

Bioinformatics
Multidisciplinary
Environment

Centro
Multiusuário
de Bioinformática



Bancos de dados de informação proteica

Objetivos

- Usar os diferentes bancos de dados de informação de proteínas.
- Caracterizar proteínas desconhecidas utilizando não apenas a sequência, mas perfis, motivos e domínios.
- Obter informações sobre a família de uma determinada proteína.
- Obter alinhamentos e matrizes de posicionamento específicas para uma família de proteínas.

Observação:

- Este tutorial foi construído ***apenas para fins didáticos para a disciplina de Bioinformática Estrutural. A reprodução dele para qualquer outro fim não é permitida e nem consentida pelos professores do curso.***

Identificação de Proteínas utilizando o UniPROT

Este é o principal banco de dados de informação de proteínas, com links e informações cruzadas para diversos outros bancos de dados.

Quando utilizar o Uniprot?

- Para obter informações gerais sobre uma proteína.
- Registro de anotação mais completo para proteínas únicas.
- Swiss-prot: proteínas anotadas e revisadas manualmente.
- Para realizar busca de similaridade com proteínas conhecidas.
- Encontrar homólogos curados e já confirmados experimentalmente.
- Encontrar dados relacionados a mutantes naturais e variações associadas a doenças ou fenótipos alterados.

- Obter informações sobre a estrutura e resíduos importantes para a atividade.

Buscas de similaridade utilizando o BLAST também poderão ser realizadas diretamente no site do UniPROT. Vamos a um exemplo?

- Abrir a Home Page do UNIPROT:

<http://www.uniprot.org>

- Clicar no link **BLAST** (Canto superior esquerdo).

The screenshot shows the UniProt homepage. At the top, there's a navigation bar with links for 'UniProtKB' (selected), 'Align', 'Retrieve/ID mapping', 'Peptide search', 'Advanced', 'Search', 'Help', and 'Contact'. A red arrow points to the 'BLAST' link. Below the navigation bar, there's a mission statement: 'The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.' The main content area is divided into several sections: 'UniProtKB' (Swiss-Prot: 557,713 entries, TrEMBL: 116,030,110 entries), 'UniRef' (Sequence clusters), 'UniParc' (Sequence archive), 'Proteomes' (represented by icons of a fly, a person, and a protein), 'News' (with links to 'Forthcoming changes' and 'UniProt release 2018_06'), 'Supporting data' (Literature citations, Taxonomy, Diseases, Subcellular locations, Keywords), and 'UniProt data' (Download latest release, Statistics, How to cite us). At the bottom, there's a 'Getting started' section with links to 'Text search', 'BLAST', and 'Sequence alignments'. A banner at the very bottom informs users about GDPR compliance and a privacy notice update.

- Copiar e colar a sequência abaixo no campo *Query*:

```
>Seq5
MASFTTTAAAASRLLPSSSSISRLSLSSSSSSSSKLCLRSSLVSHLFLRQRGGSAYVTKTRFSTKC
YASDPAQLKNAREDIKELLQSKFCHPIMVRLGWHDAGTYNKDIKEWPQRGGANGSLSFDVELRHGANAGL
VNALKLLQPIKDVKYSGVTYADLFQLASATAIEEAGGPTIPMKYGRVDAKGPEQCPPEEGLPDAGPPSPAQ
HLRDVFYRMGLDDKDIVALSGAHTLGRSRPERSGWGKPETKYTDGPGAPGGQSWTAEWLKFDNSYFKDI
KEKRADLLVLPRTDAALFEDPSFKVYAEKYAADQEAFFKDYAEAHAKLSNQGAKFDPAEGITLNGTPAGA
APEKFVAAKYSSNKRSELSDSMKEKIRAEYEGFGGSPNKLPTNYFLNIMIVIGVLAVLSYLAGN
```

- Clicar em BLAST, e após o aparecimento dos resultados, analisar a tabela de hits (Sequências similares ou iguais presentes no banco de dados).
- Verificar os 6 primeiros *Hits*.
- Clique no primeiro *hit* e verifique a estrutura das informações contidas no Uniprot.

Obtendo informações sobre estrutura proteíca no UniProt:

Abrir novamente o site do UniProt:

<http://www.uniprot.org>

- No campo Query, inserir o termo **PGH2_MOUSE**.
- Observar os resultados.
- Em uma outra janela/aba do navegador, na mesma página acima, faça a busca pelo termo **GYS2_HUMAN**.
- Observar os resultados.

Explorando o *Protein Data Bank*

O RCSB PDB é o principal banco de estruturas de proteínas resolvidas experimentalmente. É dele que retiramos as proteínas molde para realização de modelagem comparativa.

The screenshot shows the main interface of the RCSB PDB website. At the top, there is a navigation bar with links for Deposit, Search, Visualize, Analyze, Download, Learn, and More. A "MyPDB" button is also present. Below the navigation bar, the RCSB PDB logo is displayed along with a map of the world. A search bar allows users to search by PDB ID, author, macromolecule, sequence, or ligands. Below the search bar, there are links for "Advanced Search" and "Browse by Annotations". On the left side, a sidebar menu lists "Welcome", "Deposit", "Search", "Visualize", "Analyze", "Download", and "Learn". The main content area features a section titled "A Structural View of Biology" which discusses the archive of biological macromolecular structures and their applications in research and education. It also mentions the RCSB PDB's role as a member of the wwPDB and its work on antibiotic resistance. To the right, there is a "July Molecule of the Month" section featuring a 3D ribbon model of the Piezo1 Mechanosensitive Channel, colored in blue and purple. A "Contact Us" link is located on the far right edge of the page.

