Universidade Federal do Rio Grande do Sul
Centro de Biotecnologia do Estado do Rio Grande do Sul
Bacharelado em Biotecnologia - Ênfase em Bioinformática

Andrey Felipe Schoier

Análise evolutiva das Metionina Aminopeptidases

Uma abordagem bioinformática

Prof. Dr. Jose Claudio Fonseca Moreira

Orientador

Porto Alegre

2021

Andrey Felipe Schoier

Análise evolutiva das Metionina Aminopeptidases

Uma abordagem bioinformática

Trabalho de conclusão de curso apresentado como requisito parcial à obtenção do título de bacharel em Biotecnologia - Ênfase em Bioinformática do Centro de Biotecnologia da Universidade Federal do Rio Grande do Sul.

Orientador: Prof. Dr. Jose Claudio Fonseca Moreira

Porto Alegre

2021

AGRADECIMENTOS

Aos meus pais, Adir e Eliane, por terem me concebido, criado e sustentado durante tantos anos sempre com muito amor e carinho.

À minha esposa, Isadora, pelo apoio e paciência, especialmente durante o período da criação deste trabalho.

À Universidade Federal do Rio Grande do Sul, que me acolheu e acolhe por tanto tempo e me formou como cidadão consciente.

Ao meu orientador, José Claudio, pela compreensão e pelos valiosos conselhos.

“Se eu vi mais longe, foi por estar sobre ombros de gigantes.”

Isaac Newton

RESUMO

A Excisão de Metionina N-terminal (EMN) é um processo que ocorre em cerca de 30% a 60% de todas proteínas expressadas por um genoma, a depender do organismo e seu compartimento celular. Este é um processo conservado em todos domínios da vida e essencial para manutenção da funcionalidade e do crescimento celular em qualquer organismo investigado. Através do seu controle global da meia vida de proteínas, a EMN já demonstrou regular a homeostase redox global de glutationa, ao menos em plantas, leveduras e arqueas. A enzima responsável pela EMN se chama Metionina Aminopeptidase (MetAP). Assim como a EMN, as MetAPs são conservadas em bactérias, arqueas e eucariotos. Já se comprovou um papel fundamental de MetAPs humanas nos processos de angiogênese e de linfogênese. As MetAPs são alvos terapêuticos em uma vasta gama de estudos que vão desde terapias anti câncer a tratamentos de zoonoses. Estruturalmente, todas variantes de MetAP apresentam um eixo de simetria *pseudo two fold* com um sítio catalítico e um sítio de ligação a íon metálico na interface entre os domínios. Os tipos e subtipos de MetAP caracterizam-se pela presença ou ausência de inserções adicionais nas regiões C-terminal e N-terminal, respectivamente. A distribuição dos tipos de MetAPs ao longo da biodiversidade se dá de maneira que, em seus genomas, bactérias apresentam apenas genes homólogos a *MetAP* do tipo 1 (MetAP1), arqueas apresentam apenas genes homólogos a *MetAP* do tipo 2 (MetAP2) e eucariotos apresentam genes homólogos a ambos tipos. Adicionalmente, eucariotos apresentam variantes de MetAP1 localizadas no interior de suas mitocôndrias e/ou plastídeos. Apesar de se atribuir a presença de homólogos a ambos tipos de MetAP em eucariotos a eventos endossimbióticos, pouco se sabe sobre as origens evolutivas de cada uma das variantes de MetAP. Neste estudo, analisamos a distribuição dos tipos e subtipos de MetAP ao longo da biodiversidade, investigando possíveis candidatos a variantes ainda não identificadas, assim como realizamos a construção de um modelo filogenético para a evolução das MetAPs. Resultados preliminares indicam a ocorrência de membros de MetAP2 em uma família bacteriana, o que é indicativo da necessidade de uma melhor descrição da distribuição dos tipos de *MetAP* ao longo da árvore da vida.

Palavras-chave: MetAP, MNE, Filogenia, Bioinformática.

ABSTRACT

N-terminal Methionine Excision (NME) is a process that occurs in about 30% to 60% of all expressed proteins in a genome, depending on the organism and cellular compartment. It is conserved through all life domains and it is essential for normal growth and function in any organism investigated. By means of its capability of control over protein half-life, NME has demonstrated to regulate global glutathione redox homeostasis, at least in plants, yeast and Archaea. The enzyme responsible for NME carriage is Methionine Aminopeptidase (MetAP). Like NME, MetAPs are conserved in Bacteria, Archaea and Eukaryotes. It has proven to play a primordial role in human angiogenesis and lymphangiogenesis processes. MetAPs are pharmaceutical targets in a wide range of studies ranging from anti-cancer therapies to zoonose treatments. Structurally, every MetAP variant presents a pseudo twofold symmetry axis of symmetry with the catalytic site and metal ion binding located at the interfaces between the domains. The types and subtypes of MetAP are characterized by the presence or absence of additional insertions in their C-termini and N-termini regions, respectively. The distribution of types of MetAPs along biodiversity is given in a way that, in their genome, Bacteria presents only genes homologous to type 1 *MetAP* (MetAP1), Archaea presents only genes homologous to type 2 *MetAP* (MetAP2) and Eukaryotes presents genes homologous to both types. Additionally, Eukaryotes present MetAP1 variants that are located inside their mitochondria and/or plastids. Despite the presence of both homologous types of *MetAP* in eukaryotes is attributed to endosymbiotic events, little is known about the evolutionary origins of each MetAP variant. In the present study, we analyze the distribution of MetAP types and subtypes across biodiversity, investigating possible candidates for as-yet-unidentified variants, as well as construct a phylogenetic model of MetAPs evolution. Preliminary results indicate the occurrence of MetAP2 members in a bacterial family, which is indicative of the necessity of better description of the *MetAP* types distribution along the tree of life.

Keywords: MetAP, NME, Phylogeny, Bioinformatics.

Sumário

INTRODUÇÃO GERAL

1. CONTEXTUALIZAÇÃO

A vida como conhecemos possui diversas características universais. Todos organismos vivos descobertos e documentados até hoje possuem características estruturais constitutivas semelhantes em nível microscópico – todos somos formados por células (Alberts 2010; Capítulo 1). Além disso, as células por si só são construídas pelos mesmos grupos funcionais de moléculas, responsáveis não somente por moldá-las espacialmente, mas também por dotá-las de características funcionais metabólicas (Lehninger, Nelson 2017 - Capítulo 4). Desempenhando um papel fundamental neste contexto, um grupo funcional destaca-se por sua presença em todos níveis funcionais celulares, as proteínas. Pode-se observá-las na constituição do aparato molecular responsável por diversos processos. Dentre estes, estão desde a sinalização intracelular (desempenhada por peptídeos), passando pela sinalização extracelular (desempenhada por glicoproteínas), pelos polímeros estruturais que constituem o esqueleto celular, até as enzimas - que atuam como catalisadores biológicos nas reações químicas inerentes aos metabolismos indispensáveis para viabilidade da vida. (Lehninger, Nelson 2017)

Basicamente todos processos celulares envolvem algum tipo de proteína. A estrutura geral proteica é de uma cadeia repetitiva de blocos constitutivos denominados aminoácidos. Estes são todos idênticos entre si, exceto por uma porção denominada radical. Surpreendentemente, apenas 20 resíduos (aminoácidos individuais dentro de uma cadeia) de aminoácidos compõem toda proteína em uma combinação específica. Não à toa, cada célula viva possui, via de regra, a fórmula para todas proteínas que a constituem. (Alberts 2010; Lehninger, Nelson 2017)

Esta informação está armazenada em moléculas pertencentes a outro grupo funcional químico universal na vida; os ácidos nucleicos. Dentro desta categoria de moléculas, encontram-se o DNA e o RNA, que são formados também por uma cadeia repetitiva, porém de nucleosídeos. O que diferencia os nucleosídeos entre si é somente a base nitrogenada ligada a uma porção que é idêntica em todas moléculas. Esta característica torna os ácidos nucleicos capazes de codificar informações compostas por diferentes sequências destes nucleotídeos (nucleosídeo individual dentro de uma cadeia). As moléculas de RNA caracterizam-se por uma cadeia única de nucleotídeos denominada fita simples, sendo estas pouco estáveis e com ciclo de vida relativamente curto. Em contrapartida, moléculas de DNA são mais estáveis e são formadas por duas fitas complementares que assumem um formato de hélice. (Alberts 2010; Lehninger, Nelson 2017)

Os cromossomos, formados por moléculas de DNA e proteínas associadas, funcionam como arcabouços informacionais, consultados pela célula na ocasião da síntese de qualquer proteína. O trecho de um cromossomo que codifica um produto é conhecido como gene. Apesar de existirem outros tipos de produtos gênicos, o foco deste texto se dá nos genes que codificam proteínas. (Alberts 2010; Lehninger, Nelson 2017)

O fluxo da informação biológica que viabiliza a continuidade de todo ser vivo se dá através de três processos centrais preservados, sendo estes a replicação, a transcrição e a tradução. Estes três processos constituem o dogma central da biologia celular, postulado em 1958 por Francis Crick. Em suma, a replicação é responsável por produzir cópias dos cromossomos e garantir a herança genética de uma linhagem celular, a transcrição transcreve a informação de um gene em moléculas de RNA mensageiras (denominadas mRNA) e a tradução traduz estas mensagens em uma cadeia de aminoácidos.

Durante o processo de tradução e maturação de proteínas, a informação genética unidimensional contida nos códons de um mRNA maduro se transforma em um produto peptídico tridimensional. O maquinário enzimático responsável pela tradução é constituído pelo ribossomo – composto por duas subunidades que consistem de uma fita de RNA ribossômico (rRNA) cada – e por outras proteínas associadas. Mecanisticamente, as subunidades ribossômicas acoplam-se à fita de um mRNA maduro e efetuam a leitura da mensagem nela contida de três em três nucleotídeos. A esta trinca dá-se o nome códon, previamente mencionado. Este fenômeno biológico está preservado em todos domínios da vida, havendo diferenças observadas somente em seus significados (aminoácidos correspondentes) ao longo da biodiversidade. O mecanismo, porém, é idêntico, independentemente do organismo vivo. É por este motivo que a tradução compõe o dogma central da biologia celular. (Alberts 2010; Lehninger, Nelson 2017)

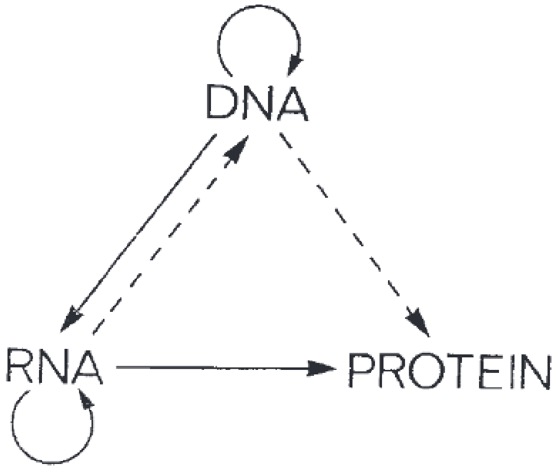


Fig. 2. The arrows show the situation as it seemed in 1958. Solid arrows represent probable transfers, dotted arrows possible transfers. The absent arrows (compare Fig. 1) represent the impossible transfers postulated by the central dogma. They are the three possible arrows starting from protein.

Figura 1 - Dogma central da biologia molecular proposto por Krick em 1970

A mensagem contida na região codificante de um mRNA maduro é constituída por um códon de início, uma sequência codante e um códon de terminação. O último é lido pelo maquinário enzimático como uma instrução de parada, o que leva o ribossomo a se desacoplar da molécula de mRNA. O primeiro, por sua vez, é traduzido para um aminoácido de metionina, via de regra. A proteína é sintetizada pelas subsequentes ligações peptídicas, a partir da metionina, entre os aminoácidos codificados na mensagem. Este produto proteico, gerado pela tradução, acaba por iniciar-se invariavelmente por um aminoácido de metionina (ao menos em uma de suas extremidades), devido às características do próprio processo. (Alberts 2010; Lehninger, Nelson 2017)

Ao analisarmos, porém, a sequência final de todas proteínas produzidas pelos organismos vivos, podemos observar que apenas uma fração destas contém uma metionina em sua extremidade N-terminal – primeira extremidade a deixar o ribossomo. Esta constatação evidencia a necessidade de uma etapa não desempenhada pelo ribossomo na síntese das proteínas até a maturação destas, a excisão da metionina correspondente ao códon de início. Esta reação ocorre de maneira cotraducional, ou seja, simultaneamente à leitura do mRNA pelo ribossomo. A enzima responsável por catalisar tal reação denomina-se Metionina aminopeptidase (MetAP). (Alberts 2010; Lehninger, Nelson 2017) As MetAPs, assim como o próprio processo de tradução, encontram-se preservadas em todos domínios da vida. (Giglione *et al.* 2004; Lowther *et al.* 2002; Ross *et al*. 2005)

A excisão da metionina N-terminal (EMN) tem impactos conhecidos sobre a meia vida de proteínas maduras. O mecanismo mais bem descrito para determinação da taxa de degradação proteica chama-se “regra do n-final”, no qual o resíduo N-terminal de uma proteína é reconhecido e direcionado por ubiquitina ligases, mediando assim a ubiquitinação e marcando-a para degradação. Através do controle global da meia-vida de proteínas, a EMN demonstrou regular a homeostase global do estado redox de glutationas, ao menos em plantas, leveduras e arqueas. (Frottin *et al.* 2009; Giglione *et al.* 2015).

1.2 OBJETO DE ESTUDO

As MetAPs são metaloproteases dinucleares evolutivamente relacionadas à creatinase, à prolidase e à aminopeptidase P. Todas são pertencentes à família *pita-bread*, também conhecida como “clan MG” ou família M24. Esta família apresenta um eixo de simetria *pseudo-2-fold* com o sítio catalítico e uma ligação a um íon metálico localizados nas interfaces entre os dois domínios (Giglione et al. 2004). Estudos de necessidade de íons metálicos indicam que as MetAPs podem ser ativadas por diversos cátions divalentes, porém, a determinação dos metais com função fisiológica ainda é alvo de debates (Lowther *et al.* 2002; Olaleye *et al.* 2009). Um nível relativamente baixo de similaridade entre suas sequências é compensado por estruturas tridimensionais do sítio ativo similares e pela preservação de um sítio de cinco resíduos específico para ligação a metais (Giglione *et al.* 2015).

A caracterização estrutural e a similaridade de sequências permitiram a classificação das MetAPs em dois tipos; 1 e 2. Enquanto organismos procariotos possuem homólogos de apenas um dos tipos - tipo 1, no caso de bactérias ou 2, no caso de arqueas -, eucariontes apresentam genes homólogos de ambos. Estruturalmente, os tipos de MetAP diferem-se pela presença de dois sub-domínios adicionais de função desconhecida na região C-terminal dos sítios catalíticos das MetAPs do tipo 2 (MetAP2). Adicionalmente, a presença ou ausência de sequências adicionais na região N-terminal divide estas enzimas em subtipos (Giglione *et al.* 2015).

