

## **Identificação de Padrões em Proteínas Utilizando a Ferramenta de Bioinformática CD-Search**

Gustavo José da Silva<sup>1</sup>, Mozart Marins<sup>2</sup> e Ana Lucia Fachin<sup>3</sup>

<sup>1</sup>Universidade de Ribeirão Preto, Ribeirão Preto, SP [gustavo@eafmuz.gov.br](mailto:gustavo@eafmuz.gov.br) <sup>2</sup>Universidade de Ribeirão Preto, Ribeirão Preto, SP [mmarins@unaerp.br](mailto:mmarins@unaerp.br) <sup>3</sup>Universidade de Ribeirão Preto, Ribeirão Preto, SP [afachin@unaerp.br](mailto:afachin@unaerp.br)

### **Introdução**

Proteínas são macromoléculas biológicas que são formadas por cadeias de moléculas orgânicas. Essas moléculas são denominadas de aminoácidos. Existem vinte aminoácidos diferentes. Cada aminoácido é formado por três nucleotídeos e cada nucleotídeo pode ser de quatro tipos: adenina, timina, citosina e guanina.

O DNA armazena o código genético. Essa informação é transcrita para o RNA, que contém a sequência específica de aminoácidos. As proteínas são sintetizadas em um processo chamado tradução do RNA. Essa série de eventos caracteriza o dogma central da biologia molecular.

A sequência de nucleotídeos determina os aminoácidos que farão parte da proteína. A sequência de aminoácidos determina a proteína. A estrutura da proteína determina sua função. A ferramenta de bioinformática CD-Search permite analisar uma proteína, utilizando sua sequência de aminoácidos visando identificar padrões. Os padrões são sequências de aminoácidos que determinam um domínio conservado. Cada domínio identificado é responsável por executar uma determinada função.

Segundo Bork (1993), uma proteína secretada por uma bactéria pode possuir padrões que são identificados também em eucariotos.

Atualmente, muitas informações biológicas estão armazenadas em bancos de dados. Marchler-Bauer et al. (2009) relata que bancos de dados de domínios conservados podem ser acessados no portal do NCBI (National Center for Biotechnology Information).

Existem bancos de sequências de nucleotídeos, como o Genbank. Bancos de sequências de proteínas, como o SWISS-PROT e bancos de estrutura tridimensional de proteínas, como o PDB. Muitas informações foram geradas, o desafio agora é analisar e tomar decisões sobre essas informações.

A ferramenta CD-Search permite entrada de arquivo fasta, que traz a sequência de aminoácidos da proteína. Pode ser analisada uma sequência por vez, gerando um arquivo de

saída disponível para download em uma variedade de formatos. A análise pode ser vista em formato gráfico, que mostra a posição de todos os domínios identificados na sequência de aminoácidos.

O objetivo deste trabalho é mostrar como a ferramenta de bioinformática CD-Search pode identificar padrões em uma sequência de aminoácidos.

### **Materiais e Métodos**

Para analisar sequências de proteínas, é necessário agrupá-las em famílias. Para cada família, relacionam-se características comuns expressas por suas sequências de aminoácidos. Essas características podem ser chamadas de padrões. As proteínas são classificadas em uma família, geralmente, por mais de um padrão. As similaridades entre as proteínas podem ser em relação às suas sequências de aminoácidos, suas formas estruturais ou por suas funcionalidades. Padrões identificados na evolução das proteínas podem estar relacionados estrutural e funcionalmente. Portanto, identificar padrões é identificar características na sequência que são comuns a um grupo de proteínas que exercem determinada atividade e ausentes em grupos que não exercem essa mesma atividade.

Muitas informações sobre sequências identificadas estão disponíveis nos depositários a seguir.

Arquivos de sequências de ácidos nucleicos:

- GenBank, situado no US National Center for Biotechnology Information (NCBI);
- EMBL Nucleotide Sequence Database, localizado no European Bioinformatics Institute (EBI);
- The Center for information Biology e DNA DataBank of Japan no National Institute of Genetics.