Quando utilizar o PDB?

- Para obter estruturas de proteínas resolvidas experimentalmente.
- Para obter informações estruturais sobre uma proteína.
- Para verificar e encontrar estruturas proteicas homólogas.

- Para obter informações sobre sítios importantes para a atividade enzimática.
- Obter informações sobre os ligantes das proteínas.

Vamos agora explorar o banco PDB.

- Abrir a Home Page do Protein Data Bank (PDB):
www.rcsb.org
- No campo “PDB ID or Text” insira o termo: 3HTB
- A partir da página aberta, obter as seguintes informações:
 - Identificação da Proteína;
 - Organismo Fonte;
 - Número de Cadeias Polipeptídicas;
 - Método Experimental pelo qual o modelo foi obtido;
 - Outras estruturas relacionadas;
 - Mutações encontradas (se existirem).
 - Clicar na estrutura e observar o modelo 3D utilizando a ferramenta Jmol.

Identificando os domínios de uma proteína

Vamos agora identificar a arquitetura dos domínios que esta proteína abaixo possui e a família a qual pertence.

```
>1smd
GRTSIVHLFEWRWVDIALECERYLAPKGFGGVQVSPPNENVAIHNPFRPWERYQPVSYK
LCTRSGNEDEFRNMVTRCNNVGVRIVDAVINHMCNAVSAGTSSTCGSYFNPGSRDFPA
VPYSGWDFNDGKCKTSGDIENYNDATQVRDCRLSGLLDLALGKDYVRSKIAEYMNHLID
IGVAGFRIDASKHMPGDIKAIQLDKLHNLSNWFPEGSKPFIYQEVIDLGGEPIKSSDYF
GNGRVTEFKYGAKLGTIVRKWNGEKMSYLKNWGEWGFMPSDRALVFVDNHDNQRGHGAG
GASILTFWDARLYKMAVGFLAHPYGFTRVMSSYRPRYFENGKDVNWDVGPPNDNGVTK
EVTINPDTCGNDWVCEHRWRQIRNMVNFRNVVGQFTNWYDNGSNQVAFGRGNRGFIV
FNNDDWTFSLTQTLQTPAGTYCDVISGDKINGNCTGIKIYVSDDGKAHFSISNSAEDPFI
AIHAESKL
```

Para isto, iremos utilizar primeiramente o banco [CDD](#) (*Conserved Domain Databases*), o qual é vinculado ao NCBI. A ferramenta que faz esta identificação é o [SPARCLE](#) (*Subfamily Protein Architecture Labeling Engine*), que é um recurso que caracteriza funcionalmente e rotula sequências de proteínas que foram agrupadas por sua arquitetura de domínio conservado característica. Uma arquitectura de domínio é definida como a ordem sequencial de domínios conservados numa sequência de proteínas (CDD-NCBI).

Quando usar o CDD?

- Para procurar domínios conservados de proteínas.
- Encontrar informações sobre famílias de proteínas.
- Inclusive superfamílias e subfamílias.
- Para obter matrizes PSSM específicas para cada família de proteína.
- Para obter alinhamentos entre representantes de cada família.
- Entre os mais distantes e os mais representativos.
- Verificar relações de proximidade entre as famílias proteicas.

A utilização do SPARCLE pode ser realizada de duas maneiras: a partir de uma sequência de aminoácidos ou por uma palavra-chave. Para utilizar a partir da sequência, utilizaremos o [CD-Search](#): Abra o [CD-Search](#). Copie a sequência fasta `1smd` na caixa de consulta, como indicado na figura abaixo:

The screenshot shows the NCBI CD-Search interface. At the top, there is a navigation bar with links for HOME, SEARCH, GUIDE, Structure Home, 3D Macromolecular Structures, Conserved Domains, Pubchem, and BioSystems. Below the navigation bar, the main title is "Search for Conserved Domains within a protein or coding nucleotide sequence". A yellow banner at the top right says "NEW! Use Batch CD-search to submit multiple query proteins at once!". The search form has a text area labeled "Enter protein or nucleotide query as accession, gi, or sequence in FASTA format" containing the sequence `>1smd` followed by the full protein sequence. To the right of the sequence is a "OPTIONS" panel with various search parameters. Below the search form is a section titled "Retrieve previous CD-search result" with a "Request ID:" input field and a "Retrieve" button. At the bottom, there is a "References:" section listing four scientific publications, and a footer with links for Help, Disclaimer, Write to the Help Desk, NCBI, NLM, and NIH.

Deixe as opções já marcadas no campo **Options**. Clique em **Submit**.

