Instituto de Ciências Biomédicas / Universidade de São Paulo
Elisabete José Vicente

Introdução à Bioinformática - do DNA à proteína

CAPÍTULO 4: Banco de Dados Secundários ("SECONDARY DATA BASES")

Roteiro Teórico-prático

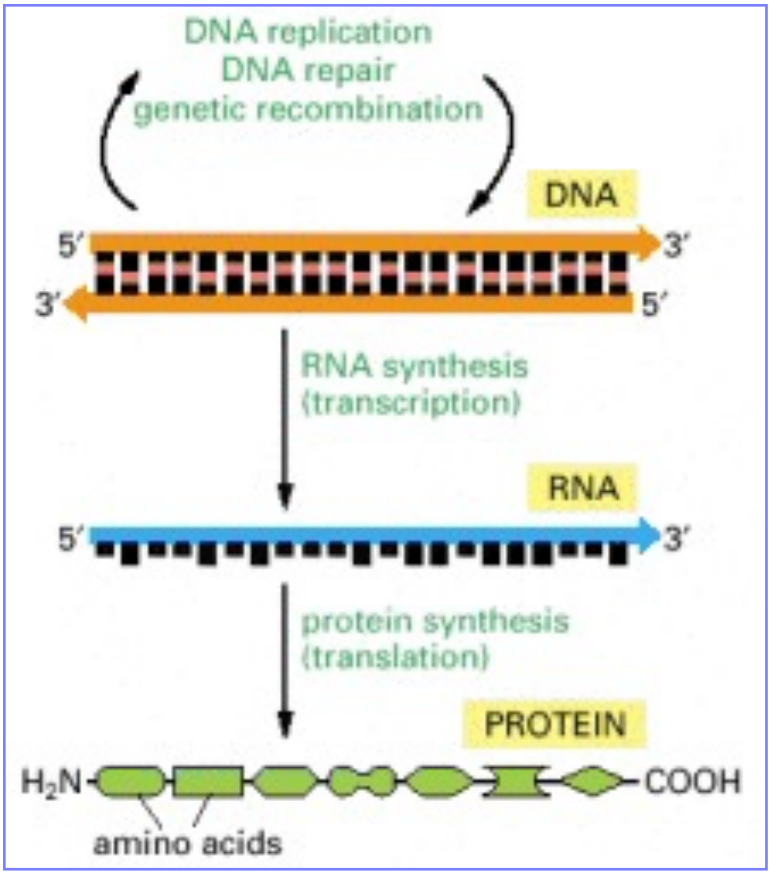


Fig 1: **Dogma central da Biologia:** O fluxo de informação genética do **DNA** para o RNA (**transcrição**) e do RNA para a **proteína** (**tradução**) ocorre em todas as células vivas. Copyright © 2002, Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter; Copyright © 1983, 1989, 1994, Bruce Alberts, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts, and James D. Watson
(Fonte: https://www.ncbi.nlm.nih.gov/books/NBK21050/figure/A974/)

Adaptado de: Anna V. Protasio, Christine Boinett, Martin Aslett, Matthew Dorman et al. Welcome Genome Campus Advanced Courses and Scientific Conferences.

2020

2020.07.21

INTRODUÇÃO À BIOINFORMÁTICA – DO DNA A PROTEÍNA
CAPÍTULO 4: Banco de Dados Secundários ("SECONDARY DATA BASES")

1. Introdução - O que genes e proteínas estão fazendo nos diferentes organismos?

Os bancos de dados biológicos são recursos centralizados que contêm representações de sequências de DNA e proteínas e suas informações associadas. Os bancos de dados primários armazenam e disponibilizam dados ao público, atuando como repositórios.

Os **Bancos de Dados Secundários** usam dados de sequência disponíveis publicamente nos **bancos de dados primários** para fornecer camadas de informações aos dados de sequência de DNA ou proteína.

Já discutimos bancos de dados primários ou repositórios para sequências de nucleotídeos, **GenBank** (**NCBI**), **ENA** (**EMBL-EBI**) e **DDBJ**. O papel dos bancos de dados primários não é restrito às sequências de nucleotídeos, sequências de proteínas e outros tipos de dados podem ser submetidos a alguns bancos de dados primários. Dois exemplos que pesquisadores em bioinformática usam regularmente incluem:

(i) WorldWide Protein Data Bank, um recurso onde os dados estruturais tridimensionais de proteína (e também ácidos nucleicos) podem ser depositados e disponibilizados ao público; e

(ii) **Uniprot** - um banco de dados primário para sequências de proteínas e anotação funcional com base em evidências experimentais - que já discutiremos na próxima etapa.

Os **bancos de dados secundários** compreendem dados derivados da análise de entradas nos bancos de dados primários. Na maioria dos casos, eles também fornecem ferramentas para investigar mais profundamente os genes e as proteínas. Eles trabalham analisando dados pré-existentes (por exemplo, **todas as sequências de proteínas já enviadas** ou a **tradução conceitual de todas as sequências de nucleotídeos**) e coletam, quando possível, informações sobre a função dessa sequência.

Os bancos de dados secundários **cobrem informações adicionais**, comumente derivadas de suas próprias análises, levando em conta (considerando) uma característica específica da proteína ou sequência; por exemplo, a ocorrência de um **sítio catalítico enzimático** ou **um sítio para modificação da proteína**.

Muitos bancos de dados secundários são aplicados às sequências de proteínas em vez de sequências de nucleotídeos e alguns exemplos são dados nas próximas etapas.

Questão 1: Muitos bancos de dados secundários são aplicados às sequências de proteínas, e não às sequências de núcleos, sabe por quê?

a) Proteínas têm estrutura 3D

b) Número de acesso à sequência é mais fácil comparar proteínas

c) Conservação da sequência de aminoácidos

d) Organismos têm diferentes preferências de uso de códons

e) Para as buscas em informática a pesquisa de aminoácidos fica reduzida à1/3 porque 3 nucleotídeos codificam 1 aminoácido

Resposta: **d)** O códon é degenerado e os organismos têm diferentes preferências de uso do códon. (Veja: uso de códons Capítulo 1)

2. Bancos de dados secundários de proteínas

2.1. Bancos de dados de proteínas – O Consórcio UNIPROT

O **Uniprot f**oi originalmente formulado como um **banco de dados primário** para sequências de proteínas e anotação funcional com base em evidências experimentais. **Atualmente**, **combina uma rede de bancos de dados irmãos centralizando todos os níveis de anotações produzidos para sequências de proteínas**. Ele **também** contém **traduções conceituais** derivadas da conclusão contínua de milhares de projetos de genoma. O **UniProt** é **atualizado regularmente**, fornecendo uma **nova atualização** a cada poucas semanas.

A página inicial do **UniProt** (UniProt homepage) fornece acesso para todas as sequências anotadas em dois bancos de dados que contêm entradas de proteínas:
- **SwissProt** – manualmente anotadas e revisadas (com curadoria manual) e,
- **TrEMBL** - anotadas automaticamente e não revisadas.

Como é formado o Consórcio UNIProt:

("**The Universal Protein Resource** (**UniProt**) is a comprehensive resource for protein sequence and annotation data. The UniProt databases are the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), and the UniProt Archive (UniParc). The **UniProt consortium** and host institutions **EMBL-EBI (London, UK)**, **SIB** (and **PIR** are committed to the long-term preservation of the UniProt databases.

UniProt is a collaboration between the European Bioinformatics Institute (EMBL-EBI), the SIB Swiss Institute of Bioinformatics and the Protein Information Resource (PIR)- EUA. Across the three institutes more than 100 people are involved through different tasks such as database curation, software development and support.

EMBL-EBI and **SIB** together used to produce **Swiss-Prot** and **TrEMBL**, while **PIR** produced the Protein Sequence Database (**PIR-PSD**). These two data sets coexisted with different protein sequence coverage and annotation priorities:
- **TrEMBL** (Translated EMBL Nucleotide Sequence Data Library) was originally created because sequence data was being generated at a pace that exceeded Swiss-Prot's ability to keep up.
- **PIR** maintained the PIR-PSD and related databases, including **iProClass**, a database of protein sequences and curated families.

In **2002** the three institutes decided to pool their resources and expertise and formed the **UniProt consortium**.")
Na página de entrada **UNIPROT**, surgem as entradas para: **Wikipedia**, **Pfam** e **Interpro.**

O vídeo tutorial UniProt help, do próprio site, ajudará muitos em suas pesquisas iniciais.

