Universidade de São Paulo
Instituto de Matemática e Estatística

Programa Interunidades em Bioinformática

Aureliano Coelho Proença Guedes

Classificação e evolução de SBPs em contexto de transdução de sinal:
Interação entre SBPs e Cache.

São Paulo, 2021

Aureliano Coelho Proença Guedes

Classificação e evolução de SBPs em contexto de transdução de sinal:
Interação entre SBPs e Cache.

Dissertação apresentada ao Programa
de Pós Graduação Interunidade em
Bioinformática do Instituto de
Matemática e Estatística da
Universidade de São Paulo para
obtenção do título de Mestre em
Ciências.

Área de Concentração: Bioinformática.
Orientador: Robson Francisco de Souza

São Paulo, 2021

Dedicatória

Dedico este trabalho a todas as pessoas que acreditam na ciência e aos profissionais da saúde pelo empenho em salvar vidas.
Dedico também esse trabalho à minha mãe, Dona Rita, por sempre acreditar e apostar em mim e por me permitir crescer dentro de uma biblioteca.

Agradecimentos

Agradeço, antes de tudo, a minha família por manter a expectativa e confiança em mim.

Agradeço imensamente a todos meus colegas/amigos do LEEP, aos que passaram e aos que estão ainda hoje lá, por todos os momentos que a descontração trouxe risos, pelas pizzas e esfihas a noite e os lanchinhos da tarde na hora do café. Em especial os 4G: Gianluca, Gilberto, Guilherme e Gabriel, por me tolerarem e me apoiarem nesse tempo todo e por ‘Se está formando el pulpo’. São tantos nomes que tenho medo de esquecer alguns, mas não posso deixar de falar da Tânia, Youvika e Day, que são praticamente minhas contemporâneas no laboratório, o que nos torna “Os Veteranos” do LEEP, e que sempre foram ótimas amigas que me expandiram o conhecimento da culinária mundial.

Agradeço aos meus amigos de fora da academia pelos laços de amizade que desenvolvi com eles e que me fez me sentir muito mais em família mesmo que eu esteja a mais 1200Km da minha família de sangue. Aos meus amigos que estão em Goiás ou na Suíça, que mantêm uma amizade tão próxima como se fossem há alguns anos atrás quando costumávamos passar os tempos de lazer juntos. Em especial agradeço a um grande amigo desde o início da graduação, o Marielton, que tem participação importante na minha decisão de fazer o mestrado neste programa.

Não menos importante, agradeço a Capes e a Fapesp que financiaram essa pesquisa.

Epigrafe

“Prefiro ser um homem de paradoxos que um homem de preconceitos.” - Jean-Jacques
Rousseau

Resumo

Neste trabalho estudamos a origem e evolução de novas combinações de domínios protéicos (arquiteturas) em proteínas bacterianas contendo domínios de três *folds* relacionados, conhecidos como proteínas ligadoras de substratos (SBPs), que participam de processos de transporte e transdução de sinal. Com base na reclassificação das famílias nestes três *folds*, identificamos famílias cujos membros participam tanto de funções de transdução de sinal como de transporte, demonstrando uma rápida mudança de contexto funcional dentro de uma mesma família. Muitas destas famílias de SBPs são codificadas por genes vizinhos a genes que codificam proteínas com um ou mais domínios da superfamília Cache. A interação física dos domínios SBP e Cache já foi demonstrada em alguns desses sistemas e nossa análise sugere que algumas das fusões de SBPs periplasmáticos com domínios efetores de transdução de sinal teriam se originado da simplificação de sistemas de três ou mais componentes contendo pares cognatos SBP-Cache. Embora os fenômenos descritos aqui estejam limitados à mudança de função recente de SBPs em famílias destes *folds* encontradas apenas em procariotos, fica claro que a flexibilidade dos SBPs, que são capazes de se ligar a muitos solutos diferentes e, em consequência, sofrer grandes rearranjos estruturais que transmitem informação para proteínas parceiras, é continuamente re-utilizada em processos que envolvem interações proteína-proteína intermediárias, que ainda são, na maior parte, desconhecidas. A análise detalhada destas interações têm potencial para prover mais conhecimento sobre a evolução de complexos protéicos e vias de sinalização, com potencial para aplicação em campos de bioengenharia e biologia de sistemas.

Palavras-chave: Transdução de sinal; Genômica comparativa; Homologia; Filogenia; SBP; Cache

ABSTRACT:

We study the origin and evolution of three related *folds*, collectively known as substrate-binding proteins (SBPs), in novel protein domain architectures, that part in transport and signal transduction process in prokaryotes. We propose a new classification of families within these three *folds*, and identified families whose members work on both signal transduction and transport that shows a repeated instances of gain and change of function within these protein families. Many of these SBP genes are found to be located close to genes that encode proteins with one domain of the Cache superfamily, most likely forming operons that combine SBP and sensors in a stable evolutionary unity. Physical interaction of some SBP and Cache domains has been demonstrated for a few of these systems and our analysis show that fusions of extracellular SBPs to cytoplasmic effector domains of two component systems and related signal transduction proteins most likely originated from the fusion of SBP and sensor proteins, accompanied by deletion of a Cache domain, leading to streamlining of a system of three or two genes containing cognate SBP-Cache pairs. While the phenomena we describe are limited to the neofunctionalization of SBPs within prokaryotic families of these *folds*, the flexibility of the SBP domains, which are able to bind many different ligands and undergo wide structural rearrangements that transmit information to partner proteins, is being reused through a process that might involve yet unknown intermediary protein interactions. The detailed analysis of these interactions is likely to reveal novel insights into the evolution of such protein complexes, providing support for advances in bioengineering of biological pathways and systems biology.

Keywords: Signal transduction; Comparative genomics; Homology; Phylogeny; SBP; Cache

Lista de figuras

Figura 1.	Arquitetura típica, simplificada, de proteínas de transdução de sinal	18
Figura 2.	Fluxograma de análise.	31
Figura 3.	Distribuição dos valores de distância e p-values, estimados pelo programa Mash, nas comparações par-a-par entre genomas	41
Figura 4.	Representação da intersecção de conjuntos de sequências identificados nas diferentes estratégias de busca por homólogos de sensores extracitoplasmáticos	43
Figura 5.	Distribuição do número de regiões extracitoplasmáticas por clã do banco de dados Pfam	47
Figura 6.	Resultados da aplicação de diferentes algoritmos de agrupamento às estruturas tridimensionais de SBPs	50
Figura 7.	Rede de similaridade das famílias de SBPs	52
Figura 8.	Relação entre a presença de assinaturas de secreção em SBPs e a frequência de genes vizinhos da superfamília Cache	55
Figura 9.	Variação do p-value em função da abundância de SBPs com localização extracitoplasmática	56
Figura 10.	Composição de domínios protéicos de homólogos de Cache e SBP codificados por genes vizinhos	58
Figura 11.	Agrupamento hierárquico das famílias da comunidade 1	60
Figura 12.	Árvore de máxima verossimilhança do domínio SBP dos membros da família XAC2383	61
Figura 13.	Agrupamento hierárquico baseado na similaridade entre as famílias da comunidade 2	63
Figura 14.	Filogenia de máxima verossimilhança para a família 2.1534	64
Figura 15.	Filogenia dos membros da família NspS	65
Figura 16.	Agrupamento hierárquico por similaridade das famílias da comunidade 4	66
Figura 17.	Filogenia, vizinhança gênica e domínios das proteínas da família CBP	67
Figura 18.	Agrupamento de similaridade para as famílias da comunidade 5	68
Figura 19.	Filogenia, vizinhança gênica e domínios das proteínas da família 5.749 e seus vizinhos	69
Figura 20.	Agrupamento de similaridade para as famílias da comunidade 3	71
Figura 21.	Filogenia, vizinhança gênica e domínios das proteínas da família LuxQ	72
Figura 22.	Filogenia, vizinhança gênica e domínios das proteínas da família 3.286 e seus vizinhos	72
Figura 23.	Agrupamento de similaridade para a comunidade 6	74

Lista de Tabelas

Tabela 1.	Tipo e abundância de domínios Cache	25
Tabela 2.	Classificação e ligantes de proteínas ligadoras de substratos	27
Tabela 3.	Comunidades de domínios periplasmáticos identificados em proteínas de transdução de sinal	44
Tabela 4.	Relação de probabilidade entre famílias do Pfam sem anotação de clã e outras famílias, do mesmo banco de dados, com clãs assinalados	47
Tabela 5.	Relação das comunidades de SBPs com os grupos descritos por Scheepers e colaboradores	52
Tabela 6.	Tabela de contingência e valores do teste χ^2 para a presença de homólogos de SBPs e Cache em uma janela deslizante de 2 genes	54
Tabela 7.	Tabela de contingência e valores do teste χ^2 para a presença de homólogos de SBPs e Cache em uma janela deslizante de 2	57

Lista de Abreviaturas

SBP	Proteína ligadora de sustrato
GPCR	Receptor acoplado a proteína G
cAMP	Adenosina monofosfato cíclica
cGMP	Guanosina monofosfato cíclica
c-di-GMP	Di-guanosina cíclica
GDP	Guanosina difosfato
GTP	Guanosina trifosfato
c-di-AMP	Diadenosina cíclica
HTH	Motivo hélice-curva-hélice
DNA	Ácido desoxorribonucleico
RNA	Ácido ribonucleico
TM	Hélice transmembrana
SP	Peptídeo señal

Sumário

Agradecimentos	4
Resumo	6
Introdução	13
Transdução de sinal	13
Transdução de sinal em eucariotos	13
Transdução de sinal em bactérias	16
Domínios sensores comuns em bactérias	19
O domínio Cache e sensores similares	22
Sensores do grupo das proteínas ligadoras de substrato (SBPs)	24
Interações entre domínios sensores	26
Objetivos	30
Metodos	31
Coleta de genomas	32
Caracterização das sequências proteicas	33
Coleta dos homólogos de SBPs	33
Identificação de famílias de SBPs	34
Busca e classificação de domínios sensores	35
Busca de vizinhanças gênicas	36
Coocorrência de SBPs e domínios sensores	37
Resultados	38
Seleção de genomas completos	38
Análise de domínios sensores	39
Busca por sensores extracitoplasmáticos	40
Domínios sensores de ChiS e MbaA: nova família de domínios Cache	43
Domínios sensores não catalogados no Pfam	43
Análise de proteínas ligadoras de substrato	45
Buscas por homólogos de SBPs	45
Classificação de SBPs	46
Estatísticas das associações entre SBPs e Cache	50
Panorama das associações Cache-SBP	55
SBPs de classe II	55
Comunidade 1	55
Comunidade 2	59
Comunidade 4	61
Comunidade 5	64
SBPs de classe I (comunidade 3)	67

SBPs de classe III (comunidade 6)	70
Discussão	73
Conclusão	78
Referências	79

1. Introdução

1.1. Transdução de sinal

A transdução de sinal pode ser definida como o conjunto de eventos moleculares que resultam em uma resposta do organismo a algum sinal interno ou do ambiente (Bradshaw and Dennis 2003). Em nível celular e molecular, esse processo depende da passagem de informação entre componentes distintos da mesma célula, incluindo a passagem de informação de um compartimento celular para outro ou entre células e tecidos. Esse processo é quase sempre mediado por proteínas, que monitoram o ambiente interno e externo da célula e transmitem a informação para outros componentes intracelulares, alterando o comportamento celular em função dos parâmetros detectados pelos sensores (Parkinson and Kofoed 1992).

As unidades funcionais elementares nas vias de transdução de sinal são os domínios proteicos, ou seja, regiões de uma proteína com história evolutiva e estrutura tridimensional independentes de outras partes da mesma proteína. De forma simplificada, é possível classificar os domínios protéicos em dois tipos de elementos funcionais: (i) domínios sensores, que monitoram sinais químicos ou físicos e (ii) domínio efetores, que traduzem a informação externa em uma alteração química dentro da célula. Um número imenso de combinações de domínios sensores e efetores é encontrado na natureza, resultando numa grande variedade de proteínas de transdução de sinal, em todas as linhagens de organismos (Galperin 2004; Parkinson and Kofoed 1992).

1.2. Transdução de sinal em eucariotos

O estudo da transdução de sinal em eucariotos começa com a descoberta dos hormônios no início do século XX, que deu origem ao campo da endocrinologia. A partir da caracterização química dos hormônios, estudos bioquímicos e genéticos da sinalização em eucariotos permitiu a identificação e caracterização de uma ampla gama de mecanismos que permitem a comunicação entre células em organismos

multicelulares e a resposta de células individuais a estímulos externos. Esses sinais extracelulares são interpretados, em sua maior parte, por receptores de membrana que ativam proteínas intracelulares responsáveis pela resposta ao estímulo, que pode ser mediada pela ativação de canais iônicos, a síntese de mensageiros como cAMP e cGMP, a fosforilação de alvos por serina-treonina ou tirosina quinases e outros mecanismos efetores.

Diversas famílias de receptores participam da transdução de sinal em eucariotos, incluindo tanto proteínas e complexos de proteínas que incluem componentes extracelulares, como receptores citoplasmáticos de sinais que atravessam membranas celulares, como o óxido nítrico, ou de outra forma se difundem entre compartimentos intracelulares. Embora uma descrição detalhada dos sistemas de transdução de sinal de eucariotos esteja fora do escopo deste trabalho, é importante mencionar alguns sistemas arquetípicos desses organismos para ilustrar conexões interessantes destes sistemas com os encontrados em procariotos, os quais, muitas vezes, deram origem aos sistemas eucarióticos (Bradshaw and Dennis 2009). Será interessante também apontar os casos em que esses sistemas usam sensores homólogos a SBPs (SBPs para *substrate binding protein*), a ampla classe de domínios protéicos que são o foco deste trabalho e que descrevemos em detalhe mais adiante.

Os receptores acoplados à proteína G estão entre os mais importantes e abundantes em eucariotos sendo codificados, em mamíferos, por mais de 1000 genes (Strachan and Read 2011) e constituem uma forma interessante de regulação. A proteína G é formada por 3 subunidades e encontra-se no lado interior da membrana plasmática ligado a um GDP e a uma GPCR no seu estado inativado. A GPCR, por sua vez, está embebida na membrana plasmática e é caracterizada pela presença de sete hélices transmembranares (7TM). Quando a GPCR se liga a uma molécula sinalizadora, específica de cada tipo de GPCR, a proteína G é liberada, trocando seu GDP por um GTP e se dividindo em duas subunidades ativadas (Strachan and Read 2011). Uma vez liberada no citoplasma, a subunidade α da proteína G, uma GTPase da superfamília TRAFAC, interage diretamente com outras proteínas dentro da célula e, a partir dessas interações, regula a ativação de proteínas efetoras intracelulares, como adenilato ciclases e a fosfolipase C, que produzem, respectivamente os sinalizadores

intracelulares cAMP e inositol-3-fosfato. Os GPCRs são ubíquos em eucariotos, constituem uma das principais vias para a transmissão de um sinal externo para o ambiente intracelular nesse grupo de organismos, participando de processos importantes, como a regulação do metabolismo energético, a percepção da luz, cheiros e na sensação de gosto, na regulação do sistema imune e na transmissão de sinais nervosos. Notavelmente, embora os sistemas heterotriméricos característicos da proteína G sejam encontrados apenas em eucariotos, proteínas com sete hélices transmembranares, incluindo homólogas de GPCRs, estão presentes em bactérias e arqueas. Um exemplo são as rodopsinas procarióticas, também encontradas em fungos, e que atuam diretamente na sinalização, via transporte fotodependente de íons de cloro ou prótons, ou pela indução de mudanças conformacionais que são transmitidas a outros domínios efetores, como os domínios quimiotáticos em cianobactérias (Anantharaman and Aravind 2003).

Um dos papéis mais importantes das proteínas G é a regulação da atividade de diferentes famílias de adenilato ciclases, enzimas que produzem cAMP, cuja concentração celular afeta a atividade de enzimas, transportadores e outras proteínas regulatórias (Bradshaw and Dennis 2009). As classes I e II de nucleotídeo ciclases são encontradas apenas em bactérias, mas enzimas da classe III são abundantes em bactérias e eucariotos. Uma superfamília de nucleotídeo ciclases da classe III, os receptores natriuréticos, são um grupo de guanilato ciclases caracterizados pela presença de uma única hélice transmembrana conectando os domínios citoplasmáticos a um ou mais domínios SBP da classe I, localizados na porção extracelular (ver classificação de SBPs mais abaixo, na seção “[Sensores do grupo das proteínas ligadoras de substrato](#)”). Os domínios SBP nessas proteínas atuam como sensores de peptídeos natriuréticos, que são hormônios parácrinos e endócrinos produzidos pelo coração, rim e outros órgãos, e que atuam na regulação da pressão sanguínea e no crescimento de ossos (Potter et al. 2009). Outros membros da famílias dos receptores natriuréticos, expressos no intestino, também respondem a peptídeos natriuréticos endógenos e são alvos de toxinas bacterianas.

Os SBPs presentes em receptores natriuréticos pertencem à mesma família de SBPs encontrados em alguns GPCRs, como os receptores metabotrópicos, e em

alguns receptores ionotrópicos, que são exemplos de canais de íons ativados por ligantes (*ligand-gated ion channels*). Na maioria dessas famílias, o domínio SBP tem papel sensor, reconhecendo e ligando aminoácidos ou derivados de aminoácidos, tais como o glutamato e outros neurotransmissores. A relação desses domínios sensores com domínios de famílias relacionadas, encontrados em proteínas sensoras e de transporte em bactérias, é um dos itens que exploramos neste trabalho.

1.3. Transdução de sinal em bactérias

Em vias de transdução de sinal de procariotos, predomina uma organização de domínios sensores e efetores na qual os domínios sensores se localizam majoritariamente na região amino-terminal das proteínas sensoras e os domínios efetores compreendem a região carboxiterminal destas proteínas. A prevalência da organização sensor-efetor impõe um fluxo característico N-para-C para a energia e a informação transmitidas no sistema (Galperin 2004), fluxo que, em linhas gerais, é preservado mesmo em versões mais complexas desses sistemas, onde proteínas ou domínios adicionais agem como intermediários ou regulam os sensores e efetores.