As MetAP do tipo 1 (MetAP1) dividem-se em MetAPs de quatro subtipos (MetAP1a-d), iniciando pelas MetAP1a que apresentam apenas o domínio catalítico conservado em sua constituição. Um exemplo clássico de MetAP1a é a proteína codificada pelo gene pertencente ao genoma da bactéria *Escherichia coli* (*E. coli)*. A seguir, MetAP1b apresentam em sua porção N-terminal uma região de aproximadamente 50 resíduos de aminoácido que liga o domínio catalítico a um domínio *zinc finger*, cuja função supõe-se ser de interação com o ribossomo. Atribui-se a presença de MetAP1b a genomas eucarióticos, em princípio. No caso das MetAP1c, também atribui-se a um domínio N-terminal uma interação com o ribossomo. Neste caso, as MetAP1c apresentam uma cadeia de aproximadamente 40 resíduos de aminoácidos que liga o domínio catalítico a um motivo P-X-X-P de ligação à SH_3, que acredita-se estar envolvido com uma ligação ao ribossomo. Considera-se como modelo para MetAP1c a proteína descoberta em bactérias do filo *Actinobacteria*. Finalmente, as MetAP1d apresentam um peptídeo sinal que é removido durante a translocação destas enzimas para organelas - plastídeos e mitocôndrias - e que, portanto, só estão presentes em eucariotos. (Giglione *et al.* 2015)

Divididas em apenas dois subtipos, as MetAP2 diferenciam-se pela presença ou ausência de um domínio composto por blocos de resíduos poliacídicos e polibásicos formando uma alfa-hélice. Sendo assim, as MetAP2 do subtipo b (MetAP2b) possuem tal inserção adicional na região N-terminal, enquanto as MetAP2 do subtipo a (MetAP2a) não possuem nenhuma inserção adicional nesta região. Além disso, um fator que diferencia ainda mais as MetAP1 de MetAP2 é a presença de uma ou duas inserções adicionais no sítio catalítico no segundo caso. (Giglione *et al.* 2015)

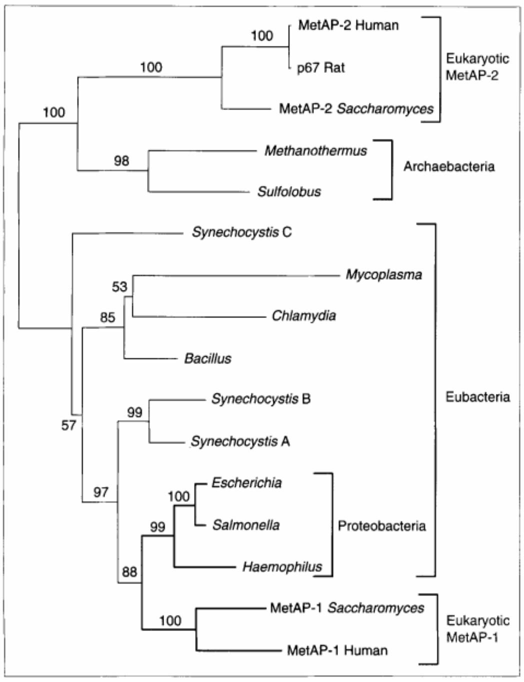


Figura 2 Árvore filogenética proposta por Francisco em 1996. Os tipos de MetAP podem ser observados nos diferentes clusters.

Resultados preliminares do presente trabalho revelam a ocorrência de membros de MetAP2 confirmados em organismos pertencentes à família *Acidobacteria* e indicam a existência de membros de MetAP1 em genomas de organismos pertencentes aos filos *Euryarchaeota* e *Crenarchaeota*. Estes resultados indicam que os dados disponíveis para análise revelam, no mínimo, exceções à regra proposta até então na literatura.

OBJETIVOS

Objetivo Geral

Como objetivo geral, desejamos compreender as origens evolutivas dos tipos e subtipos de Metionina Aminopeptidases, assim como identificar possíveis tipos e subtipos ainda desconhecidos.

Objetivos específicos

Especificamente, estabelecemos os seguintes objetivos:

1. mineração de dados - obtenção de um conjunto significativo de sequências codificantes de genes *MetAP*;

2. armazenamento - construção de um banco de dados local que dê suporte à manipulação do conjunto de dados;

3. prospecção - análise preliminar da distribuição de tipos e subtipos dos genes de *MetAP*. A árvore proposta em 1996 por Francisco se mantém?

4. filtragem - realização de análises e agrupamentos visando a obtenção de um conjunto de dados mínimo que seja representativo do conjunto total;

5. descrição - investigação de sítios e domínios das sequências presentes no conjunto mínimo; e

6. análise filogenética - alinhamento das sequências representativas de todos subgrupos presentes no conjunto mínimo.

CAPÍTULO 1

Artigo a ser submetido ao periódico AMINO ACIDS

Evolutionary analysis of Methionine Aminopeptidases.

Andrey Felipe Schoier, Jose Claudio Fonseca Moreira

Original paper

Evolutionary analysis of Methionine Aminopeptidases.

Schoier A. F.,[1,*] Fonseca Moreira J.C.[1]

[1]Centro de Estudos em Estresse Oxidativo, Departamento de Bioquímica, ICBS, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil.

*Corresponding author: E-mail: andrey@cpd.ufrgs.br

Running title: Evolutionary analysis of Methionine Aminopeptidases.

Keywords: MetAP, NME, Phylogeny, Bioinformatics

Abstract

N-terminal Methionine Excision (NME) is a process that occurs in about 30% to 60% of all expressed proteins in a genome, depending on the organism and cellular compartment. It is conserved through all life domains and it is essential for normal growth and function in any organism investigated. By means of its capability of control over protein half-life, NME has demonstrated to regulate global glutathione redox homeostasis, at least in plants, yeast and Archaea. The enzyme responsible for NME carriage is Methionine Aminopeptidase (MetAP). Like NME, MetAPs are conserved in Bacteria, Archaea and Eukaryotes. It has proven to play a primordial role in human angiogenesis and lymphangiogenesis processes. MetAPs are pharmaceutical targets in a wide range of studies ranging from anti-cancer therapies to zoonose treatments. Structurally, every MetAP variant presents a pseudo twofold symmetry axis of symmetry with the catalytic site and metal ion binding located at the interfaces between the domains. The types and subtypes of MetAP are characterized by the presence or absence of additional insertions in their C-termini and N-termini regions, respectively. The distribution of types of MetAPs along biodiversity is given in a way that, in their genome, Bacteria presents only genes homologous to type 1 *MetAP* (MetAP1), Archaea presents only genes homologous to type 2 *MetAP* (MetAP2) and Eukaryotes presents genes homologous to both types. Additionally, Eukaryotes present MetAP1 variants that are located inside their mitochondria and/or plastids. Despite the presence of both homologous types of *MetAP* in eukaryotes is attributed to endosymbiotic events, little is known about the evolutionary origins of each MetAP variant. In the present study, we analyze the distribution of MetAP types and subtypes across biodiversity, investigating possible candidates for as-yet-unidentified variants, as well as construct a phylogenetic model of MetAPs evolution. Preliminary results indicate the occurrence of MetAP2 members in a bacterial family, which is indicative of the necessity of better description of the *MetAP* types distribution along the tree of life.

Keywords: MetAP, NME, Phylogeny, Bioinformatics.

Introduction

The protein translation initiation step is unleashed when an initiator codon is read by the ribosome. Throughout all domains of life, this codon represents the same instruction, which is interpreted by the ribosome enzymatic machinery as a kickstart for reading the messenger RNA (mRNA) coding sequence. This step is always carried out with the addition of an initiator L-Methionine (Met) amino acid residue in the N-terminus of the nascent polypeptide being synthesized. The protein translation process demands, therefore, a further step consisting of the excision of initiator Met. The NME process is known to be critical in the half-life of mature proteins (Varshavsky 2011; Gibbs *et al.* 2014). The N-termini of proteins starting with an N-alpha-acetylated Met or a free Met may serve as the site of N-ubiquitination by regulation of their steric shielding (Giglione *et al.* 2015). The control of protein half life grants the N-terminal Methionine Excision (NME) process the capability of fine tuning global glutathione redox homeostasis at least in plants, yeast and Archaea (Frottin *et al.* 2009; Giglione *et al.* 2015).

NME process is carried out by Methionine Aminopeptidases (MetAP), which are essential in every life domain (Giglione *et al.* 2004; Lowther *et al.* 2002; Ross *et al*. 2005). MetAPs are dinuclear metalloenzymes evolutionarily close to creatinase, prolidase and aminopep-tidase P, which are all part of the "pita-bread" family, also known as "clan MG" M24 protease family. This family is characterized by a pseudo two-fold axis of symmetry with the catalytic site and metal ion binding located at the interfaces between the domains. It has been shown that the first N-terminal poly charged Lys-rich block of MetAP2b, stores a POEP (protection of eIF2α phosphorylation) activity (Datta, 2000), preventing the phosphorylation of the alpha subunit of eukaryotic initiation factor 2 by interacting with it (Datta, 2000; Datta *et al.* 2003; Datta *et al.* 2006). Human MetAP2 is a target for angiogenesis inhibition (Griffith *et al.* 1997) and it is also known to play a key role in cancer lymphangiogenesis (Esa *et al.* 2020).

MetAP members' type differentiation is determined by the presence or absence of C-termini additional insertions for types 1 and 2, accordingly. Eukaryotes present both MetAP1 and MetAP2 homologous genes, while Prokaryotes carry only one type in their genomes. Archaea are known to show MetAP2 homologous genes, while Bacteria display only MetAP1 members. MetAPs are further differentiated into subtypes, being determined by the presence or absence of N-termini additional insertions in the gene. MetAP1c is associated with bacterial Actinobacteria phylum and it contains a SH_3 binding domain in its N-terminal region. The eukaryotic MetAP1b contains a zinc finger domain in this region, proposed to be responsible of interaction with the ribosome (Vetro, Chang 2002). Both SH_3 binding domain and zinc finger domain are spatially positioned on the same region. An additional described variant is present in the genome of *Streptococci* and *Lactobacilli*, with two additional characteristic insertions. The structure of *S. pneumonae* MetAP1a' variant protrudes this insertion in the same region of the other MetAP types and subtypes, which is indicative of selective pressure of interfaces proposed to interact with the ribosome. The aforementioned structural observations can be noted in Figure 1. Finally, MetAP1d eukaryotic subtype contains a signal peptide and it occurs in eukaryotic mitochondria and plastids. The only insertion shown in MetAP structures that occupies a different spot is MetAP2b N-terminal portion, known to be involved in other activities than the initiator Met excision. MetAP2a subtype members display a three helix structure in its selectively-favored region.

Despite being used as a molecular marker in several evolutionary studies (Vavilova *et al.* 2015; Morgante *et al.* 2009; Zhang *et al.* 2005; Cao *et al.* 2020) and presenting some of its phylogeny resolved (Pandrea *et al.* 2005), little is known about the global distribution of MetAPs across biodiversity and the most diverse taxonomy presented to date is still the one shown by Francisco in 1996. Nowadays, the number of sequences available for analysis may reveal a different perspective. Furthermore, the origins of each type and subtype remains unknown. The importance of MetAP in several different areas of interest, ranging from drug targets in anti-cancer and zoonose treatments to analysis in glutathione redox states (Munkhjargal *et al.* 2016; Chiu *et al*. 2014; Lin *et al.* 2018), and the fact that unveiling the origins of eukaryotic MetAPs might shed light on endosymbiotic events justifies the present work. More than 59 thousands aminoacidic sequences have been retrieved and the results are shown below.

Materials and Methods

As shown in Figure 2, a query was composed in order to retrieve all members belonging to M24A subfamily, which is composed by all MetAP members (more information available at https://www.ebi.ac.uk/merops/cgi-bin/famsum?family=M24), from the UniProtKB database in XML format. The data retrieved from the file was processed and stored in a local MySQL database (Widenius *et al.* 2002). Throughout web scraping using the BeautifulSoup python library (Richardson 2016), HTTP requests were made to EBI (Kanz *et al.* 2005), NCBI (NCBI Resource Coordinators 2018) and DDBJ (Fukuda *et al.* 2021) in order to obtain the respective coding DNA sequences associated with UniProtKB data.

A composition tree was built and populated according to NCBI Taxonomy reference (Schoch *et al.* 2020). Sci-kitBio NodeTree (scikit-bio development team 2020) was used to create trees. The procedure described in Figure 3 is proposed in order to extract clusters of optimal size. A number of 72 clusters were extracted, with sizes ranging from 2,905 to 102 members each and they are all presented in Table 2.

Every cluster extracted in the aforementioned step was considered as a new starting subgroup, a Multiple Sequence Alignment (MSA) was performed with the MAFFT software (Rozewicki *et al.* 2019) using the ‘auto’ flag with each subgroup and a tree is currently being built with the IQ-TREE software (Chernomor *et al.* 2016) using default model-finder resource for each and every subgroup. The procedure described in Figure 3 is applied again in the resulting trees. A method for data filtering is proposed in Figure 4. Every subcluster extracted in the latter step was subjected to the described filtering process. Data sampling applied on the filtering process is further detailed in Figure 5.

An overall representation of the dataset composition is presented in Figure 6. A detailed analysis of each interest group is presented in Suppl. Figures 1-4.

A preliminary tree was built with cherry picked groups of interest and it is presented in Figure 11. We tried to group as many *Eukaryotes* as possible, focusing on unicellular organisms - like *Sar* members -, algae that possess plastids and bacterial clades with hypothetical involvement with the endosymbiosis hypothesis. *Archaea* members were included looking for possible MetAP2 eukaryotic origins. In this step, Clustal Omega (Madeira *et a.l* 2019) was utilized to perform MSA and BIONJ (Gascuel 1997) was used to build the tree with the neighbor-joining method.

PDB files retrieved manually from the PDB database were submitted to EBI PDBeFold SSM (Krissinel *et al.* 2004) and the Q-Scores value matrix is presented at Table 1.

Results and Discussion

Figure 8 shows overall dataset composition. A total of 59,203 aminoacidic sequences was obtained and stored locally, most of it belonging to bacterial members. This result was expected since the vast majority of sequenced genes has bacterial origins. Even though most prokaryotic genomes present only one *MetAP* homologous gene, there are many cases of clades of *Bacteria* containing two or more paralogous sequences. The expression of two different paralogous genes of *Bacillus subtilis* was confirmed (You *et al.* 2005). We have identified occurrences of up to twelve paralogous genes in the same bacterial genome in our dataset, as it is the case of *Myxococcus xanthus*, even though it is not possible to tell if every sequence is valid.

Figure 9 shows the MetAP tree of *Acidobacteria* as an example of application of the methods proposed. The insertion event detection has pointed out a specific C-terminal insertion in the cluster colored in cyan. This cluster was closely analyzed and the MSA build with the MAFFT software (Rozewicki *et al.* 2019) is presented at Figure 10, whereas every sequence denoted was automatically categorized as a MetAP2 member. This is the first time a group of bacterial MetAP members is described as MetAP2 and the presence of type 2 homologs in a *Bacteria* clade is indicative of the necessity of closer review in the MetAP types distribution across biodiversity. Another example of data extraction is given in Suppl. Figure 1.

Further analysis of Suppl. Figures 2-5 helped us to compose the group used to create Figure 11. Interest clades were fitted into a unique group and the colored Trees in Figure 11 a and b helps us to wonder the evolution of MetAPs and each of its variants. A closer look into the upper left corner of the presented topology reveals one particular cluster of interest. There, we find *Actinobacteria*, *Cyanobacteria, Sar, Viridiplantae*, *Opisthokonta*, *Haptista*, *Amoebozoa* and *Rhodophyta* members clustered together. This group is indicative of close shared evolutionary relations between the most basal known *Eukaryota* and *Bacteria*, especially in eukaryotic groups known to present plastids and *Cyanobacteria* members. This is an indication that the complete solving of MetAP evolution history might shed light on endosymbiotic events. Knowing which bacterial family fits the most in the role of eukaryotic MetAP1d variants' ancestor is of key importance, since factors like the overexpression of MetAP1d in colon cancer demonstrated by Leszczyniecka *et al.* in 2006 may be affected by this knowledge.

The split in *Actinobacteria* members was expected, since its members are known to present MetAP1a and MetAP1c paralog genes, but the clustering of a subgroup of this bacterial phylum within eukaryotic populated cluster is novel and may indicate closer relations between MetAP1c and eukaryotic MetAP variants.

One last interesting fact to note is the clustering of *Archeae* members with *Actinobacteria* and *Cyanobacteria* members. It is most likely an indicative of horizontal transfer of genes, among those species, revealing the possibility of MetAP1 members in *Archeae*.

A good candidate for rooting the tree would be the sequences obtained from deep-sea hydrothermal vents by Elsaied *et al.* in 2007, since these organisms are hypothesized as related to life origins (Martin *et al.* 2008).

The presented results are enough to justify the revision of the distribution of *MetAP* homologs along biodiversity. We have pointed out confirmed MetAP2 members in *Acidobacteria* and there is a possibility of the occurrence of MetAP1 members in *Archeae*, defying the first MetAP tree proposed by Francisco in 1996. Categorizing and filtering the whole dataset remains a challenge, though.