Arquivo de sequências de aminoácidos de proteínas, determinado exclusivamente pela tradução de sequências de genes:

- United Protein Database (UniProt);
- SWISS-PROT;
- The Protein Identification Resource (PIR);
- Translated EMBL (TrEMBL);
- Sequence Retrieval System (SRS) – seleção e recuperação de sequências.

A ferramenta CD\_Search possui uma interface amigável, sendo facilmente operacionalizada. Pode ser acessada pelo link <http://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi>.

Conforme explica Marchler-Bauer e Bryant (2004), o CD-Search é uma interface que utiliza o programa RPS-BLAST, que significa "BLAST Posição Específica Reverse", para identificar os padrões.

Segundo Marchler-Bauer et al. (2011), o CD-Search utiliza o banco de dados de domínio conservado (CDD). O CDD trabalha com modelos de domínio importados de várias bases de dados externas como PFAM, SMART, COG, PRK e TIGRFAM.

Na interface, a proteína que será analisada deve ser inserida na caixa de texto específica. É preciso inserir uma sequência fasta ou uma sequência de aminoácidos. O usuário pode, também, buscar o arquivo fasta da proteína que será submetido para a análise. Depois basta clicar em submeter para que o programa faça a análise. Após a realização dos cálculos será exibida uma página de resumo.

NCBI

HOME SEARCH GUIDE Structure Home 3D Macromolecular Structures Conserved Domains Pubchem BioSystems

Search for Conserved Domains within a protein or coding nucleotide sequence

**NEW!** Use Batch CD-search to submit multiple query proteins at once!

Enter protein or nucleotide query as accession, gi, or sequence in FASTA format

Submit Reset

**OPTIONS**

Search against database: CDD -- 43088 PSSMs

Expect Value threshold: 0.01

Apply low-complexity filter: ☒

Force live search: ☐

Maximum number of hits: 500

Result mode: ☒ Concise ☐ Full

Retrieve previous CD-search result

Request ID: Retrieve

**References:**

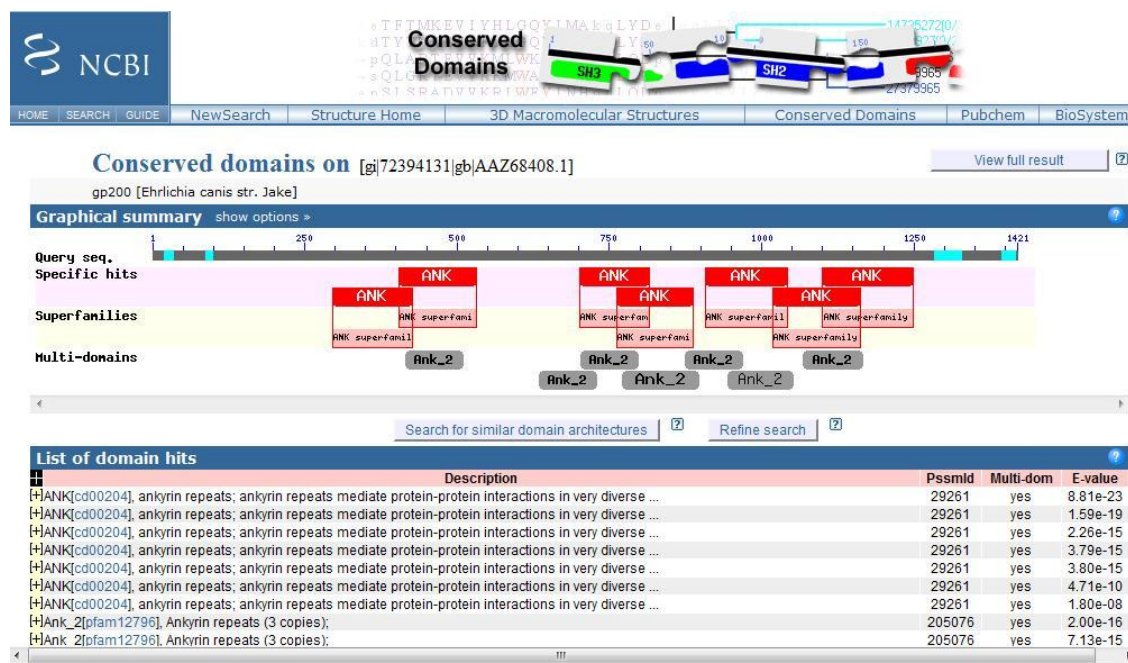
Marchler-Bauer A et al. (2011), "CDD: a Conserved Domain Database for the functional annotation of proteins.", *Nucleic Acids Res.*39(D)225-9.