Para mais informações sobre o banco de **dados UniProt**, leia a publicação: "2015, UniProt: a hub for protein information"

4. Ferramentas e recursos no UniProt
Agora vamos investigar um exemplo de entrada do UniProt e discutir a entrada.
Você usará o gene **bamE** como exemplo. Procurando por "**bamE**" na caixa de pesquisa do **UniProtKB**. Para acessar o UniProt, clique aqui: http://www.uniprot.org/ e digite "**bamE**" na caixa de pesquisa. Use o banco de dados **UniprotKB**.

A página de resultados, apresentada na imagem, mostra em uma seção de uma tabela de entradas que lista todos os genes correspondentes "**bamE**" para diferentes espécies. Usando sua pesquisa por **bamE** no **UniProtKB**: "Clique" na entrada "**P0A937**" correspondente à proteína **E. coli BamE.** Isso levará você à anotação completa desta proteína. A página está cheia de informações, com uma faixa do lado esquerdo para ajudar na navegação.

Desafio: Exploração inicial UniProt

Pesquise no UniProt informações sobre a proteína bamE (clique aqui: http://www.uniprot.org/ e digite "**bamE**" na caixa de pesquisa).

1. Quais informações você obtém?
2. Qual é a proteína da primeira entrada?

Usando os links encontrados nas várias **referências cruzadas** (como **Pfam** e **Interpro**) para acessar mais informações sobre essa proteína em outros sites, tente responder as seguintes perguntas e deixe abaixo os seus comentários:

3. Onde você pode encontrar informações sobre os domínios conservados encontrados nesta proteína?
4. Qual link você deve usar para acessar uma lista de publicações relacionadas a esta proteína
5. O que é PTM?

Algumas Respostas:

1. Entradas de milhares de proteínas, sendo que 25 na primeira página.

2. PAO937 BAME_E.coli. Outer membrane protein assembly factor BamE. Fico sabendo que:

1) é uma proteína de E. coli
2) é uma proteína típica de bactérias Gram-negativas;
3) Faz parte da membrana externa;
4) Nonessential member of the complex that stabilizes the interaction between the essential proteins BamA and BamD. May modulate the conformation of BamA, likely through interactions with BamD. Efficient substrate folding and insertion into the outer membrane requires all 5 subunits (PubMed:20378773, PubMed:21823654, PubMed:27686148).

3 FAMILY & DOMAINS

4) Devera analisar todas entradas e especificamente buscar em: "**Cross-references**"

5) PTM – post-transcriptional modification

Faça alguns alinhamentos de sequencias de proteínas.

Dica: Assista o vídeo tutorial UniProt help . Pesquise as proteínas:

- recA – Em que organismos está presente?
- recA - Comparando com a sequencia recA de E. coli, quais são os organismos onde a sequencia é mais similar?

- lexA: - Em que organismo esta presente?

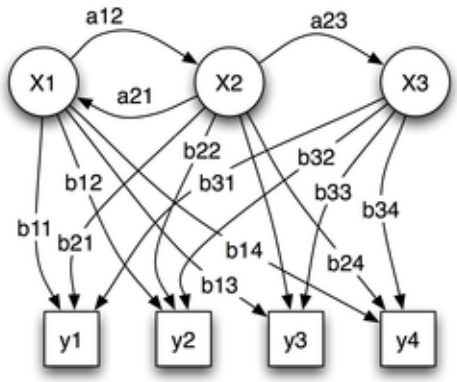
Obs.: Por que as sequencias recA e lexA são tão conservadas? Você saberia explicar?

- Compare recA de E. coli, rad51, rad51 de yeast – Que informações você obtém?

2.2. Outros bancos de dados de proteínas: **Pfam, Interpro** do EMBL-EBI

Dois dos bancos de dados secundários mais populares **reconhecem domínios proteicos conservados** dentro de uma sequência de proteínas. Esses bancos de dados são Pfam e Interpro e eles são hospedados pelo **EMBL-EBI**.

O **Pfam** é um banco de dados com **curadoria manual**, ou seja, um pesquisador pessoalmente construiu as diferentes "famílias" nas quais as proteínas com os mesmos domínios conservados foram classificadas. O **Pfam** é um grande banco de dados de **grupos de famílias de proteínas** que compartilham **domínios conservados**. Emprega como modelo estatístico Modelos Ocultos de Markov ("Hidden Markov Models" - **HMM**).



Em matemática, uma cadeia de Markov é um caso particular de processo estocástico com estados discretos com a propriedade de que a distribuição de probabilidade do próximo estado depende apenas do estado atual e não da sequencia de eventos precedentes (Markoviana), chamado assim em homenagem ao matemático Andrei Andreyevich Markov.

Ao lado:
Exemplo de parâmetros probabilísticos de um **modelo oculto de Markov**
x — estados
y — possíveis observações
a — Probabilidades de transição de estado
b — probabilidades de saídas

O **Interpro** é uma coleção muito maior de muitos bancos de dados. Ele pega um grande número (cerca de onze!) de algoritmos de reconhecimento de domínios proteicos e os centraliza em uma única ferramenta. Os algoritmos individuais são altamente especializados e diversos em suas previsões, e levaria muito tempo para passar por cada um todos eles, um de cada vez. A ferramenta **Interpro** permite economizar tempo e muito 'clicar' e 'copiar/colar', fornecendo um único portal para consultar todos esses bancos de dados valiosos.

Embora **Pfam** e **Interpro** tenham surgido de forma independente, agora eles estão interligados: o **Pfam é um dos bancos de dados usados pela Interpro quando se procura domínios conservado**s, e as entradas do **Pfam** têm uma guia ("tab containing") que contém a descrição do **Interpro** para o mesmo domínio conservado, conforme mostrado neste exemplo

2.3. Outros bancos de dados de proteínas: Phobius

Outros bancos de dados secundários pesquisam assinaturas associadas à localização **subcelular de proteínas** ou à **classificação de proteínas**. Essas assinaturas definem se uma proteína é:

- **Retida** no citoplasma,
- **Posicionada na membrana celular,** ou
- **Secretada no espaço extracelular**.

Não é de surpreender que essas assinaturas sejam conservadas e, portanto, possam ser analisadas de maneira semelhante às de **domínios conservados**. Por exemplo, comparando-se a sequência de várias proteínas secretadas, é possível descobrir qual é parte de sua sequência que determina sua secreção.

Duas assinaturas em proteínas são particularmente relevantes: "**peptídeos sinal**" e "**domínios transmembranares"**: Os "**peptídeos sinal** "são sequências curtas (~ 20 aminoácidos) localizadas no **N-terminal** das proteínas e atuam como marcadores que **direcionam a localização** de proteínas recém-sintetizadas. Nas bactérias, os "**peptídeos sinal**" direcionam a proteína através da membrana citoplasmática para o periplasma ou para o espaço extracelular.

As proteínas que ficam mergulhadas na membrana celular têm papéis críticos, uma vez que residem na interface entre a bactéria e o ambiente (ou hospedeiro). Algumas delas são **canais de transporte**, e outras, são **receptores de sinal**. Elas são comumente chamadas de **proteínas transmembranares** (**TM**), e essa característica é conferida por uma sequência específica de aminoácidos que forma a **seção transmembrana** da proteína. **Novamente, algoritmos podem ser aplicados para detectar "peptídeos sinal" e assinaturas TM.** Uma ótima ferramenta para a previsão dessas características é Phobius. Ele prevê ambos "**peptídeos de sinal**" e "**domínios da MT**" simultaneamente e possui uma saída gráfica útil.

Se você estiver interessado em aprender mais sobre os métodos por trás do **Phobius**, leia a publicação: Advantages of combined transmembrane topology and signal peptide prediction–the Phobius web server by Käll L, Krogh A, and Sonnhammer EL.

Os bancos quatro **Bancos de dados secundários de proteínas** que foram descritos são apenas uma pequena amostra das centenas de bancos de dados secundários dedicados à predição de domínios conservados em proteínas;

- **Pfam** e a **Interpro** foram escolhidos porque são bons exemplos de análises em larga escala aplicadas a conjuntos de dados pré-existentes, com o objetivo de recuperar informações para auxiliar na investigação de **proteínas de função desconhecida;**

- Phobius foi escolhido devido à sua ampla aplicabilidade na previsão da localização de proteínas.