Perspectives

In ongoing work, we are finishing the proposed pipeline in Figure 2 and will soon obtain a minimal representative dataset of MetAP members, which could be defined as the minimum group containing enough sequences to represent the different profiles of *MetAP* genes contained in the original mined data. After solving MetAP evolution, we will take a closer look into the existence of not yet described variants, taking into account the insertion events detected, and propose structural models for each variant with AlphaFold2 (Jumper *et al.* 2010). Structural positioning of insertions in the same observed sites noted to present selective pressure is expected. A good example of comparison can be observed in Figures 5 and 6, where the *Vibrio cholerae* member presents an additional insertion that might characterize a novel variant. The MetAP2 variant present in *Acidobacteria* is a good indicator of the presence of yet undescribed MetAP variants.

The creation of a theoretically sustained model for MetAP evolution is expected to be published soon.

References

Kanz C., Aldebert P., Althorpe N., Baker W., Baldwin A., Bates K., et al. . (2005). The EMBL Nucleotide Sequence Database. Nucleic Acids Res. 33 (Database issue), D29–D33. doi: 10.1093/nar/gki098

NCBI Resource Coordinators (2018). Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, *46*(D1), D8–D13. https://doi.org/10.1093/nar/gkx1095

Fukuda, A., Kodama, Y., Mashima, J., Fujisawa, T., & Ogasawara, O. (2021). DDBJ update: streamlining submission and access of human data. *Nucleic acids research*, *49*(D1), D71–D75. https://doi.org/10.1093/nar/gkaa982

Eddy S. R. (1998). Profile hidden Markov models. *Bioinformatics (Oxford, England)*, *14*(9), 755–763. https://doi.org/10.1093/bioinformatics/14.9.755

Di Franco, A., Poujol, R., Baurain, D., & Philippe, H. (2019). Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. BMC Evolutionary Biology, 19(1), 21. https://doi.org/10.1186/s12862-019-1350-2

Rozewicki, J., Li, S., Amada, K. M., Standley, D. M., & Katoh, K. (2019). MAFFT-DASH: integrated protein sequence and structural alignment. Nucleic Acids Research, 47(W1), W5–W10. https://doi.org/10.1093/nar/gkz342

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., & Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. Bioinformatics, 23(21), 2947–2948. https://doi.org/10.1093/bioinformatics/btm404

Bonissone, S., Gupta, N., Romine, M., Bradshaw, R. A., & Pevzner, P. A. (2013). N-terminal protein processing: A comparative proteogenomic analysis. *Molecular and Cellular Proteomics*, *12*(1), 14–28. https://doi.org/10.1074/mcp.M112.019075

Bradshaw, R. A., Brickey, W. W., & Walker, K. W. (1998). N-terminal processing: The methionine aminopeptidase and N(α)-acetyl transferase families. *Trends in Biochemical Sciences*, *23*(7), 263–267. https://doi.org/10.1016/S0968-0004(98)01227-4

Krissinel, E. & Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr. D. Biol. Crystallogr. 60, 2256–2268 (2004).

Cao, Y., Trivellone, V., & Dietrich, C. H. (2020). A timetree for phytoplasmas (Mollicutes) with new insights on patterns of evolution and diversification. *Molecular Phylogenetics and Evolution*, *149*, 106826. https://doi.org/10.1016/j.ympev.2020.106826

Chernomor, O., von Haeseler, A., & Minh, B. Q. (2016). Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. Systematic Biology, 65(6), 997–1008. https://doi.org/10.1093/sysbio/syw037

Lowther, W. T., & Matthews, B. W. (2002). Metalloaminopeptidases: common functional themes in disparate structural surroundings. *Chemical reviews*, *102*(12), 4581–4608. https://doi.org/10.1021/cr0101757

Ross, S., Giglione, C., Pierre, M., Espagne, C., & Meinnel, T. (2005). Functional and developmental impact of cytosolic protein N-terminal methionine excision in Arabidopsis. Plant physiology, 137(2), 623–637. https://doi.org/10.1104/pp.104.056861

Griffith, E. C., Su, Z., Turk, B. E., Chen, S., Chang, Y. H., Wu, Z., Biemann, K., & Liu, J. O. (1997). Methionine aminopeptidase (type 2) is the common target for angiogenesis inhibitors AGM-1470 and ovalicin. Chemistry & biology, 4(6), 461–471. https://doi.org/10.1016/s1074-5521(97)90198-8

Datta, B. (2000). MAPs and POEP of the roads from prokaryotic to eukaryotic kingdoms. Biochimie, 82(2), 95–107. https://doi.org/10.1016/s0300-9084(00)00383-7

Datta, R., Tammali, R., & Datta, B. (2003). Negative regulation of the protection of eIF2alpha phosphorylation activity by a unique acidic domain present at the N-terminus of p67. *Experimental cell research*, *283*(2), 237–246. https://doi.org/10.1016/s0014-4827(02)00042-3

Datta, B., Datta, R., Ghosh, A., & Majumdar, A. (2006). The binding between p67 and eukaryotic initiation factor 2 plays important roles in the protection of eIF2alpha from phosphorylation by kinases. *Archives of biochemistry and biophysics*, *452*(2), 138–148. https://doi.org/10.1016/j.abb.2006.06.009

Vetro, J. A., & Chang, Y. H. (2002). Yeast methionine aminopeptidase type 1 is ribosome-associated and requires its N-terminal zinc finger domain for normal function in vivo. *Journal of cellular biochemistry*, *85*(4), 678–688. https://doi.org/10.1002/jcb.10161

Chiu, J., Wong, J. W. H., & Hogg, P. J. (2014). Redox regulation of Methionine aminopeptidase 2 activity. *Journal of Biological Chemistry*, *289*(21), 15035–15043. https://doi.org/10.1074/jbc.M114.554253

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., … De Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/BIOINFORMATICS/BTP163

D480-D489. (2021). UniProt: the universal protein knowledgebase in 2021 The UniProt Consortium. *Nucleic Acids Research*, *49*. https://doi.org/10.1093/nar/gkaa1100

Elsaied, H., Stokes, H. W., Nakamura, T., Kitamura, K., Fuse, H., & Maruyama, A. (2007). Novel and diverse integron integrase genes and integron-like gene cassettes are prevalent in deep-sea hydrothermal vents. *Environmental Microbiology*, *9*(9), 2298–2312. https://doi.org/10.1111/j.1462-2920.2007.01344.x

Esa, R., Steinberg, E., Dror, D., Schwob, O., Khajavi, M., Maoz, M., … Benny, O. (2020). The role of methionine aminopeptidase 2 in lymphangiogenesis. *International Journal of Molecular Sciences*, *21*(14), 1–15. https://doi.org/10.3390/ijms21145148

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution, 4 4*, 406-25 .

Fedor. (2020). Supervenn. https://doi.org/10.5281/ZENODO.4012442

Gascuel, O. (1997). BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, *14*(7), 685–695. https://doi.org/10.1093/oxfordjournals.molbev.a025808

Giglione, C., Boularot, A., & Meinnel, T. (2004). Protein N-terminal methionine excision. *Cellular and Molecular Life Sciences*, *61*(12), 1455–1474. https://doi.org/10.1007/s00018-004-3466-8

Giglione, C., Fieulaine, S., & Meinnel, T. (2015, June 3). N-terminal protein modifications: Bringing back into play the ribosome. *Biochimie*. Elsevier B.V. https://doi.org/10.1016/j.biochi.2014.11.008

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, *33*(6), 1635–1638. https://doi.org/10.1093/MOLBEV/MSW046

Varshavsky, A. (2011). The N-end rule pathway and regulation by proteolysis. Protein Science : A Publication of the Protein Society, 20(8), 1298–1345. https://doi.org/10.1002/pro.666

Gibbs, D. J., Bacardit, J., Bachmair, A., & Holdsworth, M. J. (2014). The eukaryotic N-end rule pathway: conserved mechanisms and diverse functions. *Trends in cell biology*, *24*(10), 603–611. https://doi.org/10.1016/j.tcb.2014.05.001

Frottin, F., Espagne, C., Traverso, J. A., Mauve, C., Valot, B., Lelarge-Trouverie, C., Zivy, M., Noctor, G., Meinnel, T., & Giglione, C. (2009). Cotranslational proteolysis dominates glutathione homeostasis to support proper growth and development. *The Plant cell*, *21*(10), 3296–3314. https://doi.org/10.1105/tpc.109.069757

Langa, Ł. (2020). *PEP 596 -- Python 3.9 Release Schedule*. Retrieved from https://www.python.org/dev/peps/pep-0596/

Leszczyniecka, M., Bhatia, U., Cueto, M., Nirmala, N. R., Towbin, H., Vattay, A., … Phillips, P. E. (2006). MAP1D, a novel methionine aminopeptidase family member is overexpressed in colon cancer. *Oncogene*, *25*(24), 3471–3478. https://doi.org/10.1038/sj.onc.1209383

Lin, M., Zhang, X., Jia, B., & Guan, S. (2018). Suppression of glioblastoma growth and angiogenesis through molecular targeting of methionine aminopeptidase-2. *Journal of Neuro-Oncology*, *136*(2), 243–254. https://doi.org/10.1007/s11060-017-2663-x

Martin, F., & Lopez, M. C. (1996). Methionine aminopeptidase-I: the, *0004*(96), 285–286.

Morgante, C. V., Rodrigues, R. A. O., Marbach, P. A. S., Borgonovi, C. M., Moura, D. S., & Silva-Filho, M. C. (2009). Conservation of dual-targeted proteins in Arabidopsis and rice points to a similar pattern of gene-family evolution. *Molecular Genetics and Genomics*, *281*(5), 525–538. https://doi.org/10.1007/s00438-009-0429-7

Munkhjargal, T., Ishizaki, T., Guswanto, A., Takemae, H., Yokoyama, N., & Igarashi, I. (2016). Molecular and biochemical characterization of methionine aminopeptidase of Babesia bovis as a potent drug target. *Veterinary Parasitology*, *221*, 14–23. https://doi.org/10.1016/j.vetpar.2016.02.024

Omega, C. (n.d.). Resource Summary Report Clustal Omega RRID:SCR_001591 Type: Tool Proper Citation.

Olaleye, O. A., Bishai, W. R., & Liu, J. O. (2009). Targeting the role of N-terminal methionine processing enzymes in Mycobacterium tuberculosis. *Tuberculosis (Edinburgh, Scotland)*, *89 Suppl 1*, S55–S59. https://doi.org/10.1016/S1472-9792(09)70013-7

Pandrea, I., Mittleider, D., Brindley, P. J., Didier, E. S., & Robertson, D. L. (2005). Phylogenetic relationships of methionine aminopeptidase 2 among Encephalitozoon species and genotypes of microsporidia. *Molecular and Biochemical Parasitology*, *140*(2), 141–152. https://doi.org/10.1016/j.molbiopara.2004.12.006

Peterson, B. (2019). PEP 373 -- Python 2.7 Release Schedule | Python.org. Retrieved November 19, 2021, from https://www.python.org/dev/peps/pep-0373/

Richardson, L. (2016). Beautiful Soup Documentation. *Media.Readthedocs.Org*. Retrieved from https://media.readthedocs.org/pdf/beautiful-soup-4/latest/beautiful-soup-4.pdf%0Ahttp://www.crummy.com/software/BeautifulSoup/bs4/doc/

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., … Karsch-Mizrachi, I. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database : The Journal of Biological Databases and Curation*, *2020*. https://doi.org/10.1093/DATABASE/BAAA062

scikit-bio development team, T. (2020). scikit-bio: A Bioinformatics Library for Data Scientists, Students, and Developers. Retrieved from http://scikit-bio.org

Serero, A., Giglione, C., Sardini, A., Martinez-Sanz, J., & Meinnel, T. (2003). An Unusual Peptide Deformylase Features in the Human Mitochondrial N-terminal Methionine Excision Pathway. *Journal of Biological Chemistry*, *278*(52), 52953–52963. https://doi.org/10.1074/jbc.M309770200

Varshavsky, A. (1997). The N-end rule pathway of protein degradation. *Genes to Cells*, *2*(1), 13–28. https://doi.org/10.1046/j.1365-2443.1997.1020301.x

Vavilova, V., Sormacheva, I., Woyciechowski, M., Eremeeva, N., Fet, V., Strachecka, A., … Blinov, A. (2015). Distribution and diversity of Nosema bombi (Microsporidia: Nosematidae) in the natural populations of bumblebees (Bombus spp.) from West Siberia. *Parasitology Research*, *114*(9), 3373–3383. https://doi.org/10.1007/s00436-015-4562-4

Widenius, M., Axmark, D., & DuBois, P. (2002). Mysql Reference Manual.

You, C. H., Lu, H. Y., Sekowska, A., Fang, G., Wang, Y. P., Gilles, A. M., & Danchin, A. (2005). The two authentic methionine aminopeptidase genes are differentially expressed in Bacillus subtilis. *BMC Microbiology*, *5*, 1–15. https://doi.org/10.1186/1471-2180-5-57

Zhang, H., Huang, H., Cali, A., Takvorian, P. M., Feng, X., Zhou, G., & Weiss, L. M. (2005). Investigations into microsporidian methionine aminopeptidase type

2: A therapeutic target for microsporidiosis. *Folia Parasitologica*, *52*(1–2), 182–192. https://doi.org/10.14411/fp.2005.023

Caswell, T. A., Droettboom, M., Lee, A., Andrade, E. S. de, Hoffmann, T., Hunter, J., Klymak, J., Firing, E., Stansby, D., Varoquaux, N., Nielsen, J. H., Root, B., May, R., Elson, P., Seppänen, J. K., Dale, D., Lee, J.-J., McDougall, D., Straw, A., … Ivanov, P. (2021). matplotlib/matplotlib: REL: v3.5.0. https://doi.org/10.5281/ZENODO.5706396

Schrödinger (2016) The PyMOL molecular graphics system, version 1.7,6. Schrödinger, LLC

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Martin, W., Baross, J., Kelley, D., & Russell, M. J. (2008). Hydrothermal vents and the origin of life. Nature Reviews Microbiology, 6(11), 805–814. https://doi.org/10.1038/nrmicro1991

Figures

Figure 1. Structures of MetAP types and subtypes. **a** *Escherichia coli* MetAP1a (P0AE18); **b** *Homo sapiens* MetAP1b (P53582); **c** *Mycobacterium tuberculosis* MetAP1c (P9WK19); **d** *Staphylococcus pneumoniae* MetAP1a' (B2IQ22); **e** *Pyrococcus furiosus* MetAP2a (P56218) and **f** *Homo sapiens* MetAP2b (P50579).The pseudo two-fold axis can be noted preserved among MetAP variants.

Figure 2. Proposed pipeline. A MySQL database (Widenius *et al.* 2002) instance was used as local storage.

Figure 3. Flowchart of algorithm proposed for cluster extraction in preliminary trees. A descending and an ascending loop can be noted. As a result, the whole tree is covered and clusters with sizes within the lower limit and the upper limit are extracted.

Figure 4. Data Filtering process. A MSA is performed by MAFFT software (Rozewicki *et al.* 2019) with the given cluster. A Hidden Markov Model profile (Eddy, 1998) is built utilizing the hmmbuild function in HMMER$_{3.3}$ tools. As proposed by Di Franco *et al*. 2019, segments detected with HmmCleaner in regions with ≥70% of gaps are considered as linked to insertion events. Such insertions are stored in the local database. Sequences containing insertions are grouped together and are submitted to the data sampling method proposed in Figure 4 with a reduction factor of 0.3. The rest of the sequences are submitted to the data sampling method proposed in Figure 4 with a reduction factor of 0.08. The samples extracted are considered as candidate representative sequences.

Figure 5. Data sampling method. A group of sequences containing *n* members is rearranged into a group of *m* NCBI Taxonomy families in ascending order with regard to their number of sequences. A group of candidate representative sequences is populated, starting with a sequence extracted from the first NCBI Taxonomy family and following a loop over the *m* families until the number of candidate representative sequences meets the requirement set by the reduction factor. This method benefits families with smaller sizes and guarantees they will have most of their members extracted as candidate representative sequences.

