiation without the initiator tRNA. P**NAS** vol. 97. no. 4. 1325–1327.

Samayoa, J., Yildiz, F., Karplus, K. (2011). Identification of prokaryotic small proteins using a comparative genomic approach. **Bioinformatics**. v. 27, 13. 1765–1771.

Sartorius-Neef, S. and Pfeifer, F. (2004), *In vivo* studies on putative Shine–Dalgarno sequences of the halophilic archaeon *Halobacterium salinarum*. **Molecular Microbiology**, 51: 579–588. doi:10.1046/j.1365-2958.2003.03858.x

Silva-Rocha, R.; Pontelli, M. C; Furtado, G. P.; Zaramela, L. S.; Koide, T. (2015). Development of new modular genetic tools for engineering the Halophilic Archaeon *Halobacterium salinarum.* **PLoS ONE** 10(6): e0129215

Sharma, C. M, Hoffmann, S.; Darfeuille , F., Reignier , J.; Findeiß, S.; Sittka, A.; Chabas , S.; Kristin Reiche, C.; Hackermüller, J.; Reinhardt, R;. Stadler, P. F. & Voge, J. (2010). The primary transcriptome of the major human pathogen *Helicobacter pylori*. **Nature** 464, 250-255

Shi, L. & Schröder, W. (2004). The low molecular mass subunits of the photosynthetic supracomplex, photosystem II. **Biochimica et Biophysica Acta** 1608. 75 – 96.

Sigma-Aldrich, ANTI-FLAG M2 Affinity Gel. Technical Bulletin. Catalog Number A2220. Disponível em https://www.sigm. aaldrich.com/content/dam/sigma-aldrich/docs/Sigma/Bulletin/a2220bul.pdf

Slavoff, S. A., Mitchell, A. J., Schwaid, A. G., Cabili, M. N., Ma, J., Levin, J. Z., … Saghatelian, A. (2013). Peptidomic discovery of short open reading frame-encoded peptides in human cells. **Nature Chemical Biology**, *9*(1), 59–64. http://doi.org/10.1038/nchembio.1120

Shmuely H, Dinitz E, Dahan I, Eichler J, Fischer D, Shaanan B. (2004). Poorly conserved ORFs in the genome of the archaea Halobacterium sp. NRC-1 correspond to expressed proteins. **Bioinformatics**. , 20:1248-1253.

Soppa, J. (2005). From replication to cultivation: hot news from Haloarchaea. **Current Opinion in Microbiology**,8:737–744

Srinivasan, G., Krebs, M. P. and RajBhandary, U. L. (2006), Translation initiation with GUC codon in the archaeon *Halobacterium salinarum*: implications for translation of leaderless mRNA and strict correlation between translation initiation and presence of mRNA. **Molecular Microbiology**, 59: 1013–1024. doi:10.1111/j.1365-2958.2005.04992.x

Storz, G., Wolf, Y., Ramamurthi, K. (2014). Small Proteins Can No Longer Be Ignored. **Annual Review of Biochemistry**. 83. 753–77

Tarasov, V. Y.; Besir, H.; Schwaiger, R.; Klee, K.; Furtwängler, K.; Pfeiffer, F.; Oesterhelt, D. (2008). A small protein from the bop–brp intergenic region of *Halobacterium salinarum* contains a zinc finger motif and regulates bop and crt B1 transcription**. Molecular Microbiology**. 67(4).772–780.

Tebbe, A., Klein, C., Bisle, B., Siedler, F., Scheffer, B., Garcia-Rizo, C., Wolfertz, J., Hickmann, V., Pfeiffer, F. and Oesterhelt, D. (2005), Analysis of the cytosolic proteome of *Halobacterium salinarum* and its implication for genome annotation. **Proteomics**, 5: 168–179. doi: 10.1002/pmic.200400910

The UniProt Consortium. (2008). The Universal Protein Resource (UniProt). **Nucleic Acids Research**, *36*(Database issue), D190–D195. http://doi.org/10.1093/nar/gkm895

Van, P. T., Schmid, A. K., King, N. L., Kaur, A., Pan, M., Whitehead, K., Baliga, N. S. (2008). *Halobacterium salinarum NRC-1* PeptideAtlas: strategies for targeted proteomics. **Journal of Proteome Research**, *7*(9), 3755–3764.

Vanderperre, B., Lucier, J.-F., Bissonnette, C., Motard, J., Tremblay, G., Vanderperre, S., … Roucou, X. (2013). Direct Detection of Alternative Open Reading Frames Translation Products in Human Significantly Expands the Proteome. **PLoS ONE**, *8*(8), e70698. http://doi.org/10.1371/journal.pone.0070698

Vasconcelos, A. T. R. & Almeida, D. F.: Bioinformática na análise de genes e genomas. In:Biologia Molecular Básica. 4.ed. Artmed, 2012. p. 364 – 381.