O primeiro resultado que é retornado é uma tela como a seguinte:

Conserved domains on [lcl|seqsig_GRTSI_ecec251a781e176dd56b5c0c54772bb4]

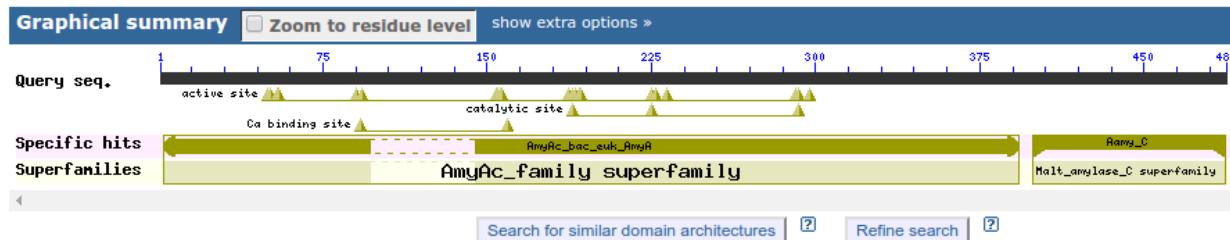
View Concise Results ▾ ?

1smd

Protein Classification

AmyAc_bac_euk_AmyA and Amy_C domain-containing protein (domain architecture ID 10183021)

AmyAc_bac_euk_AmyA and Amy_C domain-containing protein



List of domain hits

	Name	Accession	Description	Interval	E-value
[+]	AmyAc_bac_euk_AmyA	cd11317	Alpha amylase catalytic domain found in bacterial and eukaryotic Alpha amylases (also called 1, ...)	2-393	0e+00
[+]	Aamy_C	smart00632	Aamy_C domain;	399-487	1.84e-30

References:

- [1] Marchler-Bauer A et al. (2017), "CDD/SPARCLE: functional classification of proteins via subfamily domain architectures.", **Nucleic Acids Res.**45(D)200-3.
- [2] Marchler-Bauer A et al. (2015), "CDD: NCBI's conserved domain database.", **Nucleic Acids Res.**43(D)222-6.
- [3] Marchler-Bauer A et al. (2011), "CDD: a Conserved Domain Database for the functional annotation of proteins.", **Nucleic Acids Res.**39(D)225-9.
- [4] Marchler-Bauer A, Bryant SH (2004), "CD-Search: protein domain annotations on the fly.", **Nucleic Acids Res.**32(W)327-331.

[Help](#) | [Disclaimer](#) | [Write to the Help Desk](#)
[NCBI](#) | [NLM](#) | [NIH](#)

Nesta, podem ser identificados:

- A classificação da proteína (*Protein Classification*), com o link para o ID da arquitetura do domínio (do SPARCLE).
- A superfamília (*Superfamilies*) e os hits específicos (*Specific hits*) dentro desta superfamília.
- Os domínios identificados (*Domain hits*).

Pergunta: Quantos domínios esta proteína apresenta?

Na lista de domínios clique no primeiro e veja a descrição da família que contém este domínio. É uma tela como a representada abaixo:

Conserved Protein Domain Family

AmyAc_bac_euk_AmyA

HOME SEARCH SITE MAP

Entrez

CDD

Structure

Protein

Help

cd11317: AmyAc_bac_euk_AmyA ?



Alpha amylase catalytic domain found in bacterial and eukaryotic Alpha amylases (also called 1,4-alpha-D-glucan-4-glucanohydrolase)

AmyA (EC 3.2.1.1) catalyzes the hydrolysis of alpha-(1,4) glycosidic linkages of glycogen, starch, related polysaccharides, and some oligosaccharides. This group includes AmyA proteins from bacteria, fungi, mammals, insects, mollusks, and nematodes. The Alpha-amylase family comprises the largest family of glycoside hydrolases (GH), with the majority of enzymes acting on starch, glycogen, and related oligo- and polysaccharides. These proteins catalyze the transformation of alpha-1,4 and alpha-1,6 glucosidic linkages with retention of the anomeric center. The protein is described as having 3 domains: A, B, C. A is a (beta/alpha) 8-barrel; B is a loop between the beta 3 strand and alpha 3 helix of A; C is the C-terminal extension characterized by a Greek key. The majority of the enzymes have an active site cleft found between domains A and B where a triad of catalytic residues (Asp, Glu and Asp) performs catalysis. Other members of this family have lost the catalytic activity as in the case of the human 4F2hc, or only have 2 residues that serve as the catalytic nucleophile and the acid/base, such as Thermus A4 beta-galactosidase with 2 Glu residues (GH42) and human alpha-galactosidase with 2 Asp residues (GH31). The family members are quite extensive and include: alpha amylase, maltosyltransferase, cyclodextrin glycosyltransferase, maltogenic amylase, neopullulanase, isoamylase, 1,4-alpha-D-glucan maltotetrahydrolase, 4-alpha-glucotransferase, oligo-1,6-glucosidase, amylosucrase, sucrose phosphorylase, and amylo maltase.

Links	?
Source:	cd00551
Taxonomy:	root
PubMed:	19 links
Book:	1 link
Protein:	Representatives Specific Protein Related Protein Related Structure Architectures
Superfamily:	cl07893
BioSystems:	227 links

BioAssay Targets and Results	?
 CID: 444254 AID: 404692 GI: 1351933 IC ₅₀ : 0.996µM more	CID: 444254

Conserved Features/Sites

PubMed References ?

active site

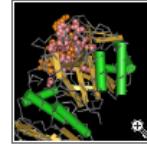
catalytic site

Ca binding site

Feature 1: active site [active site]

Evidence:

- **Structure:** 1G94: Pseudoalteromonas haloplanktis psychrophilic alpha amylase binds a hepta-saccharide and Tris, contacts at 4A
[- View structure with Cn3D](#)
- **Structure:** 1CPU: human pancreatic alpha-amylase binds acarbose, contacts at 4A
[- View structure with Cn3D](#)
- **Citation:** PMID 10769135
- **Citation:** PMID 11914073
- **Citation:** PMID 9571044

[Download Cn3D for Viewing 3D Structure](#)[Scroll to Sequence Alignment Display](#)

cd11317 is part of a hierarchy of related CD models.
 Use the graphical representation to navigate this hierarchy.
 cd11317 is a member of the superfamily cl07893.