Uma forma simples de classificar os sistemas de transdução de sinal em bactérias é dividi-los, com base no número de genes envolvidos, em sistemas de um ou dois componentes disjuntos. Os sistemas de um componente (fig. 1A) consistem de uma única proteína que contém um domínio sensor e um domínio efetor na mesma cadeia polipeptídica. A maioria dos sistemas de um componente (~84%) são proteínas citoplasmáticas, nas quais o domínio efetor é um domínio de ligação ao DNA, geralmente um HTH, que conecta o sensor diretamente ao processo de regulação transcricional. O repertório de mecanismos efetores em sistemas de um componente, porém, compreende também outras atividades, como demonstrado pelos muitos exemplos de sistemas de um componente que regulam os níveis de diferentes nucleotídeos cíclicos (Luke E. Ulrich, Koonin, and Zhulin 2005).

Sistemas de dois componentes (Fig. 1B) são definidos pela presença de duas proteínas: uma histidina quinase contendo um ou mais domínios sensores e um membro da família dos reguladores de resposta, que é caracterizada pela presença de pelo menos um domínio receptor (REC), especializado em receber o grupo fosfato da

histidina quinase para então passar a informação, por meio de rearranjos estruturais, para um domínio efetor na mesma proteína, que pode ser um domínio de ligação ao DNA ou uma nucleotídeo ciclase (Luke E. Ulrich, Koonin, and Zhulin 2005). Embora cerca de cinco famílias distintas de histidina quinases sejam catalogadas pelo banco de dados Pfam no mesmo *fold* (HisKA (PF00512), HisKA_2 (PF07568), HisKA_3 (PF07730), and HWE_HK (PF07536)), a maior parte da versatilidade das diferentes famílias de histidina quinases deriva da diversidade do domínio sensor, que pode ser extracelular ou citoplasmático (Galperin and Schultz 2019). Um componente não-universal de vias de transdução mediadas por histidina quinases é o domínio ou proteína de fosfotransferência (HPt), que tem uma estrutura de quatro α -hélices em feixe, semelhante ao domínio HisKA, e uma histidina conservada que pode receber o grupo fosfato do domínio enzimático da histidina quinase antes que esse seja transferido ao regulador de resposta, implicando variações no sistema de transferência de fosfato que tornam difícil prever o caminho e a cinética das vias de fosforilação, principalmente nos casos em que o sistema possui vários domínios REC e HPt.

Figura 1. Arquitetura típica, simplificada, de proteínas de transdução de sinal. (A) Arquitetura típica de sistemas de transdução de sinal de um componente, com domínio sensor na região N-terminal e efetor na região C-terminal, separados por uma região transmembrana, característico das proteínas bitópicas. (B) Típico sistema de transdução de sinal de dois componentes, com domínio histidina quinase na região da região C-

terminal da proteína sensora transferindo um grupo fosfato para um receptor de resposta da região N-terminal da proteína efetora.

Uma forma especializada de sistemas de dois componentes são as vias de sinalização quimiotáticas dependentes de sensores da família proteínas aceptoras de grupos metil (ou mcps, de *methyl-chemotactic proteins*). Membros dessa família são sensores com domínios extracelulares variados, conectados a um domínio citoplasmático (MCP) formado por duas longas α -hélices antiparalelas, unidas por um curva em grampo (*hairpin*). A região que une as hélices contém uma assinatura de resíduos altamente conservados, que interagem com a proteína quinase CheA regulando a atividade da mesma. CheA depende das interações com o domínio MCP dos sensores para regular sua atividade e fosforilar a proteína CheY. CheY é uma proteína composta por apenas um domínio regulador de resposta que, em sua forma ativada (fosforilada), interage com a proteína FlhM do flagelo, causando alterações conformacionais em outras proteínas do flagelo. Esse processo culmina na reversão do sentido de rotação do flagelo e na orientação do movimento da bactéria a favor ou ao contrário da maior concentração do ligante (Galperin and Schultz 2019).

Diferentes tipos de nucleotídeo ciclases podem ser encontradas em sistemas de transdução de sinal em arqueas e bactérias, onde desempenham papéis centrais na regulação de processos celulares e ocorrem tanto em sistemas de um como de dois componentes. A adenilato ciclase de *Escherichia coli*, uma enzima da classe I de adenilato ciclases, é diferente da maioria das nucleotídeo ciclases presentes em vias de transdução de sinal de bactérias, que são da classe III (Pei and Grishin 2001). Essa enzima, embora associada à membrana, não parece sentir sinais externos mas é modulada pelo transportador de glicose do sistema de fosfotransferência (proteína EIIA^{Glc}), que desativa a enzima quando a célula está num meio com glicose extracelular, levando ao fenômeno conhecido como repressão catabólica (Deutscher 2008). Adenilato ciclases da classe III foram primeiro identificadas em sistemas de um componente da cianobactéria *Arthospira platensis*, organismo no qual as variações dos níveis de cAMP e da expressão da adenilato ciclase CyaC são fortemente correlacionados com a intensidade da luz (Kashith et al. 2016). A importância de sensores com atividade de adenilato ciclases na regulação de fatores de

patogenicidade também foi demonstrada em bactérias como *Pseudomonas aeruginosa* (Smith, Wolfgang, and Lory 2004). Além de cAMP, dois outros nucleotídeos cíclicos. c-di-GMP e c-di-AMP são amplamente utilizados por bactérias como mensageiros secundários em sistemas de um ou dois componentes. Assim como, o 3'-5' di-guanilato cíclico, que é sintetizado a partir de um GTP pela nucleotídeo ciclases de classe III, GGDEF, e pode ser hidrolisado por enzimas das famílias HD-GYP e EAL. Milhares de receptores em procariotos combinam essas enzimas com domínios sensores e participam de processos como a transição de formas móveis, incluindo a formação de biofilmes e a montagem da maquinaria flagelar, e em dezenas de outros processos celulares. A presença de muitas proteínas contendo tanto domínios de síntese (GGDEF), como degradação de c-di-GMP (EAL), sugere que pode haver uma regulação do balanço entre síntese e degradação, em alguns casos, bem como uma subpopulação dessas proteínas nas quais um dos dois domínios enzimáticos é inativo e assume um papel regulatório (Galperin and Schultz 2019).

Um ponto comum na atividade de vários dos domínios efetores discutidos acima (histidina quinases, nucleotídeo ciclases como GGDEF e o domínio MCP) é a importância da oligomerização das proteínas sensores no processo de transdução de sinal. Embora a interação do receptor e ligante não seja necessária para a dimerização de várias histidina quinases e GGDEFs, o dímero é a forma ativa dessas enzimas e, portanto, um pré-requisito. Ainda que a dimerização do domínio citoplasmático seja possível na ausência do domínio sensor, fica claro que o sensor é necessário para induzir as exatas mudanças estruturais em dímeros e monômeros que efetuam a passagem de informação pelo sistema.

1.4. Domínios sensores comuns em bactérias

O universo de domínios sensores conhecidos em procariotos compreende uma ampla variedade de superfamílias de domínios protéicos, incluindo famílias que evoluíram a partir de domínios com atividade enzimática, de sistemas de transporte ou que compreendem apenas domínios especializados em atividades de sinalização. Estruturalmente, tanto domínios globulares como proteínas integrais de membrana podem atuar como sensores. Mencionamos aqui alguns dos principais exemplos de

cada uma dessas categorias estruturais e funcionais.

As rodopsinas bacterianas, mencionadas anteriormente em conexão aos GPCRs, seus homólogos eucarióticos, são um exemplo de proteínas sensoras transmembranares presentes em procariotos. Rodopsinas bacterianas têm retinol como grupo prostético e atuam como sensores luminosos em diferentes linhagens de bactérias e arqueas. Essas proteínas sofrem alterações estruturais quando o retinol salta para o estado excitado ao ser exposto à luz e tais alterações são transmitidas para os domínios efetores, que em alguns casos estão fusionados à rodopsinas, e que provocam alterações no padrão de mobilidade, atraindo ou afastando o organismo da luz (Purcell and Crosson 2008). A família 7TMR-DISM é outra família de receptores procarióticos, não relacionados às rodopsinas, mas também caracterizados pela presença de sete hélices transmembranares. Essa família está amplamente distribuída entre bactérias gram-negativas e gram-positivas, sendo especialmente abundante em espiroquetas. Membros dessa família aparecem em múltiplas combinações com domínios efetores citoplasmáticos, tais como GGDEF, histidina quinases e o domínio acceptor de metil (MCP) (Anantharaman and Aravind 2003).

Alguns domínios sensores globulares são membros inativos de famílias de enzimas que perderam os resíduos catalíticos, mas que preservam os sítios de ligação a substratos e/ou co-fatores e agem como sensores da concentração intracelular desses ligantes. Muitos fatores de transcrição em sistemas de um componente, por exemplo, combinam domínios enzimáticos inativos com domínios de ligação a DNA. Os domínios sensores dos fatores de transcrição das famílias DeoR e SocR, por exemplo, são formas inativas do *fold* ISOCOT, que inclui açúcar isomerases, acetil-CoA transferases e a enzima meteniltetrahidrofolato hidratase. Os domínios ISOCOT de membros da família DeoR atuam como sensores de formas derivadas de deoxinucleosídeos, galactosamina e L-ascorbato (Anantharaman and Aravind 2006).

Em alguns casos, porém, a relação com as versões cataliticamente ativas do domínio não é tão clara e o domínios sensor pode ter emergido, ao longo da evolução do *fold*, de uma linhagem distinta daquela que deu origem às enzimas. Nesses casos, porém, elementos estruturais compartilhados por sensores e enzimas garantem a formação dos sítios para reconhecimento de ligantes e substratos. Esse parece ser o

caso do regulador AraC, no qual um domínio do *fold* DSBH (*double-stranded β -helix*) (Iyer et al. 2010) aparece fusionado a um domínio HTH C-terminal. O domínio N-terminal de AraC reconhece o açúcar arabinose e é responsável pela dimerização da proteína. Dentre os membros cataliticamente ativos desse *fold* estão várias isomerases de carboidratos, dioxigenases, histona metilases e proteína hidroxilases. Os membros do *fold* DSBH compartilham uma estrutura de oito fitas- β dispostas em hélice, formando uma cavidade onde os resíduos catalíticos são posicionados de modo a formar um sítio para ligação a pequenas moléculas. Como em AraC, os sensores deste *fold* perderam ou nunca tiveram os resíduos catalíticos presentes em outras famílias nessa cavidade (Spiro and Dixon 2010). Outra superfamília importante de sensores do *fold* DSBH é a região regulatória dos fatores de transcrição da família CAP, que reconhecem cAMP e regulam o operon *Lac* em *Escherichia coli*. Membros da superfamília estão presentes em múltiplos fatores de transcrição de procariotos, reconhecendo cAMP e outros substratos, e também podem ser encontrados em receptores ionotrópicos de vertebrados (Yau 1994), (Busby and Ebright 1999).

Outros domínios sensores globulares importantes são membros de *folds* diversos, tais como CBS, DICT, NIT, STAS, 4HB, diferentes classes de SBPs e o grupo dos domínios similares à profilina (*Profilin-like* ou *PAS-like*). CBS e STAS são dois exemplos de domínios exclusivamente intracelulares. Os domínios CBS são constituídos por dois subdomínios $\alpha+\beta$ simétricos, cada um deles formado por folhas de três fitas- β e duas α -hélices, com a superfície das folhas empilhadas de forma a criar um domínio globular compacto. CBSs aparecem em todos os grupos de organismos e reconhecem diferentes formas do nucleosídeo adenosina, tais cAMP, AMP, ATP e S-adenosilmetionina. A partir da interação com esses nucleotídeos, eles regulam os domínios enzimáticos aos quais aparecem fusionados, tanto em vias metabólicas como de sinalização (Bateman 1997; Ponting 1997; Spiro and Dixon 2010). Os domínios STAS possuem uma estrutura α/β de 4 ou 5 fitas e foram inicialmente identificados em porções intracelulares de transportadores de sulfato e reguladores de antagonistas de fatores anti-sigma (Aravind and Koonin 2000). Versões do domínio STAS caracterizadas experimentalmente ligam-se a GTP ou ATP e aparecem em combinações com outros domínios sensores intracelulares, como PAS e GAF, e

extracelulares, como Cache, formando sensores que podem integrar múltiplos sinais, como no caso da proteína YtvA, de *Bacillus*, que combina um PAS sensível à luz com um STAS que se liga a GTP (Spiro and Dixon 2010).

O domínio DICT é um domínio sem estrutura ou ligantes conhecidos mas comum em actinobactérias, bactérias fotossintéticas e arqueas halofílicas, e que aparece quase sempre associados com domínios de sistemas de dois componentes e diguanilato ciclases (Spiro and Dixon 2010). O domínio NIT é formado por 8 a 10 α -hélices e aparece predominantemente em bactérias, como domínio sensor extracelular de sistemas de dois componentes, nucleotídeo ciclases e sensores quimiotáticos, ou como sensor intracelular de anti-terminadores de transcrição. O mesmo domínio é também encontrado em eucariotos, onde aparece como domínio extracelular de um grupo de guanilato ciclases (Shu, Ulrich, and Zhulin 2003). A proteína NasR regula a terminação da transcrição de genes de utilização de nitrato em *Klebsiella oxytoca* e é regulada por variações na concentração de de nitrato ou nitrito detectados pelo seu domínio NIT C-terminal (Shu, Ulrich, and Zhulin 2003).

Pelo menos três superfamílias de domínios contendo o motivo estrutural 4HB (*four-helical up-and-down bundle*) são encontradas em proteínas de transdução de sinal de bactéria. Este *fold* está presente nos sensores extracelulares mais comum em transdutores quimiotáticos bacterianos e ocorre também em sistemas de dois componentes e fusionado a nucleotídeo ciclases. São membros desse grupo o sensor de aspartato TarH e a família NarX/NarQ/PilJ, que inclui o domínio sensor de nitrito e nitrato da proteína quinase NarX, bem como muitos outros sensores de aminoácidos, açúcares, metais e peptídeos que aparecem fusionados a outras famílias de domínios efetores (Spiro and Dixon 2010; L. E. Ulrich 2005).

1.4.1. O domínio Cache e sensores similares

Os domínios semelhantes à profilina (*Profilin-like* ou *PAS-like*), são caracterizados por uma folha composta por 5 fitas- β , na ordem 21534, e pelo menos duas α -hélices, que formam um arcabouço de estruturas secundárias compartilhada pelos domínios de transdução de sinal deste *fold*, que inclui as superfamílias GAF, PAS, CACHE, HNOBA, PocR e GSU0582-like (Spiro and Dixon 2010). Os domínios

PAS e GAF são superfamílias de domínios sensores de localização estritamente intracelular e extremamente abundantes em bactérias, arqueas e eucariotos. Membros dessas superfamílias reconhecem uma ampla gama de ligantes, aos quais se ligam covalente- ou não-covalentemente, incluindo vários tipos de pequenas moléculas solúveis, nucleotídeos e grupos prostéticos, incluindo a coordenação de anéis tetrapirrólicos como o heme e d α ficocianobilina, que permitem sentir a concentração de gases, mudanças no potencial redox ou na intensidade da luz (Upadhyay et al. 2016; Henry and Crosson 2011). Interessantemente, alguns membros da superfamília GAF, como a região sensora de LuxR, conhecida como domínio ligador de autoindutores, têm afinidade por N-acilhomoserina lactonas e desempenham papéis em processos de *quorum sensing*, ou seja, na percepção da densidade populacional (Chai, Zhu, and Winans 2001).

Os ligantes de proteínas com domínios *Profilin-like* ainda são desconhecidos mas já foi observado que alguns membros desse grupo, especialmente os domínios PAS e GAF clássicos, formam dímeros e podem também interagir com outras proteínas (Spiro and Dixon 2010). A frequente presença desses domínios em arranjos de múltiplas cópias sugere que o papel dessas interações homotípicas, na formação de dímeros, e interações com outras proteínas pode ser muito relevante para sua função em sistemas de transdução de sinal.

O domínio Cache é o sensor extracelular mais abundante em bactérias, podendo ser encontrado em uma ampla variedade de vias de transdução de sinal (Anantharaman and Aravind 2000; Upadhyay et al. 2016). Uma característica exclusiva dos domínios Cache, não compartilhada com os outros membros do *fold* “Profilin-like”, é a presença de uma longa alfa-hélice, que alguns autores sugerem ser a principal característica estrutural que permitiu a interação desse domínio com as membranas celulares e sua adaptação para localização extracelular (Z. Zhang and Hendrickson 2010; Upadhyay et al. 2016; Cheung et al. 2008). Upadhyay e colaboradores, em 2016, introduziram os termos *sCache* (*single Cache*) e *dCache* (*double Cache*) (Upadhyay et al. 2016) para diferenciar famílias de Cache caracterizadas, respectivamente, pela presença de uma ou de duas cópias da unidade estrutural mínima, *Profilin-like*, dos domínios Cache. Diferentes famílias de *dCache* parecem ter surgido várias vezes ao

longo da evolução, a partir da duplicação de um domínio *sCache* na mesma proteína, ou da fusão de membros de famílias diferentes de *sCaches* (Upadhyay et al. 2016). Em linha com a versatilidade funcional dos domínios sensores *Profilin-like*, famílias de *sCache* e *dCache* relacionadas, ou seja, que compartilham núcleos estruturais da mesma família *sCache* (Tabela 1), podem não ter afinidade pelo mesmo tipo de ligante. Um exemplo disso é encontrado nas famílias *sCache_1* e *dCache_1*: enquanto a maioria dos *sCache_1* se ligam a ácidos orgânicos, a maior parte dos membros da família *dCache_1*, que deriva de *sCache_1* por duplicação, interagem com aminoácidos (Upadhyay et al. 2016).

Tabela 1. Tipo e abundância de domínios Cache. Os nomes dos domínios são os usados pelo banco de dados Pfam, versão 32. O número de proteínas corresponde ao total detectado no banco de dados Uniprot, versão 2012_06. Adaptado de Zhulin et. al. (2016).

