UNIVERSIDADE FEDERAL DO PAMPA

LUCAS FERREIRA MACIEL

MONTAGEM E ANOTAÇÃO FUNCIONAL DO TRANSCRIPTOMA DE *Prasiola crispa* E BIOPROSPECÇÃO DE PROTEÍNAS DE LIGAÇÃO AO GELO

São Gabriel
2017

LUCAS FERREIRA MACIEL

MONTAGEM E ANOTAÇÃO FUNCIONAL DO TRANSCRIPTOMA DE *Prasiola crispa* E BIOPROSPECÇÃO DE PROTEÍNAS DE LIGAÇÃO AO GELO

Trabalho de Conclusão de Curso apresentado ao Curso de Biotecnologia da Universidade Federal do Pampa, como requisito parcial para obtenção do Título de Bacharel em Biotecnologia.

Orientador: Prof. Dr. Paulo Marcos Pinto

Coorientador: MSc. Evelise Leis de Carvalho

São Gabriel
2017

M152m

Maciel, Lucas Ferreira

MONTAGEM E ANOTAÇÃO FUNCIONAL DO TRANSCRIPTOMA DE Prasiola crispa E BIOPROSPECÇÃO DE PROTEÍNAS DE LIGAÇÃO AO GELO / Lucas Ferreira Maciel.

77 p.

Trabalho de Conclusão de Curso (Graduação) -- Universidade Federal do Pampa, BIOTECNOLOGIA, 2017.

"Orientação: Paulo Marcos Pinto".

1. Sequenciamento de RNA. 2. Treboxiophyceae. 3. Bioinformática. 4. Transcriptômica.

LUCAS FERREIRA MACIEL

MONTAGEM E ANOTAÇÃO FUNCIONAL DO TRANSCRIPTOMA DE *Prasiola crispa* E BIOPROSPECÇÃO DE PROTEÍNAS DE LIGAÇÃO AO GELO

Trabalho de Conclusão de Curso apresentado ao Curso de Biotecnologia da Universidade Federal do Pampa, como requisito parcial para obtenção do Título de Bacharel em Biotecnologia.

Trabalho de Conclusão de Curso defendido e aprovado em: 10 de Outubro de 2017.

Banca examinadora:

_____

Prof. Dr. Paulo Marcos Pinto
Orientador
UNIPAMPA

_____

Prof. Dr. Filipe de Carvalho Victoria
UNIPAMPA

_____

Prof. Dr. Juliano Tomazzoni Boldo
UNIPAMPA

Dedico este trabalho a meu Pai, Avós e Namorada.

AGRADECIMENTO

Agradeço ao meu Pai, por ser o meu maior exemplo de vida, por estar presente em cada conquista da minha e ter me apoiado e confiado desde o primeiro instante nesta ideia de morar a 2.000km de casa, em uma cidade e universidade que nem mesmo conhecíamos. Obrigado pela oportunidade e por se esforçar tanto para que minha única preocupação seja o estudo.

Às minhas Avós Ana e Terezinha, por terem me criado e educado com tanto amor e carinho, vocês foram uma das maiores motivações de todo meu esforço ao longo da graduação. Tinham o sonho de ver eu me formar, por isso lhes prometi que sempre daria o meu máximo, e hoje olho para trás feliz, sabendo que cumpri a minha promessa. Junto com meu Avô Maciel, fico triste que hoje não possam estar aqui para me abraçarem.

A minha Namorada e Amiga Renata, por todo amor, carinho e compreensão, principalmente durante o desenvolvimento deste trabalho, no qual tem participação direta. Obrigado por aguentar as minhas crises existenciais de formando/desempregado e tornar os meus dias tão leves e felizes.

Aos meus Irmãos Luís Fernando e Thiago, por compreenderem a minha ausência, e que tanto sentem a minha falta. A minha Mãe e demais familiares também por todo apoio quando necessário.

Ao Prof. Paulo, por me acolher e me orientar em uma relação saudável e produtiva para a elaboração deste trabalho e por todo o conhecimento passado ao longo de toda a graduação, até mesmo quando a disciplina foi deixada em segundo plano para conversas sobre cidadania.

Ao Prof. Chariston, que por mais de três anos foi o meu orientador, e sempre confiou em meu trabalho, me aconselhou como se fosse o seu próprio filho e me compreendeu no momento em que acreditei que era necessária a mudança.

A todos os Professores, que forneceram toda a base para o meu crescimento e conhecimento, em especial os Profs. Andrés Cañedo, Célia Carlini, Filipe Victória, Guido Lenz, José Chies e Juliano Boldo, meu muito obrigado.

Aos meus Amigos, de São Gabriel e Porto Alegre, que tornaram esta jornada mais leve e divertida: Bárbara, Davi, Deca, Filipe, Hudson, Marcos, Nicoláz, Paulo Sérgio, Rudimar, Simone e Tainah.

E por fim a Unipampa, pois tudo isso só foi possível devido a sua existência. Eu não poderia ter escolhido uma faculdade melhor.

“NÃO ENTRE EM PÂNICO”.

O Guia de um mochileiro das Galáxias

RESUMO

Prasiola crispa é uma macroalga verde taloide, pertencente à classe *Treboxiophyceae*, de distribuição biogeográfica cosmopolita, estando presente do Ártico ao continente Antártico. Na Antártica, *P. crispa* atua como importante produtor primário, resistindo a condições ambientais extremas como baixas temperaturas, estresse osmótico, radiação ultravioleta, entre outros. Devido a sua capacidade de colonizar um ambiente tão inóspito, esta alga deve possuir mecanismos adaptativos, cujos genes envolvidos são de grande interesse na área da Biotecnologia. Entre as principais biomoléculas de interesse, as proteínas de ligação ao gelo (IBPs) destacam-se. IBPs são polipeptídeos que permitem a sobrevivência de células a baixas temperaturas. Este grupo de proteínas possui grande aplicabilidade na agricultura, biomedicina e indústria alimentícia. Assim, o objetivo deste trabalho foi sequenciar, montar e anotar funcionalmente o transcriptoma de *P. crispa* e identificar produtos gênicos que estão diretamente relacionados à capacidade de sobrevivência deste organismo no continente Antártico, principalmente IBPs. A amostra foi coletada na Ilha do Rei George, Antártica. O RNA total foi extraído, a biblioteca montada e sequenciada por Illumina HiSeq 2000, com estratégia *paired-end reads*. *Reads* de baixa qualidade e adaptadores foram removidos e as *reads* processadas submetidas a montagem pelo montador Trinity. Os transcritos montados foram anotados pelo servidor *web-based* MG-RAST em pipeline automatizado e indicaram a contaminação com a microbiota. O transcriptoma de *P. crispa* foi então descontaminado e posteriormente anotado funcionalmente através da ferramenta Blast2GO. IBPs de *P. crispa* e microbiota foram buscadas por algoritmo BLASTX, alinhando os transcritos contra as sequências de IBPs em um banco de dados local. Transcritos identificados através da busca tiveram a sequência aminoacídica predita utilizando o Trandescoder e modeladas computacionalmente através do servidor Phyre2. O sequenciamento gerou 31.563.740 *reads* de boa qualidade. A montagem gerou 376.797 *contigs*, totalizando 209.870.963 pares de bases (pb) e tamanho médio de 557 pb. Em anotação realizada pelo *web-server* MG-RAST, 93% dos *hits* foram caracterizados como pertencentes às bactérias. Após processo de descontaminação, 17.201 *contigs* foram mantidos como o transcriptoma de *P. crispa*, sendo que 52,19% destes foram identificados através do processo de anotação funcional. Da busca com o BLASTX, identificou-se 127 transcritos como possíveis IBPs, com homologia a proteínas de plantas, bactérias e fungos. A modelagem computacional da estrutura tridimensional indicou que 17 destas possíveis IBPs possuem estrutura semelhante a outras proteínas do grupo, sendo que 8 delas possuem maior potencial. Em conclusão, a montagem e

anotação funcional foram realizadas de maneira satisfatória e permitiu a identificação de potenciais IBPs. Como perspectiva, simulações de Dinâmica Molecular serão realizadas para verificar o comportamento destas proteínas em diferentes temperaturas para a escolha de quais serão expressas heterologamente.

Palavras-Chave: Sequenciamento de RNA. *Treboxiophyceae*. Bioinformática. Transcriptômica.

ABSTRACT

Prasiola crispa is a green macroalga belonging to the *Trebouxiophyceae* class, with cosmopolitan biogeographic distribution, from Arctic to Antarctic. In Antarctic continent, *P. crispa* is among the most important primary producers, surviving to extreme environmental conditions such as low temperatures, osmotic stress, ultraviolet radiation and others. Due to its ability to colonize such an inhospitable environment, this alga must have adaptive mechanisms whose genes involved are of great interest in the area of Biotechnology. Among the main biomolecules of interest, the Ice-binding proteins (IBPs) stand out. IBPs are polypeptides that allow the survival of cells at low temperatures. This group of proteins has great applicability in agriculture, biomedicine and food industry. Thus, the aim of this work was sequencing, assembly and functional annotation of *P. crispa* transcriptome and identify genic products that are directly related to the survive ability of this organism in the Antarctic continent, mainly IBPs. Samples were collected at King George Island, Antarctic. The RNA was extracted, the library built and sequenced by Illumina HiSeq 2000, with paired-end reads strategy. Low-quality reads and adapters were removed and processed reads were subdued to the assembly by Trinity assembler. Assembled transcripts were annotated by an automatized pipeline of the MG-RAST webserver and indicated contamination with the microbiota. *P. crispa* transcriptome was decontaminated and subsequently functionally annotated by BLAST2GO. IBPs of *P. crispa* and microbiota were searched by algorithm BLASTX, aligning the transcripts against IBPs sequences in a local database. Identified transcripts through search had its amino acid sequence predicted by Transdecoder and computationally modeled by Phyre 2. The sequencing generated 31,563,740 high-quality reads. The assembly resulted in 376,797 contigs, totalizing 209,870,963 base pair (bp) and an average size of 557 bp. In the annotation by MG-RAST, 93% of the hits were characterized as belonging to bacteria. After decontamination process, 17,201 contigs were maintained as belonging to *P. crispa* transcriptome, of which 52.19% were identified through functional annotation. 127 transcripts were identified by BLASTX search as potential IBPs, with homology with proteins from plants, bacteria and fungi. The computational modeling of the three-dimensional structure indicated that 17 of these possible IBPs have a similar structure to other proteins in the group, of which 8 have a higher potential. In conclusion, the assembly and functional annotation was performed in a satisfactory manner and allowed the identification of potential IBPs. As a perspective, Molecular Dynamics simulations will be performed to verify the

behavior of these proteins at different temperatures for the choice of which will be expressed heterologously.

Keywords: RNA-Sequencing. *Trebouxiophyceae.* Bioinformatics. Transcriptomics.

SUMÁRIO

1 INTRODUÇÃO E REVISÃO BIBLIOGRÁFICA

1.1 Algas verdes e *Prasiola crispa*

O termo alga compreende uma gama de organismos de grande diversidade, agrupados artificialmente, contendo espécies de eucariotos e procariotos, que filogeneticamente pouco têm em comum e não compreendem um grupo taxonômico [1].

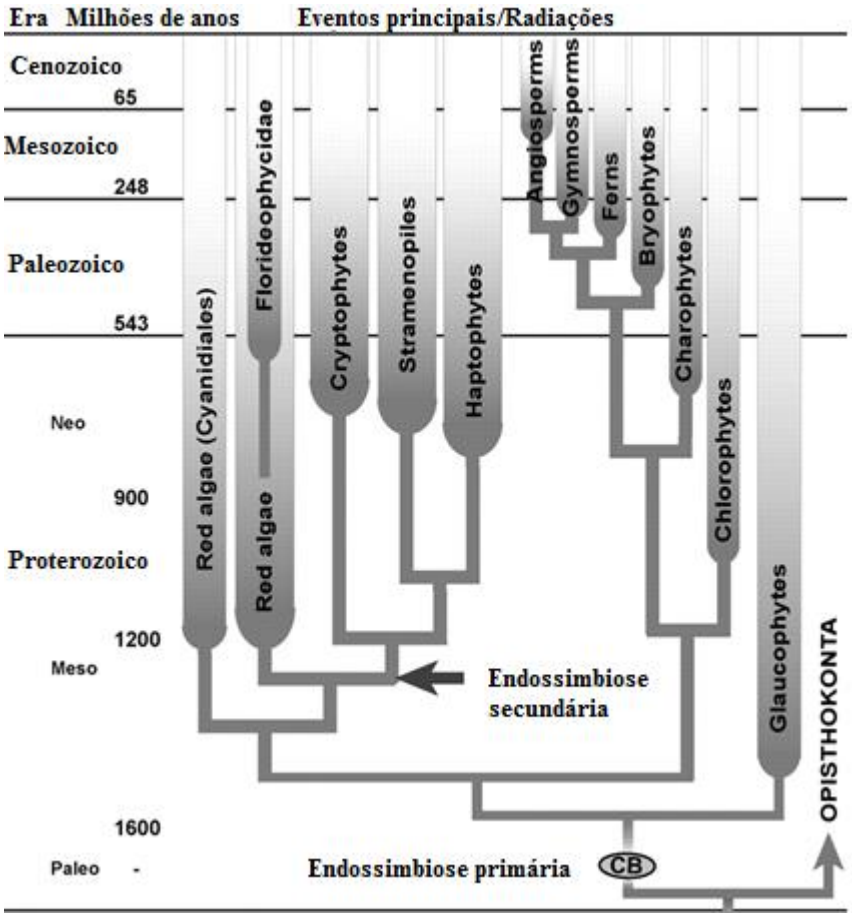
As algas verdes, uma das três linhagens de algas com plastídios primários, possuem ampla diversidade morfológica, apresentando diferentes graus de complexidade, desde organismos microscópicos unicelulares a multicelulares com mais de um metro de tamanho [2]. São componentes essenciais de ecossistemas marinhos, de água doce e terrestres. Estima-se que existam entre 6.000 e 8.000 espécies [3]. Ao longo da evolução as algas verdes construíram uma importante relação de simbiose com fungos [4] e bactérias [5].

Este grupo se evidencia dos outros pois foram as algas verdes que inicialmente colonizaram o ambiente terrestre, há 470-500 milhões de anos atrás, dando início a toda flora terrestre, sendo que plantas terrestres e algas verdes formam o clado filogenético *Viridiplantae* [6-9] (Figura 1). Entre suas aplicações biotecnológicas, algas verdes são importantes fontes de compostos para a indústria química farmacêutica, alimentícia e de biocombustíveis [2, 10].

O grupo é dividido em dois filos: carófita (*Charophyta*) e clorófita (*Chlorophyta*). As carófitas possuem poucos *taxa* mas que são bem diversificados, exibindo uma ampla gama de formas como organismos unicelulares, filamentosos ou "parenquimatosos" [7]. Junto às plantas terrestres formam o clado *Streptophyta*. Ademais, carófitas são importantes modelos para estudos de estruturas como a parede celular e outros mecanismos adaptativos que foram importantes para colonização do ambiente terrestre [11]. Já o filo *Chlorophyta* possui a maior quantidade de espécies de algas verdes descritas, com grande diversidade morfológica e ecológica [7].

Entre as espécies clorófitas, *Prasiola crispa* (Lightfoot) Kützing 1843 pode ser destacada. *P. crispa* é uma macroalga taloide, descrita pela primeira vez na Escócia em 1777 [12, 13]. Esta alga inicialmente foi nomeada *Ulva crispa* e classificada como pertencente a classe *Ulvophyceae*, entretanto, estudos posteriores e as aplicações de ferramentas de biologia molecular permitiram a sua classificação como uma alga pertencente a classe *Trebouxiophyceae* [13-15]. *P. crispa* participa do líquen *Mastodia tessellata*, através da relação simbiótica com o fungo ascomiceto *Guignardia prasiolae* (WINTER) REED [12, 16].

Figura 1 - Representação das relações evolutivas entre as algas e as plantas e dos eventos de endossimbiose ao longo das diferentes eras geológicas: Proterozoica, Paleozoica, Mesozoica e Cenozoica.



Legenda: CB: cianobactéria (CB)
Fonte: YOON *et al*., 2004. [7]

O gênero *Prasiola* possui como característica células individuais com cloroplasto axial estrelado e apenas um pirenoide, sua reprodução pode ser de modo sexuado (oogamia) ou assexuado (esporos ou fragmentação), e sua distribuição biogeográfica é cosmopolita, presentes do Ártico ao continente Antártico [12, 13].

No continente Antártico, *P. crispa* é um dos organismos mais presentes e importante produtor primário [17]. Esta alga está presente nas regiões supralitorais, formando grandes tapetes verdes, sendo encontrada principalmente em locais próximos a colônias de aves, substrato rico em guano, onde há altas concentrações de nitrogênio e ácido úrico (Figura 2) [12, 18].

O continente Antártico é inóspito à vida, sendo o mais seco, frio e ventoso do planeta, sem registros da existência de população humana nativa [19]. A precipitação anual média é de apenas 200 mm, ademais, ventos de 327 km/h e temperaturas abaixo de – 90 ºC já foram

registradas [20]. A área total é de 14.000.000 km², sendo 98 a 99,7% coberta por neve e gelo, com camadas em média de 1,6 km de espessura [20, 21]. Além disso, o buraco na camada de ozônio sobre a região Antártica, descrito pela primeira vez na década de 80, faz com que haja um alto índice de radiação ultravioleta sobre a região, que é intensificada pelo reflexo gerado pelo gelo [22, 23]. Outros fatores abióticos como alterações constantes de salinidade e baixa disponibilidade de nutrientes orgânicos também são desafios para a sobrevivência [24].

Figura 2 - *P. crispa* no continente Antártico.

Fonte: Graciele Alves de Menezes.

Devido a sua capacidade de colonizar um ambiente tão inóspito e extremo, *P. crispa* deve possuir mecanismos adaptativos naturalmente selecionados durante sua evolução. Os genes e biomoléculas envolvidos nestes importantes mecanismos são de grande interesse e potencial na área da Biotecnologia.

1.2 Potencial Biotecnológico e Proteínas de Ligação ao gelo

O potencial biotecnológico de *P. crispa* já foi demonstrado em outros trabalhos. Estudos avaliando o extrato desta alga demonstraram o efeito inseticida em espécies modelos,

Drosophila melanogaster e *Nauphoeta cinerea*, provavelmente através da modificação de sistemas antioxidantes do organismo [25]. Já o composto químico 7-ceto-estigmasterol, um esterol purificado a partir do extrato de *P. crispa*, apresentou atividade antiviral quando testado contra o Herpesvírus equino 1, o vírus causador de uma doença até o momento sem tratamento eficiente [26].

Um grande potencial, ainda não explorado neste organismo, principalmente tratando-se de um organismo Antártico, são as proteínas de ligação ao gelo (do inglês *Ice-binding protein*, IBP). IBPs são polipeptídeos expressos em uma ampla gama de organismos que permitem a sobrevivência das células à baixas temperaturas [27, 28]. A primeira IBP foi identificada em 1969, quando Arthur DeVries demonstrou que proteínas presentes no sangue de peixes teleósteos, habitantes da região Antártica, diminuem o seu ponto de congelamento [29, 30]. Estas proteínas foram inicialmente denominadas proteínas anticongelantes (do inglês *antifreeze proteins*, AFPs), um termo utilizado de maneira genérica por muitos anos, mas que atualmente compreende apenas um subconjunto das IBPs [31].

Além dos peixes, este grupo de proteínas já foram identificadas em grande número de bactérias [32], fungos [33], insetos [34, 35] e plantas [36]. Ademais, IBPs foram encontradas em algumas algas da Antártica [37-42]. A distribuição de IBPs em tantas espécies distintas ocorreu provavelmente através de eventos de convergência evolutiva e transferência horizontal de genes [43].

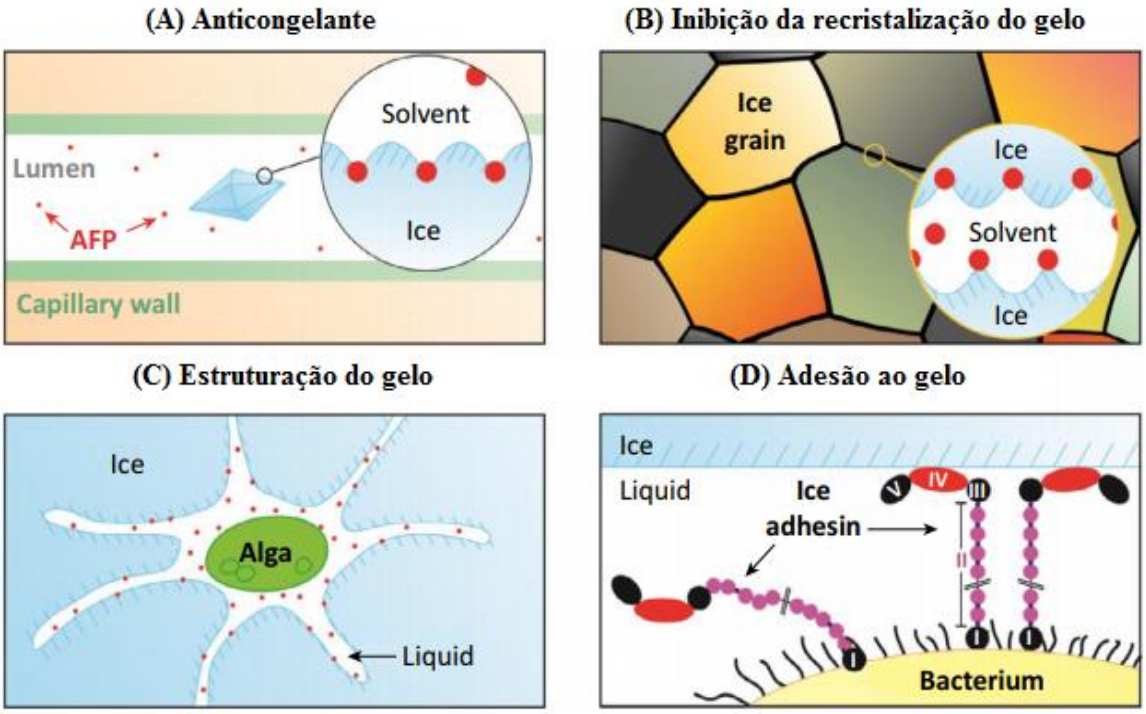
A ausência destas proteínas, e de outros agentes crioprotetores, em organismos expostos a baixas temperaturas teriam como consequência o congelamento da água e concentração dos fluidos, causando a desidratação, choque osmótico e rompimento da membrana celular pela formação de cristais de gelo [43].