Figure 6. Alignment of the different known types and subtypes of MetAPs. Additional insertions are marked in different colors in the lower bar. Sequence names of proteins which PDB structure is resolved are highlighted accordingly with the matching color in Figure 7. Sequence UniProtKB accession numbers are the following: *E. coli* MetAP1a: P0AE18, *A. baumanmii* MetAP1a: V5VCW7, *P. aeruginosa* MetAP1a: Q9HXY1, *P. maritima* MetAP1a: Q9X1I7, *R. prowazekii* MetAP1a: Q9ZCD3, *S. aureus* MetAP1a: P0A078, *V. cholerae* MetAP1a: Q9KPV1, *M. abscessus*: A0A7Y4J8D5_MYXXA, *S. pneumoniae* MetAP1a': B2IQ22, *H. sapiens* MetAP1b: P53582, *T. brucei brucei* MetAP1b: Q4FKC0, *P. falciparium* MetAP1b: Q8IJP2, *M. tuberculosis* MetAP1c: P9WK19, *P. furiosus* MetAP2a: P56218, *E. cuniculi* MetAP2b: Q8SR45, *H. sapiens* MetAP2b: P50579. Plot made with ClustalX (Larkin *et al*. 2007).

Figure 7. Structural tree of a group of solved crystallographically solved MetAP members. EBI PDBeFold SSM (Krissinel *et al.* 2004) multiple alignment tool was used to align 3D models and Q-scores results are presented in Table 1. For each pairwise comparison, (1 - Q-score) was used as reference for distance. The Neighbourhood-joining method was carried out by BIONJ (Gascuel 1997) with the distance matrix as input. The pseudo two-fold axis can be clearly noted in the cartoon representations of the MetAP members. A tendency of spatial positioning in additional insertions is noticed when analyzing the upper part of the presented view. The human variant of MetAP2b presents a N-terminal extension behind the presented view, colored in green. The colors of additional insertions are the same presented in Figure 6. RCSB PDB accession numbers are the following: *E. coli* MetAP1a: 2GG2, *A. baumanmii* MetAP1a: 6MRF, *P. aeruginosa* MetAP1a: 4FO7, *P. maritima* MetAP1a: 1O0X, *R. prowazekii* MetAP1a: 3MX6, *S. aureus* MetAP1a: 1QXY, *V. cholerae* MetAP1a: 6LH7, *M. abscessus:* 3TAV, *S. pneumoniae* MetAP1a': 4KM3, *H. sapiens* MetAP1b: 2B3H, *T. brucei brucei* MetAP1b: 4FUK, *P. falciparium* MetAP1b: 3S6B, *M. tuberculosis* MetAP1c: 3PKA, *P. furiosus* MetAP2a: 1XGS, *E. cuniculi* MetAP2b: 3FM3, *H. sapiens* MetAP2b: 5D6E. Molecules were rendered using the PyMol (Schrödinger 2016) community version.

Figure 8. Dataset composition according to taxonomic groups. Only groups with more than 100 members are shown. The Supervenn library (Fedor 2020) was used in the plot.

Figure 9. *Acidobacteria* MetAP tree containing 624 aminoacidic sequences. Clusters detected applying the procedure described in Figure 2 are highlighted in different colors. A total of 45 members with detected insertions (names in blue) and 23 additional random members (names in black) were considered representative sequences and can be visualized along the tree. The sheer amount of representative sequences extracted from the tree applying the procedure described in Figure 3 accounts for 10,9% of the original mined sequences belonging to that group. The ETE$_3$ toolkit (Huerta-Cepas *et al.* 2016) was used in the plot.

Figure 10. Multiple sequence alignment performed on the cluster highlighted in cyan on Figure 9, extracted from *Acidobacteria* MetAP tree. C-terminal insertions are noted in sequences of UniProtKB Accession numbers: A0A2V7V6R6, A0A2V7XBC8, A0A2V7Y3I3 and A0A2V7VTB7, all of which contains the MAP_2 PROSITE (Sigrist *et al.* 2013) or the met_pdase_II TIGRFAMs (Li *et al.* 2021) reference. The latter reference is available at https://www.ncbi.nlm.nih.gov/genome/annotation_prok/evidence/TIGR00501/ and the first is available at https://prosite.expasy.org/doc/PS01202. Plot made with ClustalX (Larkin *et al*. 2007).

Figure 11. Preliminary tree with prokaryotic phyla and eukaryotic groups of interest cherry picked analysing coding sequences data composition. Setting every branch to 1 facilitates topology analysis and is presented in a. Preserved tree branches according to neighbour-joining are shown in b. Clustering *Cyanobacteria* and eukaryotic members, especially from the *Viridiplantae* kingdom and the *Sar* supergroup can be noted. Clustal Omega (Madeira *et a.l* 2019) was used in MSA. Tree constructed with BIONJ (Gascuel 1997) using neighbour-joining method. The ETE$_3$ toolkit (Huerta-Cepas *et al.* 2016) was used in the plot.

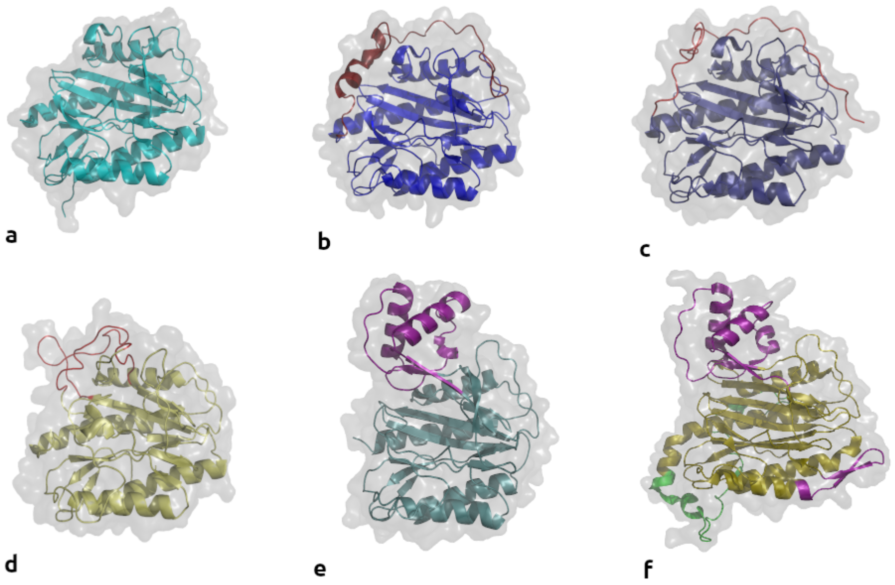


Figure 1. Structures of MetAP types and subtypes. **a** *Escherichia coli* MetAP1a (P0AE18); **b** *Homo sapiens* MetAP1b (P53582); **c** *Mycobacterium tuberculosis* MetAP1c (P9WK19); **d** *Staphylococcus pneumoniae* MetAP1a' (B2IQ22); **e** *Pyrococcus furiosus* MetAP2a (P56218) and **f** *Homo sapiens* MetAP2b (P50579).The pseudo two-fold axis can be noted preserved among MetAP variants.

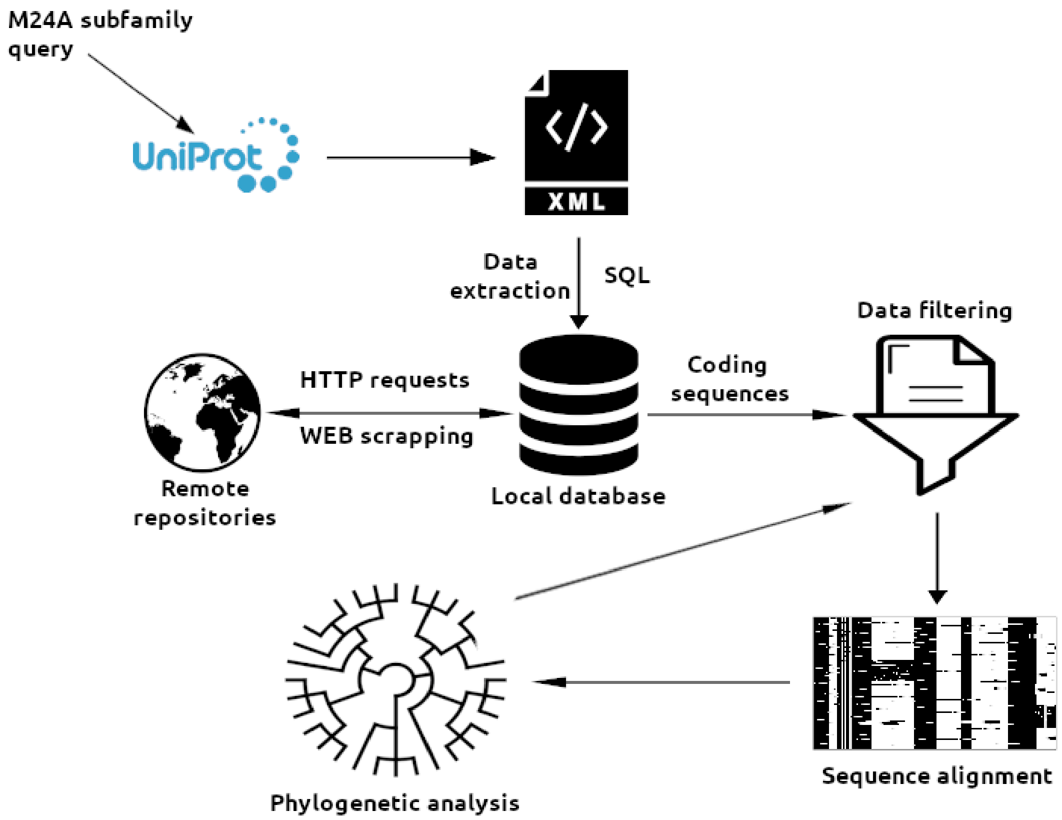


Figure 2. Proposed pipeline.

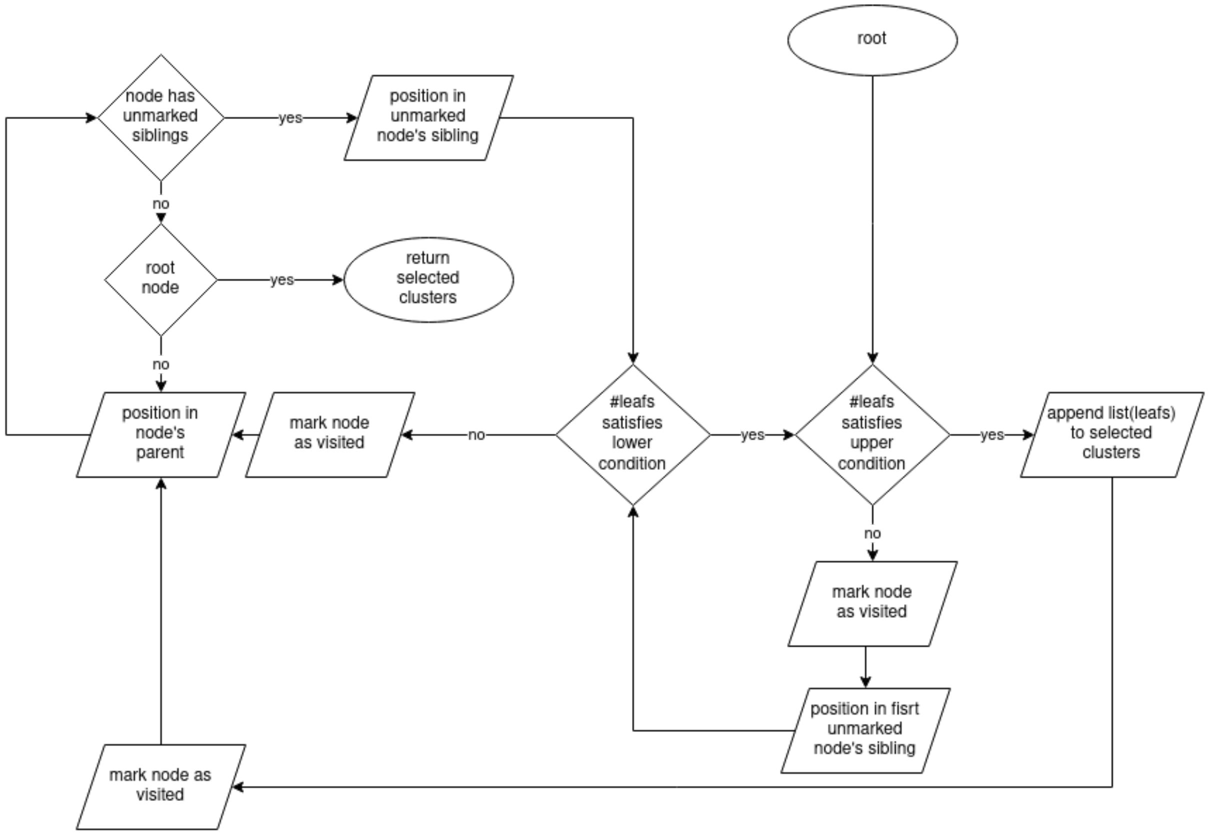

root
node has unmarked siblings
yes
position in unmarked node's sibling
no
root node
yes
return selected clusters
no
position in node's parent
mark node as visited
no
#leafs satisfies lower condition
yes
#leafs satisfies upper condition
yes
append list(leafs) to selected clusters
no
mark node as visited
position in fisrt unmarked node's sibling
mark node as visited

Figure 3. Flowchart of algorithm proposed for cluster extraction in preliminary trees. A descending and an ascending loop can be noted. As a result, the whole tree is

covered and clusters with sizes within the lower limit and the upper limit are extracted.

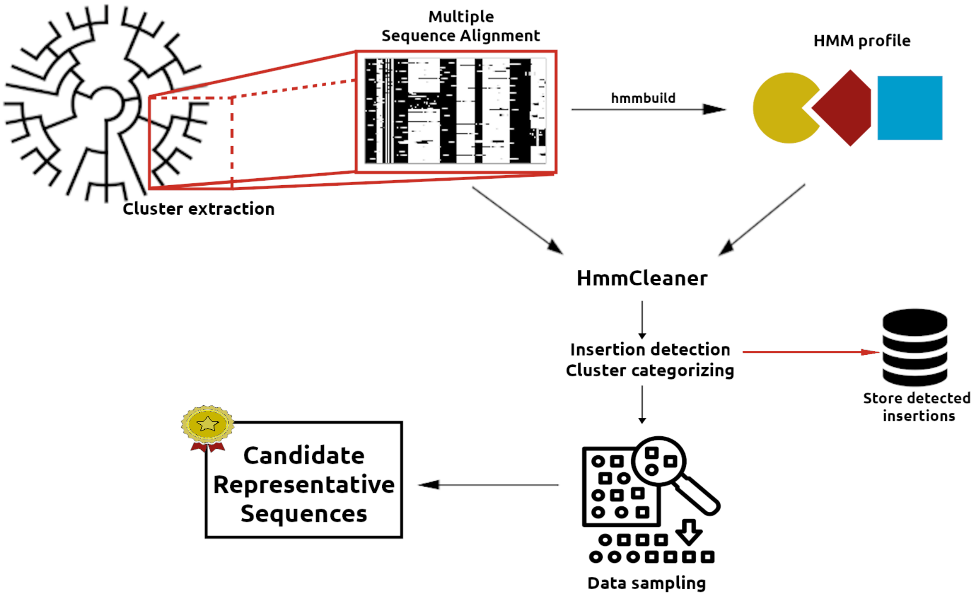


Figure 4. Data Filtering process. A MSA is performed by MAFFT software (Rozewicki *et al.* 2019), using the ‘auto’ flag, with the given cluster. A Hidden Markov Model profile (Eddy, 1998) is built utilizing the hmmbuild function in HMMER$_{3.3}$ tools. As proposed by Di Franco *et al*. 2019, segments detected with HmmCleaner in regions with ≥70% of gaps are considered as linked to insertion events. Such insertions are stored in the local database. Sequences containing insertions are grouped together and are submitted to the data sampling method proposed in Figure 4 with a reduction factor of 0.3. The rest of the sequences are submitted to the data sampling method proposed in Figure 4 with a reduction factor of 0.08. The samples extracted are considered as candidate representative sequences.