Família	Tipo	Proteínas
dCache_1	<i>dCache</i>	15569
CHASE	<i>dCache</i>	1214
CHASE4	<i>sCache</i>	529
Diacid_rec	<i>sCache</i>	1274
dCache_2	<i>dCache</i>	299
2CSK_N	<i>sCache</i>	966
PhoQ_Sensor	<i>sCache</i>	556
LuxQ-periplasm	<i>dCache</i>	115
DUF2222/sCache4	<i>sCache</i>	713
SMP_2	<i>sCache</i>	788
Ykul_C	<i>sCache</i>	277
Stimulus_sens_1	<i>sCache</i>	203
dCache_3	<i>dCache</i>	883
sCache_2	<i>sCache</i>	299
Cache_3-Cache_2	<i>dCache</i>	407
sCache_3_1	<i>sCache</i>	2854
sCache_3_3	<i>sCache</i>	276
sCache_3_2	<i>sCache</i>	2499

1.4.2. Sensores do grupo das proteínas ligadoras de substrato (SBPs)

Proteínas ligadoras de substrato ou SBPs, do inglês ***S**ubstrate-**b**inding **p**roteins*, também conhecidas como *Periplasmic Binding Proteins* (PBPs), estão presentes em todos os organismos e compreendem três grupos distintos de proteínas da classe α/β (Andreeva et al. 2014), todas elas caracterizadas pela presença de duas regiões homólogas e quase-simétricas, que se movimentam por meio de uma região flexível localizada entre os dois subdomínios. As proteínas desta classe sofrem mudanças conformacionais, ao interagirem com seus ligantes, que induzem o estreitamento da cavidade entre os dois subdomínios, levando ao aprisionamento de moléculas nessa cavidade. Tal mecanismo lembra o movimento das dionéias ou apanha-moscas (*Dionaea muscipula*), plantas carnívoras que aprisionam pequenos artrópodes, usando suas folhas em forma de lóbulos, para obtenção de nitrogênio, o que valeu o nome cognato de armadilha de Vênus (*Venus fly trap*) para o modelo de mudança conformacional na interação com o ligante aceito para este grupo de proteínas.

Os SBPs foram originalmente descritos como componentes de sistemas transportadores ABC, mas são também componentes importantes e ubíquos de vários outros sistemas, tais como: transportadores independentes de ATP (TRAP), canais de íons dependentes de ligante, guanilato ciclases que atuam como receptores de peptídeos natriuréticos atriais, sistemas de regulação de dois componentes de procariotos, receptores ligados à proteína G de eucariotos e proteínas envolvidas na regulação gênica de procariotos (Fukami-Kobayashi, Tateno, and Nishikawa 1999) .

Os três grupos de domínios protéicos denominados SBPs correspondem a três classes estruturais (*folds*) com arranjos de estruturas secundárias semelhantes, porém distintos, e que compartilham o movimento de *Venus fly-trap* na interação com o ligante: (i) o *fold* SBP-I, caracterizado pela presença de cinco fitas- β paralelas; (ii) o *fold* SBP-II, onde uma das cinco fitas- β aparece posicionada de forma antiparalela à demais fitas e (iii) a classe SPB-III, cuja suas fitas- β são paralelas porém possui apenas 4 fitas quando comparado com SBP-I, e diverge dos outros *folds* dois por incluir uma α -hélice que liga os dois subdomínios α/β (Fukami-Kobayashi, Tateno, and Nishikawa 1999; Berntsson et al. 2010). Em 2016, Scheepers e colaboradores dividiram os 3 *folds* em 7 superfamílias com base em alinhamentos pareados das estruturas disponíveis no *Protein Data Base* (PDB) (Tabela 2) (Scheepers, Lycklama A Nijeholt, and Poolman 2016).

Tabela 2. Classificação e ligantes de proteínas ligadoras de substratos (*Substrate-Binding Proteins* ou SBPs). A coluna C (classe) se refere ao *fold* e a coluna G (grupo) corresponde às superfamílias descritas por Scheepers e colaboradores, com base na similaridade estrutural dos domínios SBP. Os identificadores da última coluna são exemplos de estruturas de proteínas em cada grupo (Scheepers, Lycklama a Nijeholt, and Poolman 2016; Berntsson et al. 2010; Scheepers, Lycklama A Nijeholt, and Poolman 2016).

G	C	Características	Ligantes	PDB
A	III	Os dois subdomínios são conectados por uma α -hélice ao invés de um <i>loop</i> (região não-estruturada). Membros desse grupo fazem parte de transportadores ABC.	Metais, sideróforos, grupo heme, vitamina B12, molibdato, tungstato	1N2Z
B	I	Transportadores ABC, sensores histidina quinase e receptores guanilato-ciclase de peptídeos natriuréticos atriais	Açúcar, álcool, autoindutores, aminoácidos, ácidos aromáticos, compostos natriuréticos, glutamato	1DRJ
C	II	Possuem um domínio extra e interagem com sistema de transportadores ABC	oligopeptídeos, celulobiose, níquel e opine	3DRF
D	II	A região hinge desse domínio é formada por dois pequenos segmentos de 4 a 5 aminoácidos.	açúcar, álcool, espermina e/ou putrescina, tiamina, fosfato, molibdato, tungstato, íons de ferro	1POT

E II	Todos os integrantes desse grupo fazem parte de transportadores TRAP e possuem uma cadeia beta que flanqueiam ambos domínios assim como uma alfa hélice que atravessa os domínios	ácidos orgânicos, álcool amino, dipeptídeos, glicerol 3-fosfato, aminoácidos	3FXB
F II	A região flexível entre os dois subdomínios é formada por dois longos segmentos de 8 a 10 aminoácidos cada um	tiamina, pirimidina, sulfonato, bicarbonato, metionina, aminoácidos	1HSL
G -	Massa molecular de aproximadamente 60kDa	não determinado	1Y3P

1.5. Interações entre domínios sensores

O modelo mais simples para o funcionamento de uma via de sinalização em procariotos coloca o domínio sensor, citoplasmático ou extracelular, como única região envolvida na percepção do sinal regulatório. Muitas vias, porém, envolvem não apenas a interação do ligante com o domínio sensor, mas a regulação da atividade do domínio efetor por outras moléculas. Um exemplo disso é a autofosforilação do regulador de resposta UvrY, que ocorre durante o crescimento de *Escherichia coli* em meio com glicose (McCleary and Stock 1994). Nessas condições, as células produzem um excedente de acetil-CoA, que não é assimilado pelo ciclo de Krebs mas convertido em acetato, que se acumula no meio. Na fase exponencial tardia, a alta concentração de acetato ativa a histidina quinase BarA, que fosforila UvrY, permitindo que este último interfira na transcrição de genes reguladores do catabolismo, como CsrA/CsrB. Ao mesmo tempo, o acetato em excesso passa a ser assimilado e convertido em fosfoacetato no citoplasma, que promove autofosforila de UvrY, que passa a ser ativado por essa via, independentemente de BarA. Mecanismos como esse, de regulação citoplasmática, são um exemplo da maior complexidade que as vias de transdução de sinal podem exibir no processamento do sinal de entrada. Outro exemplo de múltiplas entradas está na capacidade de várias histidinas quinases convergirem para um mesmo alvo, como ocorre nas vias de percepção da densidade populacional (*quorum sensing*) de *Vibrio harveyi*, onde as histidina quinases sensoras LuxN, LuxQ e CqsS ligam-se a diferentes moléculas autoindutoras e fosforilam a proteína de fosfotransferência LuxU (domínio HPt no banco de dados Pfam), que transfere o grupo fosfato para o regulador de resposta LuxO, integrando os diferentes sinais na mesma resposta final (Alvarez et al. 2016).

Em *Escherichia coli*, a histidina quinase DcuS regula a expressão dos genes *dctCBA*, componentes do transportador de C₄-dicarboxilatos. Na ausência da proteína DctA, DcuS mantém uma conformação ativa, independente da concentração de dicarboxilatos e, portanto, induz a expressão dos componentes do transportador. Quando DctA estiver presente, formará um complexo com DcuS. Ao contrário de DcuS livre, o complexo DcuS-DctA só demonstra atividade de fosforilação em altas concentrações de substratos do transportador. O sistema age, portanto, como um operador lógico OR, que eleva a expressão dos genes do transportador quando há baixas concentrações de componentes do transportador ou quando a concentração do substrato for elevada (Piepenbreier, Fritz, and Gebhard 2017).

DctA é um membro da superfamília de transportadores catiónicos e, portanto, uma proteína integral de membrana. A interação de DctA e DcuS ocorre no citoplasma, com contatos entre uma hélice não conservada de DctA e um domínio PAS em DcuS (Witan et al. 2012). Notavelmente, os exemplos mais bem conhecidos de integração ou ramificação de sinais na fase extracelular envolvem a interação de SBPs com proteínas regulatórias (Piepenbreier, Fritz, and Gebhard 2017). O primeiro caso reportado corresponde à descoberta de que a histidina quinase BctE é ativada, na presença de citrato, pela proteína ligadora de substrato BctC, do grupo II-E. BctE regula a expressão do operon *bctCBA*, que codifica os componentes do transportador tripartido de tricarboxilatos (TTT), responsável pela captação de citrato em *Bordetella pertussis*. BctC também é responsável por fornecer o citrato ao complexo da permeases do sistema (BctAB) e a ativação de BctE por BctC eleva os níveis de expressão do operon *bctCBA*, aumentando a eficiência da captação de citrato. Notavelmente, o domínio periplasmático de BctE é um domínio Cache da família 2CSK_N, constituindo o primeiro exemplo de associação entre SBPs e domínios Cache.

Mais recentemente, foi demonstrada que o gene NpsS de *Vibrio cholerae*, um SBP-II, que regula a atividade da proteína MbaA, faz isso por meio da interação direta com o domínio *dCache* dessa proteína, uma dinucleotídeo fosfodiesterase da família EAL com um domínio GGDEF inativo (Cockerell et al. 2014). A intensidade da atividade de ciclase de MbaA é afetada pela concentração do ligante, espermidina ou norespermidina, mas cessa quando a proteína NpsS não está ligada a nenhuma

poliamina. NspS não tem impacto no transporte de poliaminas, que em *Vibrio* é mediado por outro SBP (PotD), mas tem efeito sobre a formação de biofilmes via atividade de fosfodiesterase de MbaA (Sobe et al. 2017). O domínio sensor de MbaA, porém, não foi classificado (ver Resultados).

Em *Vibrio cholerae*, a proteína ligadora de oligossacarídeos de quitina, conhecida como CBP (VC0620), um SBP do grupo C (ver Tabela 2), foi identificada como repressor de ChiS (VC0622), a histidina quinase que regula a expressão dos genes de degradação de quitina neste organismo (Fukamizo, Kitaoku, and Suginta 2019; Li and Roseman 2004). CBP forma um complexo estável com ChiS, independente da presença ou não de quitina, mas mantém ChiS em uma conformação na qual predomina a atividade de autofosforilação quando não há oligossacarídeos no meio. A autofosforilação do domínio REC de ChiS suprime sua atividade de ativador da transcrição do programa de degradação de quitina. Na presença de quitina, o complexo CBP-ChiS muda de conformação e passa a favorecer a atividade de fosforilase, levando à ativação do domínio de ligação a DNA desta proteína e o recrutamento da RNA polimerase para a ativação dos genes de degradação de quitina (Klancher et al. 2020b). Notavelmente, como os genes CBP e ChiS estão localizados no mesmo operon que codifica a ATPase e as permeases que transportam oligossacarídeos, bem como das enzimas da via de degradação destes compostos, cuja expressão é regulada por ChiS. Isso significa que a proteína CBP age, ao mesmo tempo, no transporte de oligossacarídeos e na regulação do sistema de degradação destes compostos.

Um sistema semelhante ao NpsS-MbaA foi também identificado em *Xanthomonas citri* pv. *citri* str. 306, porém, nesse caso, o domínio efetor tem atividade de diguanilato ciclase e o domínio sensor periplasmático é um sCache. Os autores provaram a existência de uma interação física entre a proteína codificada pelo gene XAC2382, um SBP-II, com o produto do gene vizinho, XAC2383, que codifica uma proteína da família GGDEF com domínio sensor da família sCache_3_2 (Teixeira et al. 2018). Embora esses autores não tenham conseguido encontrar um ligante de alta afinidade, eles identificaram uma preferência por compostos carregados negativamente contendo íons fosfato. Esses autores fizeram ainda a primeira análise filogenética dos domínios citoplasmáticos e extracelulares dessas proteínas, demonstrando que o par

de domínios SBP e Cache nesses genes permanece associado, como uma unidade conservada, enquanto o domínio efetor citoplasmático muda, podendo ser substituído por histidina quinases ou MCPs (Teixeira et al. 2018).

Finalmente, outros exemplos de interação de SBPs com domínios sensores foram reportados, como a interação do sensor quimiotático Tar de *Escherichia coli*, um MCP com domínios periplasmáticos do *fold* 4HB, e a proteína MBP, um SBP especializado no reconhecimento do açúcar maltose (Gardina et al. 1997). No caso do sistema Tar-MBP, a associação dos dois sensores responde à presença dos dois ligantes de forma independente, com um efeito aditivo, a metilação do domínio citoplasmático aumentando com a soma das concentrações de aspartato e maltose (Mowbray and Koshland 1987).

Apesar das evidências dispersas nesses vários estudos, nenhuma análise sistemática das associações SBP-sensor foi realizada. Os exemplos acima sugerem que domínios Cache distintos, tanto *dCache* como *sCache*, e com afinidade a diferentes substratos, podem depender de interações com SBPs para efetivar seu papel em processos de transdução de sinal. Se diferentes famílias de SBPs convergiram para uma interação análoga com membros da superfamília CACHE, talvez esteja ocorrendo a reutilização de um mesmo mecanismo de interação ou, alternativamente, uma vantagem seletiva possa advir da cinética regulatória decorrente do uso de domínios SBPs combinado com domínios sensores clássicos. Uma análise sistemática desse fenômeno tem potencial para revelar mais detalhes sobre esses mecanismos de interação e sobre a função de diferentes famílias de sensores.

Neste trabalho buscamos inferir a história evolutiva dos sistemas onde ocorreu o recrutamento de SBPs para vias de transdução de sinal, com enfoque nas proteínas que possuem o domínio Cache como intermediário neste processo. Nossos dados demonstram que o recrutamento de SBPs em possíveis processos de transdução de sinal tem origem em proteínas primariamente envolvidas com sistemas de transporte.

2. Objetivos

Múltiplas classes de domínios que ligam substratos por meio de um mecanismo do estreitamento do sítio de ligação podem ser encontrados em transportadores presentes em todas as linhagens de organismos. Tais domínios são coletivamente chamados de proteínas ligadoras de substratos e versões especializadas desses domínios podem ser identificadas em muitas vias de transdução de sinal presentes em arqueas, bactérias ou eucariotos.

No presente trabalho estudamos a origem e evolução de domínios SBP em vias de transdução de sinal de procariotos. Nosso objetivo principal foi identificar e catalogar todos os eventos de recrutamento de SBPs extracelulares para funções regulatórias em via de transdução de sinal, nas quais esse domínio passa a interagir com outro domínio presente nessas vias. Uma vez completado esse catálogo, nosso objetivo secundário foi determinar os tipos mais frequentes de alterações na arquitetura das vias afetadas, ou seja, se há aumento ou redução da complexidade desses sistemas.

Os objetivos específicos que fizeram parte deste trabalho, compreendem:

- Ampliar e atualizar a classificação dos SBPs, apresentando uma classificação natural baseada na similaridade desses domínios;
- Identificar novos domínios da superfamília Cache, que se revelou a mais frequentemente associada com SBPs em vias de transdução de sinal bacterianas;
- Revelar novos domínios sensores que possam interagir com SBPs e atuar na integração de sinais em vias de transdução de sinal.

3. Metodos

Este trabalho foi desenvolvido se baseando em classificação de famílias de domínios proteicos com métodos de homologia remota e análise de contexto dentro de um conjunto de genomas procarióticos completos e não redundantes. Nos capítulos subsequentes serão descritos: coleta, agrupamento e seleção de representativos dentro de um conjunto de genomas completos; coleta de fragmentos homólogos de SBPs e de região intracitoplasmática de proteínas com domínios efetores de transdução de sinal; classificação dos fragmentos coletados em famílias em função do agrupamento das sequências baseado em homologia; agrupamento das famílias de proteínas em comunidades baseado em alinhamento de HMMs das famílias; e por último, análise do contexto genômico das famílias de proteínas baseado em filogenia e vizinhança gênica. O fluxograma deste trabalho está descrito na figura 2.

Figura 2. Fluxograma de análise deste trabalho: Começa pela coleta de genomas procarióticos completos e não redundantes; coleta de homologos de SBP e Cache; classificação das proteínas por identidade de sequência e coleta de vizinhança gênica; análise comparativa e de

contexto genômico das proteínas coletadas

3.1. Coleta de genomas

Nossas análises focaram um conjunto representativo de genomas de procariotos anotados pelo RefSeq (*Reference Sequence Database*) (Haft et al. 2018) disponibilizados pelo NCBI (*National Center for Biotechnology Information*) em 18 de maio de 2018, e anotados como *latest* e *complete genomes*. Foram coletados destes genomas todas as sequências de proteínas, nucleotídeos e anotação gênica (*feature table*).

A anotação dos genomas foi carregada em um banco de dados relacional PostgreSQL, onde pôde ser consultado usando a linguagem SQL ou com programas auxiliares desenvolvidos especificamente para este projeto. As proteínas de cada genoma foram concatenadas em um único arquivo para uso em buscas de similaridade. As sequências de nucleotídeos de todos os genomas foram comparadas entre si usando o programa Mash (Ondov et al. 2016), com *sketch* de 5000 câmeros de 27 nucleotídeos de comprimento, resultando, para cada par de genomas, em uma distribuição de valores da distância Mash e em uma métrica de confiabilidade (*p-value*). Os dados gerados foram filtrados para manter valores de *p-value* menores que 10^{-10} e distância Mash menor que 0.05. Os genomas foram então agrupados pelo algoritmo *UnionFind*, que detectou os componentes conectados no grafo das relações de similaridade entre os genomas. Nosso critério de seleção deu preferência a genomas de referência e, na ausência destes, genomas representativos, sendo mantidos apenas genomas anotados como *full* (segundo a anotação do Refseq). Genomas que não atendiam a nenhum dos critérios citados foram removidos.

Adicionalmente, no caso de genomas não anotados como genomas de referência, foram removidos todos os genomas onde a taxa de pseudogenes, provenientes da anotação gerada pelo NCBI *Prokaryotic Annotation Pipeline* do RefSeq, correspondia a mais de 40% do total de genes preditos.