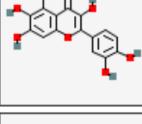
Role esta página até que o alinhamento de proteínas pertencentes a este CD seja mostrado. A opção padrão é o alinhamento dos membros mais diversos (*most diverse members*), ou seja, aqueles que apresentam sequências menos similares.

No exercício 104 foi ressaltado que 3 resíduos eram importantes para a atividade catalítica desta proteína. São os resíduos que estão marcados na sequência fasta acima. Verifique se os 3 resíduos estão conservados na sequência dos membros mais diversos no alinhamento obtido nesta página.

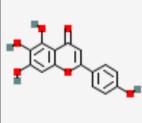
Podemos afirmar que os resíduos marcados são realmente característicos desta família?

Sim. Isto pode ser verificado ao clicar em **catalytic site**, na caixa **Conserved Features/Sites**.

A matriz de escores de posicionamento específico (PSSM) para esta família pode ser obtida na caixa **Statistics**, presente na lateral esquerda. Clique no link indicado pela seta, como representado na figura abaixo:



AID: 404692 **GI:** 1351933 **IC₅₀:** 10.2μM
[more](#)



CID: 5281697 **AID:** 404692 **GI:** 1351933 **IC₅₀:** 9.64μM
[more](#)

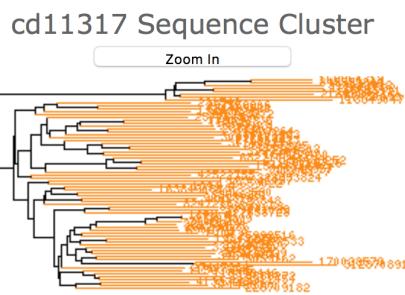
[Explore more](#)

Statistics ?

PSSM-ID: 200456
[View PSSM: cd11317](#) ←
 Aligned: 73 rows
 ThresholdBitScore: 356.488
 ThresholdSettingGi: 167535121
 Created: 25-Nov-2011
 Updated: 17-Jan-2013

Structure ?

Structure View
 Program: Cn3D
 Drawing: All Atoms
 Aligned Rows: up to 10
[Download Cn3D](#)



Sub-family Hierarchy

Interactive Display with CDTree [?](#)

cd00551	AnyRc_family
cd11313	AnyRc_arch_bac_AmyR
cd11314	AnyRc_arch_bac_plant_AmyR
cd11315	AnyRc_bac1_AmyR
cd11316	AnyRc_bac2_AmyR
cd11317	AnyRc_bac_euk_AmyR
cd11318	AnyRc_bac_fung_AmyR
cd11319	AnyRc_euk_AmyR
cd11320	AnyRc_AmyMalt_CGTase_like
cd11321	AnyRc_bac_euk_BE
cd11322	AnyRc_Glg_BE
cd11323	AnyRc_RGS
cd11324	AnyRc_Amylosucrase
cd11325	AnyRc_GTHase
cd11326	AnyRc_Glg_debranch
cd11327	AnyRc_Glg_debranch_2
cd11328	AnyRc_maltase
cd11329	AnyRc_maltase-like
cd11330	AnyRc_Oligo6Glu

Na nova página aberta, a matriz PSSM pode ser observada.

Resources

Learn Page

Amino Acid Explorer

PSSM Viewer Help

CDD Help

Show Color Key

Questions or comments

Scores

cd11317 : Alpha amylase catalytic domain found in bacterial and euk...