O conhecimento combinado **da localização e da função de uma proteína**, pode ser altamente informativo. Além disso, esses quatro exemplos são populares entre os pesquisadores e são ferramentas bem mantidas.

3. Como as proteínas podem ser analisadas computacionalmente

A plataforma do **EBI-ENA** estará com novo endereço as partir de agosto/2020. Disponibiliza um completo treinamento da ferramenta Pfam online.

As ferramentas **Interpro** e **Pfam** permitem análises de famílias de RNA (Rfam) e de metagenoma ("Metagenomics Analysis Platform") do **EBI**. Estas ferramentas permitem analisar informações da comunidade microbiana e tentar descobrir que proteínas estão lá e o que cada uma está fazendo.

A ferramenta **Pfam** é um dos primeiros recursos do ENA e já existe há cerca de 20 anos, desde o início da informática. E seu papel principal sempre foi caracterizar proteínas considerando unidades funcionais, conservadas chamadas **domínios**. Por exemplo: uma proteína que se liga ao DNA, e experimentalmente foi identificado o domínio de ligação ao DNA e agora vê esse domínio em outra sequência, isto permite inferir que provável esta nova proteína também se liga ao DNA por meio deste domínio.

Ou seja, as proteínas têm domínios conservados e estas sequencias permitem a identificação de sua função (Perfil **Markov).** O processo da **evolução** provoca mudanças nas sequencias, mas sempre deixa um sinal subjacente. Existem genes que são essenciais para a "manutenção doméstica" ("**housekeeping genes**"), por exemplo, que são usados para replicação do DNA, onde os genes foram conservados ao longo de todo tempo. A análise de sequências de vários organismos permite estudo da evolução dos organismos no tempo, que podem ser realizados construindo-se arvores filogenéticas e análise dos **clados*** **(**das ramificações).

Alvos de medicamento contra uma bactéria podem ser identificados com esta ferramenta: Caso seja encontrando um domínio específico apenas em bactérias que não esteja presente em humanos, provavelmente poderá ser alvo de um medicamento não será reativo contra humanos.

O Pfam já foi usado para identificar famílias envolvidas na patogenicidade de bactérias. Existem vários exemplos em uma família de proteínas ou de uso extensivo de uma família de proteínas em diferentes bactérias.
- Proteínas de imunidade, onde uma bactéria realmente excreta uma proteína específica que mata todas as outras bactérias ao seu redor. Esta bactéria tem uma parceira cognato que permite que ela não trabalhe contra si mesma., que ela possa existir sem se autodestruir. O Pfam permite identificar bactérias relacionadas que possuem esse parceiro cognato. A característica bacteriana descrita permite que ela invada um ambiente específico, digamos, no intestino humano, e depois cause essa patogenicidade.

*** clado:** Um **clado** ou **ramo** é um grupo de organismos originados de um único ancestral comum exclusivo. Cada um dos ramos da árvore filogenética. Por conseguinte um clado é um grupo de espécies com um ancestral comum exclusivo. Podem ser modelados em um cladograma: um diagrama dos organismos em forma de árvore. Um clado particular pode ser sustentado ou não diante de uma análise subsequente usando um conjunto diferente de dados ou de um modelo distinto de evolução.

Vídeo 1:
Entrevista com Dr Rob Finn, chefe da equipe do grupo "EMBL-EBI Sequence Families"

Nesta entrevista, Anna Protasio do Welcome Genome Institute, entrevista Rob Finn, do European Bioinformatics Institute (EMBL-ENA). O Dr Rob Finn, que explica como as famílias de proteínas são classificadas computacionalmente e como esse resultado é benéfico para a pesquisa em qualquer organismo.

	https://view.vzaar.com/14594888/adaptive.m3u8 Obs. A transcrição do vídeo em português e em inglês está no final do Capítulo.

4. Como fazer a citação de Plataformas e de suas Versões

Agora, vamos mostrar a maneira correta de **citar o trabalho de outros em bioinformática**.

Assim como é importante registrar citações para a literatura que consultamos, é igualmente importante citar os sites dos quais coletamos informações. Ao citar bancos de dados, é comum reconhecer os autores e pesquisadores que criaram o banco de dados com um link para o site apropriado; no entanto, essa nem sempre é a maneira mais apropriada de citar seu trabalho ou pesquisa.

Os bancos de dados são o resultado de pesquisas científicas e de informática abertas a outros pesquisadores e / ou ao público e, na maioria dos casos, os autores produziriam um artigo revisado por pares, descrevendo os métodos usados em conjunto com a implementação on-line. Estas são as citações que precisamos usar para referenciar seu trabalho adequadamente.

Por exemplo: o servidor da plataforma Pfam http://pfam.xfam.org/ mostra a seguinte sugestão de citação na parte inferior da página:

Citando Pfam

'Se você acha que a Pfam é útil, considere citar a referência que descreve este trabalho: O banco de dados de famílias de proteínas da Pfam: em direção a um futuro mais sustentável: R.D. Finn, P. Coggill, R.Y. Eberhardt, S.R. Eddy, J. Mistry, A.L. Mitchell, S.C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G.A. Salazar, J. Tate, A. Bateman. Nucleic Acids Research (2016) Edição 44 do banco de dados: D279-D285 '

Citing Pfam

If you find Pfam useful, please consider citing the reference that describes this work: The Pfam protein families database: towards a more sustainable future: R.D. Finn, P. Coggill, R.Y. Eberhardt, S.R. Eddy, J. Mistry, A.L. Mitchell, S.C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G.A. Salazar, J. Tate, A. Bateman. Nucleic Acids Research (2016) Database Issue 44:D279-D285'

É igualmente importante registrar a **versão ou o release** do **banco de dados** ou **software**. No caso de um banco de dados, **a data do acesso também deve ser registrada**. Isso equivale a citar a edição correta de um livro. Nos bancos de dados, novas versões ou lançamentos terão modificações significativas da mesma maneira que novas edições de um livro podem ter atualizações importantes sobre determinados tópicos. **No caso da Pfam, a versão ou o release do banco de dados é mostrado na página inicial**.

Se a as informações de como citar estiverem ausentes em um banco de dados ou servidor, é possível usar a **URL** e os **dados de acesso**, mas você também deve entrar em contato com os autores, eles terão prazer em ajudá-lo a fornecer a citação correta para o trabalho deles.

.....

Desafio: Citação do banco de dados do UNIPROT

Sugestão: Procure a citação sugerida em Uniprot database.

- Você pode encontrá-lo?

- E a versão ou release do banco de dados? Resp.: Página inicial por assunto

The UniProt Consortium
UniProt: a worldwide hub of protein knowledge
Nucleic Acids Res. 47: D506-515 (2019)
...or choose the publication that best covers the UniProt aspects or components you used in your work:

.....

5. Como usar bancos de dados públicos para coletar informações sobre uma sequência de proteínas

Para exemplificar, usaremos uma proteína da bactéria ***Salmonella enterica*** :

>NP_456741.1 hypothetical protein STY2412

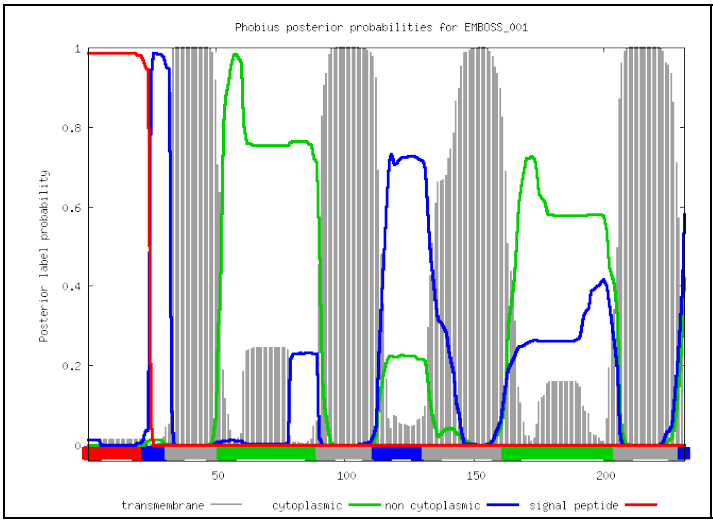
```
>NP_456741.1 hypothetical protein STY2412
MMTYIWWSLPLTLAVFFAARRLAAHFKMPLLNPLLVAMVVIIPFLLLTGIPYEHYFKGSEVLNDLLQPAV
VALAYPLYEQLHQIRARWKSIISICFVGSLVAMITGTSVALLMGATPEIAASVLPKSVTTPIAMAVGGSI
GGIPAISAVCVIFVGILGAVFGHTLLNAMHIRTKAARGLAMGTASHALGTARCAELDYQEGAFSSLALVI
CGIITSLVAPFLFPLILAVMR


MMTYIWWSLPLTLAVFFAARRLAAHFKMPLLNPLLVAMVVIIPFLLLTGIPYEHYFKGSE
VLNDLLQPAVVALAYPLYEQLHQIRARWKSIISICFVGSLVAMITGTSVALLMGATPEIA
ASVLPKSVTTPIAMAVGGSIGGIPAISAVCVIFVGILGAVFGHTLLNAMHIRTKAARGLA
MGTASHALGTARCAELDYQEGAFSSLALVICGIITSLVAPFLFPLILAVMR
```