3.2. Caracterização das sequências proteicas

Os proteomas de todos os organismos selecionados foram anotados com base nos modelos de Markov ocultos do banco de dados Pfam (versão 32.0), usando o programa *hmmsearch* do pacote HMMER (versão 3.1b). As coordenadas das regiões identificadas como “envelope”, para cada um dos resultados da análise de cada modelo, foram usadas para extração da região correspondente aos domínios SBPs em cada proteína. O programa Phobius (Ondov et al. 2016; Käll, Krogh, and Sonnhammer 2004) foi usado para predição de regiões transmembrana e de peptídeo sinal, o que permitiu identificar proteínas transmembranares, extracelulares ou periplasmáticas. O conjunto dessas predições foram combinados usando *scripts* em Perl desenvolvidos por nosso grupo e que geraram uma descrição sucinta da arquitetura de domínios de cada proteína.

3.3. Coleta dos homólogos de SBPs

Foram coletadas, no banco de dados PDB, 501 estruturas de proteínas descritas como *substrate-binding protein* (Scheepers, Lycklama A Nijeholt, and Poolman 2016), além da proteína 5UB6 referente ao gene XAC2383 (Teixeira et al. 2018). Para cada uma das estruturas coletadas, a extração das sequências de aminoácidos correspondente à estrutura foi feita usando o pacote Bio3D (Skjærven et al. 2014). Para a proteína 2E4U esse processo foi feito manualmente com auxílio do software Pymol (DeLano 2002), devido à necessidade de isolar o domínio SBP de um segundo domínio carboxi-terminal presente nessa estrutura. A redundância desse conjunto inicial de sequências com estrutura resolvida foi reduzida, usando o pipeline *easy-cluster* do pacote MMseqs2 (Steinegger and Söding 2017). Nessa análise, pares de proteínas foram considerados redundantes quando a identidade do par era maior ou igual a 80% e a cobertura, ou seja, a fração alinhável do comprimento de ambas as proteínas, fosse de, no mínimo, 90%. As sequências não-redundantes foram usadas como sementes para buscas de similaridade baseadas em 3 iterações de buscas, realizadas com o programa jackhmmer do pacote HMMER3.1b (Finn, Clements, and Eddy 2011), contra

o banco de dados de proteínas dos genomas coletados. Os resultados dessas buscas foram filtradas para valores de $e\text{-value} \leq 10^{-3}$ e comprimento da região alinhada entre 100 e 800 aminoácidos. Para recuperar possíveis falsos negativos descartados por esses critérios, fragmentos curtos foram avaliados quanto à presença de alinhamentos altamente significativos com modelos de SBPs do Pfam ($e\text{-value} \leq 10^{-10}$).

Adicionalmente, anotamos as estruturas de SBPs com os números que identificam famílias de transportadores no banco de dados TCDB (Andreeva et al. 2014; Van Tran 2004; Saier et al. 2016) escolhendo o melhor resultado em buscas de similaridade usando BLAST e os mesmos critérios de cobertura na semente e $e\text{-value}$ citados acima.

3.4. Identificação de famílias de SBPs

As regiões de envelope, identificadas pelo jackhmmer nos resultados das buscas por SBPs, foram isoladas e submetidas à análise de agrupamento hierárquico. A redundância desse conjunto foi reduzida ao selecionarmos, para as etapas posteriores, apenas sequências com cobertura menor ou igual a 80% e no máximo 90% de identidade. A redução de redundância foi feita com o programa MMseqs2 (Steinegger and Söding 2017) com o algoritmo de agrupamento *greedy-cluster* e selecionando, em cada grupo de sequências redundantes, a sequência de maior comprimento. O conjunto de sequências não-redundantes foi submetido a novas etapas de agrupamento por similaridade de forma iterativa, usando o resultado de cada iteração para uma nova rodada de agrupamento, o que gerou uma hierarquia de grupos. Para gerar essa hierarquia, a cada iteração, os alinhamentos par-a-par entre as sequências sob análise eram filtrados com valores progressivamente menores para o $e\text{-value}$ máximo, explorando todos os valores entre 10^{-3} a 10^{-200} . Em cada iteração do agrupamento hierárquico, o algoritmo *setcover* do programa MMSeqs2 foi usado para definir os grupos de sequências e para selecionar a sequência representante de cada grupo. Em todas as iterações foi também exigido um mínimo de 80% de cobertura para os alinhamentos de pares de sequências (parâmetro “-c 0.8” do MMseqs2).

As sequências dos grupos formados na última etapa do agrupamento hierárquico, ou seja, grupos definidos pelo $e\text{-value}$ máximo de 10^{-3} , foram alinhadas

pelo software Mafft (Kato et al. 2005), o algoritmo Needleman-Wunsch (opção *--globalpair*) . Nessa etapa, consideramos apenas grupos com pelo menos três sequências e selecionamos para os alinhamento múltiplos, em cada grupo, sequências com no máximo 70% de identidade entre si. Esses alinhamentos foram convertidos para o formato a3m e um banco de modelos de Markov ocultos (HMM) foi derivado desses alinhamentos usando o programa hhmake do pacote HH-suite 3.0 (Kato et al. 2005; Steinegger et al. 2019). Todos os HMMs cnstruídos por nós foram alinhados uns contra os outros, usando o software HHsearch, e o resultado, após processamento, foi submetido ao algoritmo de identificação de comunidades de Leiden (Traag, Waltman, and van Eck 2019). O algoritmo de Leiden foi executado com 100 iterações e a probabilidade de homologia, calculada pelo HHsearch, foi usada como peso para as arestas da rede.

Para obtenção das relações de similaridade entre as estruturas das proteínas usadas como semente nas buscas de SBPs, estimamos as distâncias euclidianas entre estruturas (RMSD ou *root mean square deviation*), usando o software TM-align (Y. Zhang and Skolnick 2005). O TM-align avalia os valores de RMSD usando apenas as coordenadas dos átomos de carbono- α nas estruturas tridimensionais das proteínas. Essas estimativas de distância entre estruturas foram submetidas a agrupamento hierárquico pelo algoritmo de Ward, implementado na função hclust do software R (Murtagh and Legendre 2014), resultando em um dendrograma para as relações de similaridade entre as estruturas. Paralelamente, o algoritmo k-means foi usado para obter grupos baseados nas mesmas medidas de distância.

3.5. Busca e classificação de domínios sensores

A coleta de homólogos de Cache foi feita usando o software hmmsearch, do pacote HMMER3, e os modelos do clã Cache do banco de dados Pfam. Buscas, usando esses modelos, no nosso banco de dados de genomas foram filtradas com *e-value* máximo de 10^{-3} . Executamos também buscas de 5 iterações com o software Jackhmmer, usando as regiões correspondentes as domínios Cache das seguintes proteínas: 46 a 175 de XAC2382 (WP_003486063.1) de *Xanthomonas citri* pv. *citri* (Teixeira et al. 2018), 26 a 223 de MbaA (NP_230352.1) de *Vibrio cholerae* (Teixeira et

al. 2018; Cockerell et al. 2014) e 41 a 355 de ChiS (VC0622, NP_230271.2) de *Vibrio cholerae* (Li and Roseman 2004). Os resultados das buscas iterativas foram aceitos se tinham *e-value* $\leq 10^{-3}$ e cobertura da *query* acima de 80%. Regiões de proteínas correspondentes ao envelope identificado nas buscas com *hmmsearch* e *jackhmmer* foram isoladas para análises posteriores.

Uma segunda etapa de coleta e identificação de domínios sensores extracitoplasmático se baseou na identificação de regiões extracelulares. A predição de regiões extracelulares foi baseada na identificação de hélices transmembrana (TM) e peptídeos sinal (SP) pelo software Phobius (Kall, Krogh, and Sonnhammer 2007). Apenas proteínas com no máximo duas hélices foram preservadas e as regiões extracelulares preditas foram obtidas isolando do primeiro aminoácido até a primeira hélice transmembrana, nos casos de proteínas com uma TM, ou a região entre as duas hélices transmembrana, no caso de proteínas com duas hélices.

O conjunto de regiões extracelulares preditas foi reduzido, eliminando-se as regiões absolutamente idênticas (100% de cobertura e 100% de identidade). O conjunto de fragmentos não redundantes foi classificado em dois níveis: o primeiro constituído de grupos com identidade de até 70% e cobertura de 80%, e o segundo em grupos com 80% de cobertura e *e-value* máximo de 10^{-3} foram identificados. Grupos com pelo menos de 3 sequências foram selecionados e alinhados com o software Mafft usando o algoritmo de Smith-Waterman (“--localpair”) e 1000 iterações. Os alinhamentos foram usados para construir HMMs no formato compatível com HHsuite (Traag, Waltman, and van Eck 2019; Steinegger et al. 2019) e os modelos foram comparados, todos contra todos, usando buscas com o software HHblits. Os HMMs dos grupos de regiões extracelulares foram comparados também com os modelos de famílias do Pfam, usando HHblits, e os resultados foram aceitos quando a probabilidade de homologia foi maior ou igual a 60%. Finalmente, os grupos de regiões extracitoplasmáticas foram reunidos em comunidades pelo algoritmo de Leiden e as comunidades foram curadas manualmente.

3.6. Busca de vizinhanças gênicas

Os dados de vizinhança gênica para as famílias de proteínas sob análise foram

extraídos da anotação dos genomas, depositada no NCBI (Seção 2.1). Para cada gene identificado como homólogo de SBPs, três genes codificantes de proteínas foram identificados à jusante e à montante. Em algumas análises, a vizinhança gênica foi filtrada de modo a manter apenas genes que estejam na mesma fita e fossem separados por uma distância máxima de até 150 nucleotídeos. Adicionalmente, foram identificadas também regiões genômicas na vizinhança de sistemas de transporte ABC com pelo menos uma porina e uma ATPase..

3.7. Coocorrência de SBPs e domínios sensores

Para avaliar se domínios SBPs e domínios sensores de proteínas de transdução de sinal ocorriam na mesma vizinhança um número maior de vezes do que seria esperado por chance, coletamos todos os pares de genes nos genomas analisados usando uma janela deslizante de dois genes sobre a lista de genes de cada genoma. Descartamos os pares de genes vizinhos quando um dos genes era não-codificante ou que estivessem separados por uma região intergênica com mais de 250 nucleotídeos. Uma tabela de contingência foi gerada para a presença ou ausência de SBPs e domínios sensores nos pares de genes e a desvio do valor esperado de pares SBPs e proteínas sensoras foi avaliada com um teste de chi-quadrado de Pearson usando o pacote *infer* do software R.

4. Resultados

4.1. Seleção de genomas completos

Para a construção de um banco de dados de genomas completos, fizemos o *download* de todas as sequências de DNA, de genomas completos de procariotos, disponíveis no NCBI até o mês de maio de 2018. A lista de genomas completos foi gerada filtrando as linhas contendo o termo “Complete Genome” na coluna “assembly_level” das tabelas “assembly_summary_*”, disponibilizadas no diretório ASSEMBLY_REPORTS da seção de genomas do site de FTP do NCBI. Este filtro foi construído com base nas instruções dos qualificadores descritos no arquivo ftp://ftp.ncbi.nlm.nih.gov/genomes/ASSEMBLY_REPORTS/README_assembly_summary.txt. Aplicando esse critério, selecionamos 10.467 genomas de procariotos, sendo 10.175 bactérias e 292 archaeas, totalizando 17.702.628 sequências de proteínas.

Um dos grandes problemas encontrados durante a construção deste banco de dados foi o número elevado de genomas redundantes, que poderia gerar vieses nas análises subsequentes. Para reduzir a redundância em nossos dados, utilizamos o programa Mash (Ondov et al. 2016), que usa um procedimento de redução de dimensionalidade baseado no MinHash, uma técnica de *locality-sensitive hashing*. O software Mash percorre a sequência com uma janela deslizante de tamanho k gerando uma coleção dos câmeros (k -mers) para cada genoma de interesse. Em seguida, a função de *hashing* retorna os s menores *hashes*, conjunto que é denominado *sketch* de tamanho s . A comparação de dois *sketchs* permite o cálculo do índice de Jaccard (J), uma estatística usada para comparar a similaridade entre dois conjuntos, definida como a razão entre o tamanho da intersecção e o tamanho da união de dois conjuntos. O valor $1-J$ é um estimador para a distância entre genomas que pode substituir o ANI (*average nucleotide identity*), sem exigir o cálculo explícito e custoso dessa última estatística (Ondov et al. 2016).

Buscando atingir uma alta sensibilidade, cada genoma foi convertido em um *sketch* de 5000 câmeros de comprimento de 27, resultando em estimativas da distância mash e do *p-value* para as relações de similaridade entre pares de genomas. Os

resultados foram filtrados para distância máxima de 0.05 e o p -value para 10^{-10} . O algoritmo *Union Find*, implementado em um *script* na linguagem Perl 5, foi usado para identificar os elementos conectados na rede de similaridade de genomas e um genoma representante de cada componente conectado foi escolhido, selecionando preferencialmente aqueles anotados com as propriedades “*representative genome*” e “*Complete Genome*” para as propriedades *refseq_category* e *assembly_level*, especificadas na anotação dos genomas disponível no NCBI.

Figura 3. Distribuição dos valores de distância e p -values, estimados pelo programa Mash, nas comparações par-a-par entre genomas. **Parte superior:** distribuição completa dos valores de distância e destaque do intervalo 0-0.4 (inserto menor). **Parte inferior:** distribuição completa dos valores de p -value e destaque do intervalo 10^{-100} - 10^{-10} (inserto menor).

Após a filtragem dos genomas redundantes restaram 3311 genomas, sendo 3087 de bactérias e 224 de Archaea. Esses genomas continham um total de 11.286.807 proteínas anotadas. Além do critério automático de seleção de genomas, descrito acima, o genoma de *Xanthomonas citri* pv. *citri* str. 306 foi escolhido para substituir o genoma de *Xanthomonas citri* pv. *vignicola*.

4.2. Análise de domínios sensores

4.2.1. Busca por sensores extracitoplasmáticos

A busca de proteínas com regiões homólogas a domínios Cache foi realizada em duas etapas. A primeira etapa foi baseada no software *hmmsearch* (Finn, Clements, and Eddy 2011) e nos modelos do Pfam para a superfamília Cache que, nesse banco de dados, correspondem ao clã Cache (CL0165) (Mistry et al. 2021). A segunda rodada de buscas foi implementada com o software *jackhmmer* (Finn, Clements, and Eddy 2011), usando as regiões ChiS₄₁₋₃₅₅ e MbaA₂₆₋₂₂₃, que correspondem, respectivamente, ao SBP NspS e à região sensora da fosfodiesterase MbaA, ambas de *Vibrio cholera*. Também foi usada como semente nas buscas de *jackhmmer* a região XAC2382₄₆₋₁₇₅, que corresponde ao domínio sCache dessa proteína de *Xanthomona citri* pv. *citri*. Essas buscas identificaram, em nosso banco de dados de genomas selecionados, 40.627 proteínas.

Com o intuito de identificar novos sensores, um conjunto adicional de proteínas foi obtido a partir da extração de todas as regiões presentes entre duas hélices transmembrana (TM) ou uma hélice transmembrana e um peptídeo sinalizador de secreção pelo sistema Sec ou Tat (SP), identificados, pelo programa Phobius, nas proteínas dos genomas sob análise. Esse critério identificou um total de 175.588 fragmentos de proteínas. Esse conjunto de proteínas foi filtrado, removendo-se todas as proteínas sem domínios efetores de transdução de sinal e todos os fragmentos menores que 50 aminoácidos. Foram filtrados também os fragmentos que alinhavam com domínios efetores, resultando em 119.774 sequências, das quais um subconjunto de 40.920 fragmentos foi selecionado para representar os grupos formados por proteínas com pelo menos 70% de identidade. Os 40.920 representantes formaram 3.306 famílias de proteínas definidas por valores de *e-value* menores ou iguais a 10^{-3} em alinhamentos par-a-par calculados pelo software MMseqs2. Alinhamentos múltiplos foram produzidos para cada família com pelo menos três sequências, usando o software Mafft, e HMMs foram construídos para cada alinhamento usando o software hmmer do pacote HH-suite. Aplicando o algoritmo de Leiden às comparações par-a-

par destas HMMs, geradas pelo software hhsearch, todas as 3.306 famílias foram classificadas em 28 comunidades, listadas na Tabela 3.

Figura 4. Representação da intersecção de conjuntos de sequências identificados nas diferentes estratégias de busca por homólogos de sensores extracitoplasmáticos. O conjunto st2tm é o conjunto de regiões identificados em proteínas com duas hélices transmembranares (TM) ou com um peptídeo sinalizador para secreção (SP) e uma assinatura TM. Pfam são todas as proteínas que alinharam com $e\text{-value} \leq 10^{-2}$ aos modelos do clã Cache do Pfam. XAC2382, ChiS e MbaA representam todos os resultados de buscas iterativas (software jackhmmer), iniciadas com estas proteínas e identificados em 5 iterações com $e\text{-value}$ menor que 10^{-2} .

Tabela 3. Comunidades de domínios periplasmáticos identificados em proteínas de transdução de sinal. A anotação na coluna “Domínio periplasmático” corresponde ao clã mais frequente entre os modelos mais similares no Pfam. A detecção dos modelos similares no Pfam foi baseada na análise com hmmscan e $e\text{-value} \leq 10^{-3}$.

Comunidade	Proteínas	Grupos		Domínio periplasmático	Comentários
		70% id.	$e\text{-value} \leq 10^{-3}$		
0	36775	13025	813	Cache	Cache e fusões Cache + SBP
1	22685	7618	699	4HB_MCP	<i>four helical bundle</i> (4HB)
2	22746	6974	591	Cache	
3	17301	5304	503	Cache	
4	4945	1165	142		

5	1881	1221	140	TPR	
6	4613	1676	125	TPR	
7	4013	1448	57	SBP	SBPs das classes I e II
8	619	386	36		
9	2733	1594	24		
10	727	214	23	E-set / Beta_propeller	Domínios Ig-like
11	254	86	16		
12	93	50	8		
13	110	40	6	Cache-like	
14	78	26	6	Cache-like	
15	16	14	3	OB	
16	59	14	3		
17	11	11	1		
18	6	6	1		
19	5	4	1	Beta_propeller	
20	6	3	1		
21	3	3	1		
22	8	5	1		
23	4	4	1		
24	5	3	1		
25	27	6	1		
26	48	17	1		
27	3	3	1		

Combinando a inspeção manual das comunidades de domínios extracitoplasmáticos com alinhamentos contra modelos de domínios conhecidos, do banco de dados Pfam, foi possível classificar as regiões sensoras putativas identificadas com base na arquitetura derivada da análise com o programa Phobius. Dos fragmentos coletados, 8.839 não apresentaram similaridade a nenhuma região sensora descrita por um modelo no banco de dados Pfam. Outras regiões extracitoplasmáticas são caracterizadas pela presença de regiões repetidas, que são assinaturas comuns em regiões envolvidas em interações com macromoléculas (TPR) ou formam estruturas secundárias comuns em proteínas da membrana externa de bactérias Gram-negativas (Betapropeller). A maior parte das regiões extracelulares,

porém, apresentam similaridade detectável com domínios Cache, incluindo algumas fusões de Cache e SBP. Muito comuns também são regiões extracelulares homólogas de SBPs e os feixes de quatro α -hélices (*four-helical bundle*, 4HB_MCP), característicos das proteínas de vias quimiotáticas. Domínios semelhantes a imunoglobulinas (*Ig-like*), da superfamília E-set, também foram encontrados em múltiplas proteínas, muitas vezes fusionados a domínios Betapropeller, assim como alguns exemplos de domínios do *fold* OB, um amplo grupo de proteínas que inclui famílias citoplasmáticas capazes de se ligar a oligonucleotídeos, como RNA e DNA, a oligossacarídeos e, potencialmente, a nucleotídeos.