Change PSSM/Sequence Matrix View Reset Download Table to File Tutorial

Jump to position on consensus

Draw table showing only those positions where the consensus is Any

Draw table showing only this feature: active site

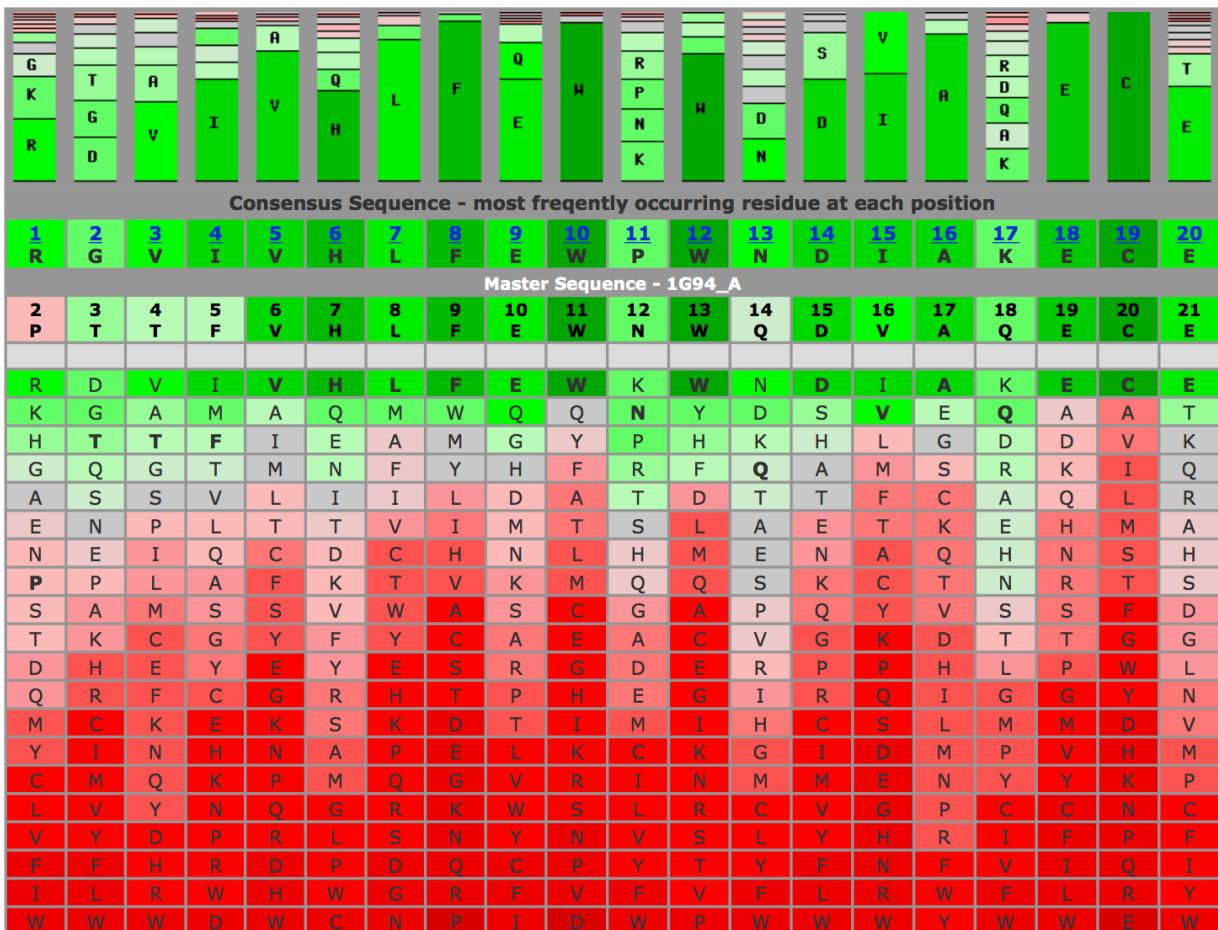
Master: 1G94_A View Master in CD

Click on any consensus position for a detailed view of that column.

Hide frequency bars

Scroll right

Click on any frequency bar to display detailed frequency data



Para cada posição da sequência das proteínas desta família, está denotada a possibilidade de mudanças entre os aminoácidos:

- Na primeira parte, a probabilidade de aquela posição conter cada aminoácido está representada na forma de barras empilhadas.
- Abaixo, temos os aminoácidos presentes na sequência consenso, ou seja, o resíduo que ocorre mais frequentemente naquela posição, quando considerado todas as proteínas desta família.
- Logo abaixo, temos a sequência *master* (mestre). A seqüência mestre é a primeira seqüência listada no alinhamento do CD. É uma proteína real, e é a seqüência à qual todas as outras seqüências no alinhamento do CD estão alinhadas. Sempre que possível, a seqüência mestre será uma seqüência com uma estrutura 3D resolvida (um PDB). No caso de CDs curados pelo NCBI, um CD ou um de seus CDs superiores sempre terá uma seqüência mestre com uma estrutura 3D resolvida.

No botão **Download Table File** (indicado por um quadro vermelho na figura acima), o download da PSSM pode ser realizado, para ser utilizado em uma busca utilizando o PSI-BLAST (Tutorial 105). Por exemplo, se o próximo passo é verificar os homólogos distantes da sequência **1smd**, presentes em sequências de amostras de metagenomas (para possíveis aplicações biotecnológicas), por meio de um PSI-BLAST, faremos assim (acompanhe também pelas figuras):

The screenshot shows the NCBI BLASTp suite interface. The search parameters are as follows:

- Enter Query Sequence:** The sequence is set to >1smd, which is highlighted in yellow.
- Query subrange:** The range is set from 1 to 1000.
- Job Title:** The title is set to 1smd.
- Choose Search Set:** The database is set to Metagenomic proteins(env_nr), which is highlighted in yellow. A red box highlights this field with the annotation: "Selecione aqui para utilizar o banco de sequências de proteínas obtidas a partir de metagenomas".
- Program Selection:** The algorithm selected is PSI-BLAST (Position-Specific Iterated BLAST), which is highlighted in yellow. A red box highlights this selection with the annotation: "Selecione aqui para utilizar o PSI-BLAST".
- Algorithm parameters:** This section is highlighted with a red box and an arrow pointing to it. A red box highlights the "Algorithm parameters" button. A red box highlights the note: "Note: Parameter values that differ from the default are highlighted in yellow and marked with + sign".
- Search results:** The search is set to use PSI-BLAST and to show results in a new window.

- Utilize o **BLASTp**, para executar uma busca contra o banco **env_nr** (*Metagenomic proteins*). Não esqueça de colar a sequência **1smd** no campo de busca.
- Na seção **Program Selection** selecione PSI-BLAST.
- Clique em **Algorithm parameters**.