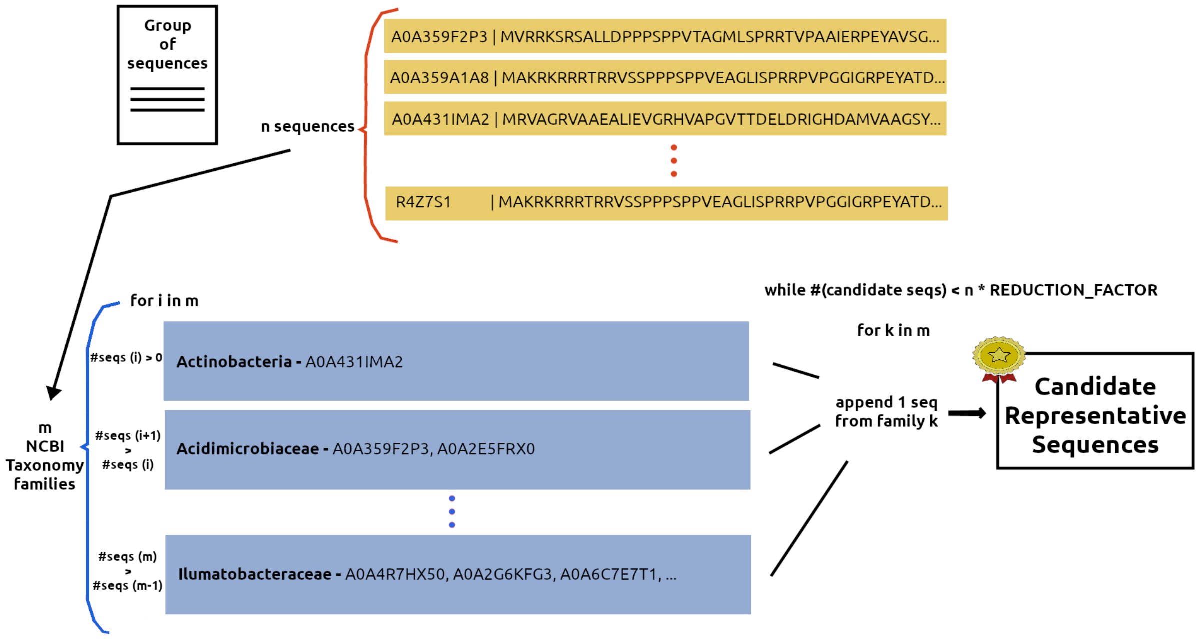

Group
of
sequences
n sequences
A0A359F2P3 | MVRRKSRSALLDPPPSPPVTAGMLSPRRTVPAAIERPEYAVSG...
A0A359A1A8 | MAKRKRRRTRRVSSPPPSPPVEAGLISPRRPVPGGIGRPEYATD...
A0A431IMA2 | MRVAGRVAAEALIEVGRHVAPGVTTDELDRIGHDAMVAAGSY...
R4Z7S1 | MAKRKRRRTRRVSSPPPSPPVEAGLISPRRPVPGGIGRPEYATD...
while #(candidate seqs) < n * REDUCTION_FACTOR
for i in m
for k in m
m
NCBI
Taxonomy
families
#seqs (i) > 0
Actinobacteria - A0A431IMA2
#seqs (i+1)
>
#seqs (i)
Acidimicrobiaceae - A0A359F2P3, A0A2E5FRX0
#seqs (m)
>
#seqs (m-1)
Ilumatobacteraceae - A0A4R7HX50, A0A2G6KFG3, A0A6C7E7T1, ...
append 1 seq
from family k
Candidate
Representative
Sequences

Figure 5. Data sampling method. A group of sequences containing *n* members is rearranged into a group of *m* NCBI Taxonomy families in ascending order with regard to their number of sequences. A group of candidate representative sequences is populated, starting with a sequence extracted from the first NCBI Taxonomy family and following a loop over the *m* families until the number of candidate representative sequences meets the requirement set by the reduction factor. This method benefits families with smaller sizes and guarantees they will have most of their members extracted as candidate representative sequences.

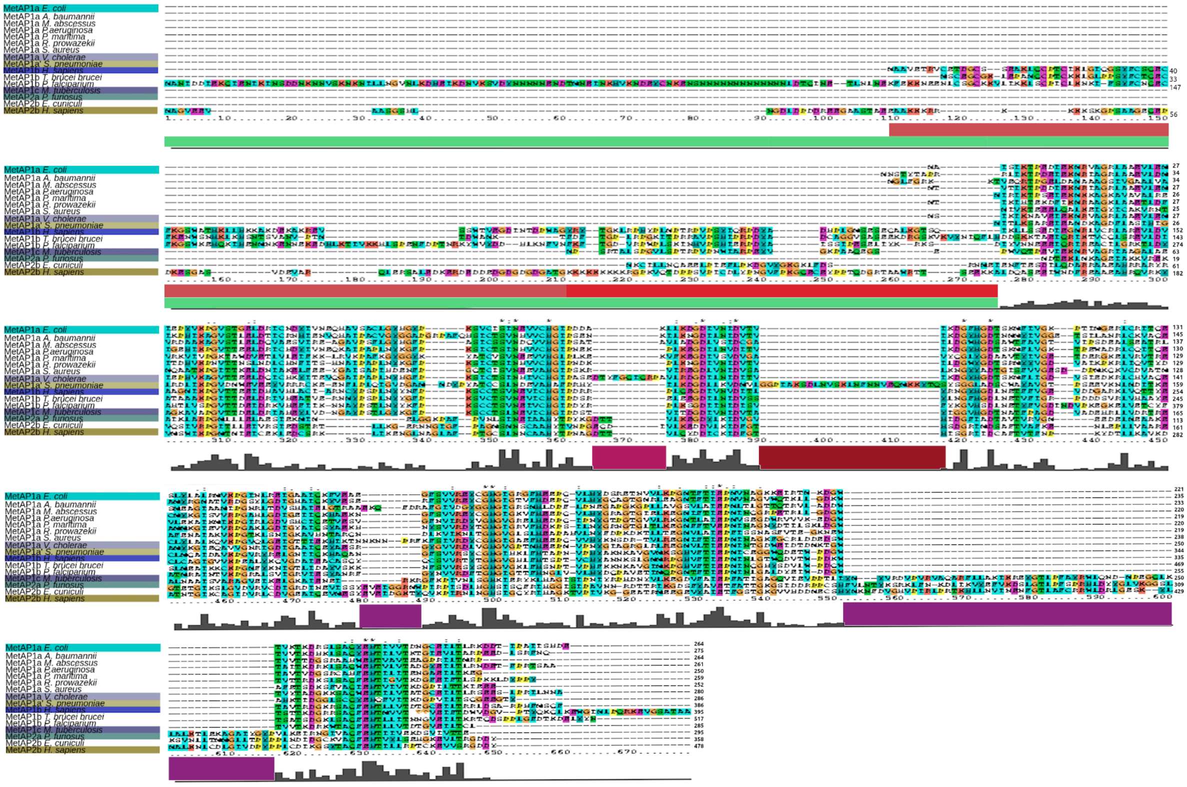


Figure 6. Alignment of the different known types and subtypes of MetAPs. Additional insertions are marked in different colors in the lower bar. Sequence names of proteins which PDB structure is resolved are highlighted accordingly with the matching color in Figure 7. Sequence UniProtKB accession numbers are the following: *E. coli* MetAP1a: P0AE18, *A. baumanmii* MetAP1a: V5VCW7, *P. aeruginosa* MetAP1a: Q9HXY1, *P. maritima* MetAP1a: Q9X1I7, *R. prowazekii* MetAP1a: Q9ZCD3, *S. aureus* MetAP1a: P0A078, *V. cholerae* MetAP1a: Q9KPV1, *M. abscessus*: A0A7Y4J8D5_MYXXA, *S. pneumoniae* MetAP1a': B2IQ22, *H. sapiens* MetAP1b: P53582, *T. brucei brucei* MetAP1b: Q4FKC0, *P. falciparium* MetAP1b: Q8IJP2, *M. tuberculosis* MetAP1c: P9WK19, *P. furiosus* MetAP2a: P56218, *E. cuniculi* MetAP2b: Q8SR45, *H. sapiens* MetAP2b: P50579. Plot made with ClustalX (Larkin *et al*. 2007).

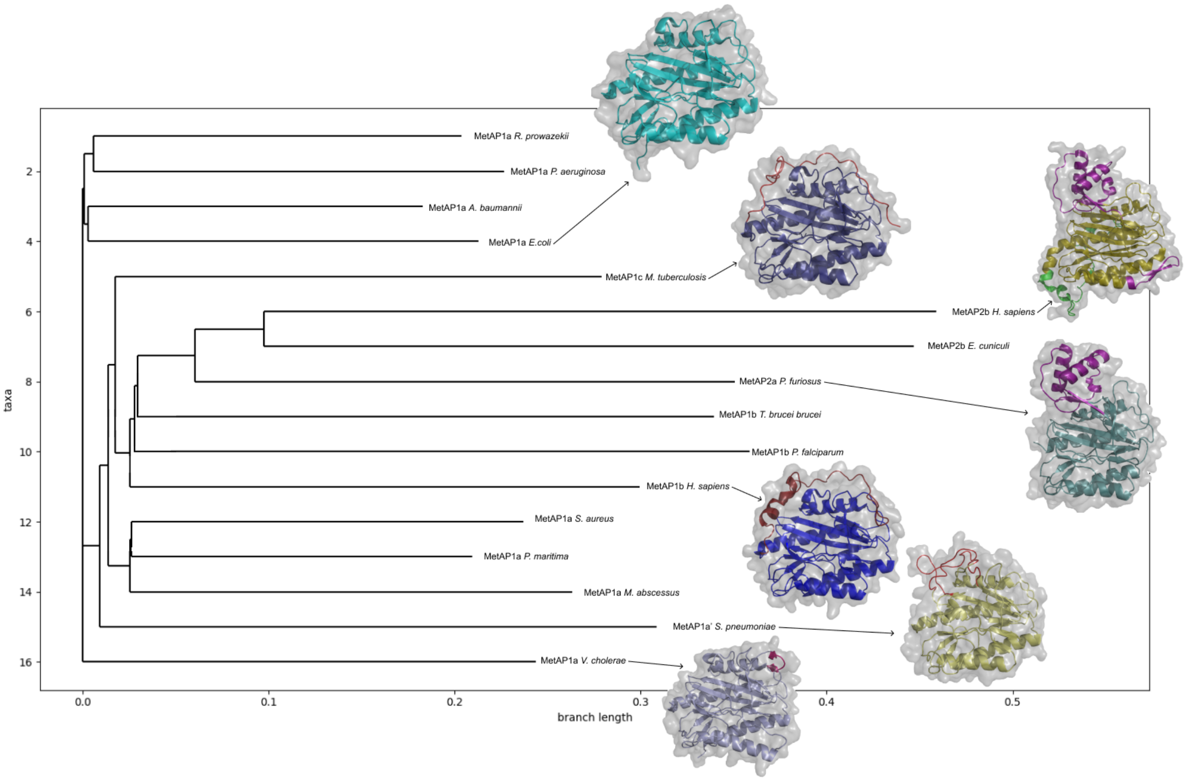


Figure 7. Structural alignment of a group of solved crystallographically solved MetAP members. EBI PDBeFold SSM (Krissinel *et al.* 2004) multiple alignment tool was used to align 3D models and Q-scores results are presented in Table 1. For each pairwise comparison, (1 - Q-score) was used as reference for distance. The Neighbourhood-joining method was carried out by BIONJ (Gascuel 1997) with the distance matrix as input. The pseudo two-fold axis can be clearly noted in the cartoon representations of the MetAP members. A tendency of spatial positioning in additional insertions is noticed when analyzing the upper part of the presented view. The human variant of MetAP2b presents a N-terminal extension behind the presented view, colored in green. The colors of additional insertions are the same presented in Figure 6. RCSB PDB accession numbers are the following: *E. coli* MetAP1a: 2GG2, *A. baumanmii* MetAP1a: 6MRF, *P. aeruginosa* MetAP1a: 4FO7, *P. maritima* MetAP1a: 1O0X, *R. prowazekii* MetAP1a: 3MX6, *S. aureus* MetAP1a: 1QXY, *V. cholerae* MetAP1a: 6LH7, *M. abscessus:* 3TAV, *S. pneumoniae* MetAP1a': 4KM3, *H. sapiens* MetAP1b: 2B3H, *T. brucei brucei* MetAP1b: 4FUK, *P. falciparium* MetAP1b: 3S6B, *M. tuberculosis* MetAP1c: 3PKA, *P. furiosus* MetAP2a: 1XGS, *E. cuniculi* MetAP2b: 3FM3, *H. sapiens* MetAP2b: 5D6E. Molecules were rendered using the PyMol (Schrödinger 2016) community version

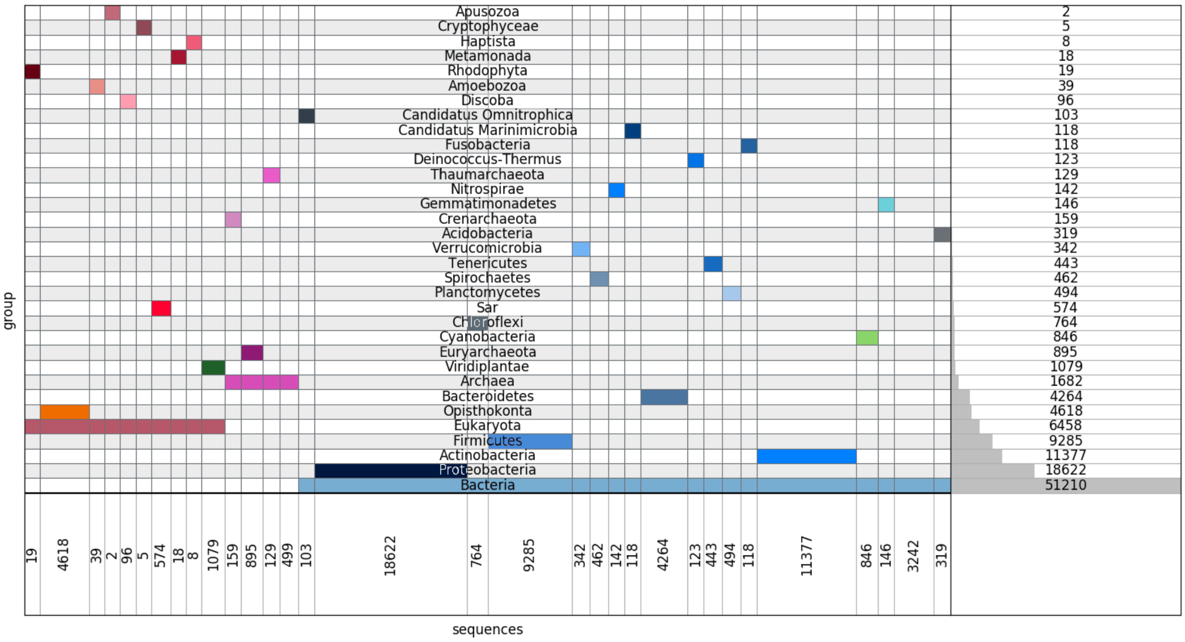


Figure 8. Dataset composition according to taxonomic groups. Only groups with more than 100 members are shown. The Supervenn library (Fedor 2020) was used in the plot.