4.2.2. Domínios sensores de ChiS e MbaA: nova família de domínios Cache

As proteínas de transdução de sinal de *Vibrio cholerae*, ChiS e MbaA, são, respectivamente, uma histidina quinase e uma diguanilato fosfodiesterase da família EAL com um domínio GGDEF inativo. As duas proteínas atuam inseridas na membrana citoplasmática, com a região efetora no citoplasma e o domínio sensor no periplasma. Em ambos os casos, a estrutura da região sensora não foi caracterizada experimentalmente e não é reconhecida pelos modelos disponíveis no banco de dados Pfam. Nossa análise, porém, permitiu reconhecer a similaridade dessas regiões com o modelo Cache3-Cache2 do Pfam, assinalando com confiança os domínios sensores das duas proteínas ao universo de dCache presentes na comunidade 0 (zero).

4.2.3. Domínios sensores não catalogados no Pfam

Durante o processo de anotação manual das famílias de domínios citoplasmáticos, frequentemente nos baseamos na similaridade entre os modelos derivados de alinhamentos dessas famílias com modelos do banco de dados Pfam. Um subconjunto dos modelos do Pfam, porém, não eram reconhecidos como membros de nenhum *fold* ou clã preexistente, o que dificultava a anotação dessas famílias. Para entender melhor a relação dessas famílias com outras mais bem caracterizadas na literatura e com estruturas tridimensionais já caracterizadas experimentalmente,

decidimos comparar os modelos de Pfam, que apareceram em nossa análise, com os demais modelos do mesmo banco de dados. Essa análise, baseada nos alinhamentos de HMMs construídos pelo hhsearch, revelou que algumas das famílias “órfãs” (sem clã) do Pfam, presentes em regiões extracitoplasmáticas de várias proteínas, correspondem a versões divergentes de famílias de domínios sensores conhecidos. Para esse resultado, tomamos o cuidado de filtrar os alinhamentos entre modelos com alta cobertura para os dois modelos alinhados, de modo a evitar artefatos de alta probabilidade e baixa cobertura, que observamos com frequências em nossas análises usando hhsearch.

Os resultados das comparações de modelos são resumidos na Tabela 4. Um exemplo interessante é o modelo DraK_HK_N, que apresentou alta probabilidade de homologia com o modelo GAPES4, mas nenhuma probabilidade acima de 60% com modelos do clã Cache. O modelo GAPES4, por sua vez, apresenta alta probabilidade contra sCache4, um exemplo sólido de um membro divergente, previamente não reconhecido, da superfamília Cache. A relação de DraK_HK_N e GAPES4 corrobora a fraca evidência indicada nos alinhamentos do primeiro contra modelos do clã Cache e permite estabelecer, por transitividade, que o modelo DraK_HK_N é um sCache.

Tabela 4. Relação de probabilidade entre famílias do Pfam sem anotação de clã e outras famílias, do mesmo banco de dados, com clãs assinalados. Valores de probabilidade calculados pelo software HHsearch.

Pfam	Membro de clã mais similar	Clã da família mais similar	Probabilidade
CHASE5	sCache_4	Cache	97.94
CSS-motif	Ykul_C	Cache	74.73
DraK_HK_N	GAPES4	Cache	86.48
GAPES3	sCache_4	Cache	98.51
GAPES4	sCache_4	Cache	76.14
HK_sensor	AbfS_sensor	Cache	96.06
HisK_sensor	sCache_like	Cache	98.94
LapD_MoxY_N	SMP_2	Cache	96.04
ECSIT	MRP-S27	TRP	99.84
DUF5112	TPR_MalT	TRP	99.82
DUF5588	PknG_TPR	TRP	99.80

Figura 5. Distribuição do número de regiões extracitoplasmáticas por clã do banco de dados Pfam. O clã foi assinalado usando o programa hmmsearch contra as sequências de proteína individuais e a anotação dos modelos.

4.3. Análise de proteínas ligadoras de substrato

4.3.1. Buscas por homólogos de SBPs

A coleta de homólogos de SBPs foi realizada por meio de buscas iterativas de 501 proteínas dos três *folds* de SBPs, as quais tinham estruturas tridimensionais disponíveis no banco de dados PDB. Essas buscas resultaram na coleta de 275.956 fragmentos de 37 a 937 aminoácidos, que foram posteriormente anotados pelo software hmmscan, usando todos os modelos do Pfam para SBPs da classe I (clã Periplas_BP), todos os modelos da classe II (clã PBP) e os modelos de quelatases (Periplas_BP_2 e ZnuA) para SBPs da classe III. Um agrupamento baseado em similaridade foi aplicado aos fragmentos coletados e, por exigir uma cobertura elevada das sequências de cada fragmento coletado, evitou a inclusão de potenciais falsos positivos, que poderiam ser coletados devido a fusões de SBPs com outros domínios. No total, os agrupamentos por similaridade dos fragmentos, construídos com 80% de cobertura e $e\text{-value} \leq 10^{-3}$, somaram 1342 grupos com 3 ou mais proteínas e 1255 fragmentos que não entraram em nenhum desses grupos.

4.3.2. Classificação de SBPs

Os SBPs descritos na literatura pertencem a 3 *folds* distintos, comumente referidos como SBPs da classe I, classe II e classe III (quelatases). Todos os três grupos compartilham a arquitetura de dois subdomínios, ligados por um conector flexível, e que interagem com ligantes por meio do movimento de estreitamento da cavidade entre os dois subdomínios. Cada um desses grupos, porém, se diferencia por um ordenamento de folhas- β em cada subdomínio. Os SBPs de classe I apresentam cinco folhas- β em cada subdomínio, os da classe III possuem apenas quatro e os SBP de classe II possuem cinco folhas mas diferem da classe I no arranjo dessas folhas, uma vez que nesse grupo a quarta folha se posiciona antiparalela à demais.

A relação evolutiva entre esses três *folds* não é clara: alguns autores tratam todos os três como membros de um mesmo *fold*, mas tendem a excluir, em suas comparações, domínios semelhantes que não se associam a transportadores ou a sistemas de transdução de sinal (Scheepers, Lycklama a Nijeholt, and Poolman 2016). Esse critério, basicamente funcional, para seleção dos alvos de análise, pode levar à exclusão de grupos intermediários e prejudicar a inferência das relações entre famílias e *folds*. Notavelmente, o modelo para a origem das classes SBP-I e SBP-II, proposto por Fukami-Kobayashi e colaboradores (Fukami-Kobayashi, Tateno, and Nishikawa 1999) não é suportado pela classificação do banco de dados ECOD (Cheng et al. 2015, 2014), que coloca os SBPs da classe II em um grupo próprio, um *fold* distinto e isolado do grupo das classes I e III. As duas últimas são classificadas como membros do mesmo *H-group* (Class I glutamine amidotransferase-like), um nível da classificação do ECOD que inclui sequências cuja homologia é derivada do revisão manual de dados estruturais, similaridade de sequência e da literatura. Curiosamente, os reguladores de resposta dos sistemas de dois componentes e proteínas similares também são membros do mesmo *H-group*, o que faz os resultados do ECOD compatíveis com a hipótese da origem da classe I de SBPs a partir de reguladores de resposta, como proposto por Fukami-Kobayashi. Nossos dados de agrupamento por similaridade das três classes de SBPs (Figura 5) são compatíveis com a classificação do ECOD, pois sugerem maior similaridade entre as classes I e III de SBPs, em detrimento da classe II. Nosso dados, porém, não nos permitem estabelecer se a classe II, cujos membros

ficam isolados no único *H-group* em seu *X-group* (nível mais alto do ECOD, onde a homologia é incerta) de fato não têm relação com o *X-group* das classes I e III (“Flavoredoxin-like”). A análise detalhada da estrutura desses e de outros *X-groups*, incluindo todos os membros dos *X-groups* “Flavoredoxin-like” e “Periplasmic binding protein-like II”, além de novas estruturas de possíveis intermediários, serão necessários para resolver a questão da origem das diferentes classes de SBPs.

Colocando de lado as incertezas associadas à relação entre as classes de SBPs, focamos na classificação detalhada das famílias de SBPs em cada classe. Nossas análises de agrupamentos de similaridade e outras análises comparativas resultaram em uma classificação hierárquica ampla e detalhada desses domínios.

Figura 6. Resultados da aplicação de diferentes algoritmos de agrupamento às estruturas tridimensionais de SBPs. As folhas, no dendrograma, e os pontos, nos painéis, correspondem a representantes das 501 estruturas usadas como semente em nossas buscas por SBPs. **(a)** O

mapa de calor (*heatmap*) e o dendrograma são baseados na distância estrutural (RMSD). A tabela de cores, entre o dendrograma e o mapa de calor, mostra, da esquerda para a direita, os grupos definidos por *k-means*, com 2 (*k*2), 3 (*k*3) e 7 (*k*7) grupos, e as comunidades formadas pelo algoritmo de Leiden, usando o resultado das comparações de HMMs (full). Os painéis na parte inferior (**b** a **e**) mostram a distribuição das estruturas representativas de acordo com os dois componentes dominantes na análise de componentes principais dos vetores de RMSD. Os painéis são: (**b**) classificação de Scheepers e colaboradores (Scheepers, Lycklama A Nijeholt, and Poolman 2016), (**c**) *k-means* com 3 grupos, (**d**) *k-means* com 7 grupos e (**e**) comunidades de Leiden.

Para definir o nível mais alto em nossa classificação hierárquica, a similaridade entre os grupos definidos pelo agrupamento de 80% de cobertura e *e-value* $\leq 10^{-3}$, que de agora em diante chamaremos de famílias, foi estimada com base em alinhamentos de HMMs. O procedimento de construção dos HMMs foi o mesmo adotado para os domínios periplasmáticos, ou seja, a combinação dos softwares Mafft e hhmake, sem intervenção manual. A similaridade entre os pares de HMMs, calculada pelo HHsearch, é a média aritmética da pontuação de colunas nos alinhamentos de pares de HMMs e reflete, satisfatoriamente, o grau de divergência entre as famílias. Usando essa estimativa de similaridade como peso para as arestas, construímos um grafo ponderado ou rede de similaridade das famílias (Figura 6). Aplicando o algoritmo de força para definir a posição dos nós, em duas dimensões, verificamos a presença de dois grandes grupos principais, que podem ser ainda subdivididos em 6 grupos, correspondentes às comunidades criadas pela aplicação do algoritmo de Leiden. Os SBPs das classes I e III formaram duas comunidades independentes, cada uma das quais inclui todos os membros de cada uma dessas classes. Os SBPs de classe II, porém, se dividiram em 4 grupos que, ao contrário das classes I e III, divergiram da classificação de Scheepers e colaboradores, baseada na similaridade de estruturas (Scheepers, Lycklama A Nijeholt, and Poolman 2016). Nossa classificação, por ser inteiramente baseada na similaridade de HMMs, é derivada apenas de informações sobre a similaridade dos resíduos e da estrutura secundária predita para cada família (Steinegger et al. 2019). Em nossa análise (Tabela 5), todos os membros dos grupos E e F, definidos por Scheepers e colaboradores, fossem inseridos na mesma comunidade

(comunidade 1). Ao mesmo tempo, o cluster D se fragmentou, a maior parte de seus membros compondo a comunidade 2, enquanto outros se misturaram com o grupo G, que não foi descrito por Scheepers, na comunidade 4.

Tabela 5. Relação das comunidades de SBPs com os grupos descritos por Scheepers e colaboradores (Scheepers, Lycklama A Nijeholt, and Poolman 2016).

Comunidade	Classe	Grupo	Proteínas
1	II	E, F	75572
2	II	D	80224
3	I	B	49361
4	II	D, G	35474
5	II	C	18900
6	III	A	16015

Figura 7. Rede de similaridade das famílias de SBPs. A rede foi construída a partir dos alinhamentos dos HMMs que representam essas famílias. O layout de força foi aplicado para

definir as posições dos nós, usando os valores de similaridade, estimados pelo HHsearch, como pesos das arestas. À esquerda, os grupos são coloridos de acordo com a classificação de Scheepers *et al.*. À direita, as cores refletem a associação de SBPs com outros domínios: a presença de pelo menos um gene da superfamília Cache, até três genes antes ou depois do *loci* SBP, é indicada com a cor verde. Famílias que incluem fusões de SBPs com domínios efetores (histidina quinases, etc.) são marcadas em vermelho e a cor laranja indica a presença de fusões de SBPs e Caches.

4.4. Estatísticas das associações entre SBPs e Cache

A conservação de um padrão de proximidade entre genes de famílias distintas, quando significativa, é um forte indício da presença de vínculos funcionais, como co-transcrição, interação física dos produtos gênicos ou participação nas mesmas vias e processos celulares (Huynen *et al.* 2000; Wolf *et al.* 2001; Koonin, Wolf, and Aravind 2001). Nessa seção, avaliamos se os SBPs, em geral, são significativamente mais abundantes na vizinhança de genes de transdução de sinal do que seria esperado por chance.

Para avaliar se a distribuição de SBPs e domínios sensores em genoma completos são correlacionadas precisamos nos assegurar de que nossa coleta de ambos os grupos de genes foi exaustiva, pois todas as ocorrências de ambos os grupos devem ser contabilizadas antes que possamos estimar a probabilidade de um grupo aparecer na vizinhança do outro por acaso. Para os SBPs, nossa coleta foi absolutamente exaustiva e complementada pela classificação detalhada do grupo. No caso dos sensores, porém, uma análise com esse nível de estringência só estava disponível para os membros da superfamília Cache (Upadhyay *et al.* 2016). Por esse motivo, decidimos focar a análise da associação dos SBPs com domínios sensores apenas nos membros da superfamília Cache.

Na Figura 6, o painel direito mostra que homólogos de Cache estão presentes na vizinhança de muitas famílias distintas de SBPs. No entanto, para essa figura, foram consideradas vizinhas de Cache toda a família de SBP com pelo menos um membro a menos de três genes de distância de um homólogo de Cache. Esse critério permissivo tem chances de produzir associações espúrias, por não levar em conta a abundância

de Cache no genoma.

Decidimos então tabular o número de SBPs codificados por genes localizados imediatamente ao lado de genes de homólogos de Cache. Para isso consideramos uma janela deslizante de dois genes, que foi aplicada a cada cromossomo ou plasmídeo em nosso conjunto de genomas. Pseudogenes e genes não-codificantes foram ignorados, assim como janelas onde o par de genes esteja a mais de 250 nucleotídeos de distância. Esse procedimento de contagem classifica os genes em cada janela como pertencendo ou não aos grupos SBP ou Cache, resultando em tabelas de contingência às quais aplicamos testes de χ^2 para avaliar se o número de janelas nas quais os dois grupos ocorrem, simultaneamente, diverge do esperado por chance. A Tabela 6 mostra o resultado desse procedimento, aplicado a todos os pares de genes dos genomas sob análise, incluindo os valores dos resíduos de Pearson.

Tabela 6. Tabela de contingência e valores do teste χ^2 para a presença de homólogos de SBPs e Cache em uma janela deslizante de 2 genes.

		Cache	
		Presente	Ausente
SBP	Presente	6156	703887
	Ausente	299184	35624735

Os dados na Tabela 6 mostram um excesso de pares SBP-Cache em janelas onde os dois grupos de genes estão presentes. Embora relativamente pequeno, esse desvio é significativo e permite descartar a hipótese de que os dois genes aparecem lado a lado por acaso. No entanto, na análise acima, todas as famílias de SBPs encontradas nos genomas de procariotos foram consideradas, incluindo famílias com localização no citoplasma e outras cuja localização é incerta. Como nossa análise está voltada para a associação de SBPs com domínios sensores extracelulares, decidimos avaliar o efeito da abundância de SBPs com assinaturas de localização extracelular na associação com homólogos de Cache.

Para estimar o número de proteínas com localização extracelular nas famílias de SBPs, usamos o software Phobius (Käll, Krogh, and Sonnhammer 2004). A Figura 7

mostra a relação da proporção de SBPs com assinaturas de secreção (eixo X) e a frequência de genes vizinhos da superfamília Cache (eixo Y). Várias das famílias de SBPs, nas quais assinaturas de secreção foram identificadas em mais de 80% dos membros, apresentaram menos de 10% de vizinhos Cache. É possível que muitas dessas famílias incluam apenas SBPs associados a processos de transporte. Por outro lado, o conjunto de SBPs com mais de 10% de vizinhos Cache mostrou uma tendência a se concentrar no intervalo de famílias com alto conteúdo de assinaturas de secreção. Interessantemente, famílias com frequências elevadas de vizinhos Cache e de assinaturas de secreção estavam amplamente distribuídas entre as comunidades de SBPs, recapitulando assim o padrão observado na Figura 6, mesmo após a aplicação de todos os esses filtros.

Figura 8. Relação entre a presença de assinaturas de secreção em SBPs e a frequência de genes vizinhos da superfamília Cache. Esses dados incluem os casos de fusão de domínios

SBP e Cache na mesma proteína. O eixo horizontal mostra a proporção de proteínas com peptídeos sinal ou hélices transmembrana, detectados pelo software Phobius, em cada família de SBPs. A presença dessas assinaturas serve como indicador de localização extracitoplasmática ou ancoragem na membrana (Gram-positivas). Os painéis correspondem às comunidades, definidas pelo algoritmo de Leiden. O tamanho de cada ponto é proporcional ao logaritmo, na base 10, do número de membros daquela família. As cores dos pontos refletem a classificação de Scheepers (Scheepers, Lycklama A Nijeholt, and Poolman 2016).