A propriedade essencial de uma IBP é a capacidade de adsorção de um ou mais planos do gelo, que tem como consequência natural a alteração do formato dos cristais de gelo, sendo esta ligação irreversível [44]. As IBPs também possuem outras propriedades, estas que podem ser utilizadas para classificação dos seus diferentes subgrupos.

A principal propriedade do subgrupo das AFPs, por exemplo, é sua atividade de histerese térmica (em inglês *thermal hysteresis*, TH), através da diminuição das temperaturas de fusão e solidificação da água (Figura 3A). Os mecanismos de ação ainda não são completamente entendidos, mas estudos indicam que a inibição de eventos de nucleação secundária tem papel essencial [34, 45]. AFPs com alta atividade TH, também denominadas de TH hiperativas, são expressas principalmente em insetos, cuja hemolinfa isolada pode baixar até 12 ºC o ponto de congelamento [45].

Em plantas tolerantes ao frio, IBPs com alta atividade de inibição da recristalização do gelo (em inglês *ice recrystallization*, IR) são mais comumente encontradas (Figura 3B) [36]. Quando os fluidos de uma planta atingem a temperatura de congelamento, cristais de gelo são formados. Por questões termodinâmicas, com o passar do tempo grandes cristais de gelo têm a tendência de aumentar de tamanho, enquanto os pequenos cristais se fundem, sendo este processo denominado IR. Assim, as IBPs possuem a capacidade de inibir a IR com grande eficiência até mesmo em baixas concentrações [46], sendo a hipótese de adsorção-inibição aos sítios de crescimento do gelo a mais aceita para o mecanismo de atuação das IBPs [47].

Figura 3 - Funções biológicas/categorias de IBPs: (A) Anticongelante; (B) Inibição da recristalização do gelo; (C) Estruturação do gelo; (D) Adesão ao gelo.



Fonte: Modificado de DAVIES, 2014. [31]

Já o subgrupo denominado proteínas de estruturação do gelo (do inglês *ice-structuring proteins*, ISPs), mais encontradas em algas unicelulares, são secretadas para o meio extracelular a fim de formar pequenos bolsões estáveis durante o congelamento da água do mar, formando um microambiente líquido (Figura 3C) [31]. ISPs também foram encontradas em outros grupos de organismos, como em bactérias [48].

A propriedade mais recentemente identificada em IBPs foi a atividade de adesão ao gelo, encontrada na bactéria Antártica *Marinomonas primoryensis* (Figura 3D) [49]. Especula-se que esta adesina possa ter a função de manter a bactéria próxima dos nutrientes e

oxigênio na zona fototrófica [45]. Diante das variadas funções biológicas e origens a estrutura tridimensional deste grupo de proteínas é bem variada.

A produção em larga escala de IBPs ainda precisa de melhorias para se tornar viável e amplamente difundida, mas as possibilidades são variadas. Entre as aplicabilidades podem ser citadas a utilização para a preservação e melhoria da qualidade de alimentos congelados [50], produção de culturas tolerantes ao frio [51], crioproteção de tecidos e órgãos [52] e até mesmo a aplicação na indústria petroleira [53].

1.3 Sequenciamento de RNAs em larga escala

Para acessar este potencial Biotecnológico, o transcriptoma de *P. crispa* foi sequenciado. Transcriptoma é o conjunto de todos os ácidos ribonucleicos (RNAs) expressos por um organismo, sendo o objeto de estudo de uma das áreas da Genômica Funcional, a Transcriptômica. O sequenciamento de RNAs em larga escala é uma abordagem recente, amplamente utilizada para a descoberta de novos genes em organismos não-modelo [54]. Esta abordagem, quando comparada a outras empregadas para análises do transcriptoma como, por exemplo, os chips de *microarray* e sequenciamento de pequenas sequências expressas (ESTs), traz grandes vantagens como: bom custo-benefício, alta sensibilidade e acurácia [55].

O sequenciamento do transcriptoma já foi realizado em outras algas da classe *Trebouxiophyceae*, objetivando a identificação de transcritos associados a resistência a estresses bióticos e abióticos, e genes com aplicabilidade na produção de biocombustível [56-60].

Para o sequenciamento do transcriptoma, o RNA total é extraído da amostra e convertido em fragmentos de ácido desoxirribonucleico complementar (cDNA), formando a biblioteca de sequenciamento. Esta biblioteca então é sequenciada, gerando pequenas *reads* que devem ser montadas e anotadas, de maneira muito semelhante aos genomas. Assim, este estudo só foi possível devido a evolução e barateamento das técnicas de sequenciamento.

Os primeiros métodos de sequenciamento de ácido desoxirribonucleico (DNA) foram criados na década de 70, quando Sanger e Coulson desenvolveram a técnica “mais e menos” [61] e Maxam e Gilbert a de clivagem química [62].

Porém, o grande avanço ocorreu quando Sanger desenvolveu o sequenciamento por terminação de cadeia. O mesmo consiste na síntese de uma cadeia nucleotídica utilizando o fragmento a ser sequenciado como molde, sendo a base correspondente para cada posição da cadeia identificada através do interrompimento da síntese, que ocorre a partir da adição de

dideoxinucleotídeos (ddNTPs) marcados radioativamente, em quatro reações paralelas [63]. A princípio, a metodologia utilizava a migração em gel de poliacrilamida para identificação das bases. Posteriormente o processo foi automatizado, sendo transformado em uma máquina com a utilização de ddNTPs marcados com fluoróforos e a migração em capilar, tornando-se a metodologia mais aplicada para o sequenciamento de DNA [64]. O sequenciador de Sanger é considerado a primeira geração, e gerava *reads* de ~1 kilobase (kb).

O sequenciamento de Sanger foi extremamente importante para as descobertas na área de Genômica estrutural, por exemplo, no Projeto Genoma Humano, porém sua aplicabilidade na área de Transcriptômica era limitada devido à baixa quantidade de dados gerados, alto custo e dificuldade da avaliação de níveis quantitativos em larga escala [55, 65].

O início da segunda geração, os denominados sequenciadores de nova geração (*Next-Generation sequencing*, NGS), foi marcado pelo desenvolvimento do pirosequenciador. Esta metodologia também utilizava o método "sequenciamento por síntese" aplicando a DNA polimerase, mas os nucleotídeos não mais eram marcados e a adição das bases acompanhadas em tempo real, sem a necessidade de migração em capilar [65].

O pirosequenciamento detecta o pirofosfato liberado durante o processo de formação da ligação fosfodiéster entre nucleotídeos, sendo a detecção realizada através de luminescência [66]. A técnica foi licenciada à 454 Life Sciences para a produção de máquinas de pirosequenciamento, que se valiam da técnica de reação em cadeia da polimerase (PCR) em emulsão para amplificação de DNA aderido em *beads* antes do sequenciamento. As *reads* geradas possuem entre 400 e 500 pares de bases (pb).

Após o sucesso dos pirosequenciadores, outras plataformas com competitividade foram desenvolvidas, destacando-se as máquinas Solexa/Illumina e SOLiD, todas possuindo como característica o sequenciamento em massa. O sistema SOLiD (*Sequencing by Oligonucleotide Ligation and Detection*) é fundamentado na hibridização e ligação de oligonucleotídeos marcados com fluoróforos, utilizando a enzima DNA ligase. Apesar das pequenas *reads* geradas e de não possuir grande profundidade, o SOLiD é competitivo por conta do menor custo por base sequenciada [65].

A plataforma Solexa/Illumina baseia-se na adição de adaptadores nas extremidades dos fragmentos de DNA, formação de *clusters* por PCR em ponte e o sequenciamento por síntese através da incorporação de nucleotídeos "terminadores reversíveis" marcados com fluoróforos [67]. As primeiras máquinas produziam *reads* de apenas 35pb, mas a introdução da estratégia de *paired-end reads* (*reads* geradas no sentido *forward* e *reverse* do fragmento sequenciado) permitiram um melhor mapeamento entre as *reads* para a montagem [65]. Os

modelos posteriormente lançados, nomeados Illumina HiSeq e Illumina MiSeq, trouxeram como vantagem a geração de *reads* maiores e menor custo.

Estes NGS, através do sequenciamento em massa por um preço mais acessível, permitiram o desenvolvimento e maior popularização da área de sequenciamento de transcriptomas, sendo a plataforma Illumina a mais empregada neste tipo de estudos. Porém, como as *reads* são pequenas, surgiram os desafios para a montagem do transcriptoma. Assim, além da evolução das técnicas de sequenciamento, o desenvolvimento das ferramentas de Bioinformática para a montagem e análise da avalanche de dados gerados também foram essenciais.

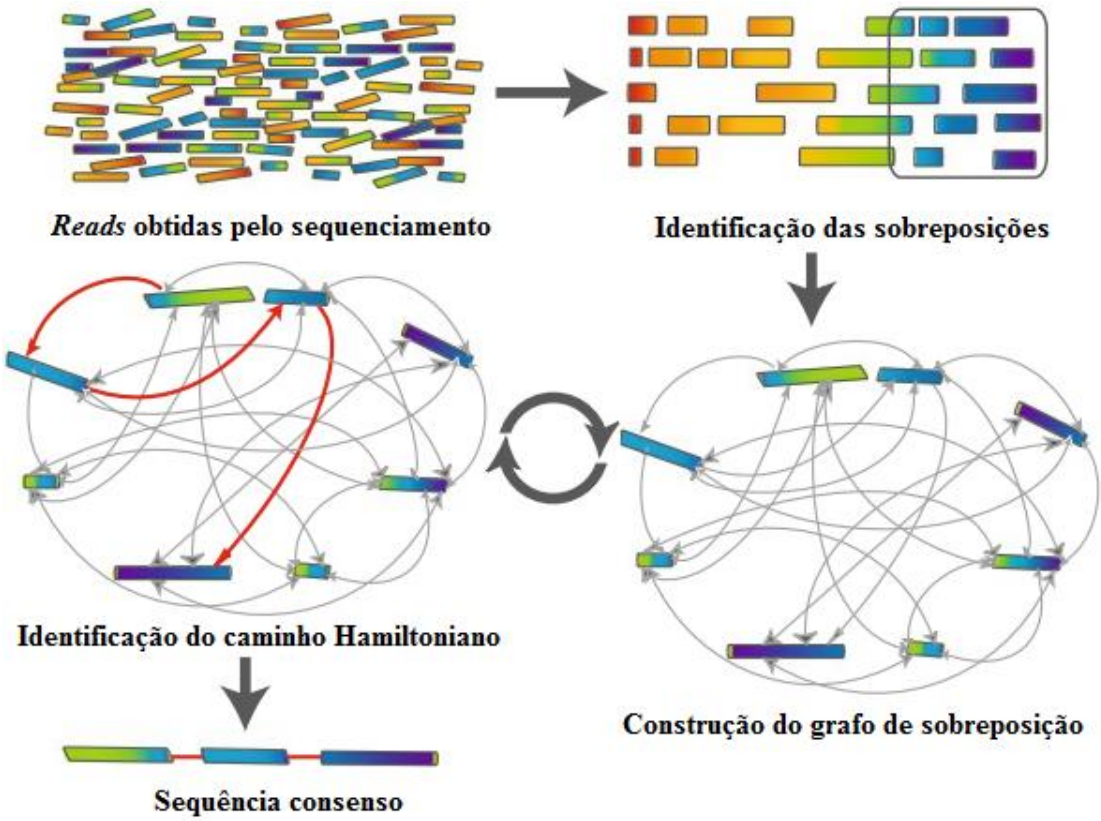
1.4 Montagem e Anotação de Transcriptomas

A montagem das pequenas *reads* geradas em transcritos pode ser realizada em duas abordagens: montagem com referência ou *de novo*. A montagem de referência é escolhida quando há um genoma para guiar a montagem, podendo ser da mesma espécie ou de espécies relacionadas. Neste tipo de abordagem as *reads* são mapeadas contra o genoma de referência utilizando programas de alinhamento de sequências [68]. Porém, a maioria das espécies não-modelo nem sempre possuem um genoma a ser utilizado como referência, como no caso de *P. crispa*, sendo necessária a montagem *de novo*.

Por isso, diversos algoritmos foram desenvolvidos para a reconstrução *de novo* e estes são classificados em duas classes: *overlap-layout-consensus* (OLC) e grafos de Bruijn (DBG). Os algoritmos OLC trabalham em três passos: (I) todas as sobreposições entre as *reads* são identificadas; (II) todas as informações são dispostas em um grafo de sobreposição e; (III) através do grafo, a sequência consenso é inferida (Figura 4) [69]. Neste grafo, cada *read* é considerada um vértice, e as arestas que conectam dois vértices correspondem as sobreposições de *reads*. O algoritmo então deve resolver o problema passando exatamente uma vez em cada vértice, caracterizando o problema do caminho Hamiltoniano [70].

Entre os softwares da categoria destacam-se o Celera Assembler [71], CAP3 [72] e Phrap [73], aplicados principalmente para a montagem de *reads* provenientes do sequenciamento de Sanger. Contudo, a solução de problemas pelo caminho Hamiltoniano são computacionalmente dispendiosos e possuem grandes dificuldades para montagem correta de regiões repetitivas e montagem de *reads* geradas por NGS (Figura 5A) [74].

Figura 4 - Funcionamento básico das ferramentas de montagem *Overlap–layout–consensus*.

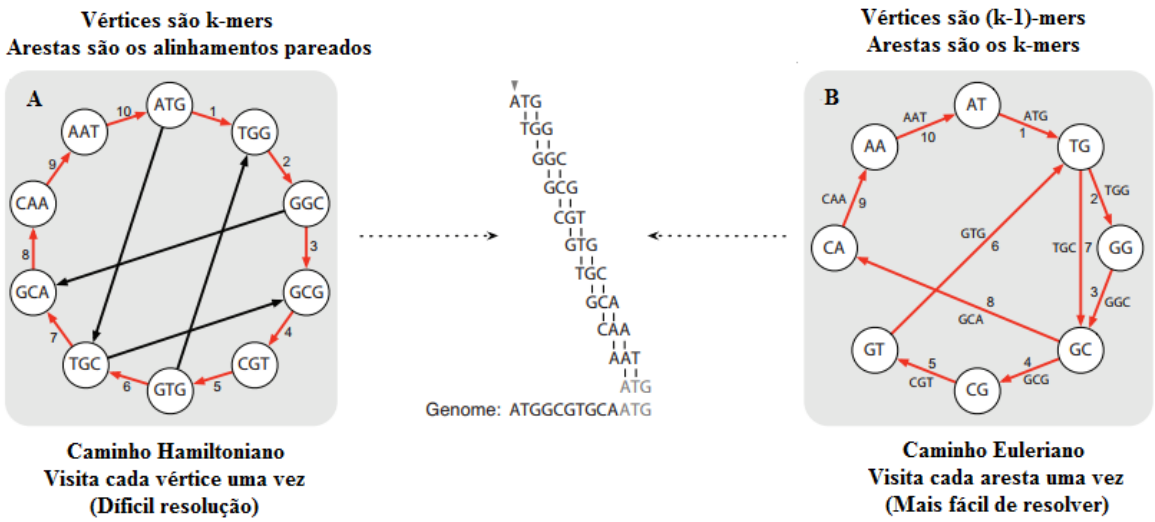


Legenda: Setas cinzas indicam as sobreposições e setas vermelhas o melhor caminho Hamiltoniano.
Fonte: Modificado de COMMINS *et al.*, 2009. [75]

E assim, os montadores que aplicam o modelo DBG foram desenvolvidos. O primeiro passo para a montagem utilizando softwares da classe DBG é a quebra virtual de todas as *reads* em fragmentos de tamanho k (k-mers) [76]. Por exemplo, quando se deseja utilizar fragmentos de 25 pb (k = 25), cada *read* de tamanho 100 (T) é quebrada em 76 fragmentos de 25 (T - k + 1), sendo k um número ímpar para reduzir o número de sequências palindrômicas [77]. Após a fragmentação, os grafos de Bruijn são formados, sendo o prefixo e sufixo de cada k-mer distinto representado como um vértice de comprimento k-1, e as arestas que conectam estes vértices são os k-mer que os compõem (Figura 5B). Então, o caminho Euleriano, aquele que visita todas as arestas uma única vez, é traçado entre todos os k-mers que se sobrepõe perfeitamente em k-1 nucleotídeos, levando à reconstrução das sequências [74]. O problema do caminho Euleriano é muito mais simples de se resolver que o Hamiltoniano. Para a montagem de transcriptomas, cada caminho no grafo representa um possível transcrito. A Figura 5 compara o funcionamento do caminho Hamiltoniano e Euleriano com 3-mers.

Entre os montadores da categoria DBG destacam-se o Trans-ABySS [78], SOAPdenovo [79], Oases [80] e Trinity [77], sendo este último o utilizado em todas as montagens realizadas de novo do transcriptoma de algas da classe *Trebouxiophyceae*.

Figura 5 - Comparação entre o caminho (A) Hamiltoniano e (B) Euleriano em um grafo.



Legenda: As setas vermelhas o melhor caminho (A) Hamiltoniano (B) Euleriano; setas pretas indicam problemas difíceis de serem resolvidos pelo algoritmo.
Fonte: Modificado de COMPEAU *et* al., 2011. [74]

Após a montagem, o último passo é a anotação do transcriptoma. A anotação consiste em agregar de informações biológicas aos transcritos, por meio da busca de sequências homólogas, havendo uma ampla gama de *pipelines* e ferramentas que podem ser aplicadas [81]. Esta metodologia de busca baseada em homologia é aplicada não apenas para anotação Transcriptomas, mas também genomas, metagenomas e metatranscriptomas [82]. Entre as ferramentas o Blast2GO é um dos mais utilizados, sendo robusto e simples [83, 84].

O primeiro passo da anotação feita pelo Blast2GO é a busca utilizando a ferramenta de alinhamento BLAST (*Basic Local Alignment Search Tool*) [85] por sequências que sejam similares aos transcritos. Os resultados são expressos através do *E-value*, que descreve o número de *hits* que se espera ao realizar um alinhamento de sequências contra um banco de dados de determinado tamanho apenas ao acaso [86]. Também é levado em consideração o tamanho do alinhamento e similaridade. Assim, quanto menor o *E-value*, mais significante é considerado aquele alinhamento entre as sequências.

Após isso, as sequências são mapeadas e anotadas para associação de termos funcionais do banco de dados *Gene Ontology* (GO) de acordo com o resultado do BLAST. O

Gene Ontology Consortium é um projeto que permite classificar os genes e seus produtos de maneira uniforme, atuando como uma linguagem universal, rotineiramente empregado na pesquisa durante o processo anotação funcional [87]. Este projeto possui três categorias principais que indicam que o produto gênico possui (I) determinada atividade ou processo a nível molecular (Função Molecular, FM), (II) que ocorre em uma localização específica celular (Componente celular, CC) e (III) que contribui para um efeito biológico (Processo biológico, PB). Além disso, existe a divisão de níveis que vão dos termos mais gerais para os mais específicos.

Outras ferramentas, como o MG-RAST [88], não possuem anotação tão detalhada quanto o Blast2GO mas tem como vantagem a grande velocidade para trabalhar com números muito grandes de *contigs*. Além disso, outros vocabulários de anotação como o COG (*Cluster of Orthologous Groups*) [89] e *Subsystems* [90] podem ser aplicados.

Ao fim do processo de anotação, o transcriptoma pode ser analisado para a busca por transcritos de interesse, como as IBPs.

1.5 Bioinformática: Modelos Tridimensionais de Proteínas

A Bioinformática é dividida em duas categorias: Bioinformática Tradicional e a Bioinformática Estrutural. A Tradicional aborda principalmente problemas envolvidos com sequências de nucleotídeos e aminoácidos, como a montagem e anotação de transcriptomas. Já a Bioinformática Estrutural, aborda as questões de um ponto de vista tridimensional, abrangendo as técnicas de modelagem molecular e química computacional [91]. A união de ambas as vertentes permite maior confiabilidade e valor aos dados analisados *in silico*.

A similaridade entre as sequências é importante durante o processo identificação da função do transcrito, porém, não é suficiente. Isso porque a função biológica de uma proteína está intimamente relacionada à sua estrutura tridimensional [92]. Contudo, a determinação experimental da estrutura de proteínas é um processo longo e caro, sendo a cristalografia de raio-X a metodologia mais eficiente e aplicada [93]. Apenas uma pequena fração das proteínas possuem estrutura conhecida. Assim, a predição da estrutura tridimensional de proteínas através de abordagens computacionais é uma importante área da Bioinformática, que está em franco desenvolvimento e possui grandes impactos sobre a Biotecnologia [91].