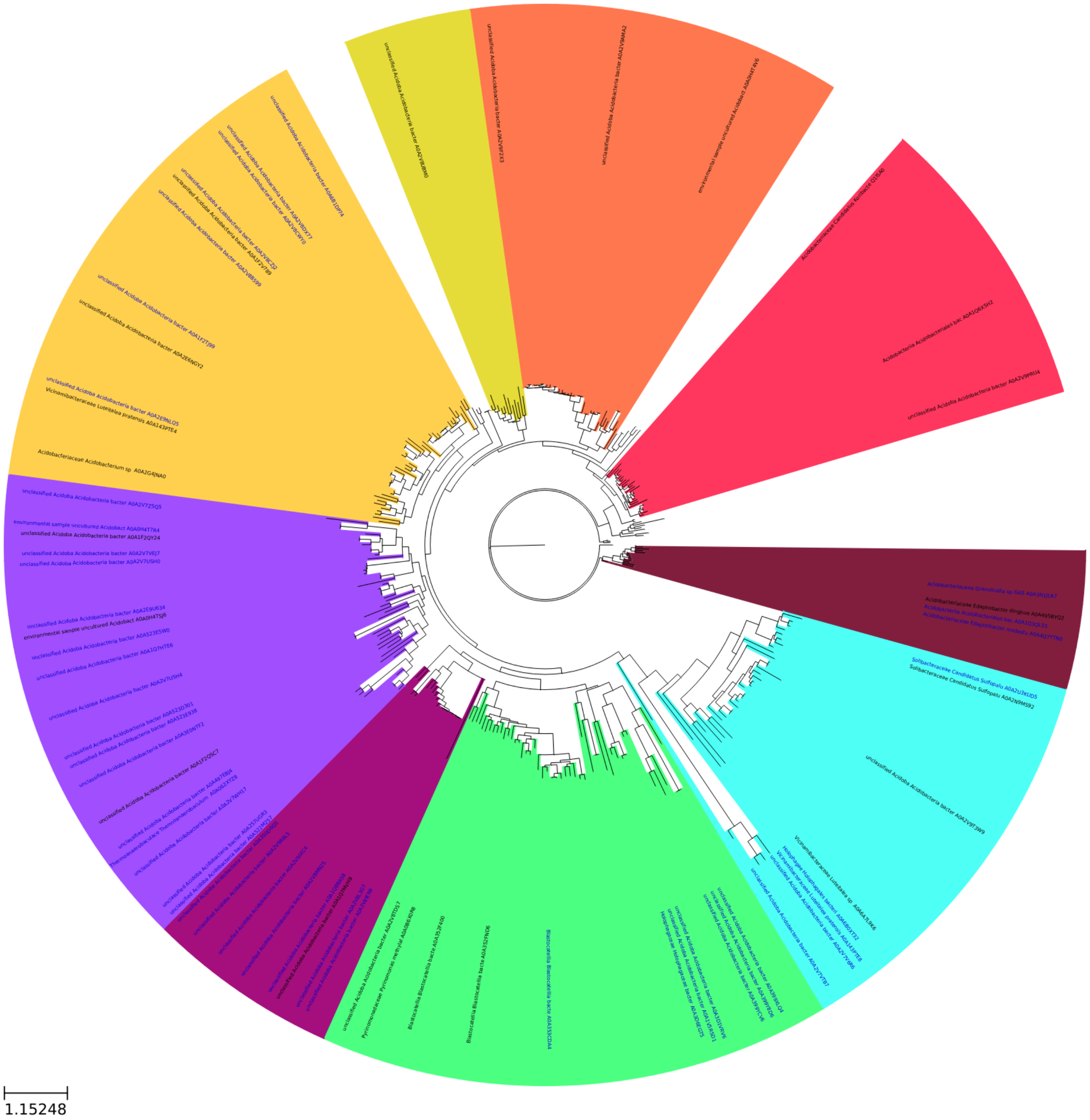

unclassified Acidoba Acidobacteria bacter A0A2V8UBM0
unclassified Acidoba Acidobacteria bacter A0A2V9F2X3
unclassified Acidoba Acidobacteria bacter A0A2V9AMA2
environmental sample uncultured Acidobact A0A0H4T4V6
Acidobacteriaceae Candidatus Koribacte Q1ISA0
Acidobacteriia Acidobacteriales bac A0A1Q6X5H2
unclassified Acidoba Acidobacteria bacter A0A2V9PRU4
Acidobacteriaceae Granulicella sp GAS A0A3N1JLR7
Acidobacteriaceae Edaphobacter dinghue A0A495BYQ2
Acidobacteriia Acidobacteriales bac A0A1Q3QL55
Acidobacteriaceae Edaphobacter modestu A0A4Q7YTN0
Solibacteraceae Candidatus Sulfopalu A0A2U3KUD5
Solibacteraceae Candidatus Sulfopalu A0A2N9MS92
unclassified Acidoba Acidobacteria bacter A0A2V9T3W9
Vicinamibacteraceae Luteitalea sp A0A6A7L9K6
Holophagae Holophagales bacteri A0A6B0XT32
Vicinamibacteraceae Luteitalea pratensis A0A143PTE8
unclassified Acidoba Acidobacteria bacter A0A2V7V6R6
unclassified Acidoba Acidobacteria bacter A0A2V7VTB7
unclassified Acidoba Acidobacteria bacter A0A399XLQ4
unclassified Acidoba Acidobacteria bacter A0A399Y8D6
unclassified Acidoba Acidobacteria bacter A0A399YCV6
unclassified Acidoba Acidobacteria bacter A0A3D1VRV6
unclassified Acidoba Acidobacteria bacter A0A1V5R5D1
Holophagaceae Holophagaceae bacter A0A3D6EGT5
Blastocatellia Blastocatellia bacte A0A353CDA4
Blastocatellia Blastocatellia bacte A0A352FND6
Blastocatellia Blastocatellia bacte A0A352F400
Pyrinomonadaceae Pyrinomonas methylal A0A0B6X0R8
unclassified Acidoba Acidobacteria bacter A0A2V8TDS7
unclassified Acidoba Acidobacteria bacter A0A2V8IEN9
unclassified Acidoba Acidobacteria bacter A0A2V8L3G7
unclassified Acidoba Acidobacteria bacter A0A1Q7MVH9
unclassified Acidoba Acidobacteria bacter A0A1Q8BA58
unclassified Acidoba Acidobacteria bacter A0A2V8MN25
unclassified Acidoba Acidobacteria bacter A0A2V9IXC4
unclassified Acidoba Acidobacteria bacter A0A2V9R6L5
unclassified Acidoba Acidobacteria bacter A0A350D6Q5
unclassified Acidoba Acidobacteria bacter A0A522M257
unclassified Acidoba Acidobacteria bacter A0A257UGR3
unclassified Acidoba Acidobacteria bacter A0A2V7WH17
Thermoanaerobaculace Thermoanaerobaculum A0A062XYZ8
unclassified Acidoba Acidobacteria bacter A0A497EBJ4
unclassified Acidoba Acidobacteria bacter A0A1F2QSC7
unclassified Acidoba Acidobacteria bacter A0A3E0NTF2
unclassified Acidoba Acidobacteria bacter A0A523E938
unclassified Acidoba Acidobacteria bacter A0A523D301
unclassified Acidoba Acidobacteria bacter A0A2V7USH4
unclassified Acidoba Acidobacteria bacter A0A1Q7HTE6
unclassified Acidoba Acidobacteria bacter A0A523E5W0
environmental sample uncultured Acidobact A0A0H4TSJ6
unclassified Acidoba Acidobacteria bacter A0A2E9U634
unclassified Acidoba Acidobacteria bacter A0A2V7USH0
unclassified Acidoba Acidobacteria bacter A0A2V7VEJ7
unclassified Acidoba Acidobacteria bacter A0A1F2QY24
environmental sample uncultured Acidobact A0A0H4T7R4
unclassified Acidoba Acidobacteria bacter A0A2V7Z5Q5
Acidobacteriaceae Acidobacterium sp A0A2G4JNA0
Vicinamibacteraceae Luteitalea pratensis A0A143PTE4
unclassified Acidoba Acidobacteria bacter A0A2E9NLQ5
unclassified Acidoba Acidobacteria bacter A0A2E6NGY2
unclassified Acidoba Acidobacteria bacter A0A1F2TJ99
unclassified Acidoba Acidobacteria bacter A0A2V8B599
unclassified Acidoba Acidobacteria bacter A0A1F2V789
unclassified Acidoba Acidobacteria bacter A0A2V8CZJ2
unclassified Acidoba Acidobacteria bacter A0A2V8CWY0
unclassified Acidoba Acidobacteria bacter A0A2V8DX77
unclassified Acidoba Acidobacteria bacter A0A6B1DP74
1.15248

Figure 9. *Acidobacteria* MetAP tree containing 624 aminoacidic sequences. Clusters detected applying the procedure described in Figure 2 are highlighted in different colors. A total of 45 members with detected insertions (names in blue) and 23 additional random members (names in black) were considered representative sequences and can be visualized along the tree. The sheer amount of representative sequences extracted from the tree applying the procedure described in Figure 3 accounts for 10,9% of the original mined sequences belonging to that group. The ETE$_3$ toolkit (Huerta-Cepas *et al.* 2016) was used in the plot.

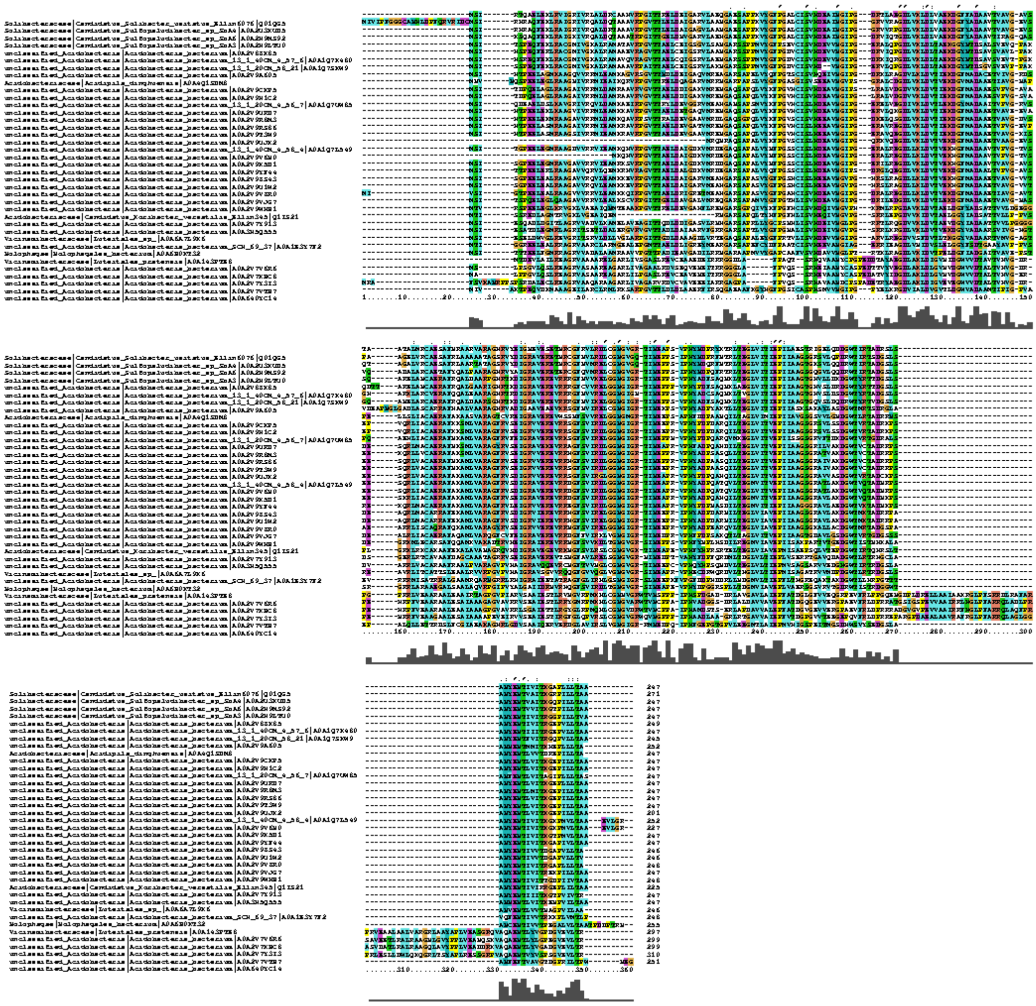


Figure 10. Multiple sequence alignment (MSA) performed on the cluster highlighted in cyan on Figure 8, extracted from *Acidobacteria* MetAP tree. C-terminal insertions are noted in sequences of UniProtKB Accession numbers: A0A2V7V6R6, A0A2V7XBC8, A0A2V7Y3I3 and A0A2V7VTB7, all of which contains the MAP_2 PROSITE (Sigrist *et al.* 2013) or the met_pdase_II TIGRFAMs (Li *et al.* 2021) reference. The latter reference is available at https://www.ncbi.nlm.nih.gov/genome/annotation_prok/evidence/TIGR00501/ and the first is available at https://prosite.expasy.org/doc/PS01202. MSA performed with the MAFFT software (Rozewicki *et al.* 2019), using the auto flag. Plot made with ClustalX (Larkin *et al*. 2007).

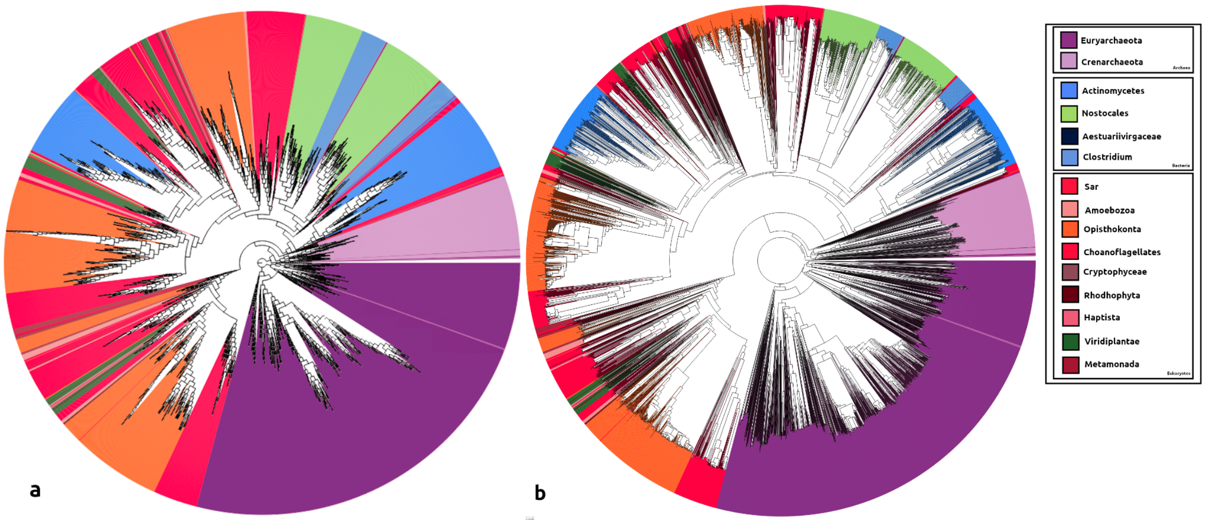


Figure 11. Preliminary tree with prokaryotic phyla and eukaryotic groups of interest cherry picked analysing coding sequences data composition. Setting every branch to 1 facilitates topology analysis and is presented in a. Preserved tree branches according to neighbour-joining are shown in b. Clustering *Cyanobacteria* and eukaryotic members, especially from the *Viridiplantae* kingdom and the *Sar* supergroup can be noted. Clustal Omega (Madeira *et a.l* 2019) was used in MSA. Tree constructed with BIONJ (Gascuel 1997) using neighbour-joining method. The ETE$_3$ toolkit (Huerta-Cepas *et al.* 2016) was used in the plot.

Tables

Table 1. Q-score results of EBI PDBeFold SSM (Krissinel *et al.* 2004) submission.

	6MRF	2GG2	4KM3	3TAV	4FO7	1O0X	3MX6	1QXY	6LH7	2B3H	Q8IJP2	4FUK	3PKA	1XGS	3FM3	5D6E
6MRF	0	0.610	0.504	0.559	0.594	0.617	0.617	0.578	0.575	0.511	0.449	0.470	0.531	0.463	0.348	0.334
2GG2	0.610	0	0.480	0.519	0.563	0.585	0.582	0.550	0.544	0.494	0.436	0.451	0.513	0.436	0.338	0.323
4KM3	0.504	0.480	0	0.449	0.465	0.502	0.493	0.462	0.446	0.406	0.354	0.372	0.431	0.364	0.283	0.268
3TAV	0.559	0.519	0.449	0	0.494	0.577	0.545	0.552	0.496	0.456	0.400	0.424	0.483	0.421	0.321	0.315
4FO7	0.594	0.563	0.465	0.494	0	0.547	0.581	0.523	0.534	0.477	0.416	0.434	0.481	0.413	0.327	0.315
1O0X	0.617	0.585	0.502	0.577	0.581	0	0.603	0.606	0.547	0.512	0.444	0.460	0.556	0.475	0.356	0.344
3MX6	0.617	0.582	0.493	0.545	0.523	0.603	0	0.563	0.547	0.504	0.443	0.466	0.524	0.455	0.352	0.345
1QXY	0.578	0.550	0.462	0.552	0.534	0.606	0.563	0	0.521	0.487	0.423	0.450	0.521	0.457	0.350	0.343
6LH7	0.575	0.544	0.446	0.496	0.477	0.547	0.547	0.521	0	0.456	0.398	0.420	0.481	0.413	0.316	0.302
2B3H	0.511	0.494	0.406	0.456	0.416	0.512	0.504	0.487	0.398	0	0.397	0.412	0.460	0.392	0.302	0.285
Q8IJP2	0.449	0.436	0.354	0.400	0.434	0.444	0.443	0.423	0.420	0.397	0	0.359	0.395	0.335	0.259	0.249
4FUK	0.470	0.451	0.372	0.424	0.481	0.460	0.466	0.450	0.481	0.412	0.359	0	0.411	0.360	0.283	0.269
3PKA	0.531	0.513	0.431	0.483	0.481	0.556	0.524	0.521	0.413	0.460	0.395	0.411	0	0.405	0.305	0.290
1XGS	0.463	0.436	0.364	0.421	0.413	0.475	0.455	0.457	0.413	0.392	0.335	0.360	0.405	0	0.323	0.312
3FM3	0.348	0.338	0.283	0.321	0.327	0.475	0.352	0.350	0.316	0.302	0.259	0.283	0.305	0.323	0	0.289
5D6E	0.334	0.323	0.268	0.315	0.315	0.344	0.345	0.343	0.302	0.285	0.249	0.269	0.290	0.312	0.289	0

Table 2. Clusters extracted from tree built with mined data utilizing the algorithm proposed in Figure 2 with minimum cluster size set to 100 and maximum cluster size set to 3,000.