Vogel, J., & Luisi, B. F. (2011). Hfq and its constellation of RNA. **Nature Reviews/Microbiology**, 9(8), 578–589. http://doi.org/10.1038/nrmicro2615

Wade, J.T., Grainger, D.C. (2014). Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. **Nat Rev Microbiol**; 12(9):647-653.

Wan, X. F.,Bridges, S. M., Boyle, J. A. (2004). Revealing gene transcription and translation initiation patterns in archaea, using an interactive clustering model. **Extremophiles** 8 (4), 291-299

Warren, A. S.; Archuleta, J.; Feng, W.; Setuba, J. C. (2010). Missing genes in the annotation of prokaryotic genomes. **BMC Bioinformatics**, 2010,11:131.

Washietl, S.; Findeiß, S.; Müller, S. A.; Kalkhof, S.; Bergen, M. V.; Hofacker, I.; Stadler, P. F.; Goldman, N. (2011). RNA code: Robust discrimination of coding and noncoding regions in comparative sequence data. **RNA**. 17. 578–594.

Watanabe, S., Arai, T., Matsumi, R., Atomi, H., Imanaka, T. and Miki, K. (2009). Crystal Structure of HypA, a Nickel-Binding Metallochaperone for [NiFe] Hydrogenase Maturation**. J. Mol. Biol.** 394, 448–459.

Wilbanks, E. G.; Larsen, D. J.; Neches, R. Y. et al. (2012). A workflow for genome-wide mapping of archaeal transcription factors with ChIP-seq. **Nucleic Acids Research** 40(10).

Wilusz, C. J., & Wilusz, J. (2005). Eukaryotic Lsm proteins: lessons from bacteria. **Nature Structural & Molecular Biology**, 12(12), 1031–1036. http://doi.org/10.1038/nsmb1037

Woese, C.R., O. Kandler, and M.L. Wheelis. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. **Proc. Natl. Acad. Sci.** U. S. A.87:4576–9.

Zaramela LS, Vêncio RZN, ten-Caten F, Baliga NS, Koide T (2014) Transcription Start Site Associated RNAs (TSSaRNAs) Are Ubiquitous in All Domains of Life. **PLoS ONE** 9(9): e107680.

8. MATERIAL COMPLEMENTAR

Tabela 10: Sequências de aminoácidos e nucleotídeos das smORFs GPF1, GPF2, GPF3, GPF4 e GPF5.

smORF GPF1 **Peptídeos identificados**: GADADVSR e ADIHAELDAR **Sequência de aminoácidos**: MTAWQTLFERGADADVSRADIHAELDARRGE **Sequência de nucleotídeos**:ATGACAGCCTGGCAGACCCTCTTCGAGCGCGGCGCC GACGCCGACGTATCGCGCGCTGACATCCACGCGGAACTCGACGCACGCCGTGG CGAGTAA
smORF GPF2 **Peptídeo identificado**: ALVDTQPGLVR **Sequência de aminoácidos**: VVLNRLRALVDTQPGLVRECRDCGTTLGEDSDDAT VCPTCGSSEIATYDL **Sequência de nucleotídeos**: GTGGTACTGAACCGTCTCCGAGCGCTCGTCGACAC CCAGCCCGGACTCGTGCGCGAATGCCGGGACTGCGGGACAACCCTGGGTGAGG ACAGCGACGACGCGACAGTGTGTCCGACCTGTGGCTCCAGCGAGATCGCGACC TACGACCTCTGA
SmORF GPF3 **Peptídeo identificado:** DLPIQQIREMYRQAAR **Sequência de aminoácidos**: LGQTGSDQILLGHAHHEHLRDLPIQQIREMYRQAAR RLLGQVSETEEFYRAGIVAIDVPESDPFTGDRAGDEDEIIGTKENTDESQYITKPLS **Sequência de nucleotídeos**: TTGGGGCAAACTGGGTCAGATCAAATACTACTCG GCCACGCGCATCACGAACATCTCCGCGACCTCCCCATCCAGCAGATACGCGAG ATGTACCGACAGGCCGCACGTCGACTCTTAGGTCAGGTTTCGGAGACTGAGGA GTTCTACCGGGCTGGCATCGTTGCTATCGACGTTCCCGAGTCCGACCCGTTCAC TGGCGATCGAGCGGGCGATGAAGACGAGATTATTGGGACCAAAGAGAACACC GACGAGTCTCAGTACATCACAAAGCCGCTTAGTTAG
smORF GFP4 **Peptídeo identificado**: LTGDSAGGFEVVTTPSPVGR **Sequência de aminoácidos**: LVASTRLTGDSAGGFEVVTTPSPVGRRCSTTRRTG CPRGRRRRVRRHRPSGCRRRFRRTRRPRSRRRRSRSRRRRASQGLPGRPGCRRPAG TRPAGCRW **Sequência de nucleotídeos:** CTGGTGGCTTCGACACGGCTGACCGGCGACAGCG CCGGTGGCTTCGAGGTCGTCACGACCCCCAGCCCGGTCGGGCGGCGGTGCTCG ACTACTCGGCGAACGGGCTGTCCTCGGGGTCGACGTCGTCGGGTTCGTCGACA