Algorithm parameters

General Parameters

- Max target sequences: 500
- Short queries: Automatically adjust parameters for short input sequences
- Expect threshold: 10
- Word size: 3
- Max matches in a query range: 0

Scoring Parameters

- Matrix: BLOSUM62
- Gap Costs: Existence: 11 Extension: 1
- Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

- Filter: Low complexity regions
- Mask: Mask for lookup table only
 Mask lower case letters

PSI/PHI/DELTA BLAST

- Upload PSSM Optional: Selecionar Arquivo
- PSI-BLAST Threshold: 0.005
- Pseudocount: 0

BLAST | Search database Metagenomic proteins(env_nr) using PSI-BLAST (Position-Specific Iterated BLAST)
 Show results in a new window

BLAST is a registered trademark of the National Library of Medicine

[Support center](#) [Mailing list](#) [YouTube](#)

- Na seção **PSI/PHI/DELTA BLAST**, existe uma opção **Upload PSSM**. Clique para selecionar o arquivo e escolha o arquivo da PSSM salvo anteriormente, que possui o nome cd11317_res.txt que foi salvo anteriormente (também pode ser obtido pelo link).
- Clique em BLAST e verifique os resultados.

Usando o Pfam e o InterPro

O [Pfam](#) é um outro banco para obtenção de informações, funcionais, de famílias proteicas e domínios. A partir dele também é possível obter alinhamentos múltiplos de sequências de proteínas de uma mesma família (assim como o CDD) e obter perfis (*profiles*) de Hidden Markov Models (HMMs), que é uma das suas principais utilidades.

HMMs: É um modelo estatístico para qualquer sistema que pode ser representado como uma sucessão de transições entre estados discretos.

Quando usar o Pfam?

- Obter informações de famílias proteicas e domínios.
- Obter informações funcionais.
- Obter alinhamentos múltiplos de sequências de proteínas de uma mesma família.
- Obter perfis (*profiles*) *Hidden Markov Models*:

- É um modelo estatístico para qualquer sistema que pode ser representado como uma sucessão de transições entre estados discretos.

Nesta parte, também utilizaremos a sequência `1smd`. Para isso siga os passos abaixo:

The screenshot shows the Pfam 31.0 interface. In the sequence search field, the string `1smd` is entered. Below it, the sequence is displayed in its entirety: `GRTSIVHLFEWRVVDIALECERYLAPKGFGGVQVSPPNENVAHNPFPRWWERYOPVSYKLCTRSGNEDFRNMVTRCNNVGVRIVYDAVINHMCNCNAVSAGTSTCGSFNPGRDFPAPVPSGWDFTNDGKCKTGSDIENYNDATQVRDCRLSGLLDLALGKDYVRSKJ`. A "Go" button is visible next to the sequence input field.

Recent Pfam blog posts

- [Pfam 31.0 is released](#) (posted 8 March 2017)
- [Pfam train online](#) (posted 8 December 2016)
- [Pfam 30.0 is available](#) (posted 1 July 2016)

Pfam 31.0 (March 2017, 16712 entries)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. [More...](#)

- Abra a página do [Pfam](#).
- Clique em **Sequence Search**.
- Copie a sequência da proteína `1smd` no campo indicado.
- Espere os resultados.

A primeira página de resultados será a seguinte:

Sequence search results

Show the detailed description of this results page.

We found **2** Pfam-A matches to your search sequence (**all** significant)

Alpha-amylase

Alpha-amylase_C

Significant Pfam-A Matches

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To					
Alpha-amylase	Alpha amylase, catalytic domain	Domain	CL0058	58	324	64	302	47	298	337	47.1	2.1e-12	n/a	Show
Alpha-amylase_C	Alpha amylase, C-terminal all-beta domain	Domain	CL0369	398	485	401	484	4	95	96	54.7	9.5e-15	n/a	Show

Comments or questions on the site? Send a mail to pffm-help@ebi.ac.uk.
European Molecular Biology Laboratory

- Clique agora no clam CL0058, correspondente as alfa-amilases.

- Observe todos os resultados, clicando no menu de navegação a direita.

*Veja principalmente os itens **Domain organisation**, **HMM Logo** e **Alignments**.*

Family: Alpha-amylase (PF00128)

Summary

Domain organisation

Clan

Alignments

HMM logo

Trees

Curation & model

Species

Interactions

Structures

Jump to... ↴

enter ID/acc **Go**

Summary: Alpha amylase, catalytic domain

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

[Wikipedia: Alpha-Amylase](#) [Wikipedia: Glycoside hydrolase family 13](#) [Pfam](#) [InterPro](#)

This is the Wikipedia entry entitled "Alpha-Amylase". [More...](#)

Alpha-Amylase [Edit Wikipedia article](#)

Redirect to:

- [Alpha-amylase](#)
- **From other capitalisation:** This is a redirect from a title with another method of capitalisation. It leads to the title in accordance with the [Wikipedia naming conventions for capitalisation](#), or it leads to a title that is associated in some way with the conventional capitalisation of this redirect title. This may help writing, searching and international language issues.
 - If this redirect is an incorrect capitalisation, then {{R from miscapitalisation}} should be used instead, and pages that use this link should be updated to link directly to the target. Miscapitalisations can be tagged in any namespace.
 - Use this rcat to tag only mainspace redirects; when other capitalisations are in other namespaces, use {{R from modification}} instead.