Marchler-Bauer A et al. (2009), "CDD: specific functional annotation with the Conserved Domain Database.", *Nucleic Acids Res.*37(D)205-10.

**Figura 1.** Ferramenta CD-Search para identificação de padrões em proteínas.

## Resultados e Discussão

Após submeter uma sequência ou um arquivo fasta, o CD-Search apresenta o resultado da análise, conforme a figura a seguir.



**Figura 2.** Resultado da análise de identificação de padrões no CD-Search.

Foi submetida uma proteína secretada pela bactéria *Ehrlichia canis*. Essa proteína está catalogada no Genbank com a identificação 72394131. Ela é conhecida por possuir padrões de anquirina. Ela foi escolhida para sabermos onde estão localizados os padrões de anquirina.

Bork (1993) relata que proteínas com padrões anquirina, secretadas por bactérias, podem imitar o comportamento de um eucarioto.

A repetição de anquirina é um dos mais comuns, modulares, motivos de interação proteína-proteína na natureza (MOSAVI; MINOR; PENG, 2002).

A parte superior da figura mostra como a ferramenta apresenta a forma gráfica para apresentação dos padrões identificados. A ferramenta CD-Search identificou os padrões de anquirina e sinalizou cada padrão encontrado dentro da sequência dos aminoácidos.

Na parte inferior da tela são colocados todos os padrões identificados e suas descrições. Pode-se também obter informações detalhadas sobre cada padrão. Para isso basta clicar no padrão desejado.

## **Conclusões**

A bioinformática necessita de ferramentas para ajudar analisar e interpretar muitas informações já levantadas em laboratórios de biologia. Muitas dessas informações estão armazenadas em base de dados disponíveis gratuitamente. A ferramenta CD-Search apresentada permite uma análise eficiente e eficaz sobre padrões de proteínas. Com isso, pode-se contribuir para fabricação de fármacos para combater doenças causadas por proteínas secretadas por microorganismos.

## **Agradecimento**

Ao Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais pelo auxílio financeiro.

## **Referências Bibliográficas**

MARCHLER-BAUER, A. et al., **CDD: a Conserved Domain Database for the functional annotation of proteins.**, *Nucleic Acids Res.*39(D)225-9, 2011.

MARCHLER-BAUER A. et al., **CDD: specific functional annotation with the Conserved Domain Database.**, *Nucleic Acids Res.*37(D)205-10, 2009.

MARCHLER-BAUER A.; BRYANT, S. H., **CD-Search: protein domain annotations on the fly.**, *Nucleic Acids Res.*32(W)327-331, 2004.

BORK, P. **Hundreds of ankyrin-like repeats in functionally diverse proteins: mobile modules that cross phyla horizontally** *Proteins* 17, 363–364, 1993.

MOSAVI, L. K., MINOR, D.L., Jr and PENG, Z.Y. **Consensus-derived structural determinants of the ankyrin repeat motif.** *Proc. Natl. Acad. Sci. U. S. A.* 99, 16029–16034, 2002.