CCGCCCTTCTGGATGCCGTCGTCGGTTTCGACGTACACGTCGTCCTCGAAGCCG GCGGCGGCGTTCTCGAAGTCGGCGTCGGCGAGCTTCTCAGGGGTTGCCGGGGC GTCCGGGATGCCGTCGGCCGGCTGGAACGCGCCCAGCGGGTTGTCGATGGTGA
smORF GPF5 **Peptídeo identificado:** QQQQQQITVSDNSSSKPK **Sequência de aminoácidos**: MGVVGCCGSKQQQQQQITVSDNSSSKPKRICPDCG MENPTEANHCGDCGFTFKSPEDTDDK **Sequência de nucleotídeos**: ATGGGGGTAGTTGGCTGTTGCGGGTCGAAACAGC AACAGCAGCAACAGATTACGGTCTCCGACAACAGTAGTTCGAAGCCGAAGCG GATTTGCCCAGATTGTGGAATGGAGAATCCGACGGAAGCGAATCACTGCGGAG ACTGCGGTTTCACTTTTAAATCTCCTGAGGATACAGATGACAAATAG

8.1 Análises *in silico* para a caracterização de possíveis pequenas proteínas

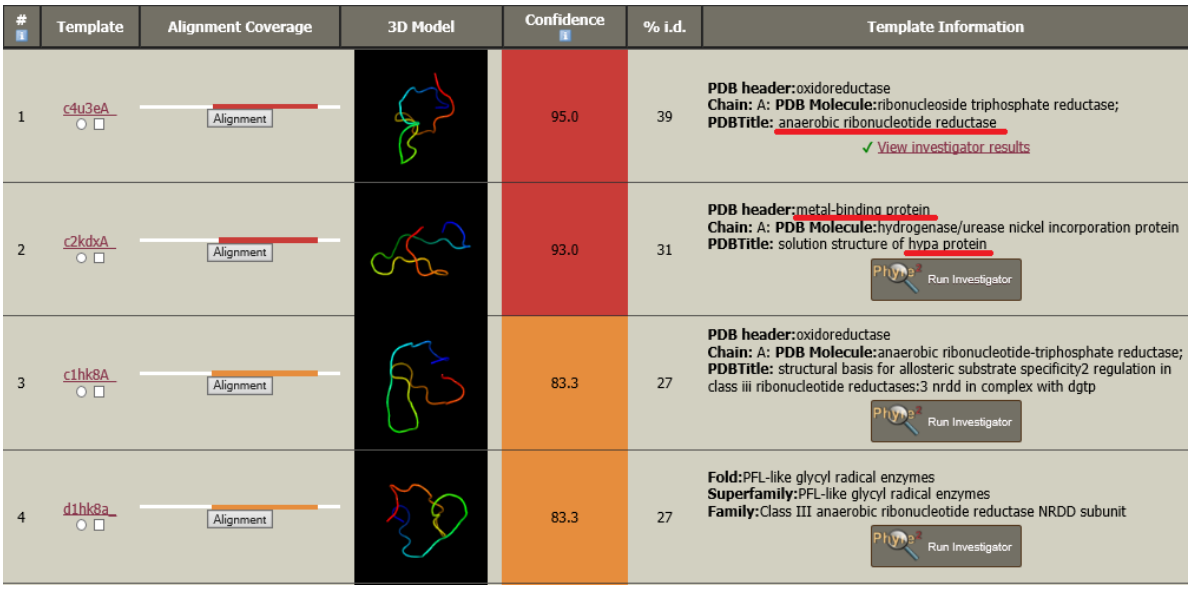

Figura 23: Software Phyre[2]: Predição de estrutura terciária e identificação homologia e domínios conservados para a possível proteína da smORF GPF2. Nota-se a identificação de domínios relacionados a ribonucleotídeo redutase anaeróbica e metaloproteína (Metal biding protein) relacionada à proteína HypA.