Figura 9. Variação do *p-value* em função da abundância de SBPs com localização extracitoplasmática.

Para entender melhor a associação entre a frequência de assinaturas de secreção e a frequência de vizinhos Cache, aplicamos nosso teste a subconjuntos de famílias selecionadas com base no número mínimo de assinaturas de secreção. A Figura 8 mostra os *p-values* resultantes da aplicação do teste de associação entre SBPs e Cache a famílias de SBPs com uma frequência mínima de assinaturas de secreção variando entre 0 e 100%. Curiosamente, resultados significativos são obtidos mesmo quando incluímos famílias com número muito reduzido de assinaturas de secreção. Isso pode estar acontecendo devido à presença de muitos falsos negativos

para essas assinaturas ou por causa da associação de SBPs citoplasmáticos com sensores contendo domínios Cache. Esse pode ser o caso, por exemplo, da família de domínios sensores LysR_substrate, que são sensores citoplasmáticos conhecidos por atuarem nos passos finais de vias de transdução de sinal que regulam níveis de expressão gênica. A curva do p -value na Figura 8 atinge um platô quando todas as famílias sob análise apresentam pelo menos 30% de membros com assinaturas de secreção. A significância da associação, porém, é atingida muito antes disso e o número de pares de SBP-Cache é o principal responsável por esse efeito como demonstrado, por exemplo, pelos resíduos normalizados do teste χ^2 para famílias de 20% ou mais de assinaturas de secreção (Tabela 7).

Tabela 7. Tabela de contingência e valores do teste χ^2 para a presença de homólogos de SBPs e Cache em uma janela deslizante de 2. Nessa análise foram considerados apenas as famílias de SBPs com assinaturas de secreção em pelo menos 20% dos membros.

		Cache	
		Presente	Ausente
SBP	Presente	2424	228856
	Ausente	32916	33099766

Ao todo, os resultados apresentados nesta seção mostram que a presença de genes de SBPs ao lado de homólogos de Cache ocorre mais frequentemente do que o esperado por chance. Essa associação é ainda mais clara quando filtramos as famílias de SBP com base na localização celular, o que suporta a noção de que SBPs de diferentes grupos, mesmo as que pertencem a famílias, comunidades ou *folds* diferentes, acabam sendo incorporadas em sistemas de transdução de sinal. Embora a classificação de SBPs construída nesse trabalho demonstre que o recrutamento dos SBPs para vias de transdução de sinal ocorreu múltiplas vezes, detalhes desse processo, tais como a função das famílias envolvidas e a natureza das interações físicas entre Caches e SBPs, requerem a análise detalhada da evolução de famílias

individuais. Na próxima seção, apresentamos um panorama das comunidades de SBPs e descrevemos alguns exemplos de análise detalhada de famílias individuais.

Figura 10. Composição de domínios proteicos de homólogos de Cache e SBP codificados por genes vizinhos. Os domínios na cor salmão, na coluna da esquerda, são da família Phosphonate-bd (comunidade 1), um SBPs da classe II. Nas duas últimas linhas, os domínios ABC_sub_bind e Peripla_BP_6, em vermelho, são exemplos de SBPs da classe I (comunidade 3). As proteínas de transdução de sinal codificadas pelos genes vizinhos estão representadas à direita e, em todos esses casos, são codificadas pelo gene transcrito após o SBP e na mesma fita do DNA. Note que diferentes famílias de domínios Cache aparecem associadas com o mesmo grupo de domínios SBP e que o mesmo ocorre com os SBPs.

4.5. Panorama das associações Cache-SBP

Nesta seção, vamos listar características gerais de cada uma das 6 comunidades de SBPs e apresentar análises mais detalhadas de algumas famílias nessas comunidades que estão associadas a genes de transdução de sinal. Na apresentação que segue, por vezes usaremos, para as famílias, uma nomenclatura composta de números separados por pontos. O primeiro número identifica a comunidade à qual pertence a família de SBPs. O segundo valor identifica as famílias de proteínas, ou seja, os grupos formados por proteínas que alinham 80% de seu comprimento, em comparações par-a-par, com *e-value* de, no máximo, 10^{-3} . Famílias que incluem proteínas com literatura relevante para esse trabalho também serão designadas pelo nome do gene da proteína.

4.5.1. SBPs de classe II

Os SBPs de classe II constituem a maior população de SBPs identificados em

nossas análises. Essas proteínas são caracterizadas pela presença de uma fita- β antiparalela no núcleo conservado de sua estrutura, formando um arranjo de estruturas secundárias com a topologia $\beta_2 \uparrow \beta_1 \uparrow \beta_3 \uparrow \beta_n \downarrow \beta_5 \uparrow$. Os SBPs da classe II compreendem, em nossa classificação, 4 comunidades distintas (1, 2, 4, e 5), com 210170 proteínas distribuídas em 1000 famílias diferentes.

4.5.1.1. Comunidade 1

A comunidade 1 corresponde aos SBPs de classe II dos grupos E e F (Scheepers, Lycklama A Nijeholt, and Poolman 2016). Compreende 364 famílias, muitas delas com evidência de associação a processos de transdução de sinal, seja pela fusão do SBP com domínios efetores ou pela presença de genes vizinhos que codificam sensores. Cerca de 81% das proteínas nessa comunidade apresentam assinaturas de secreção detectáveis, a maior parte das famílias apresentando uma proporção elevada dessas assinaturas. Na Figura 10, é possível observar as relações de similaridade entre as famílias dessa comunidade. O dendrograma representa o agrupamento hierárquico das famílias, obtido usando o algoritmo de Ward e a probabilidade de homologia dos pares de HMMs, que serviu como métrica para a distância entre as famílias. Embora várias famílias dessa comunidade incluam genes vizinhos a genes da superfamília Cache, é notório que a frequência dessas vizinhanças é, na maioria dos casos, menos de 20%. Um dos membros, a família TctC, inclui a proteína BctC, cuja interação com o domínio sensor da histidina quinase BctE e com a permease BceB já foram comprovados experimentalmente. Curiosamente, vários parálogos de TctC, presentes em *Bordetella pertussis*, não são vizinhos de permeases do sistemas ABC, ao contrário do que ocorre com a maioria dos SBPs componentes de sistemas ABC, incluindo BctC, que são codificados ao lado de suas permeases cognatas.

Figura 11. Agrupamento hierárquico das famílias da comunidade 1. O dendrograma foi construído aplicando o método de Ward à probabilidade de homologia entre HMMs, estimada pelo software HHalign. As duas colunas marcadas pelo gradiente de cores, à direita do dendrograma, correspondem à frequência relativa da proporção de genes, em cada família, que são vizinhos de Cache e a abundância de assinaturas de secreção. A última coluna à direita, antes da escala, mostra a arquitetura de domínios da família com base nos modelos mais semelhantes do Pfam, detectados pelo software hhsearch.

XAC2383 (Família 1.39)

Dentre os membros da comunidade 1, destacamos a presença dos homólogos de XAC2383, gene de *Xanthomonas citri* que interage fisicamente com a proteína de transdução de sinal XAC2382. XAC2382 contém um domínio GGDEF, na porção citoplasmática, e um sCache na extremidade amino-terminal, que fica no periplasma (Teixeira et al. 2018). Aproximadamente um terço dos homólogos de XAC2383 ($e\text{-value} \leq 10^{-3}$) podem ser encontrados ao lado de genes da superfamília Cache com arquiteturas similares a XAC2382. Outro terço é flanqueado por genes de ATPases ou permeases de transportadores ABC. Cerca 15% dos homólogos de XAC2383 são fusões do domínio SBP com as porções transmembranares e citoplasmáticas do sensor. Nenhuma dessas fusões preserva o domínio Cache (Figura 11).

Figura 12. Árvore de máxima verossimilhança do domínio SBP dos membros da família XAC2383. A filogenia acima foi inferida pelo programa FastTree2 (Price, Dehal, and Arkin 2010), com opções padrão, com base em alinhamentos gerados pelo software Mafft. Em verde estão todas as proteínas de genes vizinhos a um sensor Cache. Os ramos em azul são SBPs fusionados a domínios efetores de transdução de sinal, em sua maioria histidina quinases e diguanilato ciclases. As folhas coloridas em vermelho são SBPs de genes vizinhos de transportadores ABC.

4.5.1.2. Comunidade 2

Esta comunidade, de 80224 proteínas, é formada por SBPs do grupo D e é dominada por homólogos do domínio sensor dos reguladores de transcrição da família LTTR (*LysR-type transcriptional regulator*) (Maddocks and Oyston 2008). O agrupamento hierárquico (Figura 12) implica dois grandes subgrupos: os LTTR, na parte superior, e os PBP_like_2, no grupo da parte inferior, assim denominados devido ao modelo do Pfam mais semelhante às famílias desse grupo. Os LTTR, em sua esmagadora maioria, não aparecem no mesmo contexto gênico que os domínios sensores periplasmáticos. O grupo da parte inferior concentra a maior parte das famílias enriquecidas em assinaturas de secreção e inclui algumas famílias com um número grande de Cache nos genes vizinhos.

Família 2.1534

Nesta família, 44% dos membros são vizinhos de homólogos de Cache (Figura 13). As demais proteínas dessa família estão ao lado de sensores com regiões extracitoplasmáticas de outros *folds*, como 4HB e novos domínios, ainda não classificados. Outro ponto notável é a presença de muito transportadores na vizinhança, incluindo homólogos de transportadores de autoindutor-2 (sombra em azul), indicando possível participação em quorum sensing, assim como proteínas desta família claramente associadas a transportadores ABC, o que é evidenciado pela organização e proximidade de porinas e ATPases deste sistema. O conjunto das evidências sugere que os genes desse grupo podem atuar, ao mesmo tempo, na transdução de sinal e no transporte.

Figura 13. Agrupamento hierárquico baseado na similaridade entre as famílias da comunidade 2. O gradiente de cores nas duas colunas à direita do dendrograma indicam, respectivamente, a abundância de genes vizinhos da família Cache e frequência de assinaturas de secreção.

Figura 14. Filogenia de máxima verossimilhança para a família 2.1534. O grupo monofilético de duas OTU destacado com o fundo azul são genes vizinhos de transportadores do autoindutor-2. O destaque em amarelo identifica proteínas associadas a genes de transportadores do tipo ABC. Os nomes dos organismos aparecem à direita, na extremidade dos ramos mais externos. A família é mais comum em deltaproteobactérias e cianobactérias. Nomes de organismos em azul claro identificam genes vizinhos a sensores Cache.

4.5.1.3. Comunidade 4

Esta comunidade agrega membros dos grupos D e G e contém várias famílias relacionadas com vizinhos que codificam proteínas com o domínio Cache e domínios efetores. Nessa comunidade, são abundantes as proteínas com estrutura tridimensional resolvida contendo açúcares ou metais como ligantes. Um dos exemplos é HptA (família 4.914), um SBP de *Staphylococcus aureus*, cuja interação com HptS foi recentemente caracterizada experimentalmente (Wang et al. 2020). Na ausência de glicose, HptA forma um tetrâmero com HptS (HptA(2)-HptS(2)), no qual mantém contatos com os subdomínios distais de HptS. Ao se ligarem a uma molécula de glicose, os monômeros de HptA passam a fazer contatos com o subdomínio mais próximo da membrana na região dCache de HptS, enquanto mantém os contatos com o subdomínio distal do dCache da outra molécula de HptS. Esse rearranjo ativa a quinase. O modo de ação de HptA, porém, não pode ser generalizado para as demais

famílias dessa comunidade, que são muito diversas e incluem vizinhanças com sensores sCache e dCache. Isso sugere não apenas múltiplas origens independentes para as interações SBP-Cache na comunidade 4, mas também a possibilidade de um repertório diverso de mecanismos de interação.

NpsS (Família 4.156)

Outro exemplo de interação SBP-Cache caracterizado na comunidade 4 é a família NpsS, de *Vibrio cholera*, que ativa a atividade de fosfodiesterase de MbaA, por meio de interações com o domínio Cache periplasmático, na presença de poliaminas como espermina (Sobe et al. 2017). Nesta família, predominam contextos genômicos que sugerem fortemente a atuação em transportadores do tipo ABC, o que não é observado para NpsS e seus homólogos mais próximos.

Figura 15. Filogenia dos membros da família NpsS. O grupo marcado em azul inclui as proteínas associadas com genes Cache nessa família. Os demais membros são associados apenas com transportadores.

Figura 16. Agrupamento hierárquico por similaridade das famílias da comunidade 4.

4.5.1.4. Comunidade 5

Esta comunidade equivale ao grupo C de Scheepers *et al.*, e seus membros têm cadeias de aminoácidos mais longas que a maioria dos outros SBPs: 95% delas têm entre 400 e 600 aminoácidos. Membros desta comunidade são conhecidos por ligarem peptídeos, dipeptídeos, oligopeptídeos e níquel, mas também podem aparecer em sistemas de transporte de sacarídeos, como quitina e celulose (Li and Roseman 2004).

Figura 17. Filogenia, vizinhança gênica e domínios das proteínas da família CBP (VC0620, 5.1028) e seus vizinhos.

CBP (Família 5.1028)

A proteína CBP (VC0620, família 5.1028), regula o metabolismo de quitina em *Vibrio harveyi* (Li and Roseman 2004), por meio de interações com o domínio periplasmático da proteína ChiS, um dCache (HHpred contra PDB, 99.16% de probabilidade). Estudos recentes da interação de ChiS e CBP mostraram que a primeira é ativada pela segunda, sem que haja a dissociação das duas proteínas, quando CBP se liga aos oligossacarídeos de quitina (Klancher et al. 2020a). Isso é análogo ao comportamento de HptA-HptS, embora os dois sistemas não sejam evolutivamente relacionados. CBP também atua como SBP do sistema de transporte ABC de quitina. Apesar da evidência de interação física, a família 1028 tem menos de 10% de proteínas com um Cache na vizinhança.

Figura 18. Agrupamento de similaridade para as famílias da comunidade 5.

Família 5.749

Figura 19. Filogenia, vizinhança gênica e domínios das proteínas da família 5.749 e seus vizinhos.

A família 5.749 apresenta 75% de membros vizinhos direto de Cache e se divide em dois grupos, o maior deles presente em delta- e gamaproteobactérias, onde precedem proteínas de transdução de sinal com domínios Cache e histidina quinase. O segundo grupo é formado por terrabactérias (firmicutes e actinobactérias) onde estas proteínas são citoplasmáticas e sem proteínas de transdução de sinal na sua vizinhança. Diferentemente da família VC0622, a vizinhança das proteínas desta família não indica envolvimento com sistemas de transporte. A assinatura de envolvimento em transdução de sinal, para o subgrupo maior, porém, é reforçada pela presença de reguladores de resposta antes dos SBPs.

Um fenômeno interessante nessa família é a presença de 4 proteínas desta família em *Sulfurospirillum* sp. UCH001 associadas com proteínas que participam de processos de transdução de sinal e se organizam em 2 vizinhanças. A primeira região é a combinação de SBP (WP_082705660.1), ao lado de uma histidina quinase, seguido por um parâlogo da família 5.749 (WP_067174673.1) e uma segunda histidina quinase. A segunda região é caracterizada pela presença de dois SBP homólogos, muito próximos, seguidos de de uma histidina quinase. A proximidade dessas sequências na árvore sugere que a origem vem de duplicações recentes. Outros dois membros do gênero *Sulfurospirillum* possuem duas proteínas desta família em seus genomas.

4.5.2. SBPs de classe I (comunidade 3)

Os SBPs de classe I representam a segunda maior população de proteínas em nossa análise. Os membros deste *fold* são caracterizados pela presença de 5 fitas- β paralelas em cada subdomínio, definindo uma topologia $\beta_2 \uparrow \beta_1 \uparrow \beta_3 \uparrow \beta_4 \uparrow \beta_5 \uparrow$. Esta comunidade corresponde ao grupo B e contém várias famílias relacionadas com transdução de sinal. Alguns exemplos de SBPs de classe I já eram conhecidos pela associação com proteínas de transdução de sinal, como o quimiorreceptor Trg, cujo MCP associado contém um domínio TarH (4HB) na região periplasmática (Björkman et al. 1994). Também fazem parte desse grupo as proteínas LuxP e TorT, que regulam as histidina quinases LuxQ e TorS (Neiditch et al. 2006; Moore and Hendrickson 2012). Nossa análise, porém, identificou novos grupos onde a frequência de genes Cache e SBPs, localizados lado a lado, sugere fortemente o envolvimento desses grupos em processos de transdução de sinal (Figura 19).

LuxP (Família 3.1100)

A família LuxP está presente em gamaproteobactérias, alfabroteobactérias e em deferribactérias. É composta por dois grupos principais (Figura 20), sendo que o mais abundante deles é caracterizado pela presença de SBPs em *loci* vizinhos de histidina quinases com domínios reguladores de resposta, e com domínio periplasmático Cache da família LuxP. Esse é o grupo que inclui a proteína LuxP de *Vibrio harveyi*, que regula a histidina quinase LuxQ. O segundo grupo nesta família aparece associado a diguanilato ciclases, indicando que nesta família, assim como demonstrado para XAC2383, a recombinação do domínio Cache com diferentes domínios efetores citoplasmático preserva a unidade SBP-Cache.

Figura 20. Agrupamento de similaridade para as famílias da comunidade 3 (SBPs da classe I).

Figura 21. Filogenia, vizinhança gênica e domínios das proteínas da família LuxQ (3.1100) e seus vizinhos.

Família 3,286

Esta família é frequentemente encontrada ao lado de genes homólogos da família sCache_4 (DUF2222) e de proteínas de membrana externa da família TonB. Algumas das vizinhanças gênicas dessa família apresentam uma composição mais complexa, com outros transportadores próximos, indicando um forte vínculo desta família tanto com transporte. No dendrograma de similaridade da comunidade 3 (Figura 19), pertence a um subgrupo que aparece no alto do dendrograma, formado pelas famílias 2424, 1785, 1486, 1443, 236 e 286. Todas compartilham a maior afinidade por sistemas de transdução de sinal dentro da comunidade 3, com múltiplos exemplos de fusão do SBP com regiões transmembrana e efetores citoplasmáticas na comunidade 2424. Nenhum dos grupos, porém, foi caracterizado experimentalmente ou tem estruturas tridimensionais resolvidas.