This page is based on a [Wikipedia article](#). The text is available under the [Creative Commons Attribution/Share-Alike License](#).

Comments or questions on the site? Send a mail to pfpam-help@ebi.ac.uk.
European Molecular Biology Laboratory

Na parte **Summary** há uma aba que indica o código **InterPro** para esta família proteica, que é [IPR006047](#).

No InterPro informações similares estão também disponíveis. Neste banco, uma busca a partir da sequência também pode ser realizada.

Quando usar o InterPro

- Para obter informações gerais sobre uma proteína.
- Obter informações sobre a estrutura e resíduos importantes para a atividade.
- Obter informações ligadas com outros bancos de dados biológicos.

Qual banco utilizar?

Esta não é uma pergunta trivial. É importante que você explore as informações disponíveis em cada banco e extraia o máximo de informações sobre a proteína de interesse. Como pode ser notado, embora exista uma redundância nas informações disponíveis, cada banco tem sua especificidade e particularidade.

Usando o PROSITE

O banco de dados [PROSITE](#) é um outro banco muito útil para obter informações funcionais e de famílias proteicas e domínios.

Quando o usar o PROSITE?

- Para obter informações de famílias proteicas e domínios.

- Para obtenção de informações funcionais.
- Obter as assinaturas de sequência que caracterizam as famílias de proteínas.
- Usar para buscas de similaridade utilizando o PHI-BLAST.

Para este exemplo, iremos utilizar a sequência abaixo:

```
>Enzyme_Test_1
MVKIVTVKTQAYQDQKPGTSLRKRVKFQSSANYAENFIQSIIISTVEPAQRQEATLVVGGDGRFYMKEAIQLIARIA
AANGIGRLVIGQNGILSTPAVSCIIRKIKAIIGGIILTASHNPGGPNDFGIKFNISNGGPAPEAITDKIFQISKTIEE
YAVCPDLKVDLGVLGKQQFDLENKFKPFTVEIVDSVEAYATMLRSIFDFSALKELLSGPNRLKIRIDAMHGVVGPYVK
KILCEELGAPANSAVNCVPLEDFGGHHPDPNLTYAADLVETMKSGEHDFGAAFDGDGDRNMILGHGFFVNPSDSVAV
IAANIFSIPYFQQTGVRGFARSMPMTSGALDRVASATKIALYETPTGWKFFGNLMDASKLSLCGEESFGTGSDHIREKD
GLWAVLAWLSILATRKQSVEDILKDHWQKYGRNFFTRYDYEEVEAEGANKMMKDLEALMFDRSFVGKQFSANDKVYTV
EKADNFEYSDPVDGSI SRNQGLRLIFTDGSRIVFRLSGTGSAGATIRLYIDSYEKDVAKINQDPQVMLAPLISIALKV
SQLQERTGRTAPTVIT
```

Passos:

- Abra a página inicial do [PROSITE](#).
- Na caixa **Quick Scan mode of ScanProsite**, cole a sequência acima.
- Marque a opção **Exclude motifs with a high probability of occurrence from the scan**.

Esta opção serve para excluir motivos lineares na sequência proteica que são muito comuns, em inúmeras proteínas.

- Clique em **Scan**.
- Observe os resultados, que devem estar de acordo com a figura abaixo:



ScanProsite Results Viewer

Output format: Graphical view - this view shows ScanProsite results together with ProRule-based predicted intra-domain features [help].

Hits for all PROSITE (release 2017_04) motifs on sequence Enzyme_Test_1 :

found: 1 hit in 1 sequence

Enzyme_Test_1 (562 aa)

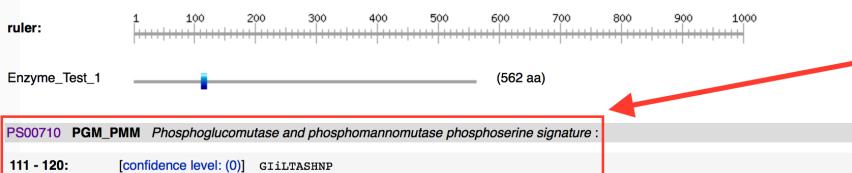
```
MVKIVTVKTQAYQQKPGTSGLRKRVFQSSANYAENFIQSIIISTVPEPAQRQEATLVVGGDGRFY  
MKFAIQLIARIAAAANGIGRLRIVIGONGILSTPAVSCIIRRKIAIGGIIITASHNPFGPNPQDFGIKFNF  
ISNGGPAPAEAITDKIFQISKTEEYAVCPDLKVLDLGVLGKQDFDENKKFPFTVEIVDSVEAYATM  
LRSIFDFSAKELLSPNRLKRKIRDAMHGVVGPYKKILCEELGAPANSAVCVPLLEDFFGGHHPPD  
NLTYAAIDLVEITMKSGEHDIFGAFAFDGGDRNMILGKGHGFVNPNSDVSVAIANISIPYFQQTFVGRG  
FARSMPGSGALDRVASATKIALYETPTGWKFPGNLMASKSLCGEESFTGSDHIREKDGLWAVL  
AWLSILATRKQSVEDILKDHWQKYGRNFFTTRYDYEVEAEAGANKMMKDLEALMFDRSFVKGQFSAN  
DKVYTVEKADNFYEVDPVGSISRNQGLRLIFTDSRIVFRLSGTGSAAGATIRLYIDSYEKDVAKI  
NQDPQVMLAPLISALKVSQLQERTGRATPTVIT
```

Legend:



Please note that the graphical representations of domains displayed hereafter are for illustrative purposes only, and that their colors and shapes are not intended to indicate homology or shared function. For more information about how these graphical representations are constructed, go to <http://prosite.expasy.org/mydomains/>.

hits by patterns: [1 hit (by 1 pattern) on 1 sequence]



Os resultados mostram que entre as posições 111 e 120 há a assinatura de enzimas do tipo fosfoglucomutase e fosfomannomutase fosfoserine. Clique no link indicado acima ([PS00710](#)) e veja as informações deste motivo.