A modelagem computacional de proteínas pode seguir por dois métodos: os independentes de estrutura molde e os baseados em estrutura molde. Os métodos independentes de proteína molde são divididos em *ab initio* e *de novo*, enquanto os baseados

em estrutura molde dividem-se em *threading* e modelagem comparativa. Estas abordagens são assim classificadas de acordo com as informações que utilizam a partir dos bancos de dados, sendo o *Protein Data Bank* (PDB) o mais utilizado [91].

A escolha do método está associada à taxa de identidade entre a sequência a ser modelada e as estruturas presentes no PDB. A abordagem *ab initio* é utilizada quando não há qualquer identidade de sequência com as proteínas do banco de dados, sendo que a mesma se baseia apenas em parâmetros físico-químicos e algoritmos desenvolvidos para a modelagem [94]. A abordagem *de novo*, também escolhida quando não há identidade encontrada, utiliza a informação da estrutura de proteínas não homólogas para modelagem. Porém, apesar de todo o esforço para o desenvolvimento de bons algoritmos *ab initio* e *de novo*, ainda não há ferramentas que possam realizar predições confiáveis através destas abordagens [95].

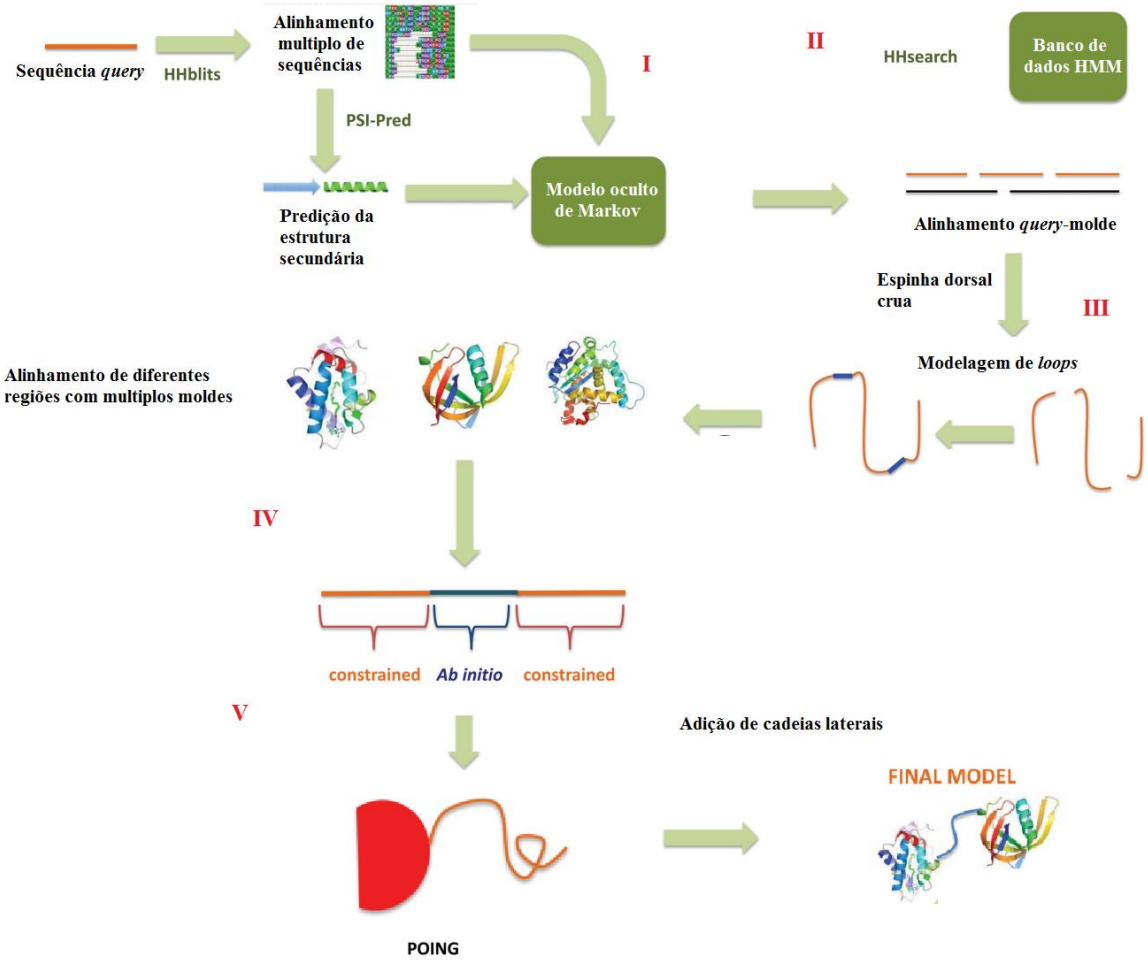
O método de *threading* baseia-se no princípio de que a estrutura tridimensional é mais conservada do que a sequência, onde sequências com pouca identidade podem possuir o mesmo tipo de enovelamento [91]. Esta abordagem é utilizada quando a sequência molde possui menos de 30% de identidade com a sequência a ser modelada [96]. Já a modelagem comparativa, a mais aplicada, é realizada quando existe uma proteína com mais de 30% de identidade que possa ser utilizada como referência, partindo do princípio que pequenas mudanças, como mutações pontuais, geralmente não causam grandes alterações conformacionais [94]. Os métodos que utilizam proteínas de estrutura conhecida como molde são os mais confiáveis e fáceis de serem realizados, contudo, possuem a limitação da ausência de informações sobre as estruturas tridimensionais da grande maioria das proteínas [96].

O evento bianual CASP (*Critical Assessment of Protein Structure Prediction*), criado para avaliar o desenvolvimento das predições de estruturas proteicas a partir da sequência de aminoácidos, reforça o grande interesse e importância do desenvolvimento da área e permite a identificação das melhores ferramentas a serem empregadas [97]. Entre as ferramentas apresentadas no CASP, o *web-server* Phyre 2 [95] tem se confirmado como um dos melhores, realizando modelagens *ab initio* e comparativas [98].

Esta ferramenta aplica um *pipeline* automatizado de predição, dividido em vários estágios (Figura 6), onde não há a necessidade de adições/modificações de parâmetros por parte do usuário. O primeiro estágio do processo é a busca por sequências homólogas e construção do perfil evolutivo de substituição de aminoácidos no grupo, feito heuristicamente pelo algoritmo HHblists [99]. A estrutura secundária então é predita utilizando o PSIPRED. O segundo estágio converte o perfil evolutivo e a estrutura secundária predita em um modelo oculto de Markov (em inglês *Hidden Markov Model*, HMM), que será analisado contra o

banco de dados contendo HMMs de proteínas de estrutura já conhecida para a formação de diversos alinhamentos *query*-molde.

Figura 6 - *Pipeline* de predição da estrutura tridimensional de proteínas utilizando Phyre 2: (I) Alinhamento e predição de estrutura secundária; (II) Alinhamento com banco de dados HMM; (III) Modelagem de *loops*; (IV) Modelagem com multiplos moldes; (V) Simulação de tradução e formação do modelo final.



Legenda: HMM – Modelos ocultos de Markov.
Fonte: Modificado de KELLEY *et al.*, 2015. [95]

O terceiro estágio é a solução de inserções e deleções presentes no alinhamento, os *indels*, que são analisados de acordo com a estrutura das regiões próximas e sequências aminoacídicas flanqueadoras. A melhor solução para a modelagem de *loops* é indicada através do cálculo de energia livre da estrutura.

O quarto estágio é a modelagem Poing, utilizando múltiplos moldes para os diferentes domínios proteicos. Poing realiza a simulação dos processos de tradução e enovelamento. Regiões modeladas por diferentes moldes formam sítios restritos, e regiões não cobertas pelo alinhamento são modeladas de forma *ab initio*. Poing considera apenas o carbono α, gerando

somente a espinha dorsal da estrutura tridimensional. No quinto e último estágio, as cadeias laterais dos aminoácidos são adicionadas para a formação do modelo final.

A estrutura final permite então a inferência mais precisa da função biológica do produto gênico.

2 OBJETIVOS

2.1 Objetivo geral

O objetivo deste trabalho foi sequenciar, montar e anotar funcionalmente o transcriptoma de *P. crispa* e identificar produtos gênicos diretamente relacionados a capacidade de sobrevivência deste organismo no continente Antártico, principalmente de resistência ao frio.

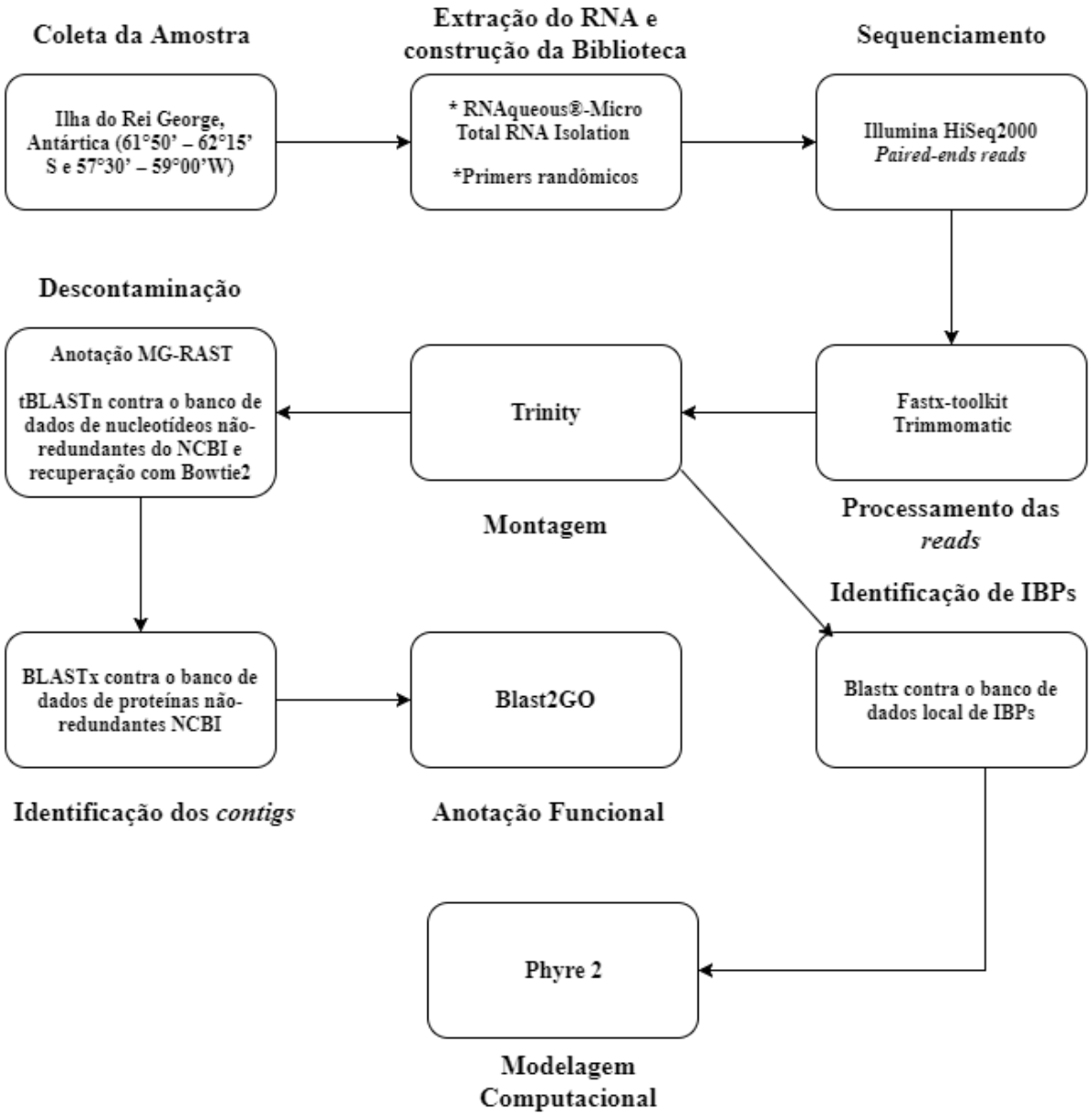
2.2 Objetivos Específicos

- Extrair e sequenciar o RNA total de *P. crispa*;
- Realizar análises de validação e qualidade dos dados;
- Reconstruir os transcritos a partir das *reads* de alta qualidade;
- Identificar e anotar funcionalmente os transcritos;
- Comparar as métricas do transcriptoma de *P. crispa* com as de outros organismos da classe *Trebouxiophyceae*;
- Identificar potenciais IBPs por alinhamento contra banco de dados local;
- Modelar a estrutura tridimensional de IBPs identificadas a partir de sua sequência aminoacídica.

3 METODOLOGIA

A Figura 7 apresenta o resumo de todo *pipeline* aplicado durante o estudo.

Figura 7 - Resumo da metodologia empregada neste trabalho.



3.1 Coleta da alga

Prasiola crispa foi coletada na Ilha do Rei George, Antártica (61°50' – 62°15' S e 57°30' – 59°00'W). A coleta foi realizada no verão Antártico, em janeiro de 2014. As

amostras foram mantidas no tampão RNAlater® (Sigma-Aldrich, USA) até a extração do RNA.

3.2 Extração do RNA total e sequenciamento

O RNA total foi obtido através extração com o kit RNAqueous®-Micro Total RNA Isolation (Thermo Fisher Scientific Inc., USA) de acordo com as instruções do fabricante. A biblioteca de sequenciamento foi preparada utilizando primers randômicos. O transcriptoma foi sequenciado pelo serviço da Macrogen, utilizando a plataforma Solexa-Illumina HiSeq 2000, manipulada de acordo com as instruções do fabricante. A estratégia de *paired-end reads* de ~100 pb, separadas por inserto de 300 pb foi aplicada.

3.3 Montagem do transcriptoma *de novo*

Reads brutas foram filtradas para a remoção dos adaptadores e das *reads* de baixa qualidade com Fastx-toolkit (*cut-off* de qualidade = 30) [100] e Trimmomatic v 0.36 em parâmetros *default* [101]. Após isso, a ferramenta Trinity [102] versão r2014-07-17 foi utilizada como montador DBG para a geração dos *contigs*, utilizando k-mers de 25 pb.

Devido à complexidade do solo Antártico e as relações simbióticas que naturalmente ocorrem, esperávamos contaminações com bactérias e fungos. Para verificação de contaminação, as sequências foram submetidas ao *pipeline* automatizado de busca por homologia do *web-server* MG-RAST, aplicado na anotação de metagenomas e metatranscriptomas, buscando a identificação da presença e origem de transcritos contaminantes [88]. A fim de remover as sequências contaminantes, o transcriptoma de *P. crispa* foi submetido a alinhamento pelo algoritmo TBLASTN com parâmetros *default* [85] contra o banco de dados de nucleotídeos não-redundantes do NCBI. Todos os *contigs* cujo melhor BLAST *hit* ocorreram com sequências de plantas e algas foram mantidos, e os demais removidos. Através do Bowtie2 [103], apenas as *reads* que constituíam os *contigs* mantidos no transcriptoma de *P. crispa* foram recuperados.

3.4 Anotação Funcional

Os *contigs* montados e recuperados foram alinhados contra o banco de proteínas não-redundantes do NCBI, sendo o *cut-off* escolhido 1e-10. Identificou-se os transcritos de acordo

com o melhor *hit* contra sequências de função conhecida. Blast2GO [84] foi utilizado para mapear e anotar as sequências, associando termos do GO e predizendo suas funções.

3.5 Criação de banco de dados locais e busca por IBPs

Os termos *Ice-binding protein*, *Anti-freezing protein*, *Thermal hysteresis* e *Ice-structuring protein* foram buscados no banco de dados de proteínas do NCBI. Todas as sequências associadas a estes termos foram baixadas e um banco de dados local foi criado. Após a criação do banco, todos os transcritos, inclusive os contaminantes, foram submetidos a busca contra o banco de dados local utilizando o algoritmo BLASTX (*cut-off*: 1e-10), a partir do alinhamento de sequências proteicas contra as seis fases de leitura dos transcritos [104].

3.6 Modelagem computacional

Todas as sequências que tiveram alinhamento por BLASTX contra as IBPs, foram submetidas a análise pela ferramenta Transdecoder [102] para a identificação de fases de leitura aberta (ORFs) e a sequência proteica correspondente. As sequências aminoacídicas dos transcritos com ORFs identificadas foram submetidas a modelagem no *web-server* Phyre 2 [95].

3.7 Depósito de dados

As *reads* brutas e o Transcriptoma montado foram depositados em repositórios públicos sob os seguintes códigos de acesso:

- Bioproject ID: PRJNA329112
- Biosample: SAMN05392062
- Sequence Read Archive (SRA): SRR5754271
- Transcriptome Shotgun Assembly (TSA): GFTS00000000

4 RESULTADOS E DISCUSSÃO

4.1 Montagem e anotação funcional do transcriptoma de *P. crispa*

O sequenciamento pela plataforma Solexa-Illumina HiSeq 2000 gerou 42.978.976 *reads*. Após o processo de remoção das *reads* de baixa qualidade, 31.563.740 foram mantidas. A montagem das mesmas gerou 354.661 *contigs*, com tamanho médio de 477 pb. Estes e outros valores da montagem são apresentados na Tabela 1.

Tabela 1 - Métricas da montagem inicial do Transcriptoma.

Atributos	Valor
Total de *reads* brutas	42.978.976
Reads de alta qualidade	31.563.740
Número de *contigs*	354.661
Tamanho total (pb)	169.085.999
GC (%)	53,00
Tamanho médio (pb)	477

Legenda: pb - pares de bases.

Este número de *contigs* é 37 vezes maior que o transcriptoma de *Coccomyxa subellipsoidea* [59], alga também pertencente a classe *Trebouxiophyceae*, indicando a provável contaminação com sequências de outros organismos. A rápida anotação através do *server* MG-RAST permitiu a confirmação da contaminação. De todos os *contigs*, 152.909 (44,11%) foram anotadas como transcritos que seriam expressos em proteínas com função conhecida e 193.077 (55,69%) com função desconhecida (Figura 8A). A distribuição taxonômica indicou que 612.522 *hits* (93%) ocorreram com sequências de bactérias, 1.163 (0,18%) de arqueias e apenas 39.414 (6,03%) de eucariotos (Figura 8B).

Após processo o de descontaminação apenas 17.201 *contigs* (4,84%), constituídos por 5.233.428 *reads*, foram recuperados. As estatísticas da montagem são resumidas na Tabela 2.

As métricas da montagem foram comparadas a todos os transcriptomas já sequenciados da classe *Trebouxiophyceae*: *Chlorella minutissima* [60], *Trebouxia gelatinosa* [57], *Coccomyxa subellipsoidea* [59], *Chlorella sorokiniana* [58] e *Botryococcus braunii*

[105] (Tabela 3). A aplicação do mesmo montador para os projetos de montagem *de novo* permite boa comparação dos dados.

Figura 8 - Gráfico do processo de anotação da metamontagem realizada pelo MG-RAST: (A) Características preditas de cada *contig* e (B) Distribuição de *hits* por Domínios.

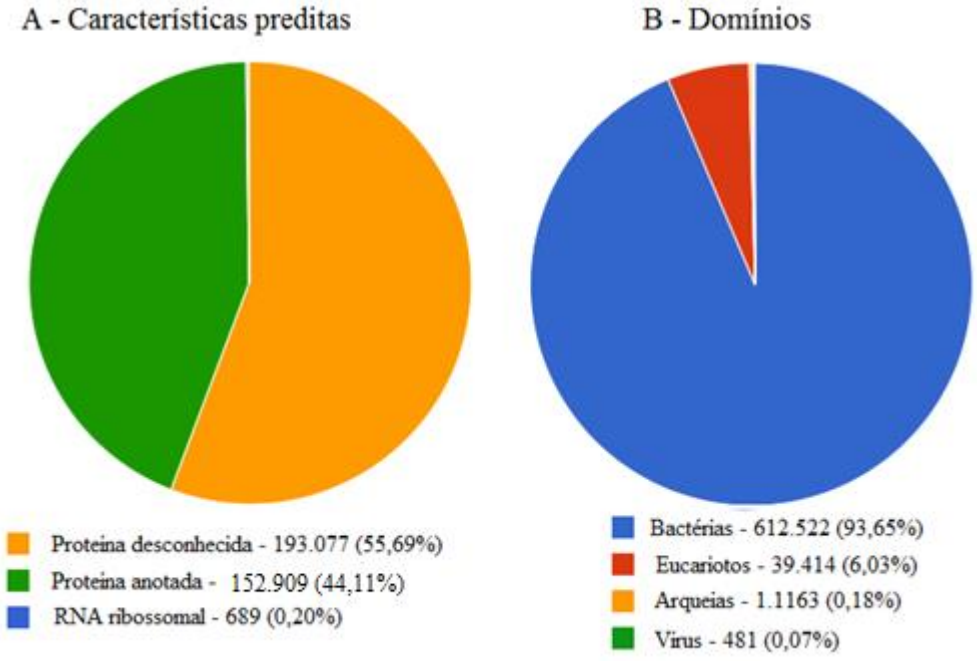


Tabela 2 - Resumo das métricas do Transcriptoma de *P. crispa*.