Figura 22. Filogenia, vizinhança gênica e domínios das proteínas da família 3.286 e seus vizinhos.

4.5.3. SBPs de classe III (comunidade 6)

A comunidade 6 inclui todas as proteínas da classe II e equivale ao grupo A. Os SBPs de classe III formam um grupo relativamente pequeno dentre os SBPs, em sua maioria especializados em reconhecer e coordenar metais. As proteínas deste fold são caracterizadas pela presença de 4 fitas- β paralelas em cada subdomínio, definindo a topologia $\beta_2\uparrow\beta_1\uparrow\beta_3\uparrow\beta_4\uparrow$, que é similar à da classe I. A região entre os dois subdomínios possui uma longa alfa-hélice, que é uma característica exclusiva desse grupo.

Esta comunidade é composta por 74 famílias, divididas em dois grandes grupos que correspondem às famílias ZnuA e PBP_like_2 do Pfam. Em sua maioria, essas famílias estão associadas com transportadores de metais e proteínas da membrana externa. Os casos observados de associação com proteínas e domínios de transdução de sinal são raros, totalizando apenas 192 genes em um universo de 14135. Não obstante o número reduzido de vizinhanças com genes Cache, a possibilidade de interações entre SBPs da comunidade 6 e Cache não deve ser imediatamente descartada, pois existem evidências experimentais para outras comunidades com baixa frequência de vizinhos Cache que suportam essa associação. Mais interessante ainda, alguns membros do subgrupo maior nessa comunidade (PBP_like_2) aparecem fusionados a serina/treonina quinases em *Mycobacterium* e outras actinobactérias e poderiam desempenhar papéis importantes na divisão celular e patogenicidade desses

organismos (Iswahyudi et al. 2019; Manuse et al. 2016).

Cluster method: ward

Agglomerative Coefficient: 0.96

Figura 23. Agrupamento de similaridade para a comunidade 6 (SBPs da classe III).

5. Discussão

Ao longo de bilhões de anos, os procariotos desenvolveram vários tipos de transportadores que se utilizam de uma proteína extracelular para reconhecimento e captação de solutos e nutrientes, conhecida como proteína ligadora de substratos (SBP). Os transportadores do tipo ABC, o sistema tripartido de transporte de tricarboxilatos (TTT) e os sistemas tripartidos independentes de ATP (TRAP) são exemplos de sistema de transporte que utilizam as SBPs para capturar solutos e apresentá-los às permeases (Scheepers, Lycklama A Nijeholt, and Poolman 2016), que constituem o canal pelo qual os solutos atravessam a membrana celular. Alguns homólogos de SBPs, porém, interagem ou são parte de outro componente importante das membranas celulares: os sistemas de transdução de sinal. A natureza do envolvimento de famílias de SBPs extracelulares com a transdução de sinal e a relação dessas famílias com transporte, no caso específico de procariotos, não foi completamente explorada, salvo em alguns casos isolados, que foram caracterizados experimentalmente.

Neste trabalho, analisamos a origem e evolução de famílias de SBPs (*Substrate-Binding Proteins*) que interagem com domínios de outros *fold*s especializados no reconhecimento de sinais extracelulares, em especial domínios da superfamília Cache, o mais abundante entre os sensores extracelulares de procariotos. Para entender a origem da associação entre SBPs e Cache, construímos uma classificação ampla de SBPs usando dados de estruturas tridimensionais e uma análise profunda das relações de similaridade dessas sequências. Nossa análise foi limitada a genomas de procariotos e a famílias de SBPs encontradas nesses organismos, pois embora o envolvimento de SBPs em processos de transdução de sinal seja bem documentado para algumas famílias de proteínas de eucariotos, o *fold* Cache é basicamente um domínio de procariotos e apenas casos isolados da interação dos dois tipos de domínios foram descritos até o momento, o que não permite generalizações sobre os processos de exaptação dos SBPs entre sistemas de transporte, onde são extremamente comuns, e sistemas de transdução de sinal.

Os casos conhecidos de SBPs que regulam proteínas de transdução de sinal

incluem exemplos das classes I e II de SBPs. Embora um autor tenha sugerido uma origem comum a essas classes (Fukami-Kobayashi, Tatenno, and Nishikawa 1999), as SBPs procarióticas já caracterizadas em vias transdução de sinal são muito divergentes, o que sugere origens independentes para a associação. No entanto, a grande divergência entre as famílias de SBPs, mesmo dentro de cada classe, tornava difícil entender se as famílias associadas a domínios sensores tinham ou não uma origem comum. Combinando a classificação de SBPs baseada na distância estrutural com os agrupamentos hierárquicos das famílias em cada comunidade, estes últimos derivados da análise de similaridade de sequências, conseguimos mapear a frequência de co-ocorrência de genes de SBPs e domínios Cache em todos os níveis de divergência observados para SBPs. A análise dessas vizinhanças mostrou que os SBPs se associam a sensores com domínios Cache e outros domínios extracitoplasmáticos em múltiplos episódios independentes, que parecem ter ocorrido ao longo de milhões de anos da evolução dos procariotos e que, provavelmente, continuam a acontecer.

Nossa estratégia de investigar a associação entre a superfamília Cache e os SBPs a partir da vizinhança gênica é baseada nos vários estudos que demonstraram que a conservação da proximidade de dois ou mais genes em múltiplos genomas e em linhagens diferentes de organismos é um bom preditor de interações funcionais entre esses genes (Huynen et al. 2000)(Wolf et al. 2001)(Burroughs et al. 2015)(Aravind 2000). Na análise estatística global dessas associações, mostramos que pares de genes SBP e Cache são encontrados em genes vizinhos mais frequentemente do que seria esperado por chance. Mais importante, a aplicação do mesmo teste estatístico a proteínas com localização extracelular predita mostrou um nível de suporte ainda maior para essa associação, tanto ao nível de famílias, quanto para as comunidades de SBPs em nossa classificação. Curiosamente, quando olhamos a abundância de vizinhanças contendo SBPs e Cache, a maioria das comunidades apresentaram uma frequência relativamente baixa de associações, ainda que significativas. Como demonstrado em outros estudos (Iyer et al. 2010), um número pequeno de vizinhanças gênicas conservadas é suficiente para a previsão de correlações funcionais, muito embora a previsão de detalhes dessas interações necessitem dados adicionais, como a análise

de homologia, e a homogeneidade dessas interações só seja verificada quando dominam uma população diversa de genomas (Huynen et al. 2000).

O poder dos métodos de contexto genômico para a predição de interações é consideravelmente maior quando se observa a presença de fusões de genes, ou seja, casos em que genes diferentes de uma via metabólica ou outro sistema aparecem como um único gene em outro organismo (Promponas, Ouzounis, and Iliopoulos 2014). Em nossas análises, em um universo de 1338 famílias de SBPs, detectamos fusões de Cache e SBP em apenas 8 famílias, mas observamos fusões de SBPs com os mesmos domínios efetores citoplasmáticos e regiões transmembranares, típicos dos genes sensores da superfamília Cache. Ao todo, as fusões foram observadas em 84 famílias, 5 das quais também incluíam as fusões SBP-Cache. Embora todas as comunidades contenham pelo menos um exemplo, as comunidades 1 e 3 se destacam, por apresentarem, respectivamente, 50 e 22 famílias de SBPs com pelo menos uma fusão a domínios efetores. Essa distribuição das fusões reforça a noção de que muitas das famílias de SBPs onde não se observam fusões podem interagir fisicamente com homólogos dos sensores, embora as análises necessárias para prever e generalizar essas interações ainda estejam em andamento. Especificamente, embora a previsão de interações SBP-sensor possa, em nossa abordagem, ser feita para os membros de uma mesma família, a previsão para as famílias onde não se observa a associação por vizinhança ou fusão vai exigir suporte filogenético adicional ou a análise de outras evidências de contexto, como o uso de perfis filogenéticos (Pellegrini 2012).

As evidências experimentais disponíveis para algumas das famílias analisadas nesse trabalho demonstram a existência de interações físicas diretas, entre o domínio extracitoplasmático do sensor e o SBP. Qual seria então a origem e as consequências da fusão de SBPs e domínios efetores, em especial na ausência de outros domínios sensores extracelulares? Uma possibilidade é que estas versões derivam da fusão de dois genes, que teria ocorrido concomitantemente à deleção do domínio extracelular do sensor. Nesse caso, a função do domínio extracelular, se for além da de interagir com o SBP, será perdida ou terá que ser provida pelo próprio SBP. Em nossos dados, as fusões de SBPs e domínios efetores foram observada em múltiplas famílias onde também encontramos a associação com vizinhos Cache. No entanto, a família

XAC2383 é o único exemplo no qual podemos propor uma resposta a essa pergunta, porque nesse caso já temos a filogenia da família (Figura 11). Assumindo que a raiz da árvore fica na linhagem que leva aos genes associados a transportadores, a linhagem contendo as fusões SBP-efetor teria se originado dentro da linhagem de genes associados ao domínio Cache e, portanto, a hipótese mais parcimoniosa é a de que as fusões representam um único evento de simplificação da arquitetura dos sistemas pela fusão dos genes vizinhos e perda do domínio Cache.

Se a filogenia de outras famílias onde transportadores, sensores com domínios Cache e fusões SBP-efetor tiverem topologias similares e suportarem a mesma trajetória evolutiva, em detrimento da hipótese alternativa, de fissão gênica (Kummerfeld and Teichmann 2005), então a próxima questão será: qual a possível vantagem adaptativa na convergência de múltiplas famílias para uma simplificação da arquitetura de via regulatória? Note que falamos em “vantagem adaptativa” porque o número de famílias onde fusões são observadas é grande o suficiente para que possamos descartar a ideia de que são produto do acaso. Uma possível explicação para a convergência para arquiteturas mais simples pode estar na diferença cinética entre arquiteturas regulatórias compostas por um único domínio sensor, ligado covalentemente ao domínio efetor, e uma via regulada por um domínio sensor solúvel. Como o último tipo de via vai depender da concentração dos SBPs, modelos matemáticos para as duas arquiteturas serão diferentes em características importantes, como o intervalo de concentrações do soluto que altera o comportamento da célula, o tempo de resposta ao estímulo e o efeito da concentração basal de transportadores (Piepenbreier, Fritz, and Gebhard 2017). O efeito desses parâmetros na aptidão do organismo vai depender do modo de vida do organismo, mas pode explicar ajudar a explicar, de forma unificada, a convergência de diferentes linhagens de SBPs para vias de sinalização com essas arquiteturas.

A observação de famílias de SBPs no mesmo *locus* onde estão genes de sistemas de transdução de sinal e transportadores, sugere ainda que muitos desses SBPs operam de forma análoga ao SBP BctC, ou seja, que interagem fisicamente com os dois sistemas. Não aplicamos, porém, métodos de análise *in silico* para detectar as assinaturas de interação entre SBPs e transportadores. Outro fator que não

investigamos em detalhes foi o mecanismo da interação SBP-sensor, incluindo se as múltiplas famílias de SBPs e Cache usam as mesmas superfícies e mecanismos de interação. Uma revisão cuidadosa das evidências genéticas e estruturais, aliada à aplicação de métodos de análise de coevolução de resíduos poderão ajudar a esclarecer os detalhes dessas interações.

Finalmente, é importante ressaltar o potencial da nova classificação hierárquica de SBPs procarióticos, introduzida neste trabalho, para a continuação dos estudos da interação de SBPs e outras proteínas. Nossa classificação foi baseada em métodos de comparação de modelos de Markov, que serviram para identificar e mensurar a similaridade entre grupos mais divergentes de proteínas do que os métodos clássicos, baseados na comparação direta entre sequências, são capazes de detectar. O agrupamento hierárquico das famílias de SBPs e a derivação de comunidades a partir da análise topológica da rede de similaridade das famílias, baseada no algoritmo de Leiden, produziram uma classificação hierárquica com grande potencial para uso na propagação de anotações por homologia. Em seus níveis mais altos (classes e comunidades), nossa classificação é congruente com as classificações já propostas na literatura.

Outro resultado interessante da nossa análise é a contribuição para a história evolutiva dos SBPs. A literatura especula a muito tempo que SBP de classe I e II são homólogos remotos, com um autor postulando a origem das duas classes a partir da fusão de reguladores de resposta (Fukami-Kobayashi, Tateno, and Nishikawa 1999). Esses autores, porém, não analisaram o que hoje chamamos de SBPs de classe III, que foi descrito no mesmo ano de 1999 em *Haemophilus ducreyi* (Lewis et al. 1999). Recentemente, foi demonstrada a dimerização de CheY (Lewis et al. 1999; Paithankar et al. 2019), o que dá suporte à hipótese de Fukami-Kobayashi. Nossos resultados sugerem uma maior similaridade entre estruturas de SBPs de classe I e III, quando comparados com SBPs de classe II. Essa conclusão é suportada tanto pela análise de similaridade dos HMMs das famílias de SBP I e III, que se mostram mais semelhantes entre si, quanto pelo agrupamento hierárquico das estruturas baseado na distância euclidiana das coordenadas de átomos. Uma diferença relevante que separa SBPs de Classe I e III dos SBPs de classe II é a presença de uma fita beta antiparalela em cada

um dos dois subdomínios dos SBPs de classe II. Essa sinapomorfia não existe nas outras classes.

Transporte e sinalização são funções que muitas vezes integram uma mesma via ou sistema. Outras proteínas de transporte participam de vias de sinalização, seja devido à sua capacidade de reconhecer moléculas específicas ou de sentir perturbações na membrana celular, como é o caso dos transportadores de amônia e de aminoácidos (Guedes A.C.P., resultados não mostrados). Contudo, nas versões procarióticas desses transportadores, fusões com domínios efetores de sistemas de um e/ou dois componentes aparecem em grupos monofiléticos, sugerindo que o recrutamento do transportador, para uma atividade sensora, ocorreu apenas uma vez. Isso pode implicar que as vantagens adaptativas deste evento são distintas das envolvidas no recrutamento de SBPs, talvez por causa da versatilidade dos SBPs, que aparecem em múltiplas versões, cada uma delas capaz de ligar um repertório distinto de solutos.

6. Conclusão

Neste trabalho, construímos uma nova classificação de proteínas ligadoras de substratos (SBPs) e identificamos as famílias participando de sistemas de transdução de sinal, em particular as associadas ao domínio sensor Cache, por meio de métodos *in silico*. Nossos resultados mostram múltiplos episódios de recrutamento de domínios SBP, em procariotos, para contextos de transdução de sinal. Identificamos dezenas de famílias de SBPs de todas as classes onde é comum a associação ao domínio Cache, mas para as quais nenhum estudo anterior apontou o envolvimento em transdução de sinal. Iniciamos a análise da evolução dessas famílias, identificando famílias onde a composição de domínios e a filogenia sugerem a fusão de genes de SBPs e sensores como a origem mais provável para o repertórios de SBPs fusionados a sistemas de um e/ou dois componentes e a outras proteínas de transdução de sinal, como sensores quimiotáticos e enzimas que produzem mensageiros secundários.

A presença de famílias distantes de SBPs em contextos de transdução de sinal e, simultaneamente, de transportadores, indica que a transição do domínio entre os sistemas de transporte e de transdução de sinal pode estar associada com a ampla presença de interações de SBP de transportadores com proteínas de outros sistemas, incluindo, potencialmente vários sensores de sistemas de um ou dois componentes. A presença, em famílias de SBPs que atuam em transporte, de interações físicas com outros sistemas ainda carece de confirmação experimental, tendo sido demonstrada apenas para algumas famílias. Se confirmada a natureza promíscua da interações proteína-proteína dos SBPs procarióticos, a emergência de famílias especializadas em transdução de sinal poderá ser entendida como uma especialização e a recorrência do fenômeno, em múltiplas famílias, um evento provável.