1. Inicialmente, entre no **InterPro.** Na caixa de entrada insira a sequência **FASTA** acima (copie e cole). Após alguns segundos ou poucos minutos surge a informação da sequência. Clique nela. Surge um quadro com vários intervalos coloridos de informações indicando:
- Nome da Família de Proteína
- Regiões/Programas que identificaram domínios conservados. Por exemplo:
    a) Phobius - que a proteína tem "signal peptide" (indicando que possa ser exportada)
    - que é transmembrana; etc

2. Entre em Pfam . Clique em "**Search Sequence**, e insira na caixa a sequência **FASTA** acima (copie e cole). Em segundos, abrem-se informações. A esquerda clique nas caixas: **Alignments**, **HMM LOGO** ("Hidden Markov Models"), **Trees**, **Species** - explore, pois é muito interessante! Explore o LrgB domain. ele faz arte de uma família de proteínas**.** Em que organismos ele está presente?

3. Entre em Phobius (clique no link ou simplesmente escreva no "browser" Phobius). Step 1: insira na caixa de busca a sequência **FASTA** acima (copie e cole). Step2: Escolha "**Long with Graphics**. Clique **Submit**. Analise o Lindo gráfico que surge. Ele mostra em cores as diferentes regiões da proteína.



NP_456741.1
Gráfico de Phobius – Análise:

Observa-se:
- Logo no início (N-teminal) sequência sinal (indicando que é secretada);

- depois se alternam:
Região azul (não citoplasmática),
região cinza (transmembrana),
região verde (não citoplasmática) ,
indicando dobramentos que atravessam a membrana citoplasmática e que a proteína alcança o exterior da célula bacteriana.

4. Agora vá para UNIPROT. Basta inserir o número: NP_456741.1. Clique em pesquisar, e surgem as informações, que clicando a esquerda podem ser selecionadas:

- **Name**: CidB/LrgB family autolysis modulato
- **Subcellular location:** surgem 4 “Transmembrane domains” (conforme abaixo) permitindo análise pode de cada um deles em detalhes

Feature key	Position(s)	DescriptionActions	Length
Transmembrane	33 – 52	HelicalSequence analysisAdd BLAST	20
Transmembrane	91 – 112	HelicalSequence analysisAdd BLAST	22
Transmembrane	132 – 162	HelicalSequence analysisAdd BLAST	31
Transmembrane	206 – 230	HelicalSequence analysisAdd BLAST	25

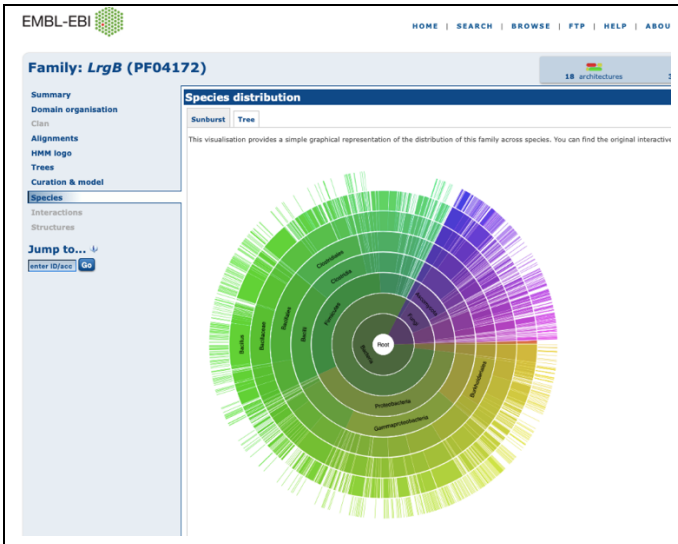
Mais no **UNIPROT**:
- Use o **basket** do UNIPROT: coloque lá uma sequencia, por exemplo “recA E.coli” e depois adicione a sequencia na basket recA de outros organismos. Analise todos dados.
- Analise proteomas de *S. cerevisiae*

Conclusão: Combinado as quatro páginas de busca **InterPro, Pfam**, **Phobius**, **UniProt** podemos fazer uma boa análise de uma proteína. Esta análise pode já resultar em muitas informações valiosas sobre a localização e da função da proteína e, assim, informar sobre sua possível contribuição em patogenicidade.

Vídeo – Análise de Proteínas usando várias plataformas

Neste vídeo, é mostrado como, usando pesquisas on-line de recursos como **UNIPROT**, InterPro, **Pfam**, **Phobius**, evidencias da provável função de uma proteína podem ser acumuladas. Neste exemplo, usaremos uma proteína da bactéria ***Salmonella enterica***. Você pode encontrar a sequência da proteína abaixo:

```
>NP_456741.1 hypothetical protein STY2412
MMTYIWWSLPLTLAVFFAARRLAAHFKMPLLNPLLVAMVVIIPFLLLTGIPYEHYFKGSEVLNDLLQPAV
VALAYPLYEQLHQIRARWKSIISICFVGSLVAMITGTSVALLMGATPEIAASVLPKSVTTPIAMAVGGSI
GGIPAISAVCVIFVGILGAVFGHTLLNAMHIRTKAARGLAMGTASHALGTARCAELDYQEGAFSSLALVI
CGIITSLVAPFLFPLILAVMR
```



Neste vídeo ,
Martin Aslett (Wellcome Genome Campus Advanced Course) demonstra como usar bancos de dados públicos para coletar informações sobre uma sequência de proteínas. Estas informações ajudam a inferir a função potencial de uma proteína.

(A Transcrição do Vídeo encontra-se no final do capítulo)

6. Sistemas de anotação automatizados

Neste artigo, você aprenderá como a anotação de proteínas é concluída em larga escala.

O processo de anotação manual de sequências de proteínas é muito trabalhoso. Com um grande número de sequências de proteínas encontradas em bancos de dados, seria quase impossível fornecer anotações manuais para todas elas.

Em vez disso, os cientistas da computação projetaram "software-pipelines" que usam **ferramentas de similaridade de sequência** e **ferramentas de previsão de domínio** de proteínas (como **Pfam** e **Phobius**) para prever automaticamente funções putativas de sequências de proteínas. Os resultados de pesquisas individuais são combinados para fornecer um resultado simplificado e uma frase de descrição da proteína. Uma ferramenta popular para anotação automática é o mecanismo de pesquisa da **Interpro** chamado **InterproScan**.

Os "**pipelines de anotação automática**" têm a vantagem de serem rápidos e sistemáticos. No entanto, sua precisão às vezes pode ser comprometida pela qualidade dos bancos de dados originais usados para as comparações. Portanto, os resultados dos **pipelines de anotação automática** devem ser tomados com cuidado.

A maioria dos **bancos de dados** possui um sistema para pontuar a anotação fornecida. Por exemplo, a **Uniprot** usa um **logotipo azul** para mostrar entradas com **anotação automática** e um **logotipo dourado** para destacar as entradas que foram **anotadas manualmente**.