Atributos	**Valor**
Total de *reads* brutas	42.978.976
Reads de alta qualidade	31.563.740
Total de *reads* processadas	5.233.428
Número de *contigs*	17.201
Tamanho total (pb)	13.127.645
N50	1.036
GC (%)	49,66
Tamanho médio (pb)	763,19
Espectro de tamanho (pb)	200 – 12.802

Legenda: pb - pares de bases.

Tabela 3 - Comparação das montagens de Transcriptomas de algas da classe *Trebouxiophyceae*.

Atributos	***Prasiola crispa***	***Chlorella minutissima***	***Trebouxia gelatinosa***	***Coccomyxa subellipsoidea***	***Chlorella sorokiniana***	***Botryococcus braunii***
Total de *reads* brutas	42.978.976	69.011.712	243.763.578	-	244.291.069	-
Total de *reads* processadas	5.233.428	67.559.338	237.404.631	46.000.000	229.228.757	-
Número de *contigs*	17.201	14.905	19.601	9.409	63.811	61.220
Tamanho médio (pb)	763,1	2.998,04	1.605	-	1.022	-
Sequences com BLAST *hits* (%)	52,19	53,50	53,60	-	36,80	-
Maior *contig* (pb)	12.802	-	31.749	-	15.932	-
N50	1.036	-	3.594	-	2.502	-
Montador	Trinity	Trinity	Trinity	Genoma como referência	Trinity	Trinity

Legenda: pb - pares de bases; (-) - Dado não reportado.

O número de *contigs* de *P. crispa* teve posição intermediária. Isto indica que o processo de descontaminação do transcriptoma foi satisfatório, compreendendo um número plausível de transcritos. Além disso, os transcritos depositados no NCBI passaram por verificação de contaminantes como requisito para aprovação.

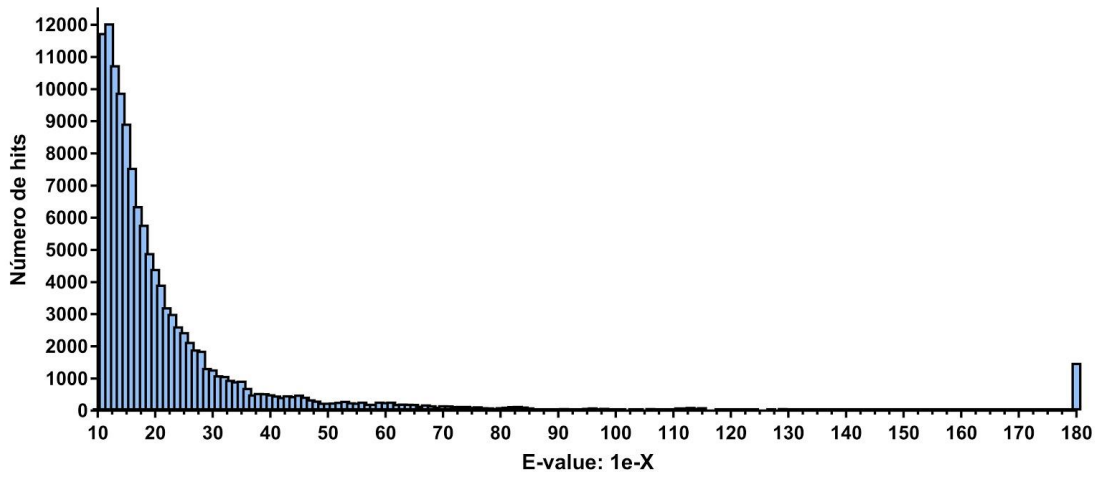
Contudo, a remoção de 88% das *reads* sequenciadas diminuiu consideravelmente o número de *reads* processadas, comprometendo os outros valores da métrica. *P. crispa* teve o menor tamanho médio e máximo de *contigs*, reflexo da baixa cobertura de sequenciamento [106]. Quanto maior o número de *reads*, mais completa seria a reconstituição dos transcritos.

Apesar dos dados de N50 serem apresentados e estarem presentes na maioria dos artigos, a tendência é que caia em desuso devido à pouca informação agregada na montagem de transcriptomas, sendo uma métrica importante para a montagem de genomas [106].

A busca de sequências dentro do banco de dados de proteínas não-redundantes do NCBI teve 52,19% dos *contigs* com pelo menos um BLAST *hit*. Esse valor, quando comparado aos de *Chlorella minutissima* e *Trebouxia gelatinosa* com 53,5% e 53,6%, respectivamente, não apresentou diferenças significativas. Isto indica a baixa quantidade de informação de espécies relacionadas nos bancos de dados. Além disso, podem compreender um grupo de transcritos característicos da classe, de grande importância biológica, mas que computacionalmente ainda não se pode inferir as funções [83].

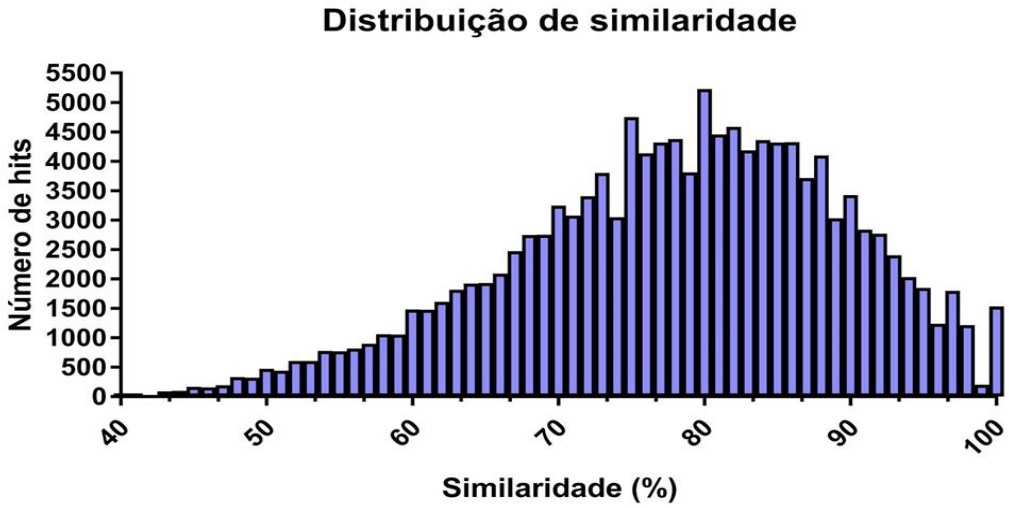
A Figura 9 demonstra a distribuição dos *hits* de acordo com seus E-*values*. De todos os *hits*, 62,93% estão entre 1e-10 a 1e-20 e apenas 10% eram menores que 1e-50. Isto é compreensível devido à baixa média de tamanho dos *contigs*, tendo em vista que maiores transcritos são mais sujeitos a conter domínios proteicos, que refletem em um melhor alinhamento [107].

Figura 9 - Distribuição de *hits* de acordo com seu *e-value*.



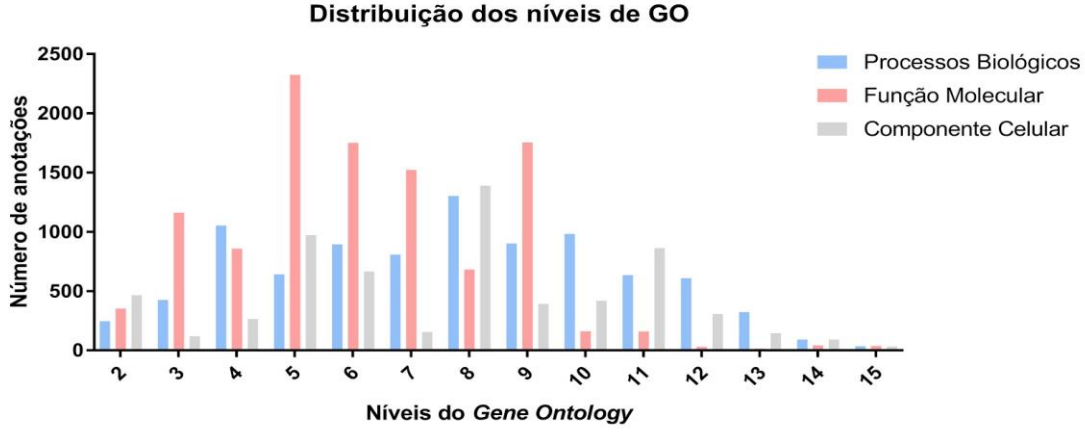
A distribuição da porcentagem de similaridade entre as sequências variou bastante, sendo a mínima de 40%, com maior concentração de *hits* entre 75 e 90% (Figura 10). Destaca-se que 1.534 *hits* que tiveram similaridade de 100%.

Figura 10 - Distribuição da porcentagem de similaridade dos *hits*.



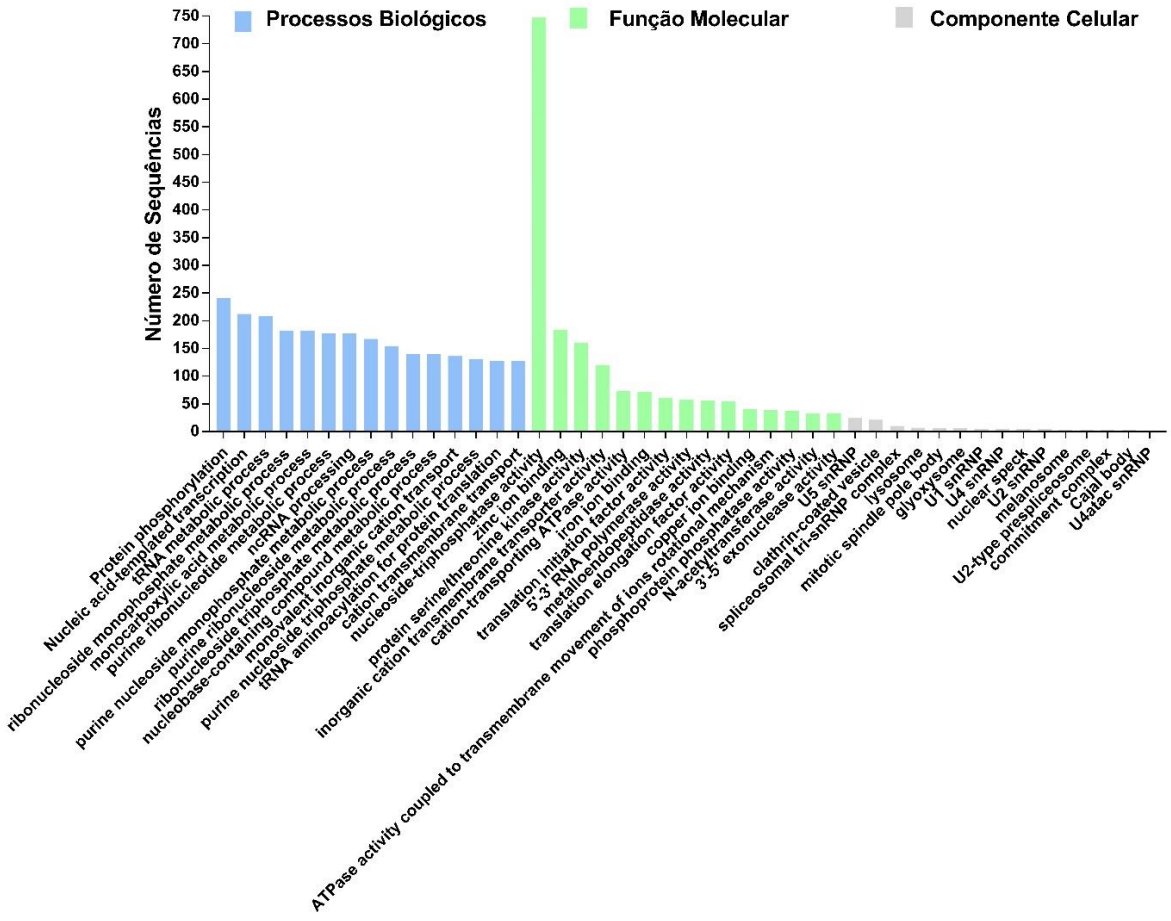
Os *contigs* com BLAST *hits* foram submetidos a mapeamento para associação de termos do GO. Destes, 7.009 foram mapeados com ao menos um termo do GO, sendo que 3.639 tiveram termos da categoria CC associados, 5.343 da categoria FM e 4.782 da categoria PB. A Figura 11 demonstra a distribuição do número de anotações ao longo de cada um dos níveis, sendo o nível 2 o mais geral e o nível 15 o mais específico, para as três diferentes categorias.

Figura 11 - Distribuição das anotações em todos os níveis do GO para as três categorias: Processos Biológicos, Função Molecular e Componente Celular.



A distribuição dos termos no nível 7, de caráter intermediário, demonstra que a fosforilação proteica foi o PB com maior número de anotações (Figura 12). A fosforilação proteica é uma modificação pós-traducional cuja dinâmica é essencial para a regulação de diversos processos celulares. Entre os já demonstrados, em plantas ou algas, pode-se citar as regulações de respostas ao estresse oxidativo [108, 109] e osmótico [110, 111], a seca [112] e o frio [113, 114]. Estudos recentes em murinos demonstraram que a fosforilação proteica está associada a capacidade anticongelante dos embriões [115].

Figura 12 - Top 15 termos do nível 7 do GO para as três categorias: Processos Biológicos, Função Molecular e Componente Celular.



Os três principais termos da categoria de FM foram os de Nucleosídeo-Trifosfatase, ligação ao íon zinco e a atividade serina/treonina quinase. A função molecular de ligação ao zinco está relacionada a formação do domínio proteico *zinc finger*, que também está associado a resposta à diversos estresses como desidratação, frio e salinidade [116]. Já os outros dois estão associados ao PB de fosforilação proteica.

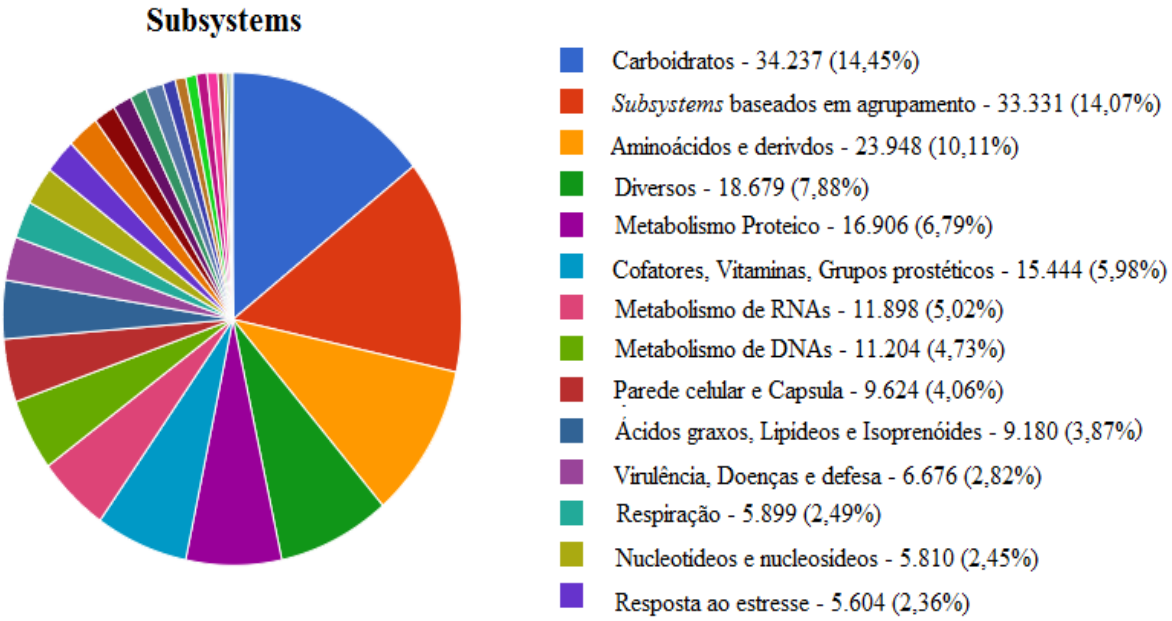
Os CC com maior número de anotações no nível 7 foram o complexo ribonucleico U5 snRNP e vesículas revestidas com clatrina, associados a *small nuclear* RNAs e o processo endocitose, respectivamente.

Os dados da montagem e anotação do transcriptoma de *P. crispa* foram submetidos ao periódico online GigaScience (ISSN 2047-217X), o artigo é apresentado no Apêndice A.

4.2 Bioprospecção de Proteínas de Ligação ao Gelo

Com o transcriptoma montado e anotado, o segundo passo foi o processo de busca por IBPs. No primeiro momento, a contaminação com transcritos de outras espécies não era desejada e por isso estas sequências foram removidas. No entanto, a microbiota associada a *P. crispa* está exposta às mesmas condições ambientais extremas e também devem possuir mecanismos que permitam a sua sobrevivência, podendo haver até mesmo relações mutualísticas com o hospedeiro. A anotação por MG-RAST aplicando o vocabulário *Subsystem* apresenta que 5.604 *hits* estão associados a resposta ao estresse (Figura 13).

Figura 13 - Distribuição dos *hits* anotadas pelo MG-RAST para as diferentes categorias do vocabulário *Subsystem*.



Desta forma, a microbiota associada também possui potencial para a presença de IBPs. Assim, todos os *contigs* gerados na metamontagem foram submetidos à busca por BLASTX contra o banco de dados de local, construído com os diferentes grupos de IBPs. A busca

identificou 127 sequências com ao menos um BLAST *hit*, com E-*value* variando entre 1e-10 e 1e-111. A origem das proteínas apresentou grande variação, contendo sequências oriundas de plantas, algas, bactérias e fungos, sendo o maior número de *hits* com proteínas bacterianas. A Tabela 4 apresenta 10 *contigs* com os menores *E-values* de representantes de cada um dos grupos citados anteriormente. Os dados completos são apresentados no Anexo A.

Tabela 4 - Top 10 *hits* BLASTX de algas, bactérias, plantas e fungos do banco de IBPs.

Descrição	***E-value***	**Similaridade (%)**	**Espécie**
Ice antifreeze protein	8,44E-23	57,4	*Flagilariopsis curta*
Ice antifreeze protein	3,18E-23	92,6	*Flagilariopsis cylindrus*
Ice-binding Protein	2,00E-80	79,5	Bactéria NI
Antifreeze protein	1,45E-101	74,1	*Chitinophaga eiseniae*
Antifreeze protein	1,02E-111	78,9	*Neisseria bacilliformis*
29 kDa chitinase-like thermal hysteresis	8,98E-59	60,6	*Solanum dulcamara*
29 kDa chitinase-like thermal hysteresis	1,38E-59	60,6	*Solanum dulcamara*
29 kDa chitinase-like thermal hysteresis	4,88E-62	60,6	*Solanum dulcamara*
Antifreeze protein	4,20E-29	74,2	*Fibularhizoctonia sp.*
Antifreeze protein	9,65E-37	74,2	*Stereum hirsutum*

Legenda: NI – Não identificada.

Os 127 *contigs* foram submetidos para a análise de presença de ORFs e obtenção da sequência aminoacídica. Apenas 11 *contigs* apresentaram ORF completa e 45 não continham ORFs. Isto é um reflexo da baixa cobertura de sequenciamento, principalmente levando-se em consideração que o experimento não foi planejado para a identificação de uma quantidade tão grande de transcritos. As sequências aminoacídicas de *contigs* com ORFs, completas ou parciais, foram submetidas a modelagem pelo *web-server* Phyre 2.

Ao todo, 81 modelagens foram realizadas, sendo que em 27 destas a abordagem *ab initio* foi aplicada. É possível que estas 27 sequências representem IBPs, contudo, devido à baixa quantidade de estruturas tridimensionais conhecidas e a não confiabilidade dos

algoritmos *ab initio* é equivocada qualquer inferência que possa ser feita atualmente e, por isso, estas sequências não foram consideradas para análises posteriores.

Entre as modelagens comparativas, 38 tiveram como molde principal proteínas de diversas funções como imunoglobulinas, fibronectinas e moléculas de sinalização celular. Apesar da conhecida dualidade funcional de IBPs [36], ainda não há estudos que indiquem qualquer associação direta das mesmas com as proteínas que foram utilizadas como molde.

Por fim, 17 *contigs* foram modelados utilizando como molde IBPs ou proteínas com clara associação de dualidade demonstrada na literatura, como as quitinases em plantas [36]. Na Tabela 5 os dados de confiança na modelagem, identidade em relação à sequência molde, proteína molde e espécie de origem são apresentadas juntamente com o resultado do BLASTX. A confiabilidade foi descrita como a porcentagem de resíduos que foram modelados com precisão > 90%, sendo que alta confiabilidade significava mais de 90% dos resíduos, confiabilidade média entre 50 e 90% e baixa confiabilidade < 50%.

A IBP de *Flavobacterium frigoris*, uma bactéria Antártica, foi o principal molde utilizado em 7 modelagens de alta e média confiabilidade, com uma das sequências chegando a identidade de 58% (Tabela 5 e Figura 14). O enovelamento B-hélice, identificado através de estudos de cristalografia de raio-x, mostrou-se essencial para a interação os cristais de gelo [117]. Já a IBP da bactéria marinha Antártica *Colwellia sp*. foi molde principal de outras 4 proteínas (Tabela 5 e Figura 15), de estrutura muito similar à de *F. frigoris*, sendo a grande diferença as regiões *motifs*.