7. Referências

- Alvarez, Adrian F., Carlos Barba-Ostria, Hortencia Silva-Jiménez, and Dimitris Georgellis. 2016. "Organization and Mode of Action of Two Component System Signaling Circuits from the Various Kingdoms of Life." *Environmental Microbiology* 18 (10): 3210–26.
- Anantharaman, Vivek, and L. Aravind. 2000. "Cache – a Signaling Domain Common to Animal Ca²⁺-Channel Subunits and a Class of Prokaryotic Chemotaxis Receptors." *Trends in Biochemical Sciences* 25 (11): 535–37.
- . 2003. "Application of Comparative Genomics in the Identification and Analysis of Novel Families of Membrane-Associated Receptors in Bacteria." *BMC Genomics* 4 (1): 34.
- . 2006. "Diversification of Catalytic Activities and Ligand Interactions in the Protein Fold Shared by the Sugar Isomerases, eIF2B, DeoR Transcription Factors, Acyl-CoA Transferases and Methenyltetrahydrofolate Synthetase." *Journal of Molecular Biology* 356 (3): 823–42.
- Andreeva, Antonina, Dave Howorth, Cyrus Chothia, Eugene Kulesha, and Alexey G. Murzin. 2014. "SCOP2 Prototype: A New Approach to Protein Structure Mining." *Nucleic Acids Research* 42 (Database issue): D310–14.
- Aravind, L. 2000. "Guilt by Association: Contextual Information in Genome Analysis." *Genome Research* 10 (8): 1074–77.
- Aravind, L., and E. V. Koonin. 2000. "The STAS Domain - a Link between Anion Transporters and Antisigma-Factor Antagonists." *Current Biology: CB* 10 (2): R53–55.
- Bateman, A. 1997. "The Structure of a Domain Common to Archaeobacteria and the Homocystinuria Disease Protein." *Trends in Biochemical Sciences* 22 (1): 12–13.
- Berntsson, Ronnie P-A, Sander H. J. Smits, Lutz Schmitt, Dirk-Jan Slotboom, and Bert Poolman. 2010. "A Structural Classification of Substrate-Binding Proteins." *FEBS Letters* 584 (12): 2606–17.
- Björkman, A. J., R. A. Binnie, H. Zhang, L. B. Cole, M. A. Hermodson, and S. L. Mowbray. 1994. "Probing Protein-Protein Interactions. The Ribose-Binding Protein in Bacterial Transport and Chemotaxis." *The Journal of Biological Chemistry* 269 (48): 30206–11.
- Bradshaw, Ralph A., and Edward A. Dennis. 2003. *Handbook of Cell Signaling, Three-Volume Set*. Academic Press.
- . 2009. *Handbook of Cell Signaling*. Academic Press.
- Burroughs, A. Maxwell, Dapeng Zhang, Daniel E. Schäffer, Lakshminarayan M. Iyer, and L. Aravind. 2015. "Comparative Genomic Analyses Reveal a Vast, Novel Network of Nucleotide-Centric Systems in Biological Conflicts, Immunity and Signaling." *Nucleic Acids Research* 43 (22): 10633–54.
- Busby, S., and R. H. Ebright. 1999. "Transcription Activation by Catabolite Activator Protein (CAP)." *Journal of Molecular Biology* 293 (2). <https://doi.org/10.1006/jmbi.1999.3161>.
- Chai, Y., J. Zhu, and S. C. Winans. 2001. "TrlR, a Defective TraR-like Protein of

- Agrobacterium Tumefaciens, Blocks TraR Function in Vitro by Forming Inactive TrIR:TraR Dimers." *Molecular Microbiology* 40 (2): 414–21.
- Cheng, Hua, Yuxing Liao, R. Dustin Schaeffer, and Nick V. Grishin. 2015. "Manual Classification Strategies in the ECOD Database." *Proteins* 83 (7): 1238–51.
- Cheng, Hua, R. Dustin Schaeffer, Yuxing Liao, Lisa N. Kinch, Jimin Pei, Shuoyong Shi, Bong-Hyun Kim, and Nick V. Grishin. 2014. "ECOD: An Evolutionary Classification of Protein Domains." *PLoS Computational Biology* 10 (12): e1003926.
- Cheung, Jonah, Craig A. Bingman, Marsha Reyngold, Wayne A. Hendrickson, and Carey D. Waldburger. 2008. "Crystal Structure of a Functional Dimer of the PhoQ Sensor Domain." *The Journal of Biological Chemistry* 283 (20): 13762–70.
- Cockerell, Steven R., Alex C. Rutkovsky, Josiah P. Zayner, Rebecca E. Cooper, Lindsay R. Porter, Sam S. Pendergraft, Zach M. Parker, Marcus W. McGinnis, and Ece Karatan. 2014. "Vibrio Cholerae NspS, a Homologue of ABC-Type Periplasmic Solute Binding Proteins, Facilitates Transduction of Polyamine Signals Independent of Their Transport." *Microbiology* 160 (Pt 5): 832–43.
- DeLano, Warren Lyford. 2002. "PyMOL." http://virology.wisc.edu/acp/Classes/DropFolders/Drop660_lectures/2013_660/L01_PyMOL_2013r.pdf.
- Deutscher, Josef. 2008. "The Mechanisms of Carbon Catabolite Repression in Bacteria." *Current Opinion in Microbiology* 11 (2): 87–93.
- Finn, Robert D., Jody Clements, and Sean R. Eddy. 2011. "HMMER Web Server: Interactive Sequence Similarity Searching." *Nucleic Acids Research* 39 (Web Server issue): W29–37.
- Fukami-Kobayashi, Kaoru, Yoshio Tateno, and Ken Nishikawa. 1999. "Domain Dislocation: A Change of Core Structure in Periplasmic Binding Proteins in Their Evolutionary History." *Journal of Molecular Biology* 286 (1): 279–90.
- Fukamizo, Tamo, Yoshihito Kitaoku, and Wipa Suginta. 2019. "Periplasmic Solute-Binding Proteins: Structure Classification and Chitooligosaccharide Recognition." *International Journal of Biological Macromolecules* 128 (May): 985–93.
- Galperin, Michael Y. 2004. "Bacterial Signal Transduction Network in a Genomic Perspective." *Environmental Microbiology* 6 (6): 552–67.
- Galperin, Michael Y., and Joachim E. Schultz. 2019. "Sensory Transduction in Bacteria." In *Encyclopedia of Microbiology (Fourth Edition)*, edited by Thomas M. Schmidt, 224–39. Oxford: Academic Press.
- Gardina, P. J., A. F. Bormans, M. A. Hawkins, J. W. Meeker, and M. D. Manson. 1997. "Maltose-Binding Protein Interacts Simultaneously and Asymmetrically with Both Subunits of the Tar Chemoreceptor." *Molecular Microbiology* 23 (6): 1181–91.
- Haft, Daniel H., Michael DiCuccio, Azat Badretdin, Vyacheslav Brover, Vyacheslav Chetvernin, Kathleen O'Neill, Wenjun Li, et al. 2018. "RefSeq: An Update on Prokaryotic Genome Annotation and Curation." *Nucleic Acids Research* 46 (D1): D851–60.
- Henry, Jonathan T., and Sean Crosson. 2011. "Ligand-Binding PAS Domains in a Genomic, Cellular, and Structural Context." *Annual Review of Microbiology* 65 (1): 261–86.
- Huynen, M., B. Snel, W. Lathe 3rd, and P. Bork. 2000. "Predicting Protein Function by Genomic Context: Quantitative Evaluation and Qualitative Inferences." *Genome*

- Research* 10 (8): 1204–10.
- Iswahyudi, Galina V. Mukamolova, Anna A. Straatman-Iwanowska, Natalie Allcock, Paul Ajuh, Obolbek Turapov, and Helen M. O'Hare. 2019. "Mycobacterial Phosphatase PstP Regulates Global Serine Threonine Phosphorylation and Cell Division." *Scientific Reports* 9 (1): 8337.
- Iyer, Lakshminarayan M., Saraswathi Abhiman, Robson F. de Souza, and L. Aravind. 2010. "Origin and Evolution of Peptide-Modifying Dioxygenases and Identification of the Wybutosine Hydroxylase/hydroperoxidase." *Nucleic Acids Research* 38 (16): 5261–79.
- Kall, L., A. Krogh, and E. L. L. Sonnhammer. 2007. "Advantages of Combined Transmembrane Topology and Signal Peptide Prediction--the Phobius Web Server." *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkm256>.
- Käll, Lukas, Anders Krogh, and Erik L. L. Sonnhammer. 2004. "A Combined Transmembrane Topology and Signal Peptide Prediction Method." *Journal of Molecular Biology* 338 (5): 1027–36.
- Kashith, M., B. Keerthana, S. Sriram, and V. Ramamurthy. 2016. "Adenylate Cyclase in *Arthrosira Platensis* Responds to Light through Transcription." *Biochemical and Biophysical Research Communications* 477 (2): 297–301.
- Katoh, Kazutaka, Kei-Ichi Kuma, Hiroyuki Toh, and Takashi Miyata. 2005. "MAFFT Version 5: Improvement in Accuracy of Multiple Sequence Alignment." *Nucleic Acids Research* 33 (2): 511–18.
- Klancher, Catherine A., Shouji Yamamoto, Triana N. Dalia, and Ankur B. Dalia. 2020a. "ChiS Is a Noncanonical DNA-Binding Hybrid Sensor Kinase That Directly Regulates the Chitin Utilization Program in *Vibrio Cholerae*." <https://doi.org/10.1101/2020.01.10.902320>.
- . 2020b. "ChiS Is a Noncanonical DNA-Binding Hybrid Sensor Kinase That Directly Regulates the Chitin Utilization Program in." *Proceedings of the National Academy of Sciences of the United States of America* 117 (33): 20180–89.
- Koonin, E. V., Y. I. Wolf, and L. Aravind. 2001. "Prediction of the Archaeal Exosome and Its Connections with the Proteasome and the Translation and Transcription Machineries by a Comparative-Genomic Approach." *Genome Research* 11 (2): 240–52.
- Kummerfeld, Sarah K., and Sarah A. Teichmann. 2005. "Relative Rates of Gene Fusion and Fission in Multi-Domain Proteins." *Trends in Genetics: TIG* 21 (1): 25–30.
- Lewis, D. A., J. Klesney-Tait, S. R. Lumbley, C. K. Ward, J. L. Latimer, C. A. Ison, and E. J. Hansen. 1999. "Identification of the *znuA*-Encoded Periplasmic Zinc Transport Protein of *Haemophilus Ducreyi*." *Infection and Immunity* 67 (10): 5060–68.
- Li, X., and S. Roseman. 2004. "The Chitinolytic Cascade in *Vibrios* Is Regulated by Chitin Oligosaccharides and a Two-Component Chitin Catabolic Sensor/kinase." *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.0307645100>.
- Maddocks, Sarah E., and Petra C. F. Oyston. 2008. "Structure and Function of the LysR-Type Transcriptional Regulator (LTTR) Family Proteins." *Microbiology* 154 (Pt 12): 3609–23.
- Manuse, Sylvie, Aurore Fleurie, Laure Zucchini, Christian Lesterlin, and Christophe Grangeasse. 2016. "Role of Eukaryotic-like Serine/threonine Kinases in Bacterial

- Cell Division and Morphogenesis." *FEMS Microbiology Reviews* 40 (1): 41–56.
- McCleary, W. R., and J. B. Stock. 1994. "Acetyl Phosphate and the Activation of Two-Component Response Regulators." *The Journal of Biological Chemistry* 269 (50): 31567–72.
- Mistry, Jaina, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A. Salazar, Erik L. L. Sonnhammer, Silvio C. E. Tosatto, et al. 2021. "Pfam: The Protein Families Database in 2021." *Nucleic Acids Research* 49 (D1): D412–19.
- Moore, Jason O., and Wayne A. Hendrickson. 2012. "An Asymmetry-to-Symmetry Switch in Signal Transmission by the Histidine Kinase Receptor for TMAO." *Structure* 20 (4): 729–41.
- Mowbray, Sherry L., and Daniel E. Koshland Jr. 1987. "Additive and Independent Responses in a Single Receptor: Aspartate and Maltose Stimuli on the Tar Protein." *Cell* 50 (2): 171–80.
- Murtagh, Fionn, and Pierre Legendre. 2014. "Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?" *Journal of Classification* 31 (3): 274–95.
- Neiditch, Matthew B., Michael J. Federle, Audra J. Pompeani, Robert C. Kelly, Danielle L. Swem, Philip D. Jeffrey, Bonnie L. Bassler, and Frederick M. Hughson. 2006. "Ligand-Induced Asymmetry in Histidine Sensor Kinase Complex Regulates Quorum Sensing." *Cell* 126 (6): 1095–1108.
- Ondov, Brian D., Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. 2016. "Mash: Fast Genome and Metagenome Distance Estimation Using MinHash." *Genome Biology* 17 (1): 132.
- Paithankar, Karthik Shivaji, Mathias Enderle, David C. Wirthensohn, Arthur Miller, Matthias Schlesner, Friedhelm Pfeiffer, Alexander Rittner, Martin Grininger, and Dieter Oesterhelt. 2019. "Structure of the Archaeal Chemotaxis Protein CheY in a Domain-Swapped Dimeric Conformation." *Acta Crystallographica Section F Structural Biology Communications*. <https://doi.org/10.1107/s2053230x19010896>.
- Parkinson, J. S., and E. C. Kofoed. 1992. "Communication Modules in Bacterial Signaling Proteins." *Annual Review of Genetics* 26: 71–112.
- Pei, J., and N. V. Grishin. 2001. "GGDEF Domain Is Homologous to Adenylyl Cyclase." *Proteins* 42 (2): 210–16.
- Pellegrini, Matteo. 2012. "Using Phylogenetic Profiles to Predict Functional Relationships." *Methods in Molecular Biology* 804: 167–77.
- Piepenbreier, Hannah, Georg Fritz, and Susanne Gebhard. 2017. "Transporters as Information Processors in Bacterial Signalling Pathways." *Molecular Microbiology* 104 (1): 1–15.
- Ponting, C. P. 1997. "CBS Domains in CIC Chloride Channels Implicated in Myotonia and Nephrolithiasis (kidney Stones)." *Journal of Molecular Medicine* 75 (3): 160–63.
- Potter, Lincoln R., Andrea R. Yoder, Darcy R. Flora, Laura K. Antos, and Deborah M. Dickey. 2009. "Natriuretic Peptides: Their Structures, Receptors, Physiologic Functions and Therapeutic Applications." *Handbook of Experimental Pharmacology*, no. 191: 341–66.
- Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2010. "FastTree 2--Approximately Maximum-Likelihood Trees for Large Alignments." *PloS One* 5 (3):

- e9490.
- Promponas, Vasilis J., Christos A. Ouzounis, and Ioannis Iliopoulos. 2014. "Experimental Evidence Validating the Computational Inference of Functional Associations from Gene Fusion Events: A Critical Survey." *Briefings in Bioinformatics* 15 (3): 443–54.
- Purcell, Erin B., and Sean Crosson. 2008. "Photoregulation in Prokaryotes." *Current Opinion in Microbiology* 11 (2): 168–78.
- Saier, Milton H., Jr, Vamsee S. Reddy, Brian V. Tsu, Muhammad Saad Ahmed, Chun Li, and Gabriel Moreno-Hagelsieb. 2016. "The Transporter Classification Database (TCDB): Recent Advances." *Nucleic Acids Research* 44 (D1): D372–79.
- Scheepers, Giel H., Jelger A. Lycklama a Nijeholt, and Bert Poolman. 2016. "An Updated Structural Classification of Substrate-binding Proteins." *Febs Letters* 590 (23): 4393–4401.
- Scheepers, Giel H., Jelger A. Lycklama A Nijeholt, and Bert Poolman. 2016. "An Updated Structural Classification of Substrate-Binding Proteins." *FEBS Letters* 590 (23): 4393–4401.
- Shu, Chengyi J., Luke E. Ulrich, and Igor B. Zhulin. 2003. "The NIT Domain: A Predicted Nitrate-Responsive Module in Bacterial Sensory Receptors." *Trends in Biochemical Sciences* 28 (3): 121–24.
- Skjærven, Lars, Xin-Qiu Yao, Guido Scarabelli, and Barry J. Grant. 2014. "Integrating Protein Structural Dynamics and Evolutionary Analysis with Bio3D." *BMC Bioinformatics* 15 (December): 399.
- Smith, Roger S., Matthew C. Wolfgang, and Stephen Lory. 2004. "An Adenylate Cyclase-Controlled Signaling Network Regulates *Pseudomonas Aeruginosa* Virulence in a Mouse Model of Acute Pneumonia." *Infection and Immunity* 72 (3): 1677–84.
- Sobe, Richard C., Whitney G. Bond, Caitlin K. Wotanis, Josiah P. Zayner, Marybeth A. Burriss, Nicolas Fernandez, Eric L. Bruger, Christopher M. Waters, Howard S. Neufeld, and Ece Karatan. 2017. "Spermine Inhibits Biofilm Formation through the NspS-MbaA Polyamine Signaling System." *The Journal of Biological Chemistry* 292 (41): 17025–36.
- Spiro, Stephen, and Ray Dixon. 2010. *Sensory Mechanisms in Bacteria: Molecular Aspects of Signal Recognition*. Horizon Scientific Press.
- Steinegger, Martin, Markus Meier, Milot Mirdita, Harald Vöhringer, Stephan J. Haunsberger, and Johannes Söding. 2019. "HH-suite3 for Fast Remote Homology Detection and Deep Protein Annotation." *BMC Bioinformatics* 20 (1): 473.
- Steinegger, Martin, and Johannes Söding. 2017. "MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets." *Nature Biotechnology* 35 (11): 1026–28.
- Strachan, Tom, and Andrew P. Read. 2011. *Human Molecular Genetics* 4. Garland Science/Taylor & Francis Group.
- Teixeira, Raphael D., Cristiane R. Guzzo, Santiago Justo Arévalo, Maxuel O. Andrade, Josielle Abrahão, Robson F. de Souza, and Chuck S. Farah. 2018. "A Bipartite Periplasmic Receptor-Diguanylate Cyclase Pair (XAC2383-XAC2382) in the Bacterium." *The Journal of Biological Chemistry* 293 (27): 10767–81.
- Traag, V. A., L. Waltman, and N. J. van Eck. 2019. "From Louvain to Leiden:

- Guaranteeing Well-Connected Communities.” *Scientific Reports* 9 (1): 5233.
- Ulrich, L. E. 2005. “Four-Helix Bundle: A Ubiquitous Sensory Module in Prokaryotic Signal Transduction.” *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/bti1204>.
- Ulrich, Luke E., Eugene V. Koonin, and Igor B. Zhulin. 2005. “One-Component Systems Dominate Signal Transduction in Prokaryotes.” *Trends in Microbiology* 13 (2): 52–56.
- Upadhyay, Amit A., Aaron D. Fleetwood, Ogun Adebali, Robert D. Finn, and Igor B. Zhulin. 2016. “Cache Domains That Are Homologous To, but Different from PAS Domains Comprise the Largest Superfamily of Extracellular Sensors in Prokaryotes.” *PLoS Computational Biology* 12 (4): e1004862.
- Van Tran, Can. 2004. *TCDB: A Membrane Transport Protein Classification Database*.
- Wang, Mingxing, Qiong Guo, Kongfu Zhu, Bo Fang, Yifan Yang, Maikun Teng, Xu Li, and Yuyong Tao. 2020. “Interface Switch Mediates Signal Transmission in a Two-Component System.” *Proceedings of the National Academy of Sciences of the United States of America* 117 (48): 30433–40.
- Witan, Julian, Julia Bauer, Ilka Wittig, Philipp Aloysius Steinmetz, Wolfgang Erker, and Gottfried Unden. 2012. “Interaction of the Escherichia Coli Transporter DctA with the Sensor Kinase DcuS: Presence of Functional DctA/DcuS Sensor Units.” *Molecular Microbiology* 85 (5): 846–61.
- Wolf, Y. I., I. B. Rogozin, A. S. Kondrashov, and E. V. Koonin. 2001. “Genome Alignment, Evolution of Prokaryotic Genome Organization, and Prediction of Gene Function Using Genomic Context.” *Genome Research* 11 (3): 356–72.
- Yau, K. W. 1994. “Cyclic Nucleotide-Gated Channels: An Expanding New Family of Ion Channels.” *Proceedings of the National Academy of Sciences of the United States of America* 91 (9): 3481–83.
- Zhang, Yang, and Jeffrey Skolnick. 2005. “TM-Align: A Protein Structure Alignment Algorithm Based on the TM-Score.” *Nucleic Acids Research* 33 (7): 2302–9.
- Zhang, Zhen, and Wayne A. Hendrickson. 2010. “Structural Characterization of the Predominant Family of Histidine Kinase Sensor Domains.” *Journal of Molecular Biology* 400 (3): 335–53.