7. DESAFIO - Anotação de uma proteína.

Explore os vários aspectos de sequências e apresente suas observações:

1. **Investigue a proteína curta (sequência mostrada abaixo) isolada de uma bactéria Gram-negativa.**

```
MISRVTEALSKVKGSMGSHERHALPGVIGDDLLRFGKLPLCLFICIIL
TAVTVVTTAHHTRLLTAQREQLVLERDALDIEWRNLILEENALGDHS
RVERIATEKLQMQHVDPSQENIVVQK
```

a) Faça uma pesquisa **BLAST** para identificar a bactéria da qual foi recuperada.

b) Use o seguinte **UniProt link** para recuperar uma sequência de proteína e descobrir:
b1) Domínios proteicos previstos com **Interpro** e **Pfam**
b2)) Compare os resultados dos dois bancos de dados

c) Esta proteína pode ser enviada para fora da célula e/ou ser uma parte constituinte da membrana.

Resultados:

a) BLAST: cell division protein FtsL [Escherichia coli]
Sequence ID: WP_103767641.1 Length: 123
FASTA: >WP_103767641.1 cell division protein FtsL [Escherichia coli]
MISRVTEALSKVKGSMGSHERHALPGVIGDDLLRFGKLPLCLFICIILTAVTVVTTAHHTRLLTAQREQL VLERDALDIEWRNLILEENALGDHSRVERIATEKLQMQHVDPSQENIVVQKKG

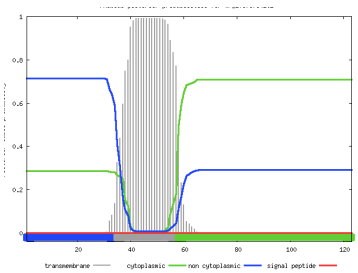
b1) Domínios proteicos (**Interpro** e **Pfam)**

b2) Compare os resultados dos dois bancos de dados

c) INTERPRO: Transmembrane protein

d) UNIPROT: 1-34 Citoplasmic; 35-57 Helical; 58-121 Periplasmic, Cell Inner Membrane

e) Phobius – veja que legal - o gráfico confirma que é uma proteína transmembranar



2. O seguinte fragmento de DNA foi recuperado de uma a amostra clínica:

```
CGGCCAGTTGGCGTCGCTGTCGACGAGCACGGCAGCGGGGCTTTCCACCGCCGCG
AGCGGCGTCGCGTCGCTGTCGACGTCGCTGCTCGGCGCGGCGGGCGATCTGGCGT
CACTGTCGACGAGCGCATCGACGGGGCTCGCCACTGCGGATAGCGGCATCGCGTC
GTTGTCCACGTCGCTGCTCGGCACCGCGGACAACGTGACGTCGCTGTCGACGAGC
CTCAGCACGGTCAACGCGAATCTGGCCGGCCTGCAGACCTCGGTGGACAACGTCG
TGTCATACGACGATCCGTCGAAGTCGGCGATCACCCTCGGCGGCGCGGGCGTCGC
GACGCCCGTCCTGCTGACGAACGTGGCTGCGGGGAAGATCGCCGCGACCAGCACG
GACGCGGTGAACGGTTCGCAGCTTTACACGCTCCAGCAGGAGTTCTCGCAGCAGT
ACGATCTGCTGACGTCGCAAGTCTCGTCGCTCAGCACTTCGGTGTCGGGTCTCCA
AGGCAGCGTCTCGGCAAATACGGGAACCGCGTCGGGTGACAACAGCACGGCGAGC
GGTGACAACGCGACCGCGTCGGGCACGAACAGCACGGCCAACGGGACGAACTCGA
CCGCGTCGGGTGATAACAGCACGGCAAGCGGGAC
```

Use qualquer um dos métodos aprendidos para descobrir:

a) a sequência completa desse gene e qual organismo pertence.
b) sua sequência protética prevista.
c) Revise a entrada **UniProt** para esta proteína (você pode usar o servidor **UniProt BLAST**) e discuta sua função potencial.

Resultados

a) Pesquisa **BLAST** : em **BLASTn** :

- Pertence a **Burkholderia pseudomallei strain BPs112 chromosome 1, complete sequence**

Então foi obtido de um genoma – que gene é ?.....

- **BLASTX** : (para pesquisa de sequencia de nucleotídeos em bancos de proteínas:

1) **hypothetical protein DM78_1570 [Burkholderia mallei]**
Sequence ID: AIO58921.1Length: 240
Cromossomo 1 – Tradução conceitual
FASTA:
>AIO58921.1 hypothetical protein DM78_1570 [Burkholderia mallei]
MLVAAIFPAATFVSRTGVATPAPPRVIADFDGSSYDTTLSTEVCRPAR-
FALTVLRLVDSDVTLSAVPSSD
VDNDAMPLSAVASPVDALVDSDARSPAAPSSDVDSDATPLAAVESPAAVLVDSDANWPT-
LTASVALTPAA
TPVSVRGPWPPAKSTVRPGELAPTLIAPLEASCVTTPTPSDSMLVESEAM-
LVAVDVDNEVNWDMLTASVG
LTPAARPLSNTPPVAFDIVNTLPFMPPVRK

2) **hypothetical protein DP57_2904 [Burkholderia pseudomallei]**
Sequence ID: KGC66910.1Length: 251

1B) UNIPROT: Run BLAST (BLASTx) com a sequência de DNA: selecionando Data Bank – Bacterias, Resultado: Burkolderia (Pseudomonas)

linha 1: B. mallei BPAC_BURMA - Autotransporter adhesin BpaC
linha 2: B. mallei - A0A3N4DGC6_BURML - Uncharacterized protein
linha 3: B. pseudomallei C4AXZ1_BURML - Hemagglutinin family protein
linha 4: B. pseudomallei B1HJ97_BURPE - Haemagglutinin family protein
linha 5: B. pseudomallei A0A3B6W557_BURPE - Hep_Hag family protein

1B) UNIPROT - RUN BLASTp: ate cerca de 50 alinhamentos: B. malei e B. pseudomalei,

Parabens!

8. Exemplo de uso dos Bancos de Dados para pesquisar uma Doença tropical – 1. EXEMPLO DE CASO: Melioidose

O que é Melioidose

A **melioidose** é uma doença infecciosa causada pela bactéria ***Burkholderia pseudomallei***, um bacilo Gram-negativo. A **doença tem alta taxa de letalidade** e **exige rápido diagnóstico e início de tratamento precoce,** visando redução desse risco. É considerada endêmica de regiões do sudoeste da Ásia e nordeste da Austrália. Há relatos de casos na América Central e do Sul, Oriente Médio, Pacífico e países da África. No Brasil, foi diagnosticada pela primeira vez no **Ceará,** em 2003, e vem provocando a ocorrência de óbitos. Nos EUA, o interesse em torno desta doença tem aumentado, pois a bactéria causadora tem potencial para ser utilizada no desenvolvimento de **armas biológicas**.

A patogênese bacteriana

A maioria dos pacientes, podem adquirir a bactéria através de inoculação, ingestão ou inalação. As taxas de mortalidade são muito altas. Mesmo com tratamento a mortalidade pode chegar a 40%; e de pacientes não tratados, estima-se possa chegar a 90%. Atualmente, **não há vacinas** disponíveis.

Genética bacteriana

A bactéria ***B. pseudomallei*** possui um genoma imenso, pelo menos dois cromossomos, ao todo 7-8 mega pb. Essa bactéria pode ocupar uma ampla variedade de ambientes: solo, água contaminada, plantas, mamíferos e no hospedeiro humano. Há uma hipótese de que alguns dos subconjuntos de genes podem permitir que a bactéria se adapte e colonize um nicho em particular, incluindo o hospedeiro humano. Assim, podem surgir variedades (recombinantes ou mutantes) genéticas deste desse patógeno que sejam capazes de se adaptar melhor à causar infecção em humanos.

Como Bancos de dados podem oferecer ajuda para a construção de uma estratégia de tratamento

Uma técnica chamada "genoma pelo estudo de associação" permite "**minerar**" os genomas buscando pela "**prevalência de genes específicos"** ou "**variações genéticas**" que sejam mais prevalentes em **isolados clínicos** versus **isolados ambientais**.

Como já existem muitos genomas disponíveis depositados em "**Bancos de dados Públicos"**, conseguimos **identificar alguns genes potenciais candidatos à virulência**. No entanto, apenas a análise computacional não é suficiente. É necessária uma validação experimental para fazer a validação de cada gene candidato que surge desta hipótese.