Estudos de interação com a proteína de *Colwellia sp*. demonstraram que, ao contrário de *F. frigoris*, esta não possui sequência *motif* específica e somente as mutações que alteram a sua estrutura tridimensional, principalmente em treoninas presentes nas folhas B, causam perda de atividade [118]. É importante acrescentar que todas as proteínas que tiveram a IBP de *F. frigoris* como molde principal, utilizaram a IBP de *Colwellia sp.* como principal molde secundário, e vice-versa.

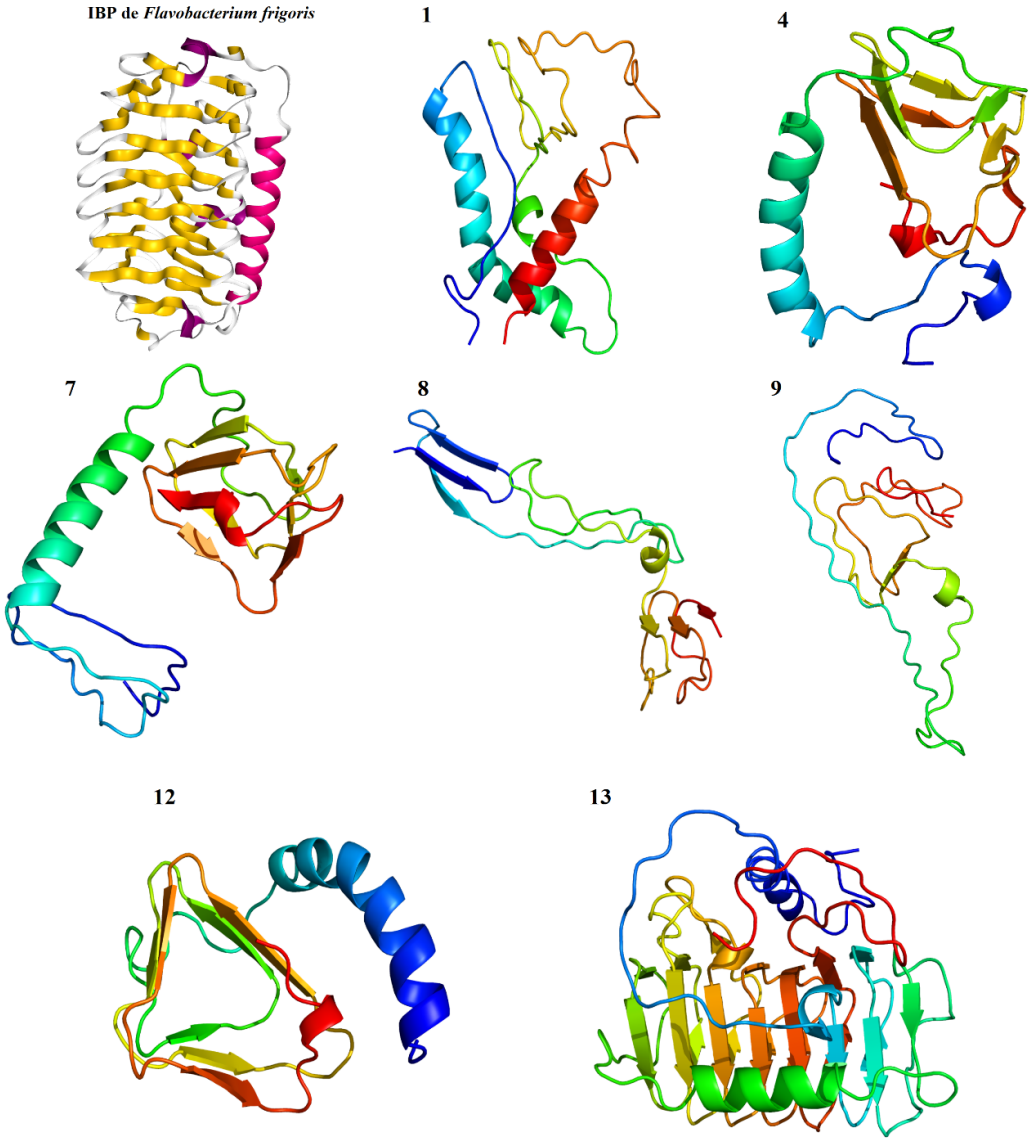
Este enovelamento foi encontrado em algumas proteínas modeladas, sendo que o tamanho da ORF refletiu em maior (modelos 3, 6, 10, 11 e 13) ou menor (1, 4, 7 e 13) similaridade. Este grupo ainda não possui um domínio proteico específico, então o próximo passo natural seria a comparação do alinhamento de sequências para avaliar a possível conservação da sequência *motif*. Entretanto, devido ao caráter parcial de todas as sequências modeladas, algumas destas não continham a região onde poderiam ter as sequências *motifs*. As proteínas 3, 6, 10, 11 e 13 foram as únicas que continham a região da sequência *motif*, mas assim como a IBP de *Colwellia sp.*, não apresentaram conservação dos resíduos T-G/A-X-T.

Tabela 5 - BLAST *hit* e dados de modelagem por Phyre 2 de transcritos que tiveram IBPs como molde principal.

ID	Descrição	E-*value*	Sim (%)	Espécie	Confiança	ID molde	Molde	Origem da proteína molde
1	*Ice-binding protein*	2,25E-14	67,60%	Bactéria NI	Média	35,00%	IBP	*Flavobacterium frigoris*
2	*antifreeze protein*	1,05E-15	70,90%	*Stigmatella aurantiaca*	Baixa	31,00%	IBP	*Leucosporidium sp.*
3	*Antifreeze*	1,51E-17	57,50%	*Fibularhizoctonia sp.*	Alta	30,00%	IBP	*Colwellia sp.*
4	*ice antifreeze*	2,11E-19	59,60%	*Fragilariopsis curta*	Alta	43,00%	IBP	*Flavobacterium frigoris*
5	*Antifreeze*	7,18E-20	56,40%	*Fibularhizoctonia sp.*	Alta	31,00%	IBP	*Leucosporidium sp.*
6	*Antifreeze*	5,13E-20	53,20%	*Fibularhizoctonia sp.*	Alta	45,00%	IBP	*Colwellia sp.*
7	*ice antifreeze*	8,44E-23	57,40%	*Fragilariopsis curta*	Alta	48,00%	IBP	*Flavobacterium frigoris*
8	*Ice-binding protein*	1,30E-24	76,70%	Bactéria NI	Média	36,00%	IBP	*Flavobacterium frigoris*
9	*Ice-binding protein*	3,11E-26	74,80%	Bactéria NI	Média	19,00%	IBP	*Flavobacterium frigoris*
10	*Antifreeze*	4,20E-29	74,40%	*Fibularhizoctonia sp.*	Média	49,00%	IBP	*Colwellia sp.*
11	*Ice-binding protein*	5,59E-36	67,20%	Bactéria NI	Média	33,00%	IBP	*Colwellia sp.*
12	*Antifreeze*	9,65E-37	74,20%	*Stereum hirsutum*	Alta	58,00%	IBP	*Flavobacterium frigoris*
13	*Antifreeze*	4,52E-39	69,80%	*Fimbriimonas ginsengisoli*	Média	39,00%	IBP	*Flavobacterium frigoris*
14	*Antifreeze*	1,69E-47	67,00%	*Stigmatella aurantiaca*	Baixa	35,00%	IBP	*Typhula ishikariensis*
15	*29 kDa chitinase-like thermal hysteresis*	8,98E-59	60,60%	*Solanum dulcamara*	Alta	49,00%	Quitinase	*Oryza sativa*
16	*29 kDa chitinase-like thermal hysteresis*	1,38E-59	60,60%	*Solanum dulcamara*	Alta	46,00%	Quitinase	*Oryza sativa*
17	*29 kDa chitinase-like thermal hysteresis*	4,88E-62	60,60%	*Solanum dulcamara*	Alta	45,00%	Quitinase	*Oryza sativa*

Legenda: ID – Identidade; Sim – Similaridade; NI – Não identificada. Áreas cinzas indicam dados de modelagem e brancas os BLAST *hits*.

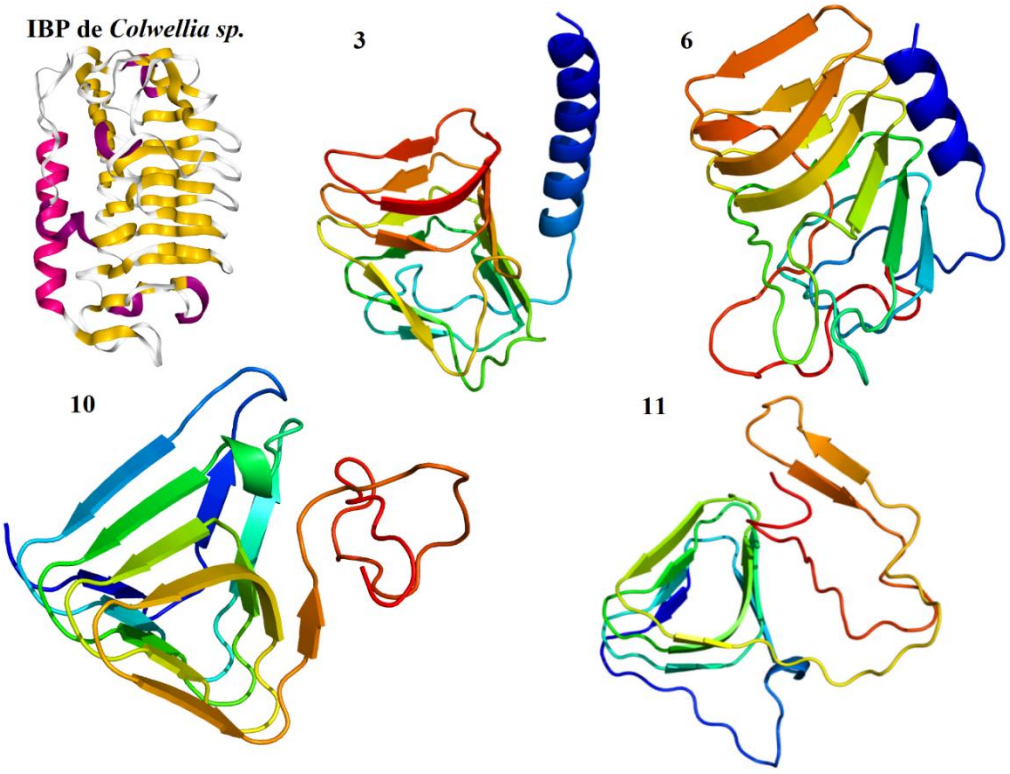
Figura 14 – Predição de estrutura tridimensional de proteínas que utilizaram como molde a IBP de *Flavobacterium frigoris*. As cores e as setas indicam a orientação, sendo o azul a região amino-terminal e o gradiente de cores até o vermelho a região carboxi-terminal. Os números de cada modelo estão associados a identificação adotada na Tabela 5.



Proteínas dos fungos *Typhula ishikariensis* [119] e *Leucosporidium sp.* [120] também foram utilizadas como moldes, porém as precisões da estrutura de ambas foram baixas (Figura 16).

Por fim, três isoformas foram identificadas com identidade entre 45 e 49% a estrutura tridimensional semelhante à quitinase de *Oryza sativa* (Figura 17). O alinhamento de sequências com outras quitinases confirmou a presença do domínio da família 19 das Glicohidrolases.

Figura 15 – Predição de estrutura tridimensional de proteínas que utilizaram como molde a IBP de *Colwellia sp*. As cores e as setas indicam a orientação, sendo o azul a região amino-terminal e o gradiente de cores até o vermelho a região carboxi-terminal. Os números de cada modelo estão associados a identificação adotada na Tabela 5.



Algumas quitinases de plantas já foram identificas com dualidade de função, possuindo atividades de IBPs [36, 121-123]. Isto porque proteínas associadas a respostas a patógenos, como quitinases e proteínas tipo-taumatina, precisam resistir a baixas temperaturas para proteger as plantas durante o inverno [124]. Assim, estas proteínas também desenvolveram a capacidade de adsorção dos cristais de gelo, apresentando atividade anticongelante. Estes três transcritos são os únicos pertencentes ao transcriptoma de *P. crispa* que foram identificados com o potencial de atividade IBP.

Sendo assim, 8 transcritos apresentaram maior potencial, sendo três deles oriundos diretamente de *P. crispa* e o restante a partir da microbiota.

Figura 16 – Predição de estrutura tridimensional de proteínas que utilizaram como molde a IBP de (A) *Leucosporidium sp.* e (B) *Typhula ishikariensis*. As cores e as setas indicam a orientação, sendo o azul a região amino-terminal e o gradiente de cores até o vermelho a região carboxi-terminal. Os números de cada modelo estão associados a identificação adotada na Tabela 5.

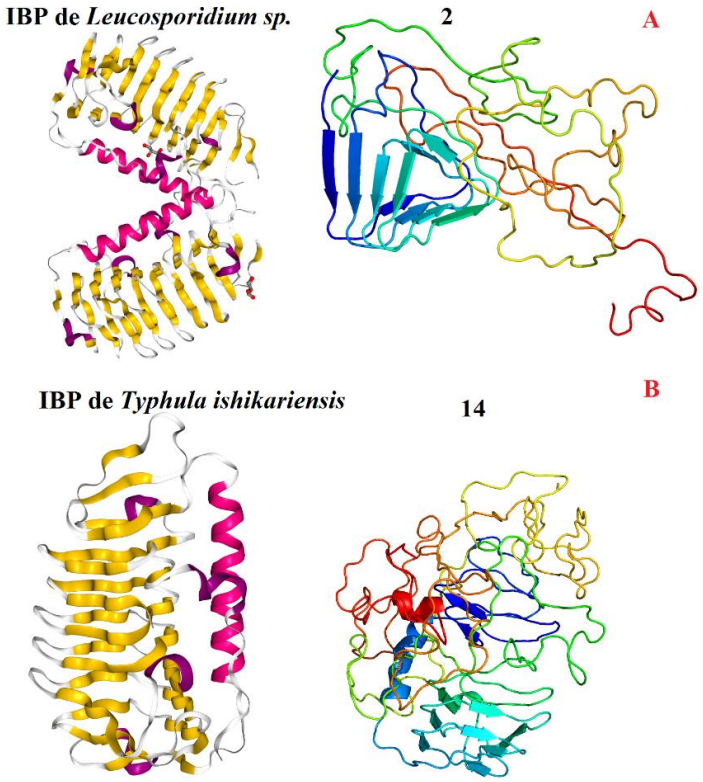
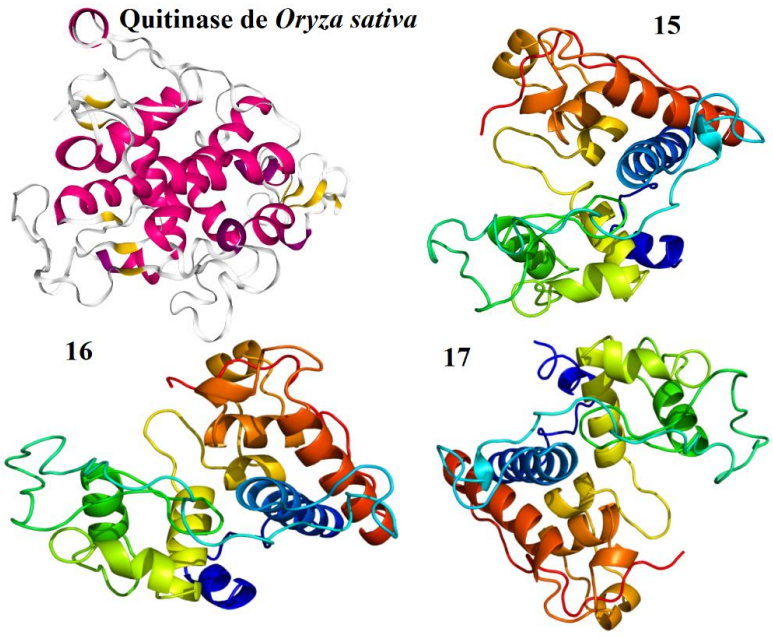


Figura 17 – Predição de estrutura tridimensional de proteínas que utilizaram como molde quitinase de *Oryza sativa*. As cores e as setas indicam a orientação, sendo o azul a região amino-terminal e o gradiente de cores até o vermelho a região carboxi-terminal. Os números de cada modelo estão associados a identificação adotada na Tabela 5.



5 CONCLUSÃO E PERSPECTIVAS

Em relação aos objetivos específicos apresentados é possível concluir que:

- O RNA de *P. crispa* foi extraído e sequenciado, porém houve enorme contaminação com a microbiota;
- As análises de validação e qualidade permitiram a remoção de *reads* de baixa qualidade e contaminantes, trazendo maior confiabilidade aos dados;
- De acordo com a comparação com as métricas de outras algas da mesma classe possível concluir que o transcriptoma foi montado de maneira satisfatória;
- 52,19% dos transcritos foram identificados e anotados funcionalmente;
- A comparação dos dados de organismos da classe *Trebouxiophyceae* demonstraram certa uniformidade e estão de acordo com os valores apresentados neste trabalho;
- O alinhamento de sequências inicialmente indicou 127 possíveis IBPs;
- A modelagem da estrutura tridimensional permitiu a identificação de 17 possíveis IBPs, sendo que 8 delas apresentaram o maior potencial;

Este trabalho abre perspectiva para estudos posteriores, sendo o próximo passo *in silico* a simulação de Dinâmica Molecular para avaliar o comportamento destas estruturas proteicas em diferentes temperaturas e sua interação com os cristais de gelo. Também se espera amplificar, utilizando primers degenerados, e sequenciar os genes de maior potencial, para identificação da sequência completa. Ademais, futuramente espera-se expressar e purificar para confirmação de suas atividades experimentalmente.

6 REFERÊNCIAS

1. PEDRINI, A.D.G. **Macroalgas**: Uma introdução à taxonomia. 1. ed., Rio de Janeiro: Techincal Books, 2010.

2. LEWIS, L.A., HALL, J.D., ZECHMAN, F.W. **Green Algae**. John Wiley & Sons, Ltd: Encyclopedia of Life Sciences, 2011. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0000333.pub2/full>. Acesso em: 10 abr. 2017, 13:45:30.

3. CHAPMAN, R.L. Algae: the world's most important "plants"- an introduction. **Mitigation and Adaptation Strategies for Global Change**, v. 18 (1), p. 5-12, 2013.

4. LUTZONI, F., PAGEL, M., REEB, V. Major fungal lineages are derived from lichen symbiotic ancestors. **Nature**, v. 411 (6840), p. 937-940, 2001.

5. RAMANAN, R., KIM, B.H., CHO, D.H., OH, H.M., KIM, H.S. Algae–bacteria interactions: Evolution, ecology and emerging applications. **Biotechnology Advances**, v. 34 (1), p. 14-29, 2016.

6. KEELING, P.J. Diversity and evolutionary history of plastids and their hosts. **American Journal of Botany**, v. 91 (10), p. 1481-1493, 2004.

7. LEWIS, L.A., MCCOURT, R.M. Green algae and the origin of land plants. **American Journal of Botany**, v. 91 (10), p. 1535-1556, 2004.

8. UMEN, J.G. Green Algae and the Origins of Multicellularity in the Plant Kingdom. **Cold Spring Harbor Perspectives in Biology**, v. 6 (11), p. 1-27, 2014.

9. YOON, H.S., HACKETT, J.D., CINIGLIA, C., PINTO, G., BHATTACHARYA, D. A Molecular Timeline for the Origin of Photosynthetic Eukaryotes. **Molecular Biology and Evolution**, v. 21 (5), p. 809-818, 2004.

10. TANG, H., CHEN, M., GARCIA, M.E.D., ABUNASSER, N., NG, K.Y.S., SALLEY, S.O. Culture of microalgae *Chlorella minutissima* for biodiesel feedstock production. **Biotechnology and Bioengineering**, v. 108 (10), p. 2280-2287, 2011.

11. SØRENSEN, I., ROSE, J.K.C., DOYLE, J.J., DOMOZYCH, D.S., WILLATS, W.G.T. The *Charophycean* green algae as model systems to study plant cell walls and other evolutionary adaptations that gave rise to land plants. **Plant Signaling & Behavior**, v. 7 (1), p. 1-3, 2012.

12. KOVÁCIK, L., PEREIRA, A.B. Green alga *Prasiola crispa* and its lichenized form *Mastodia tesselata* in Antarctic environment: General aspects. **Nova Hedwigia**, v. 123, p. 465-478, 2001.

13. MONIZ, M.B., RINDI, F., NOVIS, P.M., BROADY, P.A., GUIRY, M.D. Molecular Phylogeny of Antarctic *Prasiola* (*Prasiolales, Treboxiophyceae*) reveals extensive cryptic diversity. **Journal of Phycology**, v. 48 (4), p. 940-955, 2012.

14. CARVALHO, E.L., WALLAU, G.L., RANGEL, D.L., MACHADO, L.C., PEREIRA, A.B., VICTORIA, F.D.C., BOLDO, J.T., PINTO, P.M. Phylogenetic positioning of the Antarctic alga *Prasiola crispa* (*Trebouxiophyceae*) using organellar genomes and their structural analysis. **Journal of Phycology**, v. 53 (4), p. 908-915, 2017.

15. SHERWOOD, A.R., GARBARY, D.J., SHEATH, R.G. Assessing the phylogenetic position of the *Prasiolales* (*Chlorophyta*) using rbcL and 18S rRNA gene sequence data. **Phycologia**, v. 39 (2), p. 139-146, 2000.