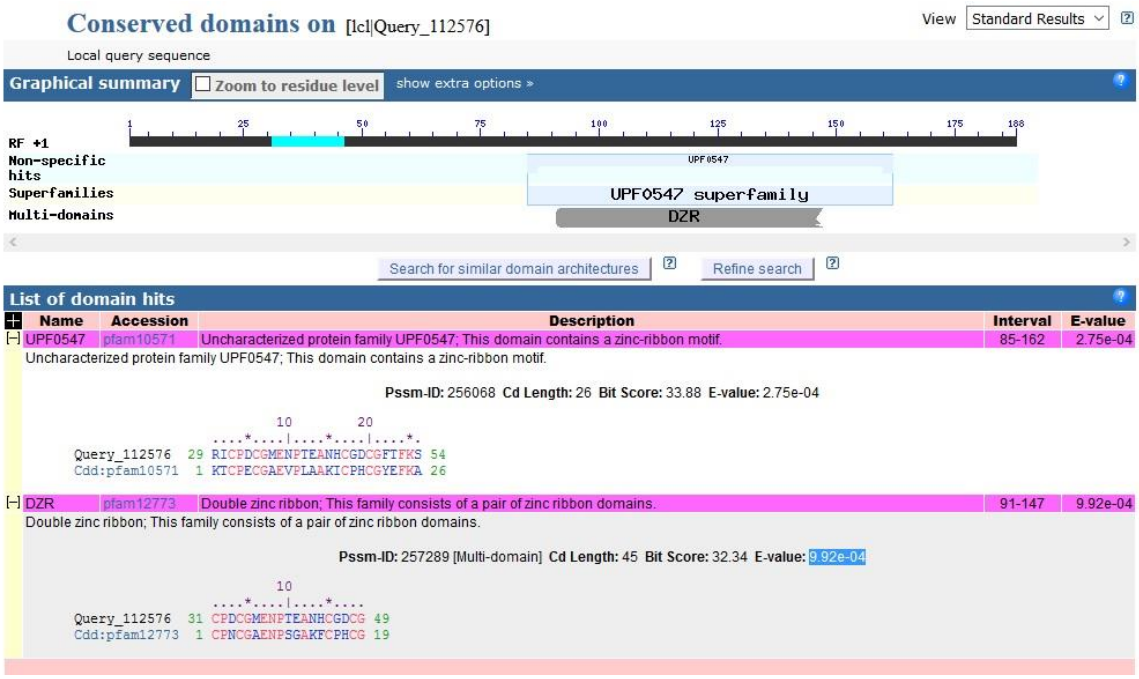

Figura 24: BLASTx no NCBI: Identificação de domínios conservados para a sequência de nucleotídeos da smORF GPF5.