Todavia tudo isto é bastante complicado, porque ***B. pseudomallei*** **é altamente virulenta**. Está enquadrada na **categoria três** e só pode ser manuseada em laboratório de contenção de segurança especial. Assim, para facilitar as pesquisas, em trabalhos experimentais é empregada outra espécie, que está intimamente relacionada, mas menos virulenta, chamada ***Burkholderia thailandensis*** como proxy. Nestes trabalhos, genomas de ***B. thailandensis*** é analisado em busca de genes homólogos. São realizadas buscas no "**Banco de Dados** do **NCBI**. Uma vez que os genes são identificados, podemos excluir e manipular os genes na ***B. thailandensis*** para investigar o que descobrimos.

Estes genes são sempre os mesmos nos isolados obtidos em diferentes regiões?

Estes genes são muito variáveis. E uma das características interessantes é que parece haver fortes sinais geográficos. Ou seja, **genes, candidatos a virulência detectados na Tailândia** são **bem diferentes daqueles identificados na Austrália.** Então, são necessários mais dados ainda para aprofundar o estudo.

Esta pesquisa pode ser aproveitada em saúde pública?

Uma boa compreensão da biologia básica é sempre fundamental para, pelo menos, orientar uma melhor escolha de tratamento. Dados obtidos até o momento indicam que **não há um alvo geral em todos os achados**, que seja útil em termos de escolha de **tratamento** e também de **vacina**. Portanto, o design da **vacina** precisa levar em conta a **diversidade bacteriana**, geograficamente e, em alguns casos, até localmente.

9. Resumo

Confira se você aprendeu tudo....
Se tiverem ficado dúvidas, revise o assunto específico.

Neste Capítulo você aprendeu sobre a homologia nas sequências de proteínas e como isso pode ser usado para inferir sua função.

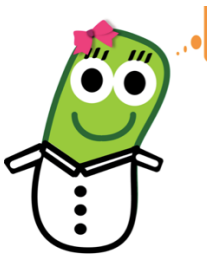
Você pode explicar os conceitos por trás da poderosa ferramenta de pesquisa de similaridade **BLAST** e usar essa ferramenta para procurar sequências semelhantes em bancos de dados.

Você aprendeu sobre os **bancos de dados primários** e **bancos de dados secundários** e seu papel importante no campo de pesquisa biológica atual.

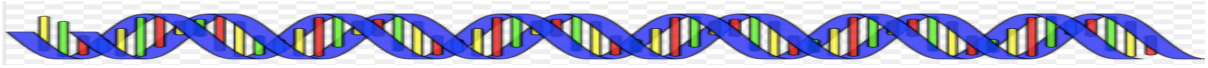
Agora**, você pode usar os bancos de dados secundários para descobrir domínios** c**onservados** nas **sequências de proteínas** e **inferir a função da sequência**.

No final, você pode:

- Fazer pesquisas BLAST
- identificar/Escolher **bancos de dados primários** e **bancos de dados secundários** e explicar o conceito por trás deles
- Usar **bancos de dados secundários** para encontrar **domínios** e **motivos conservados** em sequências de proteínas
- Integrar esses dados para inferir a função de uma proteína a partir de sua sequência.



Parabéns! Você chegou ao final do Capítulo 4 e já sabe muito!



Vídeo 1:

Entrevista de Anna Protasio do Welcome Genome Institute com Rob Finn, chefe da equipe do grupo "Sequence Families" do European Bioinformatics Institute (EMBL-ENA). (Transcrição em Português)

0:05

Estamos aqui com Rob Finn, que trabalha no **Instituto Europeu de Bioinformática** (EMBL-ENA) e vai falar conosco sobre famílias de proteínas. Olá, Rob. Olá Anna. Você pode nos contar sobre o seu papel? Sim claro. Dirijo uma equipe grande no **EBI** chamada "Equipe de Famílias de Sequências" ("**Sequence Families Team** "). Então, na verdade, é um guarda-chuva para vários recursos diferentes. Temos recursos para famílias de proteínas e há a **Interpro** e a **Pfam** lá. Temos famílias de RNA, que são Rfam e RNA central. E também executo a "Metagenomics Analysis Platform" do EBI, que realmente usa esses recursos todos para analisar o DNA da comunidade microbiana e tentar descobrir que proteínas estão lá e o que cada uma está fazendo.

0:48

Você mencionou vários bancos de dados que sua equipe cobre. Você pode nos contar mais sobre a **Pfam**? A Pfam é um dos primeiros recursos que já existe há bastante tempo, existe há cerca de 20 anos. Está lá desde o início da informática. E seu papel principal sempre foi o de transferir as informações das poucas proteínas experimentalmente caracterizadas para muitas outras plataformas de informações que saem de projetos de sequenciamento em larga escala. Portanto, as proteínas podem ser analisadas considerando unidades funcionais, chamadas **domínios**.

1:28

E o que fazemos é tentar modelar essas unidades individuais para que possamos ter esses pequenos modelos e depois digitalizar novas sequências novamente e dizer: OK, já vimos esse exemplo antes. Rotule assim. E isso nos dá uma ideia da função. Então, se você viu algo que, por exemplo, se liga ao DNA, e agora vê esse domínio em outra sequência, daí você pode inferir que é provável que ele se ligue ao DNA. Então, os domínios encontrados nessas sequências são conservados? Sim, então confiamos nisso. A **evolução**, com o tempo, provoca mudanças, mas sempre deixa um sinal subjacente.

2:05

E é exatamente isso que estamos tentando encapsular em nossos modelos matemáticos, chamados **modelos Markov** ocultos por perfil, onde realmente capsulamos os tipos evolutivos da variação de sequência: de modo a identificar quais partes são iguais e quais partes diferentes. E, o mais importante, **modelamos** as inserções e exclusões nessas sequências que nos dão modelos muito sensíveis. Como você constrói um Perfil **Markov** escondido no perfil? A maneira como trabalhamos é dar alguns exemplos em que sabemos que eles estão relacionados; exemplos de pequenas sequências. E então procuramos. Nós pegamos isso, construímos um **perfil HMM** em torno deles e depois procuramos repetidamente para expandir o conjunto de sequências. E é por isso que a **Pfam** é muito boa.

2:49

Portanto, esse conjunto de treinamento, desde que seja razoavelmente representativo desse espaço, é muito bom. Então o que descobrimos - e é isso que crescemos lineares ao longo do tempo - é que esses **HMMs** de perfis são muito bons em modelar a evolução. Então, o que encontramos na evolução é que você recebe mudanças sutis ao longo do tempo. E desde que você tenha bons representantes em toda a árvore filogenética de sequências, basta que esses **HMMs** de perfis detectem as etapas intermediárias, porque, como aludi, a evolução é um contínuo. E, de repente, você não encontra grandes mudanças. Você vê mudanças discretas nas quais, dependendo da variação natural ou da pressão de seleção, apenas vê uma ligeira mudança de família ao longo do tempo.

3:38

O mesmo perfil **HMM** pode ser usado para identificar sequências semelhantes em bactérias e em humanos? Sim, existem genes que são essenciais para a "manutenção doméstica" ("housekeeping genes"), por exemplo, que são usados para replicação do DNA, onde os genes foram conservados ao longo de todo tempo. E um único **HMM** é capaz de detectar as semelhanças ao longo do tempo. Existem outras famílias há domínios de proteínas que são específicos para um **clado*** específico. E é isso que torna essas coisas diferentes. Portanto, esta é uma das aplicações da ferramenta **Pfam**. Se você estiver interessado em ter um alvo de medicamento contra uma bactéria, e conseguir encontrar um domínio específico encontrado apenas em bactérias e não em humanos, é provável que isso signifique que o medicamento não será reativo contra humanos.

4:24

O **Pfam** foi usado para identificar famílias envolvidas na patogenicidade de bactérias? Sim, existem **vários exemplos** em que descobrimos uma família ou descobrimos o uso extensivo dessa família em diferentes bactérias. Exemplos incluem proteínas de imunidade, onde uma bactéria realmente excreta uma proteína específica que mata todas as outras bactérias ao seu redor. E então esta bactéria tem uma parceira cognato que permite que ela não trabalhe contra si mesma. Assim, ela pode existir sem se autodestruir. Então você pode identificar bactérias relacionadas que possuem esse parceiro cognato. Portanto, isso permite que uma bactéria invada um ambiente específicas, digamos, no intestino humano, e depois levar à essa patogenicidade.

Obrigado, Rob. Isso foi fascinante.

5:12

O prazer foi meu.