Cluster name	**Number of members**	**Cluster name**	**Number of members**	**Cluster name**	**Number of members**	**Cluster name**	**Number of members**
Clostridia	2,905	*Paenibacillaceae*	1,058	*Oceanospirillales*	397	*Erysipelotrichia*	177
Pseudomonadales	2,831	*PVC group*	1,026	*Vibrionales*	369	*Cellulomonadaceae*	172
Betaproteobacteria	2,788	*Rhodobacterales*	1,018	*Streptosporangiales*	341	*Rickettsiales*	171
Dikarya	2,743	*Cyanobacteria Melainabacteria* group	893	*Coriobacteriia*	326	*Cellvibrionales*	169
Streptomycetales	2,472	*Micrococcaceae*	806	*Acidobacteria*	313	*Intrasporangiaceae*	163
Bacillaceae	2,401	*Chloroflexi*	750	*Actinomycetales*	287	*Chromatiales*	148
Corynebacteriales	2,300	*Sphingomonadales*	721	*Tenericutes*	283	*Gemmatimonadetes*	142
Rhizobiales	2,178	*Xanthomonadales*	656	*Fungi incertae sedis*	260	*Legionellales*	140
Bacteria incertae sedis	2,091	*Alteromonadales*	635	*Bifidobacteriales*	250	*Nitrospirae*	139
Metazoa	2,026	*Pseudonocardiales*	587	unclassified *Bacteria*	248	unclassified *Firmicutes sensu stricto*	130
Enterobacterales	1,892	*Propionibacteriales*	570	unclassified *Bacteroidetes*	231	*Gammaproteobacteria incertae sedis*	128
delta epsilon subdivisions	1,813	*Rhodospirillales*	569	*Chitinophagia*	223	*Thiotrichales*	126
Archaea	1,677	*Sphingobacteriia*	489	*Pasteurellales*	211	*Geodermatophilales*	122
Lactobacillales	1,673	*Spirochaetes*	460	unclassified *Actinobacteria class*	205	*Fusobacteria*	118
Flavobacteriia	1,478	*Micromonosporales*	453	unclassified *Alphaproteobacteria*	200	*Candidatus Marinimicrobia*	118
Microbacteriaceae	1,441	*Cytophagia*	438	*Caulobacterales*	199	*Acidimicrobiia*	115
Bacteroidia	1,336	*Sar*	437	*Negativicutes*	195	*Deinococcus Thermus*	115
Viridiplantae	1,193	unclassified *Gammaproteobacteria*	411	*Staphylococcaceae*	192	*Brevibacteriaceae*	102

Supplementary Information

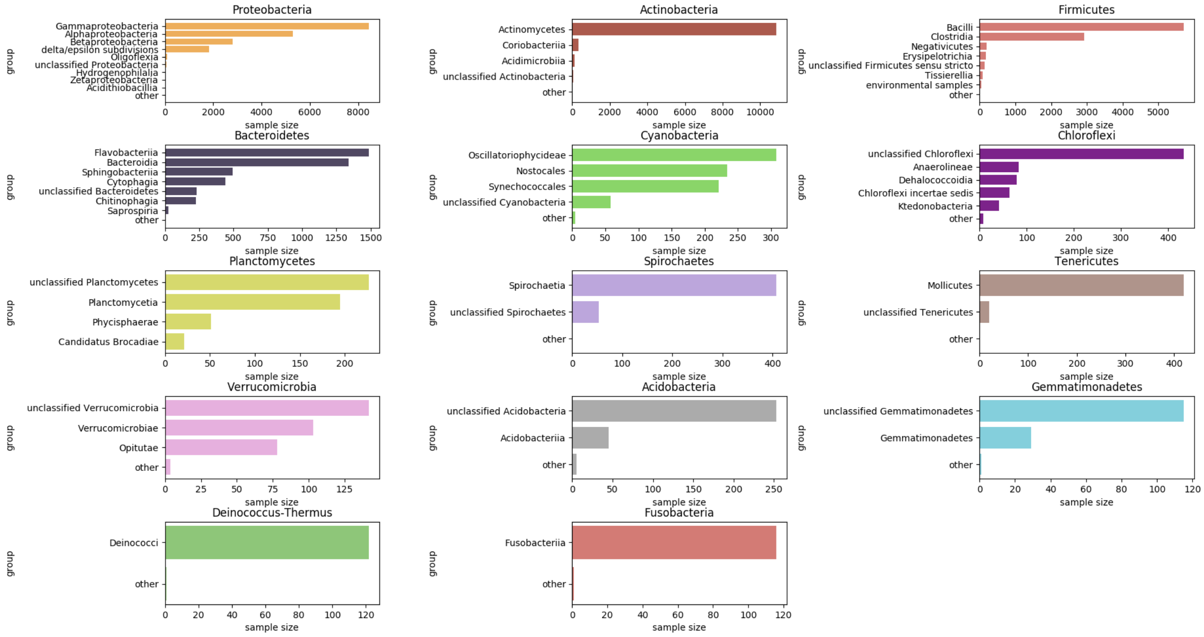
Supplementary Figure 1. Composition of bacterial members in dataset according to phylum. Plots made with the matplotlib library (Caswell *et al.* 2021).

Supplementary Figure 2. Composition of bacterial members in dataset according to class. Plots made with the matplotlib library (Caswell *et al.* 2021).

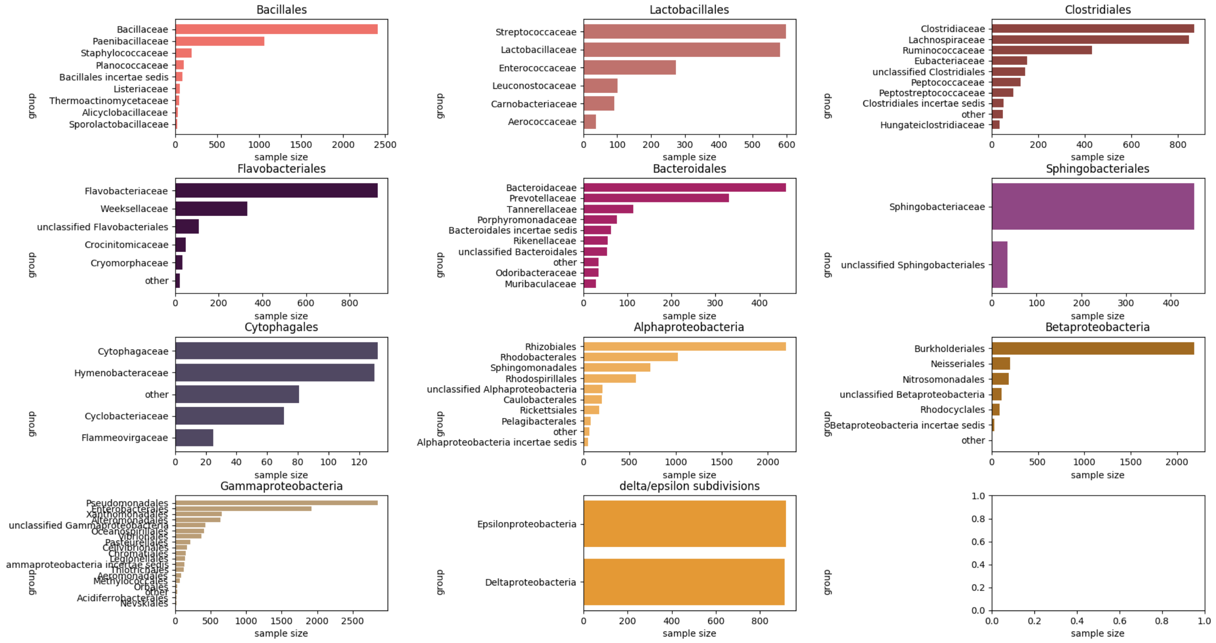
Supplementary Figure 3. Composition of eukaryotic members in dataset according to interest group. Plots made with the matplotlib library (Caswell *et al.* 2021).

Supplementary Figure 4. Composition of eukaryotic members in dataset according to lower interest group. Plots made with the matplotlib library (Caswell *et al.* 2021).

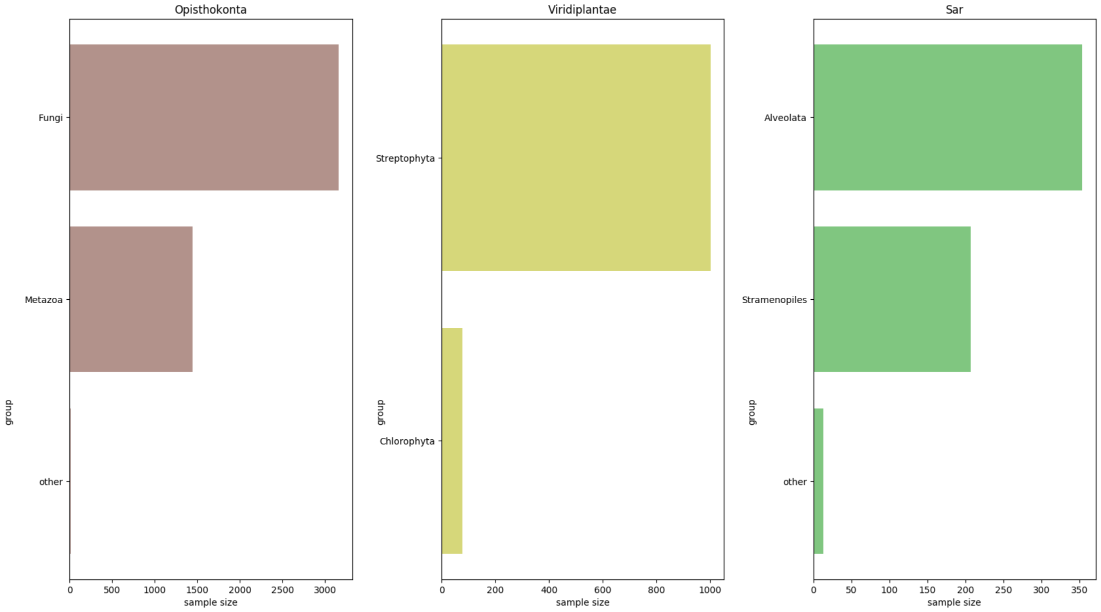
Supplementary Figure 5. Alteromonadales MetAP tree containing 1,268 aminoacidic sequences. Clusters detected applying the procedure described in Figure are highlighted in different colors. A total of 131 members with detected insertions (names in blue) and 41 additional random members (names in black) were considered representative sequences and can be visualized along the tree. The sheer amount of representative sequences extracted from the tree applying the procedure described in Figure 3 accounts for 21,45% of the original mined sequences belonging to that group. The ETE_3 toolkit (Huerta-Cepas *et al.* 2016) was used in the plot.



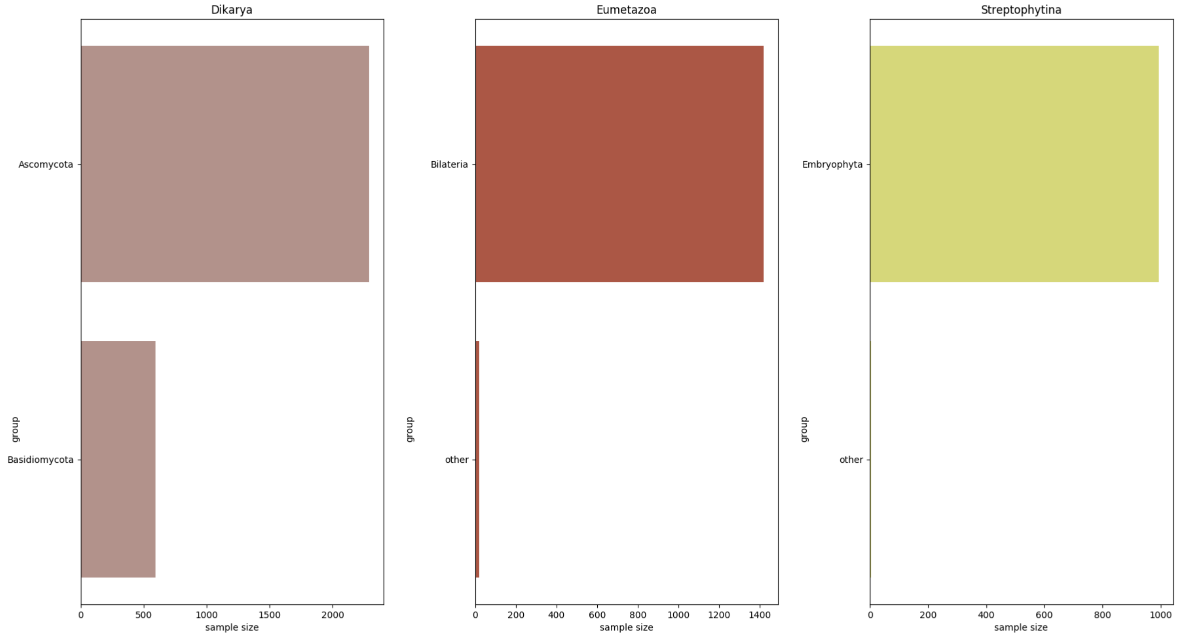
Supplementary Figure 1. Composition of bacterial members in dataset according to phylum. Plots made with the matplotlib library (Caswell *et al.* 2021).