Nesta mesma página, mais abaixo, temos um quadro denominado **PGM_PMM, PS00710; Phosphoglucomutase and phosphomannomutase phosphoserine signature (PATTERN)**. Nele temos o seguinte consenso padrão:

Consensus pattern:
[GSA]-[LIVMF]-x-[LIVM]-[ST]-[PGA]-S-H-[NIC]-P

Esta é assinatura PROSITE deste tipo de proteínas. Ela pode ser usada em uma busca PHI-BLAST (ver tutorial 105) para identificar proteínas que tenham esta assinatura em buscas de similaridade utilizando o BLAST.

Vamos a um exemplo?

Você quer verificar quais proteínas obtidas a partir de amostras ambientais de metagenomas possuem esta assinatura de fosfoglucomutase, para uma possível aplicação biotecnológica. Para isso, siga os passos abaixo (acompanhe também pela figura):

U.S. National Library of Medicine > NCBI National Center for Biotechnology Information

Sign in to NCBI

BLAST® > blastp suite

Home Recent Results Saved Strategies Help

Standard Protein BLAST

blastn blastp blastx tblastn tblastx

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

>Enzyme_Test_1
>MVKIVTVKTKQYQDQKPGTSGLRKRVKFQSSANYAENFIQSIIISTVEPAQRQEATLVGGDGRFYMK
KEAQIOLARI
AANGIGRLVIGQNGLSTPAVSCIRKIAIGGIILTASHNPGGPNGDFGIKFNISNGGPAPEITDKIFQI
SKTIEE

Or, upload file nenhum arquivo selecionado

Job Title Enzyme_Test_1
Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database Metagenomic proteins(env_nr)

Organism Optional Enter organism name or id—completions will be suggested Exclude

Exclude Optional Models (XM/XP) Uncultured/environmental sample sequences

Entrez Query Optional Enter an Entrez query to limit search

Program Selection

Algorithm blast (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST) Cole aqui a assinatura obtida no PROSITE
Enter a PHI pattern
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm

BLAST Search database Metagenomic proteins(env_nr) using PHI-BLAST (Pattern Hit Initiated BLAST) Show results in a new window

- Utilize o [BLASTp](#), para executar uma busca contra o banco **env_nr** (*Metagenomic proteins*). Não esqueça de colar a sequência **Enzyme_Test_1** no campo de busca.
- Na seção **Program Selection** selecione PHI-BLAST. Ao clicar em PSI-BLAST, uma caixa abaixo é aberta. Nela você insere a assinatura PROSITE acima.
- Clique em BLAST e espere os resultados (Pode demorar!!!).
- Verifique os resultados.

Pelos resultados, há algo promissor para aplicação?

One More Thing

O site [Jena Library](#) agrega informações de proteínas conhecidas de vários bancos de dados (alguns vistos neste próprio tutorial).

jenalib.leibniz-fli.de

Jena Library
of Biological Macromolecules

collapse expand

QuickSearch: go
by PDB,NDB,UniProt,PROSITE Code or Search Term(s)

Home Contact Hetero DB Site DB GO2PDB Genus/Species Classification Trees Entry Lists Search

Atlas of Macromolecule Structures

■ Search
■ Database Subsections
■ Entry Lists
■ Protein Domain Classification
■ Analysis Tools
■ Cross References
■ Help

The Jena Library of Biological Macromolecules (JenaLib) is aimed at a better dissemination of information on three-dimensional biopolymer structures with an emphasis on visualization and analysis.
It provides access to all structure entries deposited at the Protein Data Bank ([PDB](#)) or at the Nucleic Acid Database ([NDB](#)).
In addition, basic information on the architecture of biopolymer structures is available.
The JenaLib intends to fulfill both scientific and educational needs.

Basic Information on Biological Macromolecules

■ General
■ Proteins
■ Nucleic Acids

News | Gallery

Copyright 1993-September 2005, Institute of Molecular Biotechnology (IMB), Jena / Germany; October 2005-September 2015, Leibniz Institute for Age Research - Fritz Lipmann Institute (FLI) [formerly IMB Jena]; since October 2015, Leibniz Institute on Aging - Fritz Lipmann Institute (FLI).



Beutenbergstraße 11 Phone: +49 3641 65-6000
D-07745 Jena • Germany Fax: +49 3641 65-6351

e-mail: info@leibniz-fli.de
www.leibniz-fli.de



Para fazer um teste, coloque o código `1smd` no campo **QuickSearch** no canto superior direito da página e clique em **Go**. Na página seguinte, informações presentes em outros bancos de dados de proteínas serão retornadas, com os links para a página específica da entrada.