Vídeo 1:

(Tanscrição em Inglês) Anna Protasio/ Welcome Genome Interview with Dr Rob Finn, Team Leader of EMBL-EBI Sequence Families Group.

In this interview, Anna interviews Dr Rob Finn, who explains how protein families are classified computationally and how this outcome is beneficial for research on any organism.

0:05
We are here with Rob Finn, who works at the European Bioinformatics Institute and is going to talk to us about protein families. Hello, Rob. Hi, Anna. Can you tell us about your role? Yeah, sure. So, I run a large team at the EBI called the Sequence Families Team. So, it actually is an umbrella for a number of different resources. So, we have protein family's resources, and there's Interpro and Pfam there. We have RNA families, which is Rfam and RNA central. And then I also run the EBI's Metagenomics Analysis Platform, which actually uses those resources to actually analyse microbial community DNA to try and work out who's there and what they're doing.

0:48
You've mentioned a number of databases that your team covers. Can you tell us more about Pfam? So Pfam is the grandfather resource. It's been around for quite a long time. So, it's been around for about 20 years. So, it's been there since the start of informatics. And its primary role has always been to able us to transfer the information over from the few experimentally characterized proteins to many others that come off these large-scale sequencing projects. And so, proteins can be considered as being built up of these functional units, called domains.

1:28
And what we do is we try and model those individual units such that we can have these little models that we can then scan new sequences again and say, OK, I've seen that instance before. Label it with that. And that gives us an idea of function. So, if you've seen something that, say, binds DNA, and I see that domain in another sequence, I know that's likely to bind DNA. So, are the domains found in these sequences conserved? Yeah, so we rely on that. Evolution, over time, they make changes. But there's always an underlying signal.

2:05
And that's really what we're trying to encapsulate in our mathematical models, called Profile hidden Markov models, where we actually capsulate evolutionary types of the sequence variation-- so the parts that are the same and the parts that are different. And importantly, we model the inserts and deletions within those sequences that give us very sensitive models. How do you build a Profile hidden Markov of model? The way we work is we take a few examples where we know that they're related-- a few sequence examples. And then we search. We take those, build a profile HMM around them, and then search over and over again to expand the set of sequences. And this is really why Pfam is very good.

2:49
So that training set, as long as it's reasonably representative of that space, is very good. So, what we find-- and this is what we've grown linear over time-- is actually these profile **HMMs** are very good at modelling evolution. So, what we find in evolution is you get subtle changes over time. And so as long as you've got good representatives across the phylogenetic tree of sequences, that's enough for these profile HMMs to detect the intervening steps, because as I've alluded, evolution is a continuum. And so, you don't suddenly find massive changes. You see discrete changes where, depending on natural variation or selection pressure, you just see a slight change of a family over the course of time.

3:38
Can the same profile HMM be used to identify similar sequences in bacteria and in humans? Yeah, so there are core housekeeping genes that, for example, that you use for replicating DNA, where the genes are conserved all the way across. And a single HMM is capable of detecting those similarities all the way. There are other protein families and domains that are specific to a particular clade. And that's what makes those things different. So this the applications of Pfam. If you are interested in having a drug target against a bacterium, if you can find a particular domain that's only found in bacteria and not in humans, then that's likely to mean that the drug won't be cross-reactive.

4:24
Has Pfam been used to identify any families involved in pathogenicity in bacteria? So there are a number of examples where, certainly, we've discovered a family or we've found extensive use of that family in different bacteria. Examples include immunity proteins, where a bacterium actually excretes a particular protein that then kills all the other bacteria around it. And then it has a cognate partner that allows it to not work against itself. And so therefore, it can exist without destroying itself. And then you can have related bacteria that have this cognate partner. So that allows a bacterium to invade a particular environment for, let's say, the human gut, and then lead to that pathogenicity. Thank you, Rob. That was fascinating.

5:12
It was my pleasure.

Vídeo 2 – Análise de Proteínas usando várias plataformas **(English Transcription)**

0:05

Hello, I'm Martin Aslett. And I work for the Wellcome Genome Campus Advanced Courses [and Scientific Conferences] Team, based at the Welcome Genome Campus. In this activity, we will use a public database and protein amino acid sequences to find out more about potential protein function. Let's use this Salmonella enterica entry as an example. The short description line of the header of the FASTA formatted sequence suggests that this sequence has no known function. I have this sequence saved in the Notepad file as this is plain text format on a PC. And this will be best for cutting and pasting into web pages.

0:41

Let's now use this sequence in different secondary databases that we've already seen to see whether we can learn more about its function. I'm now going to use **InterPro Scan to look for conserved domaines in my sequence of interest**. I open my internet browser and can either search for **InterPro** Scan or type in the URL that is now appearing on the screen.

1:16

This is the **InterPro** page and then in this search box, I can cut and paste my sequence and then search for protein domains and other features of interest.

1:30

I go to the file I have opened in Notepad and I copy the sequence and then paste it into the search box. Note that I included the header line starting with the arrow symbol. Most internet sites have learned to ignore this. But you might find with some searches that you need to remove this line. **Interpro Scan utilizes searches of a large set of databases**. I will click Submit to start the search and then come back to the results later. While this is running, I'll explore other databases so I can later compare the results for all of them. The first database I will look at is **Pfam**. This is a conserved protein domain database.

2:17

I'll open a new tab and then I can either search for **Pfam** or again, type in the address that is appearing on your screen now.

2:38

This is the Pfam page. As you can see, there are many options. I will go to the Sequence Search. Again, I will paste my protein sequence into the search box, and I hit Go to search. This may take a while, depending on your internet connection or how many jobs are in the queue. **On average, this search should take around 15 to 30 seconds.** The results page shows one hit. In **Domain Organization**, is show the **LrgB domain** (LrgB-like family). That's represented by this **green box**, which extends over almost the entire protein sequence. The grey bar behind it represents the whole of our protein sequence. You'll note the conserved domain is slightly shorter than the entire protein.

3:22

An important parameter to look at is the expected value or **E-value**. This is a measurement of how good the match is **between** the **conserved domain** and **my protein sequence**. We're not discussing **E-values** in any detail at the moment. Suffice it to say that the lower the **E-value** is, the better the match. Now I want to find out more about this conserved domain as it might give me some more information about my sequence of interest. I'll click on the name of the domain. On this page, we find the **Pfam** description. This includes a detailed description of the domain, a link to **InterPro** and sometimes literature references. These are commonly the papers that the original curator used to curate this domain.

4:10

The description of this domain is interesting. The thing to highlight is that it is involved in both murein hydrolase activity and penicillin tolerance. This is interesting because in certain bacteria, it may be involved in **penicillin resistance**. One other thing to note is that, according to this description, proteins with this domain **are potential membrane proteins**. This means that they are likely to have contact with either the extracellular space or the host itself. Now I will

use another secondary database **to find out** whether my sequence of interest has **transmembrane domains**. This will indicate whether it is likely to be a **membrane protein**. I open another browser tab and can either search for Phobius or type the address that appears on screen.

5:09

This is the **Phobius** page. In the submission section I can either paste my sequence or choose a file from my computer. This is useful if the sequence is long or if you're using multiple sequences. You'll notice there are **three output formats**: short, long without graphics, or long with graphics. I'll choose the default option of long with graphics and click Submit. The results appear very quickly. Here we found the name of the submitted sequence, whether **it has transmembrane domains**, and whether **it has signal peptides**. We also find the coordinates of the signal peptides or transmembrane domains. In the graphical display, you'll notice that there are grey bars. These represent potential transmembrane domains.

6:00

Some of these are very low, indicating an unlikely probability of them being transmembrane domains. These won't be counted in our search. The red line indicates the probability of each residue being part of a signal peptides. This shows that the **first 23 amino acids** are part of a signal peptide. After this, the search quickly drops to zero. The **green** and **blue** lines respectively represent the probability of these regions being **cytoplasmic** or **non-cytoplasmic**. We'll now go back to the **InterPro** query and compare the results.

The **InterPro** search has now finished. Similar to the **Pfam** output, the conserved domains are represented with bars that spanned the length of the conserved domain with respect to the full length the protein sequence.