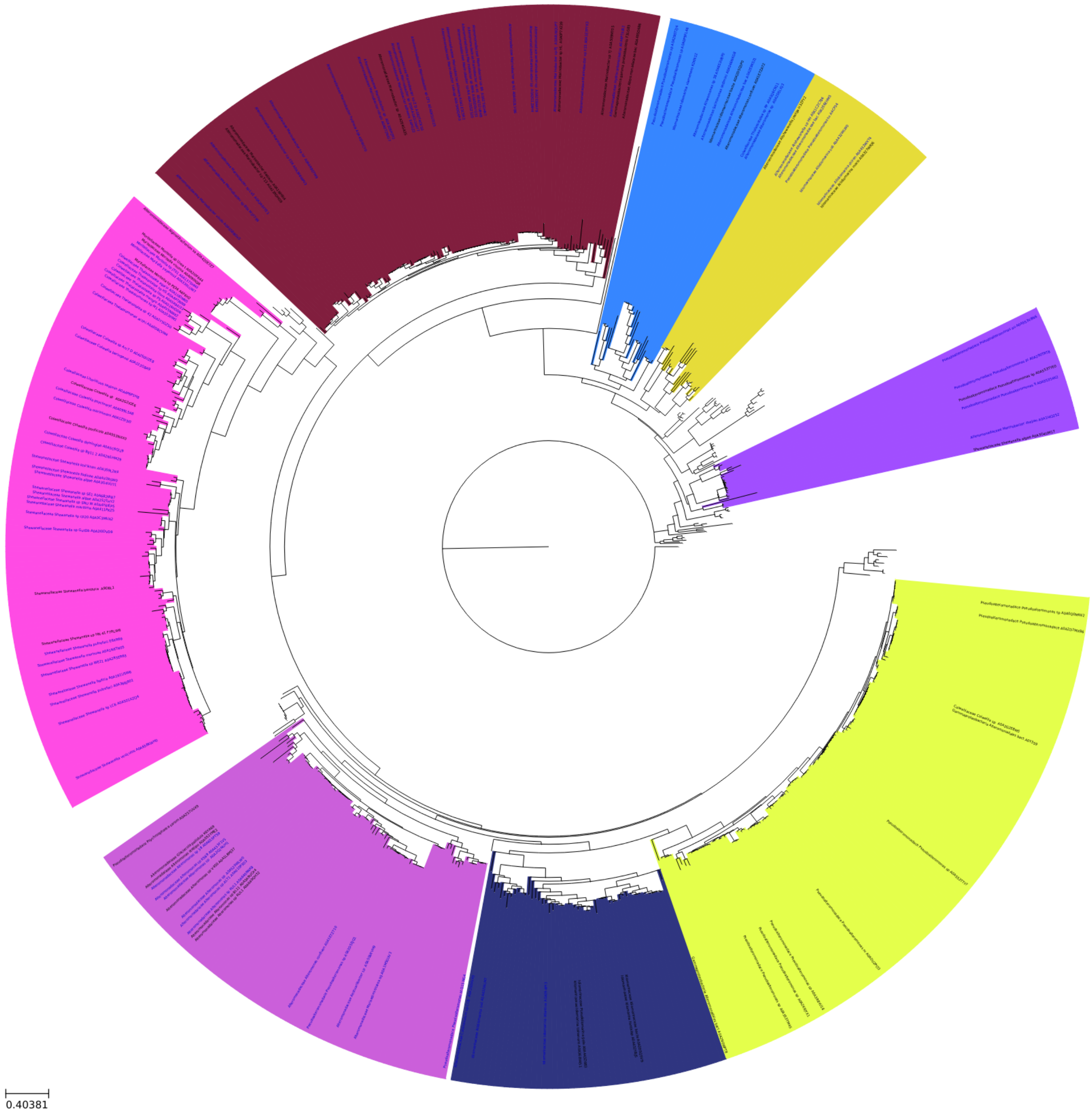
Supplementary Figure 2. Composition of bacterial members in dataset according to class. Plots made with the matplotlib library (Caswell *et al.* 2021).



Supplementary Figure 3. Composition of eukaryotic members in dataset according to interest group. Plots made with the matplotlib library (Caswell *et al.* 2021).



Supplementary Figure 4. Composition of eukaryotic members according to lower interest group. Plots made with the matplotlib library (Caswell *et al.* 2021).



Supplementary Figure 5. Alteromonadales MetAP tree containing 1,268 aminoacidic sequences. Clusters detected applying the procedure described in Figure are highlighted in different colors. A total of 131 members with detected insertions (names in blue) and 41 additional random members (names in black) were considered representative sequences and can be visualized along the tree. The sheer amount of representative sequences extracted from the tree applying the procedure described in Figure 3 accounts for 21,45% of the original mined sequences belonging to that group. The ETE$_3$ toolkit (Huerta-Cepas *et al.* 2016) was used in the plot.

DISCUSSÃO GERAL

Durante o desenvolvimento deste trabalho, fez-se necessário o uso de diversas ferramentas computacionais. O autor fez uso do domínio dos recursos disponíveis na configuração dos ambientes necessários para executar os diferentes softwares. É importante notar que sem a aplicação de técnicas de bancos de dados relacionais, visualização de dados e *data science*, requisições HTTP e *web scraping*, algoritmos de programação, estruturas de dados e paradigmas de orientação a objeto, este trabalho não seria possível.

Tanto escolha das MetAPs como objeto de estudo quanto as técnicas aplicadas nele advém das experiências que o autor vivenciou desde o início da graduação em bolsas de iniciação científica e de apoio técnico (como desenvolvedor de sistemas WEB), aulas das disciplinas e até mesmo de fatores externos relacionados ao aprendizado de outras tecnologias no ambiente Linux.

A fim de compreender a evolução geral das MetAPs, um grande conjunto de dados de sequências codificantes de genes de MetAP foi minerado. A visualização da constituição taxonômica dos organismos que expressam tais sequências foi fundamental para o desenvolvimento e proposição da metodologia. O procedimento descrito para realizar a clusterização inicial dos dados de acordo com sua referência *tax id* NCBI permitiu a extração de grandes subconjuntos que se alinharam demonstrando grande similaridade de sequências, o que representa uma oportunidade para eleger apenas uma pequena amostra representativa para cada *cluster*. Assim como a mineração dos dados e o tratamento dos mesmos, a etapa de clusterização foi automatizada utilizando *scripts* na linguagem Python.

Espera-se que a conclusão das análises iniciais apresentadas aqui permitam avaliar diversos fatores de alto impacto já discutidos no corpo do artigo a ser submetido para publicação. Ademais, o código gerado durante o desenvolvimento do trabalho permite que outras análises filogenéticas sejam realizadas utilizando o mesmo *pipeline* construído.

REFERÊNCIAS BIBLIOGRÁFICAS

(Caps. Introdução e Discussão)

Alberts, B. (2010). Cell biology: the endless frontier. *Molecular Biology of the Cell*, *21*(22), 3785. https://doi.org/10.1091/mbc.e10-04-0334

Nelson, D. L., & Cox, M. M. (2017). *Lehninger principles of biochemistry* (7th ed.). W.H. Freeman.

Bonissone, S., Gupta, N., Romine, M., Bradshaw, R. A., & Pevzner, P. A. (2013). N-terminal protein processing: A comparative proteogenomic analysis. *Molecular and Cellular Proteomics*, *12*(1), 14–28. https://doi.org/10.1074/mcp.M112.019075

Bradshaw, R. A., Brickey, W. W., & Walker, K. W. (1998). N-terminal processing: The methionine aminopeptidase and N(α)-acetyl transferase families. *Trends in Biochemical Sciences*, *23*(7), 263–267. https://doi.org/10.1016/S0968-0004(98)01227-4

Cao, Y., Trivellone, V., & Dietrich, C. H. (2020). A timetree for phytoplasmas (Mollicutes) with new insights on patterns of evolution and diversification. *Molecular Phylogenetics and Evolution*, *149*, 106826. https://doi.org/10.1016/j.ympev.2020.106826

Lowther, W. T., & Matthews, B. W. (2002). Metalloaminopeptidases: common functional themes in disparate structural surroundings. *Chemical reviews*, *102*(12), 4581–4608. https://doi.org/10.1021/cr0101757

Ross, S., Giglione, C., Pierre, M., Espagne, C., & Meinnel, T. (2005). Functional and developmental impact of cytosolic protein N-terminal methionine excision in Arabidopsis. Plant physiology, 137(2), 623–637. https://doi.org/10.1104/pp.104.056861

Griffith, E. C., Su, Z., Turk, B. E., Chen, S., Chang, Y. H., Wu, Z., Biemann, K., & Liu, J. O. (1997). Methionine aminopeptidase (type 2) is the common target for angiogenesis inhibitors AGM-1470 and ovalicin. Chemistry & biology, 4(6), 461–471. https://doi.org/10.1016/s1074-5521(97)90198-8

Datta, B. (2000). MAPs and POEP of the roads from prokaryotic to eukaryotic kingdoms. Biochimie, 82(2), 95–107. https://doi.org/10.1016/s0300-9084(00)00383-7

Datta, R., Tammali, R., & Datta, B. (2003). Negative regulation of the protection of eIF2alpha phosphorylation activity by a unique acidic domain present at the N-terminus of p67. *Experimental cell research*, *283*(2), 237–246. https://doi.org/10.1016/s0014-4827(02)00042-3

Olaleye, O. A., Bishai, W. R., & Liu, J. O. (2009). Targeting the role of N-terminal methionine processing enzymes in Mycobacterium tuberculosis. *Tuberculosis (Edinburgh, Scotland)*, *89 Suppl 1*, S55–S59. https://doi.org/10.1016/S1472-9792(09)70013-7

Datta, B., Datta, R., Ghosh, A., & Majumdar, A. (2006). The binding between p67 and eukaryotic initiation factor 2 plays important roles in the protection of eIF2alpha from phosphorylation by kinases. *Archives of biochemistry and biophysics*, *452*(2), 138–148. https://doi.org/10.1016/j.abb.2006.06.009

Vetro, J. A., & Chang, Y. H. (2002). Yeast methionine aminopeptidase type 1 is ribosome-associated and requires its N-terminal zinc finger domain for normal function in vivo. *Journal of cellular biochemistry*, *85*(4), 678–688. https://doi.org/10.1002/jcb.10161

Chiu, J., Wong, J. W. H., & Hogg, P. J. (2014). Redox regulation of Methionine aminopeptidase 2 activity. *Journal of Biological Chemistry*, *289*(21), 15035–15043. https://doi.org/10.1074/jbc.M114.554253

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., … De Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/BIOINFORMATICS/BTP163

D480-D489. (2021). UniProt: the universal protein knowledgebase in 2021 The UniProt Consortium. *Nucleic Acids Research*, *49*. https://doi.org/10.1093/nar/gkaa1100

Elsaied, H., Stokes, H. W., Nakamura, T., Kitamura, K., Fuse, H., & Maruyama, A. (2007). Novel and diverse integron integrase genes and integron-like gene cassettes are prevalent in deep-sea hydrothermal vents. *Environmental Microbiology*, *9*(9), 2298–2312. https://doi.org/10.1111/j.1462-2920.2007.01344.x

Esa, R., Steinberg, E., Dror, D., Schwob, O., Khajavi, M., Maoz, M., … Benny, O. (2020). The role of methionine aminopeptidase 2 in lymphangiogenesis. *International Journal of Molecular Sciences*, *21*(14), 1–15. https://doi.org/10.3390/ijms21145148

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution, 4 4*, 406-25 .

Fedor. (2020). Supervenn. https://doi.org/10.5281/ZENODO.4012442

Gascuel, O. (1997). BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, *14*(7), 685–695. https://doi.org/10.1093/oxfordjournals.molbev.a025808

Giglione, C., Boularot, A., & Meinnel, T. (2004). Protein N-terminal methionine excision. *Cellular and Molecular Life Sciences*, *61*(12), 1455–1474. https://doi.org/10.1007/s00018-004-3466-8

Giglione, C., Fieulaine, S., & Meinnel, T. (2015, June 3). N-terminal protein modifications: Bringing back into play the ribosome. *Biochimie*. Elsevier B.V. https://doi.org/10.1016/j.biochi.2014.11.008

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, *33*(6), 1635–1638. https://doi.org/10.1093/MOLBEV/MSW046

Varshavsky, A. (2011). The N-end rule pathway and regulation by proteolysis. Protein Science : A Publication of the Protein Society, 20(8), 1298–1345. https://doi.org/10.1002/pro.666

Gibbs, D. J., Bacardit, J., Bachmair, A., & Holdsworth, M. J. (2014). The eukaryotic N-end rule pathway: conserved mechanisms and diverse functions. *Trends in cell biology*, *24*(10), 603–611. https://doi.org/10.1016/j.tcb.2014.05.001

Frottin, F., Espagne, C., Traverso, J. A., Mauve, C., Valot, B., Lelarge-Trouverie, C., Zivy, M., Noctor, G., Meinnel, T., & Giglione, C. (2009). Cotranslational proteolysis dominates glutathione homeostasis to support proper growth and development. *The Plant cell*, *21*(10), 3296–3314. https://doi.org/10.1105/tpc.109.069757

Langa, Ł. (2020). *PEP 596 -- Python 3.9 Release Schedule*. Retrieved from https://www.python.org/dev/peps/pep-0596/

Leszczyniecka, M., Bhatia, U., Cueto, M., Nirmala, N. R., Towbin, H., Vattay, A., … Phillips, P. E. (2006). MAP1D, a novel methionine aminopeptidase family member is overexpressed in colon cancer. *Oncogene*, *25*(24), 3471–3478. https://doi.org/10.1038/sj.onc.1209383

Lin, M., Zhang, X., Jia, B., & Guan, S. (2018). Suppression of glioblastoma growth and angiogenesis through molecular targeting of methionine aminopeptidase-2. *Journal of Neuro-Oncology*, *136*(2), 243–254. https://doi.org/10.1007/s11060-017-2663-x

Martin, F., & Lopez, M. C. (1996). Methionine aminopeptidase-I: the, *0004*(96), 285–286.

Morgante, C. V., Rodrigues, R. A. O., Marbach, P. A. S., Borgonovi, C. M., Moura, D. S., & Silva-Filho, M. C. (2009). Conservation of dual-targeted proteins in Arabidopsis and rice points to a similar pattern of gene-family evolution. *Molecular Genetics and Genomics*, *281*(5), 525–538. https://doi.org/10.1007/s00438-009-0429-7

Munkhjargal, T., Ishizaki, T., Guswanto, A., Takemae, H., Yokoyama, N., & Igarashi, I. (2016). Molecular and biochemical characterization of methionine aminopeptidase of Babesia bovis as a potent drug target. *Veterinary Parasitology*, *221*, 14–23. https://doi.org/10.1016/j.vetpar.2016.02.024

Omega, C. (n.d.). Resource Summary Report Clustal Omega RRID:SCR_001591 Type: Tool Proper Citation.

Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A. R. N., Potter, S. C., Finn, R. D., & Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*, *47*(W1), W636–W641. https://doi.org/10.1093/nar/gkz268

Pandrea, I., Mittleider, D., Brindley, P. J., Didier, E. S., & Robertson, D. L. (2005). Phylogenetic relationships of methionine aminopeptidase 2 among Encephalitozoon species and genotypes of microsporidia. *Molecular and Biochemical Parasitology*, *140*(2), 141–152. https://doi.org/10.1016/j.molbiopara.2004.12.006

Peterson, B. (2019). PEP 373 -- Python 2.7 Release Schedule | Python.org. Retrieved November 19, 2021, from https://www.python.org/dev/peps/pep-0373/

Richardson, L. (2016). Beautiful Soup Documentation. *Media.Readthedocs.Org*. Retrieved from https://media.readthedocs.org/pdf/beautiful-soup-4/latest/beautiful-soup-4.pdf%0Ahttp://www.crummy.com/software/BeautifulSoup/bs4/doc/

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., … Karsch-Mizrachi, I. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database : The Journal of Biological Databases and Curation*, *2020*. https://doi.org/10.1093/DATABASE/BAAA062

scikit-bio development team, T. (2020). scikit-bio: A Bioinformatics Library for Data Scientists, Students, and Developers. Retrieved from http://scikit-bio.org

Serero, A., Giglione, C., Sardini, A., Martinez-Sanz, J., & Meinnel, T. (2003). An Unusual Peptide Deformylase Features in the Human Mitochondrial N-terminal Methionine Excision Pathway. *Journal of Biological Chemistry*, *278*(52), 52953–52963. https://doi.org/10.1074/jbc.M309770200

Varshavsky, A. (1997). The N-end rule pathway of protein degradation. *Genes to Cells*, *2*(1), 13–28. https://doi.org/10.1046/j.1365-2443.1997.1020301.x

Vavilova, V., Sormacheva, I., Woyciechowski, M., Eremeeva, N., Fet, V., Strachecka, A., … Blinov, A. (2015). Distribution and diversity of Nosema bombi (Microsporidia: Nosematidae) in the natural populations of bumblebees

(Bombus spp.) from West Siberia. *Parasitology Research*, *114*(9), 3373–3383. https://doi.org/10.1007/s00436-015-4562-4

Widenius, M., Axmark, D., & DuBois, P. (2002). Mysql Reference Manual.

You, C. H., Lu, H. Y., Sekowska, A., Fang, G., Wang, Y. P., Gilles, A. M., & Danchin, A. (2005). The two authentic methionine aminopeptidase genes are differentially expressed in Bacillus subtilis. *BMC Microbiology*, *5*, 1–15. https://doi.org/10.1186/1471-2180-5-57

Zhang, H., Huang, H., Cali, A., Takvorian, P. M., Feng, X., Zhou, G., & Weiss, L. M. (2005). Investigations into microsporidian methionine aminopeptidase type 2: A therapeutic target for microsporidiosis. *Folia Parasitologica*, *52*(1–2), 182–192. https://doi.org/10.14411/fp.2005.023