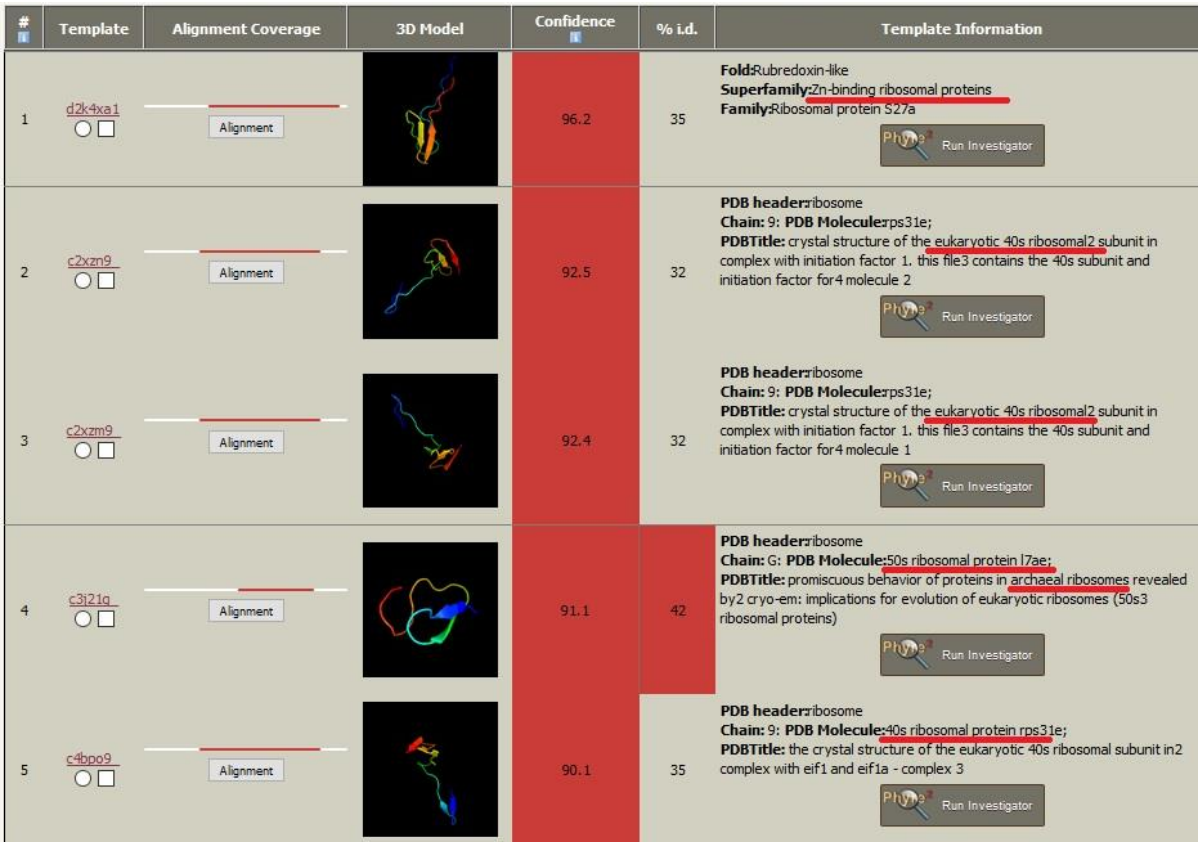

Figura 25: Software Phyre[2]: Predição de estrutura terciária e identificação homologia e domínios conservados para a possível proteína da smORF GPF5. Observa-se a identificação de domínios

relacionados a proteínas ribossomais e proteínas ligantes de zinco.

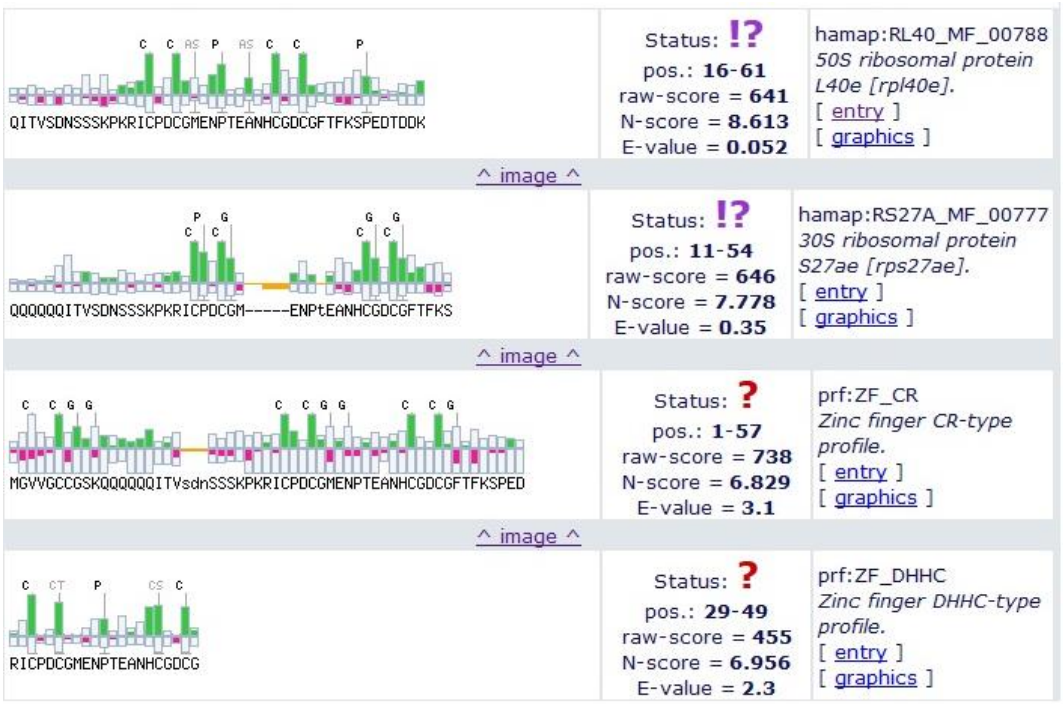

Figura 26: Ferramenta Motif scan: Myhits: Identificação de motivos estruturais para a possível proteína da smORF GPF5. Os símbolos "!?" significam que este resultado é muito improvável que seja um falso positivo, porém a determinação de sua família requer evidências biológicas adicionais; Já o símbolo "?" significa que o pareamento entre as sequências é fraco ou questionável.