16. PÉREZ-ORTEGA, S., RÍOS, A.D.L., CRESPO, A., SANCHO, L.G. Symbiotic lifestyle and phylogenetic relationships of the bionts of *Mastodia tessellata* (Ascomycota, incertae sedis). **American Journal of Botany**, v. 97 (5), p. 738-752, 2010.

17. CONVEY, P. **Encyclopedia of the Antarctic**. 1ed, Nova Iorque: Routledge, 2006.

18. GRAHAM, L.E., GRAHAM, J.M., WILCOX, L.W. **Algae**. 2. ed., São Francisco: Pearson Education, 2009.

19. HINCE, B. **The Antarctic Dictionary**: A Complete Guide to Antarctic English. CSIRO Publishing, 2000. Disponível em: < http://www.publish.csiro.au/book/2536/>. Acesso em: 17 abr. 2017, 16:23:00.

20. MARTÍNEZ-ROSALES, C., FULLANA, N., MUSTO, H., CASTRO-SOWINSKI, S. Antarctic DNA moving forward: genomic plasticity and biotechnological potential. **FEMS Microbiology Letters**, v. 331 (1), p. 1-9, 2012.

21. CONVEY, P., GIBSON, J.A.E., HILLENBRAND, C.D., HODGSON, D.A., PUGH, P.J.A., SMELLIE, J.L., STEVENS, M.I. Antarctic terrestrial life – challenging the history of the frozen continent? **Biological Reviews**, v. 83 (2), p. 103-117, 2008.

22. KUTTIPPURATH, J., NAIR, P.J. The signs of Antarctic ozone hole recovery. **Scientific Reports**, v. 7 (1), 2017.

23. MARIZCURRENA, J.J., MOREL, M.A., BRANA, V., MORALES, D., MARTINEZ-LOPEZ, W., CASTRO-SOWINSKI, S. Searching for novel photolyases in UVC-resistant Antarctic bacteria. **Extremophiles**, v. 21 (2), p. 409-418, 2017.

24. FURBINO, L.E., GODINHO, V.M., SANTIAGO, I.F., PELLIZARI, F.M., ALVES, T.M.A., ZANI, C.L., JUNIOR, P.A.S., ROMANHA, A.J., CARVALHO, A.G.O., GIL, L.H.V.G., ROSA, C.A., MINNIS, A.M., ROSA, L.H. Diversity Patterns, Ecology and Biological Activities of Fungal Communities Associated with the Endemic Macroalgae Across the Antarctic Peninsula. **Microbial Ecology**, v. 67 (4), p. 775-787, 2014.

25. ZEMOLIN, A.P.P., CRUZ, L.C., PAULA, M.T., PEREIRA, B.K., ALBUQUERQUE, M.P., VICTORIA, F.C., PEREIRA, A.B., POSSER, T., FRANCO, J.L. Toxicity Induced by *Prasiola crispa* to Fruit Fly *Drosophila melanogaster* and Cockroach

Nauphoeta cinerea: Evidence for Bioinsecticide Action. **Journal of Toxicology and Environmental Health,** v. 77 (1), p. 115-124, 2014.

26. MARINHO, R.S.S., RAMOS, C.J.B., LEITE, J.P.G., TEIXEIRA, V.L., PAIXÃO, I.C.N.P., BELO, C.A.D., PEREIRA, A.B., PINTO, A.M.V. Antiviral activity of 7-keto-stigmasterol obtained from green Antarctic algae *Prasiola crispa* against equine herpesvirus 1. **Journal of Applied Phycology**, v. 29 (1), p. 555-562, 2017.

27. USTUN, N.S., TURHAN, S. Antifreeze Proteins: Characteristics, Function, Mechanism of Action, Sources and Application to Foods. **Journal of Food Processing and Preservation**, v. 39 (6), p. 3189-3197, 2015.

28. VENKETESH, S., DAYANANDA, C. Properties, Potentials, and Prospects of Antifreeze Proteins. **Critical Reviews in Biotechnology**, v. 28 (1), p. 57-82, 2008.

29. DEVRIES, A.L. Glycoproteins as Biological Antifreeze Agents in Antarctic Fishes. **Science**, v. 172 (3988), p. 1152-1155, 1971.

30. DEVRIES, A.L., WOHLSCHLAG, D.E. Freezing Resistance in Some Antarctic Fishes. **Science**, v. 163 (3871), p. 1073-1075, 1969.

31. DAVIES, P.L. Ice-binding proteins: a remarkable diversity of structures for stopping and starting ice growth. **Trends in Biochemical Sciences**, v. 39 (11), p. 548-555, 2014.

32. CHATTOPADHYAY, M.K. Antifreeze proteins of bacteria. **Resonance**, v. 12 (12), p. 25-30, 2007.

33. XIAO, N., INABA, S., TOJO, M., DEGAWA, Y., FUJIU, S., KUDOH, S., HOSHINO, T. Antifreeze activities of various fungi and Stramenopila isolated from Antarctica. **North American Fungi**, v. 5, p. 215-220, 2010.

34. DUMAN, J.G. Animal ice-binding (antifreeze) proteins and glycolipids: an overview with emphasis on physiological function. **The Journal of Experimental Biology**, v. 218 (12), p. 1846-1855, 2015.

35. LI JIN-YAO, M.J., ZHANG, F.C. Recent Advances in Research of Antifreeze Proteins. **Chinese Journal of Biochemistry and Molecular Biology**, v. 21 (6), p. 717-722, 2005.

36. GUPTA, R., DESWAL, R. Antifreeze proteins enable plants to survive in freezing conditions. **Journal of Biosciences**, v. 39 (5), p. 931-944, 2014.

37. BAYER-GIRALDI, M., UHLIG, C., JOHN, U., MOCK, T., VALENTIN, K. Antifreeze proteins in polar sea ice diatoms: diversity and gene expression in the genus *Fragilariopsis*. **Environmental Microbiology**, v. 12 (4), p. 1041-1052, 2010.

38. JANECH, M.G., KRELL, A., MOCK, T., KANG, J.S., RAYMOND, J.A. Ice-binding proteins from sea ice diatoms (*Bacillariophyceae*). **Journal of Phycology**, v. 42 (2), p. 410-416, 2006.

39. RAYMOND, J.A. The ice-binding proteins of a snow alga, *Chloromonas brevispina*: probable acquisition by horizontal gene transfer. **Extremophiles**, v. 18 (6), p. 987-994, 2014.

40. RAYMOND, J.A., JANECH, M.G., FRITSEN, C.H. Novel Ice-binding proteins from a psychrophilic Antarctic alga (*Chlamydomonadaceae*, *Chlorophyceae*). **Journal of Phycology**, v. 45 (1), p. 130-136, 2009.

41. RAYMOND, J.A., KIM, H.J. Possible Role of Horizontal Gene Transfer in the Colonization of Sea Ice by Algae. **PLoS ONE**, v. 7 (5:35968), p. 1-9, 2012.

42. RAYMOND, J.A., MORGAN-KISS, R. Separate Origins of Ice-Binding Proteins in Antarctic Chlamydomonas Species. **PLoS ONE**, v. 8 (3:59186), p. 1-6, 2013.

43. RANDY CHI FAI, C., TZI BUN, N., JACK HO, W. Antifreeze Proteins from Diverse Organisms and their Applications: An Overview. **Current Protein & Peptide Science**, v. 18 (3), p. 262-283, 2017.

44. KUIPER, M.J., MORTON, C.J., ABRAHAM, S.E., GRAY-WEALE, A. The biological function of an insect antifreeze protein simulated by molecular dynamics. **eLife**, v. 4 (5142), p. 1-14, 2015.

45. DOLEV, M.B., BRASLAVSKY, I., DAVIES, P.L. Ice-Binding Proteins and Their Function. **Annual Review of Biochemistry**, v. 85 (1), p. 515-542, 2016.

46. VOETS, I.K. From ice-binding proteins to bio-inspired antifreeze materials. **Soft Matter**, v. 13 (28), p. 4808-4823, 2017.

47. RAYMOND, J.A., DEVRIES, A.L. Adsorption inhibition as a mechanism of freezing resistance in polar fishes. **Proceedings of the National Academy of Sciences of United States of America**, v. 74 (6), p. 2589-2593, 1977.

48. RAYMOND, J.A., FRITSEN, C., SHEN, K. An ice-binding protein from an Antarctic sea ice bacterium. **FEMS Microbiology Ecology**, v. 61 (2), p. 214-221, 2007.

49. GUO, S., GARNHAM, C.P., WHITNEY, J.C., GRAHAM, L.A., DAVIES, P.L. Re-Evaluation of a Bacterial Antifreeze Protein as an Adhesin with Ice-Binding Activity. **PLoS ONE**, v. 7 (11: 48805), p. 1-10, 2012.

50. HASSAS-ROUDSARI, M., GOFF, H.D. Ice structuring proteins from plants: Mechanism of action and food application. **Food Research International**, v. 46 (1), p. 425-436, 2012.

51. DUMAN, J.G., WISNIEWSKI, M.J. The use of antifreeze proteins for frost protection in sensitive crop plants. **Environmental and Experimental Botany**, v. 106, p. 60-69, 2014.

52. BROCKBANK, K.G.M., CAMPBELL, L.H., GREENE, E.D., BROCKBANK, M.C.G., DUMAN, J.G. Lessons from nature for preservation of mammalian cells,

tissues, and organs. **In Vitro Cellular & Developmental Biology - Animal**, v. 47 (3), p. 210-217, 2011.

53. PERFELDT, C.M., CHUA, P.C., DARABOINA, N., FRIIS, D., KRISTIANSEN, E., RAMLØV, H., WOODLEY, J.M., KELLAND, M.A., VON SOLMS, N. Inhibition of Gas Hydrate Nucleation and Growth: Efficacy of an Antifreeze Protein from the Longhorn Beetle *Rhagium mordax*. **Energy & Fuels**, v. 28 (6), p. 3666-3672, 2014.

54. LU, X., LI, J., YANG, J., LIU, X., MA, J. De novo transcriptome of the desert beetle *Microdera punctipennis* (Coleoptera: *Tenebrionidae*) using illumina RNA-seq technology. **Molecular Biology Reports**, v. 41 (11), p. 7293-7303, 2014.

55. WANG, L., SI, Y., DEDOW, L.K., SHAO, Y., LIU, P., BRUTNELL, T.P. A Low-Cost Library Construction Protocol and Data Analysis Pipeline for Illumina-Based Strand-Specific Multiplex RNA-Seq. **PLoS ONE**, v. 6 (10:26426), p. 1-12, 2011.

56. BABA, M., IOKI, M., NAKAJIMA, N., SHIRAIWA, Y., WATANABE, M.M. Transcriptome analysis of an oil-rich race A strain of *Botryococcus braunii* (BOT-88-2) by *de novo* assembly of pyrosequencing cDNA reads. **Bioresource Technology**, v. 109, p. 282-286, 2012.

57. CARNIEL, F.C., GERDOL, M., MONTAGNER, A., BANCHI, E., DE MORO, G., MANFRIN, C., MUGGIA, L., PALLAVICINI, A., TRETIACH, M. New features of desiccation tolerance in the lichen photobiont *Trebouxia gelatinosa* are revealed by a transcriptomic approach. **Plant Molecular Biology**, v. 91 (3), p. 319-339, 2016.

58. LI, L., ZHANG, G., WANG, Q. *De novo* transcriptomic analysis of *Chlorella sorokiniana* reveals differential genes expression in photosynthetic carbon fixation and lipid production. **BMC Microbiology**, v. 16 (1:223), p. 1-12, 2016.

59. PENG, H., WEI, D., CHEN, G., CHEN, F. Transcriptome analysis reveals global regulation in response to CO2 supplementation in oleaginous microalga *Coccomyxa subellipsoidea* C-169. **Biotechnology for Biofuels**, v. 9 (1:151), p. 1-17, 2016.

60. YU, M., YANG, S., LIN, X. *De-novo* assembly and characterization of *Chlorella minutissima* UTEX2341 transcriptome by paired-end sequencing and the identification of genes related to the biosynthesis of lipids for biodiesel. **Marine Genomics**, v. 25, p. 69-74, 2016.

61. SANGER, F., COULSON, A.R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. **Journal of Molecular Biology**, v. 94 (3), p. 441-448, 1975.

62. MAXAM, A.M., GILBERT, W. A new method for sequencing DNA. **Proceedings of the National Academy of Sciences of United States of America**, v. 74 (2), p. 560-564, 1977.

63. SANGER, F., NICKLEN, S., COULSON, A.R. DNA sequencing with chain-terminating inhibitors. **Proceedings of the National Academy of Sciences of United States of America**, v. 74 (12), p. 5463-5467, 1977.

64. LUCKEY, J.A., DROSSMAN, H., KOSTICHKA, A.J., MEAD, D.A., D'CUNHA, J., NORRIS, T.B., SMITH, L.M. High speed DNA sequencing by capillary electrophoresis. **Nucleic Acids Research**, v. 18 (15), p. 4417-4421, 1990.

65. HEATHER, J.M., CHAIN, B. The sequence of sequencers: The history of sequencing DNA. **Genomics**, v. 107 (1), p. 1-8, 2016.

66. RONAGHI, M., UHLÉN, M., NYRÉN, P. A Sequencing Method Based on Real-Time Pyrophosphate. **Science**, v. 281 (5375), p. 363-365, 1998.

67. VOELKERDING, K.V., DAMES, S.A., DURTSCHI, J.D. Next-Generation Sequencing: From Basic Research to Diagnostics. **Clinical Chemistry**, v. 55 (4), p. 641-658, 2009.

68. MARCHANT, A., MOUGEL, F., MENDONÇA, V., QUARTIER, M., JACQUIN-JOLY, E., DA ROSA, J.A., PETIT, E., HARRY, M. Comparing *de novo* and reference-based transcriptome assembly strategies by applying them to the blood-sucking bug *Rhodnius prolixus*. **Insect Biochemistry and Molecular Biology**, v. 69, p. 25-33, 2016.

69. LI, Z., CHEN, Y., MU, D., YUAN, J., SHI, Y., ZHANG, H., GAN, J., LI, N., HU, X., LIU, B., YANG, B., FAN, W. Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. **Briefings in Functional Genomics**, v. 11 (1), p. 25-37, 2012.

70. PEVZNER, P.A., TANG, H., WATERMAN, M.S. An Eulerian path approach to DNA fragment assembly. **Proceedings of the National Academy of Sciences of United States of America**, v. 98 (17), p. 9748-9753, 2001.

71. MYERS, E.W., SUTTON, G.G., DELCHER, A.L., DEW, I.M., FASULO, D.P., FLANIGAN, M.J., KRAVITZ, S.A., MOBARRY, C.M., REINERT, K.H.J., REMINGTON, K.A., ANSON, E.L., BOLANOS, R.A., CHOU, H.-H., JORDAN, C.M., HALPERN, A.L., LONARDI, S., BEASLEY, E.M., BRANDON, R.C., CHEN, L., DUNN, P.J., LAI, Z., LIANG, Y., NUSSKERN, D.R., ZHAN, M., ZHANG, Q., ZHENG, X., RUBIN, G.M., ADAMS, M.D., VENTER, J.C. A Whole-Genome Assembly of *Drosophila*. **Science**, v. 287 (5461), p. 2196-2204, 2000.

72. HUANG, X., MADAN, A. CAP3: A DNA Sequence Assembly Program. **Genome Research**, v. 9 (9), p. 868-877, 1999.

73. DE LA BASTIDE, M., MCCOMBIE, W.R. **Assembling Genomic DNA Sequences with PHRAP**. John Wiley & Sons: Current Protocols in Bioinformatics, 2002. Disponível em:
< http://onlinelibrary.wiley.com/doi/10.1002/0471250953.bi1104s17/abstract>.
Acesso em: 25 abr. 2017, 18:10:00.

74. COMPEAU, P.E.C., PEVZNER, P.A., TESLER, G. How to apply de Bruijn graphs to genome assembly. **Nature Biotechnology**, v. 29 (11), p. 987-991, 2011.

75. COMMINS, J., TOFT, C., FARES, M.A. Computational Biology Methods and Their Application to the Comparative Genomics of Endocellular Symbiotic Bacteria of Insects. **Biological Procedures Online**, v. 11, p. 52-78, 2009.

76. KUNDETI, V.K., RAJASEKARAN, S., DINH, H., VAUGHN, M., THAPAR, V. Efficient parallel and out of core algorithms for constructing large bi-directed de Bruijn graphs. **BMC Bioinformatics**, v. 11 (1:560), p. 1-12, 2010.

77. GRABHERR, M.G., HAAS, B.J., YASSOUR, M., LEVIN, J.Z., THOMPSON, D.A., AMIT, I., ADICONIS, X., FAN, L., RAYCHOWDHURY, R., ZENG, Q., CHEN, Z., MAUCELI, E., HACOHEN, N., GNIRKE, A., RHIND, N., DI PALMA, F., BIRREN, B.W., NUSBAUM, C., LINDBLAD-TOH, K., FRIEDMAN, N., REGEV, A. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. **Nature Biotechnology**, v. 29 (7), p. 644-652, 2011.

78. ROBERTSON, G., SCHEIN, J., CHIU, R., CORBETT, R., FIELD, M., JACKMAN, S.D., MUNGALL, K., LEE, S., OKADA, H.M., QIAN, J.Q., GRIFFITH, M., RAYMOND, A., THIESSEN, N., CEZARD, T., BUTTERFIELD, Y.S., NEWSOME, R., CHAN, S.K., SHE, R., VARHOL, R., KAMOH, B., PRABHU, A.-L., TAM, A., ZHAO, Y., MOORE, R.A., HIRST, M., MARRA, M.A., JONES, S.J.M., HOODLESS, P.A., BIROL, I. *De novo* assembly and analysis of RNA-seq data. **Nature Methods**, v. 7 (11), p. 909-912, 2010.

79. XIE, Y., WU, G., TANG, J., LUO, R., PATTERSON, J., LIU, S., HUANG, W., HE, G., GU, S., LI, S., ZHOU, X., LAM, T.-W., LI, Y., XU, X., WONG, G.K.-S., WANG, J. SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. **Bioinformatics**, v. 30 (12), p. 1660-1666, 2014.

80. SCHULZ, M.H., ZERBINO, D.R., VINGRON, M., BIRNEY, E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. **Bioinformatics**, v. 28 (8), p. 1086-1092, 2012.

81. GARBER, M., GRABHERR, M.G., GUTTMAN, M., TRAPNELL, C. Computational methods for transcriptome annotation and quantification using RNA-seq. **Nature Methods**, v. 8 (6), p. 469-477, 2011.

82. WESTREICH, S.T., KORF, I., MILLS, D.A., LEMAY, D.G. SAMSA: a comprehensive metatranscriptome analysis pipeline. **BMC Bioinformatics**, v. 17 (1:399), p. 1-12, 2016.

83. BOLGER, M.E., ARSOVA, B., USADEL, B. Plant genome and transcriptome annotations: from misconceptions to simple solutions. **Briefings in Bioinformatics**, p. 1-13, 2017.

84. CONESA, A., GÖTZ, S. Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics. **International Journal of Plant Genomics**, v. 2008 (619832), p. 1-13, 2008.

85. ALTSCHUL, S.F., GISH, W., MILLER, W., MYERS, E.W., LIPMAN, D.J. Basic local alignment search tool. **Journal of Molecular Biology**, v. 215 (3), p. 403-410, 1990.

86. KERFELD, C.A., SCOTT, K.M. Using BLAST to Teach "E-value-tionary" Concepts. **PLoS Biology**, v. 9 (2: 1001014), p. 1-14, 2011.

87. HUNTLEY, R.P., SAWFORD, T., MARTIN, M.J., O'DONOVAN, C. Understanding how and why the Gene Ontology and its annotations evolve: the GO within UniProt. **GigaScience**, v. 3 (1), p. 1-9, 2014.

88. KEEGAN, K.P., GLASS, E.M., MEYER, F. MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. In: **Microbial Environmental Genomics (MEG)**. Nova Iorque: Springer, 2016. p. 207-233.

89. TATUSOV, R.L., GALPERIN, M.Y., NATALE, D.A., KOONIN, E.V. The COG database: a tool for genome-scale analysis of protein functions and evolution. **Nucleic Acids Research**, v. 28 (1), p. 33-36, 2000.

90. OVERBEEK, R., BEGLEY, T., BUTLER, R.M., CHOUDHURI, J.V., CHUANG, H.-Y., COHOON, M., DE CRÉCY-LAGARD, V., DIAZ, N., DISZ, T., EDWARDS, R., FONSTEIN, M., FRANK, E.D., GERDES, S., GLASS, E.M., GOESMANN, A., HANSON, A., IWATA-REUYL, D., JENSEN, R., JAMSHIDI, N., KRAUSE, L., KUBAL, M., LARSEN, N., LINKE, B., MCHARDY, A.C., MEYER, F., NEUWEGER, H., OLSEN, G., OLSON, R., OSTERMAN, A., PORTNOY, V., PUSCH, G.D., RODIONOV, D.A., RÜCKERT, C., STEINER, J., STEVENS, R., THIELE, I., VASSIEVA, O., YE, Y., ZAGNITKO, O., VONSTEIN, V. The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes. **Nucleic Acids Research**, v. 33 (17), p. 5691-5702, 2005.

91. VERLI, H., **Bioinformática**: da Biologia à flexibilidade. 1. ed., São Paulo: SBBq, 2014.

92. DORN, M., E SILVA, M.B., BURIOL, L.S., LAMB, L.C. Three-dimensional protein structure prediction: Methods and computational strategies. **Computational Biology and Chemistry**, v. 53, p. 251-276, 2014.

93. SLABINSKI, L., JAROSZEWSKI, L., RODRIGUES, A.P.C., RYCHLEWSKI, L., WILSON, I.A., LESLEY, S.A., GODZIK, A. The challenge of protein structure determination - lessons from structural genomics. **Protein Science: A Publication of the Protein Society**, v. 16 (11), p. 2472-2482, 2007.

94. TIANYUN, L., GRACE, W.T., EMIDIO, C. Comparative Modeling: The State of the Art and Protein Drug Target Structure Prediction. **Combinatorial Chemistry & High Throughput Screening**, v. 14 (6), p. 532-547, 2011.

95. KELLEY, L.A., MEZULIS, S., YATES, C.M., WASS, M.N., STERNBERG, M.J.E. The Phyre2 web portal for protein modeling, prediction and analysis. **Nature Protocols**, v. 10 (6), p. 845-858, 2015.

96. MA, J., PENG, J., WANG, S., XU, J. A conditional neural fields model for protein threading. **Bioinformatics**, v. 28 (12), p. 59-66, 2012.

97. MOULT, J., FIDELIS, K., KRYSHTAFOVYCH, A., SCHWEDE, T., TRAMONTANO, A. Critical assessment of methods of protein structure prediction (CASP) - round x. **Proteins**, v. 82 (2), p. 1-6, 2014.

98. KMIECIK, S., GRONT, D., KOLINSKI, M., WIETESKA, L., DAWID, A.E., KOLINSKI, A. Coarse-Grained Protein Models and Their Applications. **Chemical Reviews**, v. 116 (14), p. 7898-7936, 2016.

99. REMMERT, M., BIEGERT, A., HAUSER, A., SODING, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. **Nature Methods**, v. 9 (2), p. 173-175, 2012.

100. GORDON, A., HANNON, G.J. **Fastx-Toolkit**: FASTQ/A Short-Reads Pre-Processing Tools. Disponível em: < http://hannonlab.Cshl.Edu/fastx_toolkit/>. Acesso em: 20 ago. 2017, 12:00:00.

101. BOLGER, A.M., LOHSE, M., USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30 (15), p. 2114-2120, 2014.

102. HAAS, B.J., PAPANICOLAOU, A., YASSOUR, M., GRABHERR, M., BLOOD, P.D., BOWDEN, J., COUGER, M.B., ECCLES, D., LI, B., LIEBER, M., MACMANES, M.D., OTT, M., ORVIS, J., POCHET, N., STROZZI, F., WEEKS, N., WESTERMAN, R., WILLIAM, T., DEWEY, C.N., HENSCHEL, R., LEDUC, R.D., FRIEDMAN, N., REGEV, A. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. **Nature Protocols**, v. 8 (8), p. 1494-1512, 2013.

103. LANGMEAD, B., SALZBERG, S.L. Fast gapped-read alignment with Bowtie 2. **Nature Methods**, v. 9 (4), p. 357-359, 2012.

104. CAMACHO, C., COULOURIS, G., AVAGYAN, V., MA, N., PAPADOPOULOS, J., BEALER, K., MADDEN, T.L. BLAST+: architecture and applications. **BMC Bioinformatics**, v. 10 (1:421), p. 1-9, 2009.

105. FANG, L., SUN, D., XU, Z., HE, J., QI, S., CHEN, X., CHEW, W., LIU, J. Transcriptomic analysis of a moderately growing subisolate *Botryococcus braunii* 779 (*Chlorophyta*) in response to nitrogen deprivation. **Biotechnology for Biofuels**, v. 8, p. 130, 2015.

106. CONESA, A., MADRIGAL, P., TARAZONA, S., GOMEZ-CABRERO, D., CERVERA, A., MCPHERSON, A., SZCZEŚNIAK, M.W., GAFFNEY, D.J., ELO, L.L., ZHANG, X., MORTAZAVI, A. A survey of best practices for RNA-seq data analysis. **Genome Biology**, v. 17 (13), 2016.

107. PATNAIK, B.B., WANG, T.H., KANG, S.W., HWANG, H.-J., PARK, S.Y., PARK, E.B., CHUNG, J.M., SONG, D.K., KIM, C., KIM, S., LEE, J.S., HAN, Y.S., PARK, H.S., LEE, Y.S. Sequencing, *De Novo* Assembly, and Annotation of the Transcriptome of the Endangered Freshwater Pearl Bivalve, Cristaria plicata, Provides

Novel Insights into Functional Genes and Marker Discovery. **PLoS ONE**, v. 11 (2: 0148622), p. 1-28, 2016.

108. HUBER, S.C. Exploring the role of protein phosphorylation in plants: from signalling to metabolism. **Biochemical Society Transactions**, v. 35 (1), p. 28-32, 2007.

109. TIKKANEN, M., ARO, E.-M. Thylakoid protein phosphorylation in dynamic regulation of photosystem II in higher plants. **Biochimica et Biophysica Acta (BBA) - Bioenergetics**, v. 1817 (1), p. 232-238, 2012.

110. KODAMA, Y., TAMURA, T., HIRASAWA, W., NAKAMURA, K., SANO, H. A novel protein phosphorylation pathway involved in osmotic-stress response in tobacco plants. **Biochimie**, v. 91 (4), p. 533-539, 2009.

111. VIALARET, J., DI PIETRO, M., HEM, S., MAUREL, C., ROSSIGNOL, M., SANTONI, V. Phosphorylation dynamics of membrane proteins from Arabidopsis roots submitted to salt stress. **Proteomics**, v. 14 (9), p. 1058-1070, 2014.

112. YUAN, L.-L., ZHANG, M., YAN, X., BIAN, Y.-W., ZHEN, S.-M., YAN, Y.-M. Dynamic Phosphoproteome Analysis of Seedling Leaves in *Brachypodium distachyon* Reveals Central Phosphorylated Proteins Involved in the Drought Stress Response. **Scientific Reports**, v. 6 (35280), p. 1-14, 2016.

113. MIURA, K., FURUMOTO, T. Cold Signaling and Cold Response in Plants. **International Journal of Molecular Sciences**, v. 14 (3), p. 5312-5337, 2013.

114. MONROY, A.F., SARHAN, F., DHINDSA, R.S. Cold-Induced Changes in Freezing Tolerance, Protein Phosphorylation, and Gene Expression (Evidence for a Role of Calcium). **Plant Physiology**, v. 102 (4), p. 1227-1235, 1993.

115. GU, M., NI, H., SHENG, X., PAUCIULLO, A., LIU, Y., GUO, Y. RhoA phosphorylation mediated by Rho/RhoA-associated kinase pathway improves the anti-freezing potentiality of murine hatched and diapaused blastocysts. **Scientific Reports**, v. 7 (6705), p. 1-7, 2017.

116. GIRI, J., VIJ, S., DANSANA, P.K., TYAGI, A.K. Rice A20/AN1 zinc-finger containing stress-associated proteins (SAP1/11) and a receptor-like cytoplasmic kinase (OsRLCK253) interact via A20 zinc-finger and confer abiotic stress tolerance in transgenic *Arabidopsis* plants. **New Phytologist**, v. 191 (3), p. 721-732, 2011.

117. DO, H., KIM, S.-J., KIM, H.J., LEE, J.H. Structure-based characterization and antifreeze properties of a hyperactive ice-binding protein from the Antarctic bacterium *Flavobacterium frigoris* PS1. **Acta Crystallographica Section D**, v. 70 (4), p. 1061-1073, 2014.

118. HANADA, Y., NISHIMIYA, Y., MIURA, A., TSUDA, S., KONDO, H. Hyperactive antifreeze protein from an Antarctic sea ice bacterium *Colwellia* sp. has a compound ice-binding site without repetitive sequences. **FEBS Journal**, v. 281 (16), p. 3576-3590, 2014.

119. KONDO, H., HANADA, Y., SUGIMOTO, H., HOSHINO, T., GARNHAM, C.P., DAVIES, P.L., TSUDA, S. Ice-binding site of snow mold fungus antifreeze protein deviates from structural regularity and high conservation. **Proceedings of the National Academy of Sciences of United States of America**, v. 109 (24), p. 9360-9365, 2012.

120. PARK, K.S., DO, H., LEE, J.H., PARK, S.I., KIM, E.J., KIM, S.-J., KANG, S.-H., KIM, H.J. Characterization of the ice-binding protein from Arctic yeast *Leucosporidium sp*. AY30. **Cryobiology**, v. 64 (3), p. 286-296, 2012.

121. HON, W.C., GRIFFITH, M., CHONG, P., YANG, D.S.C. Extraction and Isolation of Antifreeze Proteins from Winter Rye (*Secale cereale L*.) Leaves. **Plant Physiology**, v. 104 (3), p. 971-980, 1994.

122. HUANG, T., DUMAN, J.G. Cloning and characterization of a thermal hysteresis (antifreeze) protein with DNA-binding activity from winter bittersweet nightshade, *Solanum dulcamara*. **Plant Molecular Biology**, v. 48 (4), p. 339-350, 2002.

123. NAKAMURA, T., ISHIKAWA, M., NAKATANI, H., ODA, A. Characterization of Cold-Responsive Extracellular Chitinase in Bromegrass Cell Cultures and Its Relationship to Antifreeze Activity. **Plant Physiology**, v. 147 (1), p. 391-401, 2008.

124. STRESSMANN, M., KITAO, S., GRIFFITH, M., MORESOLI, C., BRAVO, L.A., MARANGONI, A.G. Calcium Interacts with Antifreeze Proteins and Chitinase from Cold-Acclimated Winter Rye. **Plant Physiology**, v. 135 (1), p. 364-376, 2004.

7 APÊNDICES

Apêndice A – Artigo submetido ao periódico online GigaScience (ISSN 2047-217X) com os dados de montagem e anotação do transcriptoma de *P. crispa*.

De novo **assembly and annotation of the Antarctic alga *Prasiola crispa* transcriptome**

Evelise Leis Carvalho1*, Lucas Ferreira Maciel1*, Pablo Echeverria Macedo1, Filipe Zimmer Dezordi[1], Filipe de Carvalho Victória[2], Antônio Batista Pereira[2], Juliano Tomazzoni Boldo1, Gabriel da Luz Wallau3, Paulo Marcos Pinto1
(1) Applied Proteomics Laboratory, University of Pampa. Avenida Antônio Trilha, 1847, São Gabriel - RS - Brazil. Zip code: 97300-000.
(2) NEVA, University of Pampa. Avenida Antônio Trilha, 1847, São Gabriel - RS - Brazil. Zip code: 97300-000.
(3) Department of Entomology, Instituto Aggeu Magalhães-IAM, Fiocruz, Recife - PE - Brazil. Zip code: 52171-011.
*indicates equal contribution
E-mail addresses (in the order of appearance): eveliseleis@gmail.com; lucasmacielbiotec@gmail.com; pabloecheverriamacedo@gmail.com; zimmer.filipe@gmail.com; filipevictoria@unipampa.edu.br; antoniopereira@unipampa.edu.br; julianoboldo@unipampa.edu.br; gabriel.wallau@cpqam.fiocruz.br; paulopinto@unipampa.edu.br (corresponding author).

Abstract

Background: The environment plays a key role in the structural and functional changes that an organism must make to survive. The environmental conditions of the Antarctic continent, such as low temperatures, physiological drought, ultraviolet radiation, and nutrient limitation are limiting factors for the development of organisms that have mechanisms of adaptation to survive in extreme environments. Among the algae present on the Antarctic continent, *Prasiola crispa* is the organism most commonly found in the supralittoral zones of the maritime and continental Antarctic. However, the molecular mechanisms involved in the adaptive potential of *P. crispa* in this environment are unknown. Therefore, we report the *de novo* assembly and annotation of the *P. crispa* transcriptome sampled from King George Island, Antarctica.

Findings: Sequencing using Illumina HiSeq 2000 produced 42,978,976 reads. Transcriptome assembly was performed using Trinity, producing a total of 17,201 contigs and an N50 of 1,036. Gene ontology generated by Blast2GO resulted in a total of 5,343 sequences associated with a molecular function, 4,782 sequences with biological processes, and 3,639 sequences with cellular components.

Conclusions: The data generated from the transcriptome of *P. crispa* may elucidate the molecular mechanisms that allow the survival of this organism in the Antarctic environment, as well as to help future studies of *P. crispa* and other organisms Antarctic.

Keywords: RNA Sequencing, *Trebouxiophyceae*, *Prasiolales*, transcriptome, extreme environments, anti-freeze proteins.

Data Description

Context

The Antarctic, located on the South Pole of the Earth and isolated by the meeting of the Atlantic, Pacific and Indian oceans, is considered a continent with severe environmental conditions for the development of life, thus limiting the Antarctic fauna and flora to specific organisms that have survival adaptation mechanisms [1]. The average annual precipitation of Antarctic is only 200 mm with winds of 327 km/h and temperatures below -90ºC have already been recorded [2]. The total area is 14,000,000 km², with 98 to 99.7% covered by snow and ice, with layers averaging 1.6 km thick [2,3]. In addition, the ozone hole over the Antarctic region, first described in the 1980s, causes a high rate of ultraviolet radiation over the region, which is intensified by the ice-generated reflection [4,5].

Among the algae present on the Antarctic ice-free areas, *Prasiola crispa* (Lightfoot) Kützing is the most commonly found organism. P. crispa is a green macroalga belonging to the *Trebouxiophyceae* class and is among the most important primary producers in the Antarctic territory. *P. crispa* occurs in hydro-terrestrial soils, in the supralittoral zones of the maritime and continental Antarctica, where they form large and green carpets on the humid soil. *P. crispa* is found close to bird populations, mainly adjacent to penguin colonies, where the soil is rich in guano, a substrate with a high incidence of uric acid and nitrogen compounds [6].

The morphology of these algae varies from unisserated filaments to stalks in the form of a tape, expanded blades or packages as colonies, which are characterized by a large phenotypic plasticity related to environmental factors [7].

During the course of the seasons in the Antarctic territory, *P. crispa* needs to tolerate extreme environments, such as repeated freeze and thaw cycles, physiological drought, salinity stress, and high levels of UV radiation [1, 8]. However, the genes associated with these adaptive characteristics in *P. crispa* remain unknown. Therefore, to better understand the genetic and metabolic adaptations that allow this organism to survive in harsh environments, we sequenced its transcriptome.

Transcriptomes represent all the expressed fractions of genomes and are a viable alternative to understand and characterize genome wide genetic information of organisms since it simplifies genetic analyses, as compared to whole genome sequencing [9].

High-throughput sequencing of transcriptomes (RNA-Seq) has provided new routes to study the genetic and functional information stored within any organism at an unprecedented scale and speed. Transcriptome approaches have been employed in a large number of the studies involving non-model organisms, which normally lack reference genomes [10, 11].

Among these organisms are the algae group. The available data consists of organisms belonging to different phylum, such as *Prymnesiophyte*, *Chlorophyta*, *Haptophyta*, *Stramenopiles* and *Rhodophyta* [12, 13, 14, 15, 16].

P. crispa represents the first organism of the *Prasiolales* order with an available transcriptome since, until this work, the mitochondrial and plastid genomes were the only molecular data available for this species [17, 18]. Therefore, the purpose of this study was to sequence the transcriptome of *P. crispa*. The identification of transcripts will help to identify genes that are responsible for organism survival in this environment, as well as assisting in future genetic, phylogenetic, and biotechnological studies of *P. crispa* and other Antarctic organisms.

Methods

Algae Collection

P. crispa was collected in areas near the Arctowski Polish Station Region, Admiralty Bay, King George Island (61°50’ – 62°15’ S and 57°30’ – 59°00’W), Antartic. The collection was carried out in the Antarctic summer, in January of 2014 Austral summer, with temperature ranging from 0.5 to 2.0ºC. The samples were maintained in RNAlater® (Sigma-Aldrich, USA) until RNA extraction.

Total RNA Extraction and RNA-Seq

Total RNA was extracted from three pools of samples using an RNAqueous®-Micro Total RNA Isolation Kit (Thermo Fisher Scientific Inc., USA) according to the manufacturer’s

instructions. The RNA-Seq library was prepared using random primers. The transcriptome was sequenced by Macrogen Service using the Solexa-Illumina HiSeq 2000 next-generation sequencing platform device according to the manufacturer's instructions. A paired-end reads with a read size of ~100 bp separated by insert size of 300 bp was employed.

De novo Transcriptome assembly

Raw reads from data sets were filtered to remove the adapter sequences, and low quality reads with Fastx-toolkit (quality cut-off = 30) [19] and Trimmomatic v 0.36 using default parameters [20]. Next, we used Trinity version r2014-07-17 [10] as Bruijn graph assembler with 25 kmer size. Due to the sequencing of a complex sample extracted from the Antarctic soil, we expected some amount of bacterial and fungal contamination. Therefore, to clean our transcriptome, we performed a tblastx with default parameters [21] searching against all of the NCBI nucleotide non-redundant database and recovered all contigs in which the best blast hit occurred with algae and plant sequences. Next, we used Bowtie2 [22] with default parameters to recover only the reads that mapped against those *P. crispa* contigs.

After a stringent filtering process, the processed reads were assembled into 17,201 contigs. Statistics of the assembly are summarized in Table 1. The metrics of *P. crispa* were compared with others transcriptomes of the organism from the *Trebouxiophyceae class*, including *Chlorella minutissima* [23], *Trebouxia gelatinosa* [24], *Coccomyxa subellipsoidea* [25], *Chlorella sorokiniana* [26], *Botryococcus braunii* [27], and was found to have the lowest number of total reads. In relation to the number of contigs, *P. crispa* is in an intermediary position, with *C. sorokiniana* being the organism with the highest number of contigs (63,811) and *C. subellipsoidea* having the lowest number of contigs (9,409). More information on the metrics of transcriptomes is given in Table 2.

Functional Annotation

The assembled and recovered contigs were searched against the NCBI protein non-redundant database using the BLASTX algorithm; the E-value cut-off was set at 1e-10. Genes were tentatively identified based on the best hits against known sequences. Blast2GO [28] was used for mapping and annotation, associating Gene Ontology (GO) terms and predicting their function.

The search of these contigs against the NCBI protein non-redundant database with BLASTX demonstrates that 8,980 (52.19%) sequences had at least one hit. The mapping of the sequences against the GO database retrieved 7,009 sequences mapped, and all assigned GO

terms were classified into three main categories: cellular component (3,639 sequences), molecular function (5,343 sequences), and biological process (4,782 sequences). The top 15 GO terms are represented in Figure 1.

Availability of the supporting data

The BioProject ID of our data is PRJNA329112, and the BioSample accession number is SAMN05392062. All raw reads were deposited into the Sequencing Read Archive (SRA) of NCBI with accession number SRR5754271. This Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GFTS00000000 and is available in the Gigascience repository Giga DB [29].

Data Validation and quality control

The reading quality of the data of this transcriptomic analysis was evaluated through FastQC software (Babraham Bioinformatics) [RRID: SCR_014583]. The paired-end reads results were merged using MultiQC (http://multiqc.info) [29] and are displayed in Figure 2. Per base quality phred scores range from 32.78 to 40.06, indicating base call accuracies of >99.9% (Figure 2A). Per sequence quality shows that 99.62% of reads had a mean phred score of 30 or above (Figure 2B) and per base N content was low, with a maximum value 0.18% (Figure 2C).

Re-use potential

The current data set of RNA-Seq may help in the identification of genes and molecular mechanisms associated with survival of *P. crispa* as well as other Antarctic organisms. Through the data of this transcriptome, it is possible to perform searches for complete genes, aiming the heterologous expression of proteins with biotechnological potential, such as antifreeze proteins, which act to inhibit freezing of intracellular fluids [31], Heat-Shock proteins that play an important role in maintain biological activities in algae present in these acclimatization process [32] and mycosporine-like amino acids responsive to high incidence of ultraviolet radiation [6]. Proteomics approaches may also be employed, aiming at the confirmation of gene expression at the transcriptional level.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

This work was supported by the National Council for Scientific and Technological Development (CNPq-Brazil), the Coordination for the Improvement of Higher Education Personnel (CAPES-Brazil), the Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS-Brazil) and National Institute of Science and Technology - Antarctic Environmental Research (INCT-APA).

Author Contributions

PMP: conducted the experiment; ELC, LFM, PEM, FZD, GLW: performed analysis on the data; FCV, ABP, JTB, PMP: conceived the project and acquired funding; ELC, LFM, PEM, GLW, PMP: wrote the manuscript.

References

[1] Jackson AE, Seppelt RD. Physiological adaptations to freezing and UV radiation exposure in Prasiola crispa, an Antarctic terrestrial alga. In Battaglia B, Valencia J, Walton DWH, editors. Antarctic Communities: Species, Structure, and Survival, Cambridge: University Press; 1997. p. 226-233.

[2] Martínez-Rosales C, Fullana N, Musto H, Castro-Sowinski S. Antarctic DNA moving forward: genomic plasticity and biotechnological potential. FEMS Microbiol Lett. 2012; doi: 10.1111/j.1574-6968.2012.02531.x.

[3] Convey P, Gibson JAE, Hillenbrand CD, Hodgson DA, Pugh PJA, Smellie JL, Stevens MI. Antarctic terrestrial life – challenging the history of the frozen continent? Biol Rev Camb Philos Soc. 2008; doi: 10.1111/j.1469-185X.2008.00034.x.

[4] Kuttippurath J, Nair PJ. The signs of Antarctic ozone hole recovery. Sci Repor. 2017; doi: 10.1038/s41598-017-00722-7.

[5] Marizcurrena JJ, Morel MA, Braña V, Morales D, Martinez-López W, Castro-Sowinsk S. Searching for novel photolyases in UVC-resistant Antarctic bacteria. Extremophiles. 2017; doi: 10.1007/s00792-016-0914-y.

[6] Kováčik L, Pereira AB. Green alga Prasiola crispa and its lichenized form Mastodia tesselata in Antartic. In: Elster J, Seckbach J, Vincent WF, Lhotský O, editors. Algae and extreme environments. Czech Republic: Nova Hedwigia 123; 2001. p. 465-478.

[7] Rindi F, McIvor L, Sherwood AR, Friedl T, Guiry MD, Sheath RG. Molecular phylogeny of the green algal order Prasiolales (Trebouxiophyceae, Chlorophyta). J. Phycol. 2007; doi: 10.1111/j.1529-8817.2007.00372.x.

[8] Jacob AC, Wiencke HL, Kirst GO. Physiology and ultrastructure of desiccation in the green-alga Prasiola crispa from Antarctica. Botanica Marina. 1992; doi: 10.1515/botm.1992.35.4.297.
[9] Riesgo A, Andrade SCS, Sharma PP, Novo M, Pérez-Porro AR, Vahtera V, González VL, Kawauchi GY, Giribet G. Comparative description of ten transcriptomes of newly sequenced invertebrates and efficiency estimation of genomic sampling in non-model taxa. Front Zool. 2012; doi: 10.1186/1742-9994-9-33.
[10] Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, Regev A. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013; doi: 10.1038/nprot.2013.084.
[11] Ekblom R, Galindo J, 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity. 2011; doi: 10.1038/hdy.2010.152.
[12] Koid AE, Liu Z, Terrado R, Jones AC, Caron AD, Heidelberg KB. 2014. Comparative transcriptome analysis of four Prymnesiophyte algae. PLoS ONE. 2014; doi: 10.1371/journal.pone.0097801.
[13] Rismani-Yazdi H, Haznedaroglu BZ, Bibby K, Peccia J, 2011. Transcriptome sequencing and annotation of the microalgae Dunaliella tertiolecta: pathway description and gene discovery for production of next-generation biofuels. BMC Genomics. 2011; doi: 10.1186/1471-2164-12-148.
[14] Talarski A, Manning RS, La Claire II JW. Transcriptome analysis of the euryhaline alga, Prymnesium parvum (Prymnesiophyceae): effects of salinity on differential gene expression. Phycol. 2016; doi: 10.2216/15-74.1.
[15] Im S, Choi S, Hwang MS, Park EJ, Jeong WJ, Choi DW. De novo assembly of transcriptome from the gametophyte of the marine red algae Pyropia seriata and identification of abiotic stress response genes. J. Appl. Phycol. 2015; doi: 10.1007/s10811-014-0406-3.
[16] Shuangxiu W, Sun J, Chi S, Wang L, Wang X, Liu C, Li X, Yin J, Liu T, Yu J, 2014. Transcriptome sequencing of essential marine brown and red algal species in China and its significance in algal biology and phylogeny. Acta Oceanol Sin. 2014; doi: 10.1007/s13131-014-0435-4.
[17] Carvalho EL, Wallau GL, Rangel DL, Machado LC, Silva AF, Silva LFD, Macedo PE, Pereira AB, Victoria FC, Boldo JT, Dal Belo CA, Pinto PM. Draft plastid and mitochondrial

genome sequences from Antarctic alga Prasiola crispa. Genome Announc. 2015; doi: 10.1128/genomeA.01151-15.

[18] Carvalho EL, Wallau GL, Rangel DL, Machado LC, Pereira AB, Victoria FDC, Boldo JT, Pinto PM. Phylogenetic positioning of the Antarctic alga Prasiola crispa (Trebouxiophyceae) using organellar genomes and their structural analysis. J Phycol. 2017; doi:10.1111/jpy.12541

[19] Gordon A, Hannon GJ, 2010. Fastx-Toolkit. FASTQ/A Short-Reads Pre-Processing Tools. http://hannonlab.Cshl.Edu/fastx_toolkit/. Accessed 10 July 2017.

[20] Bolger AM, Lohse M, Usadel M. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinform 2014; doi: 10.1093/bioinformatics/btu170.

[21] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, 1990. Basic Local Alignment Search Toll. J Mol Biology. 1990; doi: 10.1016/S0022-2836(05)80360-2.

[22] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods; 2012. doi:10.1038/nmeth.1923.

[23] Yu M, Shanjun Y, Lin X, 2016. De-novo assembly and characterization of Chlorella minutissima UTEX2341 transcriptome by paired-end sequencing and the identification of genes related to the biosynthesis of lipids for biodiesel. Mar Genomics. 2016; doi: 10.1016/j.margen.2015.11.005.

[24] Carniel FB, Gerdol M, Montagner A, Banchi E, De Moro G, Manfrin C, Muggia L, Pallavicini A, Tretiach M. New features of desiccation tolerance in the lichen photobiont Trebouxia gelatinosa are revealed by a transcriptomic approach. Plant Mol Biol. 2016; doi: 10.1007/s11103-016-0468-5.

[25] Peng E, Wei D, Chen G, Chen F. Transcriptome analysis reveals global regulation in response to CO2 supplementation in oleaginous microalga Coccomyxa subellipsoidea C-169. Biotechnol for Biofuels. 2016; doi: 10.1186/s13068-016-0571-5.

[26] Li L, Zhang G, Wang Q. 2016. De novo transcriptomic analysis of Chlorella sorokiniana reveals differential genes expression in photosynthetic carbon fixation and lipid production. BMC Microbiol. 2016; doi: 10.1186/s12866-016-0839-8.

[27] Xu Z, He J, Qi S, Liua J. Nitrogen deprivation-induced de novo transcriptomic profiling of the oleaginous green alga Botryococcus braunii 779. Genom Data. 2015; doi: 10.1016/j.gdata.2015.09.019.

[28] Conesa A, Gotz S. 2008. Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics. 2008; doi: 10.1155/2008/619832.

[29] GigaDB – adicionar

[30] Ewels M, Magnusson M, Lundin S, Kaller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016; doi: 10.1093/bioinformatics/btw354.
[31] Nath A, Radha C, Subbiah K. An insight in to the molecular basis for convergent evolution in fish antifreeze proteins. Comput Bio Med. 2013; doi: 10.1016/j.compbiomed.2013.04.013.
[32] Li R, Brawley SH. 2004. Improved survival under heat stress in intertidal embryos (Fucus spp.) simultaneously exposed to hypersalinity and the effect of parental thermal history. Marine Biology. 2004; doi: 10.1007/s00227-003-1190-9.

Figure captions

Figure 1. Gene Ontology distribution to the Top 15 terms at level 4 in the three main categories: Biological Process, Molecular Function and Cellular Component.

Figure 2. Graphs about sequencing read quality gerateds with FastQC. (A) Per base quality phred. (B) Per sequence quality.

Table 1
Summary of Prasiola crispa assembly.

Attributes	**Value**
Total raw reads	42,978,976
Total processed reads	5,233,428
Number of contigs	17,201
Total length (pb)	13,127,645
N50	1,036
GC (%)	49.66
Average size (pb)	763.19
Size range (pb)	200 - 12,802
Contigs > 1kb	3,973 (22,05 %)

Table 2

Comparison between *Prasiola crispa* and organisms from the Trebouxiophyceae class with transcriptome sequenced.

(-) Data not reported.

Attributes	***Prasiola crispa***	***Chlorella minutissima***	***Trebouxia gelatinosa***	***Coccomyxa subellipsoidea***	***Chlorella sorokiniana***	***Botryococcus braunii***
Total raw reads	42,978,976	69,011,712	243,763,578	-	244,291,069	-
Total processed reads	5,233,428	67,559,338	237,404,631	46,000,000	229,228,757	-
Number of contigs	17,201	14,905	19,601	9,409	63,811	61,220
Mean length	763.1	2,998.04	1,605	-	1,022	-
Sequences with at least one blast hit (%)	52.19	53.50	53.60	-	36.80	-
Largest contig (pb)	12,802	-	31,749	-	15,932	-
N50	1,036	-	3,594	-	2,502	-
Assembler	Trinity	Trinity	Trinity	Reference genome	Trinity	Trinity

Figure 1

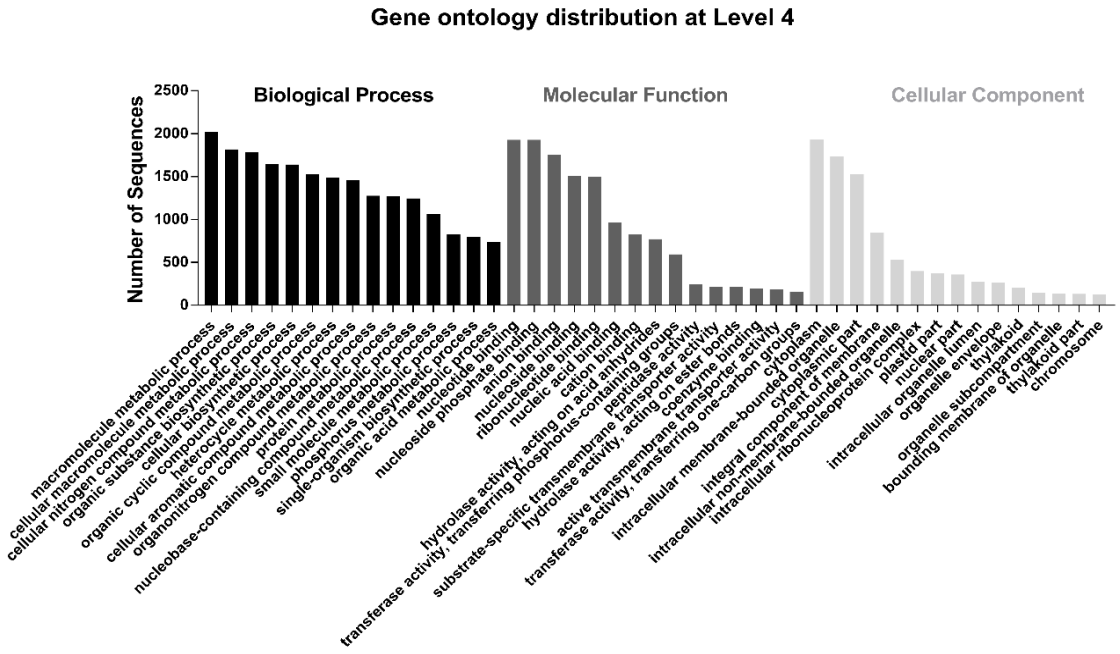


Figure 2

A

B

8 ANEXOS

Anexo A – BLASTX *hits* da metamontagem contra o banco de dados de IBPs.

Descrição	E-value	Similaridade (%)	Espécie
antifreeze protein	4,21E-11	48,10%	*Daucus carota*
antifreeze protein	2,20E-11	65,50%	*Fibularhizoctonia sp. CB 109695*
kDa class I endochitinase-antifreeze	1,06E-14	64,21%	*Secale cereale*
29 kDa chitinase-like thermal hysteresis	1,55E-15	65,50%	*Solanum dulcamara*
Antifreeze	1,51E-17	57,50%	*Fibularhizoctonia sp. CB 109695*
antifreeze	2,44E-18	69,70%	*Fragilariopsis cylindrus*
kDa class I endochitinase-antifreeze	7,3E-19	62,80%	*Secale cereale*
ice antifreeze	2,11E-19	59,60%	*Fragilariopsis curta*
antifreeze	7,18E-20	56,40%	*Fibularhizoctonia sp. CB 109695*
antifreeze	5,13E-20	53,20%	*Fibularhizoctonia sp. CB 109695*
antifreeze protein	3,84E-21	49,60%	*Saussurea involucrata*
antifreeze protein	2,68E-21	49,60%	*Saussurea involucrata*
antifreeze protein	2,53E-21	49,60%	*Saussurea involucrata*
ice antifreeze	8,44E-23	57,40%	*Fragilariopsis curta*
ice antifreeze	3,18E-23	92,60%	*Fragilariopsis cylindrus*
antifreeze	1,70E-28	80,30%	*Stereum hirsutum*
antifreeze	4,20E-29	74,40%	*Fibularhizoctonia sp. CB 109695*
antifreeze	9,65E-37	74,20%	*Stereum hirsutum*
29 kDa chitinase-like thermal hysteresis	8,98E-59	60,60%	*Solanum dulcamara*
29 kDa chitinase-like thermal hysteresis	1,38E-59	60,60%	*Solanum dulcamara*
29 kDa chitinase-like thermal hysteresis	4,88E-62	60,60%	*Solanum dulcamara*
antifreeze protein	8,74E-11	59,20%	*Halobacillus sp. BAB-2008*
type I antifreeze	1,35E-11	57,30%	*Rubrivivax benzoatilyticus*
Type II antifreeze	8,13E-11	56,60%	*Stigmatella aurantiaca Dw4/3-1*
antifreeze	1,19E-11	56,40%	*Pseudozobellia thermophila*
antifreeze protein	3,46E-12	81,80%	*Plesiocystis pacifica*

antifreeze protein	1,25E-12	46,50%	*Pseudomonas putida*
type I antifreeze	1,01E-12	86,50%	*Rhodococcus sp. B7740*
Type III antifreeze	9,93E-13	70,00%	*Nitrosomonas europaea ATCC 19718*
antifreeze protein	1,73E-13	64,20%	*Flavobacterium xanthum*
type I antifreeze	6,89E-14	71,40%	*Gamma proteobacterium NOR5-3*
type I antifreeze	2,81E-13	81,10%	*Mycobacterium abscessus subsp. Massiliense*
type I antifreeze	5,49E-15	65,30%	*Firmicutes bacterium CAG:41*
type I antifreeze	1,43E-15	58,80%	*Butyricimonas synergistica*
antifreeze protein	1,05E-15	70,90%	*Stigmatella aurantiaca Dw4/3-1*
type I antifreeze	6,02E-18	89,50%	*Nocardioidaceae bacterium Broad-1*
antifreeze	4,05E-18	67,70%	*Bacteroidetes bacterium 4572_117*
antifreeze	1,48E-18	69,30%	*Flavobacterium columnare*
antifreeze	2,19E-18	94,40%	*Chitinophaga eiseniae*
antifreeze	1,93E-18	77,90%	*Neisseria bacilliformis ATCC BAA-1200*
antifreeze	1,32E-18	78,10%	*Nonlabens ulvanivorans*
type I antifreeze	1,30E-18	78,90%	*Bifidobacterium adolescentis*
type I antifreeze	1,00E-18	78,90%	*Bifidobacterium adolescentis*
type I antifreeze	9,74E-19	81,00%	*Rhodococcus sp. B7740*
type I antifreeze	8,60E-20	62,50%	*Burkholderia mallei GB8 horse 4*
antifreeze	3,71E-20	77,90%	*Neisseria bacilliformis ATCC BAA-1200*
antifreeze	2,26E-20	85,10%	*Stigmatella aurantiaca Dw4/3-1*
antifreeze	1,68E-20	84,70%	*Desulfatibacillum aliphaticivorans*
type I antifreeze	3,31E-21	69,60%	*Salinisphaera shabanensis E1L3A*
type I antifreeze	7,98E-22	81,10%	*Bifidobacterium adolescentis*
type I antifreeze	3,50E-22	74,30%	*Salinisphaera shabanensis E1L3A*
type I antifreeze	3,22E-22	74,30%	*Salinisphaera shabanensis E1L3A*
type I antifreeze	2,07E-22	91,30%	*Ideanella sakaiensis*
type I antifreeze	9,13E-23	51,30%	*Herbaspirillum frisingense GSF30*
type I antifreeze	2,82E-23	76,90%	*Salinisphaera shabanensis E1L3A*
type I antifreeze	2,19E-23	72,90%	*Salinisphaera shabanensis E1L3A*
type I antifreeze	1,16E-23	76,10%	*Gamma proteobacterium NOR5-3*

type I antifreeze	9,34E-24	54,80%	*Clostridium symbiosum WAL-14163*
antifreeze	1,13E-24	86,20%	*Psychrobacter sp. 1501 (2011)*
type I antifreeze	1,06E-24	76,10%	*Gamma proteobacterium NOR5-3*
antifreeze	3,52E-25	61,20%	*Cystobacter ferrugineus*
type I antifreeze	1,01E-25	79,40%	*Sutterella sp. CAG:351*
AFGP	2,06E-27	84,90%	*Leifsonia syli subsp. Cynodontis DSM 46306*
type I antifreeze	1,81E-28	91,50%	*Achromobacter xylosoxidans*
type I antifreeze	1,00E-33	81,30%	*Dethiosulfatarculus sandiegensis*
antifreeze	4,61E-34	81,70%	*Psychrobacter sp. 1501 (2011)*
antifreeze	8,36E-35	89,50%	*Flavobacterium columnare*
antifreeze	4,52E-39	69,80%	*Fimbriimonas ginsengisoli Gsoil 348*
antifreeze	1,69E-47	67,00%	*Stigmatella aurantiaca Dw4/3-1*
antifreeze	3,41E-51	87,00%	*Neisseria bacilliformis ATCC BAA-1200*
antifreeze	2,93E-51	95,60%	*Williamsia sp. 1135*
antifreeze	1,75E-51	68,80%	*Cystobacter ferrugineus*
antifreeze	1,47E-68	85,80%	*Neisseria mucosa C102*
antifreeze	1,45E-101	74,10%	*Chitinophaga eiseniae*
antifreeze	1,02E-111	78,90%	*Neisseria bacilliformis ATCC BAA-1200*
Ice-structuring glycoprotein isoform X3	8,32E-11	60,00%	*Zeugodacus cucurbitae*
antifreeze	4,99E-11	48.,7%	NI
ice-structuring glycoprotein	1,82E-11	57,90%	*Drosophila busckii*
ice-structuring glycoprotein isoform X10	1,79E-11	59,30%	*Drosophila biarmipes*
ice-structuring glycoprotein isoform X9	1,02E-11	58,20%	*Ceratitis capitata*
ice-structuring glycoprotein isoform X14	8,68E-12	58,10%	*Drosophila biarmipes*
antifreeze	2,17E-12	56,80%	NI
ice-structuring glycoprotein isoform X1	2,60E-13	56,00%	*Zeugodacus cucurbitae*
Ice-binding protein	4,21E-11	58,00%	Bactéria NI
Ice-binding protein	4,21E-11	51,00%	Bactéria NI
Ice-binding protein	4,19E-11	57,00%	Bactéria NI
Ice-binding protein	3,64E-11	57,00%	Bactéria NI
Ice-binding protein	3,49E-11	64,00%	*Psychroflexus torquis ATCC 700755*

secreted Ice-binding protein cell surface	3,02E-11	47,00%	Bactéria NI
Ice-binding protein	1,55E-11	59,00%	Bactéria NI
Ice-binding protein	1,35E-11	68,00%	Bactéria NI
Ice-binding protein	1,01E-11	52,00%	Bactéria NI
Ice-binding protein	1,72E-12	60,00%	Bactéria NI
Ice-binding protein	6,97E-13	62,00%	Bactéria NI
Ice-binding protein	2,66E-13	58,00%	Bactéria NI
Ice-binding protein	2,19E-13	72,60%	Bactéria NI
Ice-binding protein	1,03E-13	55,00%	Bactéria NI
Ice-binding protein	2,25E-14	67,60%	Bactéria NI
Ice-binding protein	1,94E-14	49,00%	Bactéria NI
Ice-binding protein	1,86E-14	56,00%	Bactéria NI
Ice-binding protein	2,60E-15	55,00%	Bactéria NI
Ice-binding protein	2,49E-15	57,00%	Bactéria NI
Ice-binding protein	2,63E-16	62,20%	Bactéria NI
Ice-binding protein	1,19E-16	57,00%	Bactéria NI
Ice-binding protein	1,06E-16	68,00%	Bactéria NI
Ice-binding protein	4,14E-17	53,00%	Bactéria NI
Ice-binding protein	2,25E-17	70,80%	Bactéria NI
Ice-binding protein	8,87E-18	57,00%	Bactéria NI
Ice-binding protein	6,63E-18	69,70%	Bactéria NI
Ice-binding protein	4,08E-18	55,00%	Bactéria NI
Ice-binding protein	1,42E-18	58,90%	Bactéria NI
Ice-binding protein	6,88E-19	55,00%	Bactéria NI
Ice-binding protein	4,02E-21	81,70%	Bactéria NI
Ice-binding protein	2,80E-23	61,00%	Bactéria NI
Ice-binding protein	1,64E-23	77,30%	Bactéria NI
Ice-binding protein	8,42E-24	74,00%	Bactéria NI
secreted Ice-binding protein cell surface	1,79E-24	44,70%	*Psychroflexus torquis ATCC 700755*
Ice-binding protein	1,30E-24	76,70%	Bactéria NI
secreted Ice-binding protein cell surface	9,68E-25	53,80%	*Psychroflexus torquis ATCC 700755*
Ice-binding protein	3,11E-26	74,80%	Bactéria NI

Ice-binding protein	4,70E-27	77,80%	Bactéria NI
Ice-binding protein	2,48E-27	77,40%	Bactéria NI
Ice-binding protein	5,50E-32	86,70%	Bactéria NI
Ice-binding protein	2,75E-35	79,30%	*Flavobacterium sp*
Ice-binding protein	5,59E-36	67,20%	Bactéria NI
Ice-binding protein	3,90E-54	78,30%	Bactéria NI
Ice-binding protein	2,00E-80	79,50%	Bactéria NI

Legenda: NI – Não identificada.