6:46

Not surprisingly, the **InterProSan** results include **Pfam** matches, such as the first one here but also matches to other databases such as Panther, and TIGRFAMS You're encouraged to investigate these databases on your own. In the unintergrated signatures panel, we found some results that back up our previous searches. We find that **InterProSan** uses TM Helix, but also integrates **Phobius** results. In addition, there are a number of integrated algorithms which predict a signal peptide at the **N-teminus**. Finally, we can get to **UniProt** and search for this protein based on a succession number. Again, I will open a new tab in my browser. You can either search for **UniProt** or, as I'm doing, type in the URL.

7:48

UniProt is the central reference database for protein sequences and their functional information. I will type in the accession number for our protein, **NP_456741** and hit Search. Our search shows us the **UniProt accession code**. If I click on this, we come to the full page for our protein. As you can see, this shows that the search results that are found backed up in this entry, such as **transmembrane domains** and **family domains like Pfam**. In summary, this activity shows that we can use publicly available secondary databases to find out clues about the potential function of an amino acid sequence.

8:31

We've used four different pages: **UniProt**, **Phobius**, **Pfam**, and firstly, **InterPro** to make searches find clues as to what the function of our previously unknown functioned protein will be. The databases we've used are merely a small selection of those available online. We encourage you to do your own searches to find other databases which may be more relevant to the proteins that you are searching for functionality for. We hope you found this video enjoyable. Please add ideas, suggestions, or questions in the comments section. We look forward to hearing back from you.

Vídeo 3: Entrevista com a Dra. Claire Chewapreecha: Doença tropical, sequenciamento e evolução (Transcrição em português)

Neste vídeo, a Dra. Claire Chewapreecha explica como ela usa o sequenciamento "next-generation" para entender a evolução bacteriana e os alvos genéticos associados à **virulência**. Ela está se concentrando em uma doença infecciosa tropical chamada "**melioidose**", que é pouco estudada, mas causa uma carga substancial em saúde pública nos países afetados.

VÍDEO 3

0:06

Gostaria de dar as boas-vindas à **Dra. Claire Chewapreecha, da Universidade de Cambridge.** E ela vai falar conosco sobre **patogênese bacteriana**. Claire, conte-nos sobre você. Bem, eu sou originalmente de Bangkok, Tailândia. Estou interessado em minerar os genomas das bactérias. E, recentemente, tive muita sorte de receber a bolsa Sir Henry Wellcome. Portanto, isso permite minhas atividades de pesquisa na Tailândia e no Reino Unido. **E você poderia nos contar mais sobre o seu trabalho?** Eu trabalho em uma doença tropical específica chamada melioidose. Não sei se você já ouviu falar do nome. Bem, a maioria das pessoas não. Mas esta doença é uma carga de saúde pública em países tropicais. E os "hot spots" conhecidos da doença são o nordeste da Tailândia e o norte da Austrália.

0:50

Mas nosso entendimento global da epidemiologia da doença mudou recentemente, à medida que a infraestrutura de microbiologia melhorava. E isso expandiu as zonas endêmicas para o resto do sudeste da Ásia e para o sul da Ásia, parte da África e também na América Central e do Sul. A doença é causada pela bactéria chamada ***Burkholderia pseudomallei***. E a maioria dos pacientes, eles são agricultores, trabalhando no arrozal. E eles podem adquirir a bactéria através de inoculação, ingestão ou inalação. As taxas de mortalidade são muito altas, no entanto. Na Tailândia, cerca de 40% dos casos morreram. E isso é mesmo com tratamento. Com pacientes não tratados, estimamos que a taxa de mortalidade possa chegar a 90%. E atualmente, não há vacinas disponíveis. Isto é o que faz a bactéria ***Burkholderia pseudomallei***. Esse é um nome de espécie bastante prolixo.

1:51

Esta é uma bactéria Gram-negativa. Portanto, essa bactéria ocupa uma ampla variedade de nichos hospedeiros. **Ela vive no solo e na água contaminada. Ela pode sobreviver nas plantas, nos mamíferos e no hospedeiro humano**. **E como o sequenciamento do genoma melhorou nosso entendimento sobre essa bactéria?** O sequenciamento do genoma ajuda maciçamente. Para esta bactéria em particular, ***Burkholderia pseudomallei***, ela possui um genoma imenso, pelo menos dois cromossomos e um tamanho combinado de aproximadamente **sete a oito megabases**. Então você pode imaginar que há muitos genes e muitas inflamações lá.

2:32

Isso combinado com o fato de as bactérias ocuparem amplas faixas de nicho do hospedeiro, então levantamos a hipótese de que alguns dos subconjuntos de genes podem permitir que a bactéria se adapte e colonize um nicho em particular, incluindo o hospedeiro humano. **E como você identificou os genes que influenciam o potencial da doença?** Essa é uma pergunta muito importante. Por isso, pensamos que alguns dos genes ou variações genéticas dentro desse patógeno podem permitir que parte deles se adapte melhor à infecção humana. Então, **usamos uma técnica chamada genoma pelo estudo de associaçã**o. Isso é minar os genomas para a prevalência de genes específicos ou variações genéticas que são mais prevalentes em isolados clínicos versus isolados ambientais.

3:25

E temos muita sorte porque já existem muitos genomas disponíveis no banco de dados público. Então isso é o poder de nossos estudos. Então, conseguimos **identificar alguns potenciais candidatos à virulência**. No entanto, apenas a análise computacional não seria adequada. Precisamos de alguma **validação experimental** para fazer backup de nossa hipótese. No entanto, é bastante complicado, porque ***Burkholderia pseudomallei*** **é altamente virulenta**. E é categorizado na **categoria três**. Portanto, precisamos de contenção espacial em laboratório para realizar experimentos envolvendo esses organismos. E **essas instalações não estão disponíveis em todos os lugares**. Portanto, para evitar esse problema, usamos outra espécie, que está intimamente relacionada, mas menos virulenta, chamada ***Burkholderia thailandensis*** como proxy. E, para fazer isso, extraímos os genomas da ***Burkholderia thailandensis*** em busca de genes homólogos.

4:30

Para isso é feita uma busca rápida no **banco de dados NCBI**. E uma vez que o identificamos, podemos excluir e manipular os genes na ***Burkholderia thailandensis*** para investigar o que descobrimos. **Quão variáveis são esses genes?** Muito, muito variáveis. E uma das características interessantes que vimos é que parece haver fortes sinais geográficos. Ou seja, **genes, candidatos a virulência detectados na Tailândia** são **bem diferentes do que vimos na Austrália.** Então, acho que precisamos obter mais dados ainda. Portanto, precisamos de mais informações. **E como sua pesquisa se traduz em saúde pública?** Penso que uma boa compreensão da biologia básica é fundamental para informar uma melhor escolha de tratamento.

5:24

E pelo que descobrimos até agora nos genomas, vimos muitas diversidades entre diferentes geografias. E isso sugere que **não haverá nenhuma regra que sirva para todos**, em termos de escolha de **tratamento** e também de **vacina**. Portanto, o design da **vacina** precisa levar em conta a diversidade bacteriana, geograficamente e também, em alguns casos, até localmente. **Mais uma vez obrigado, Claire. Isso foi realmente fascinante.** O prazer é meu. Muito obrigado.

Video 3: Interview with Dr Claire Chewapreecha: Tropical disease, sequencing, and evolution (English Transcription)

In this video, Dr Claire Chewapreecha explains how she uses next-generation sequencing to understand bacterial evolution and genetic culprits associated with **virulence**. She is focusing on a tropical infectious disease called "**melioidosis**", which is understudied but causes a substantial public health burden in the countries affected.

VIDEO 3

0:06

I'd like to welcome **Dr. Claire Chewapreecha from the University of Cambridge**. And she's going to talk to us about bacterial pathogenesis. Claire, tell us about yourself. Well, I'm originally from Bangkok, Thailand. I'm interested in mining the genomes of the bacteria. And recently, I've been very fortunate to be awarded Sir Henry Wellcome Fellowship. So this allows my research activities in Thailand and UK. **And could you tell us more about your work?** I work on a particular tropical disease called melioidosis. I'm not sure you've heard of the name. Well, most people don't. But this disease is a public health burden in tropical countries. And the known hot spots for the disease is Northeast Thailand and Northern Australia.

0:50

But our global understanding of the disease epidemiologic changed recently as the microbiology infrastructure improved. And this expanded the endemic zones to the rest of Southeast Asia to South Asia, part of Africa, and also in the Central and South America. The disease is caused by the bacterium called ***Burkholderia pseudomallei.*** And most of the patients, they are farmer working in the rice paddy field. And they can acquire the bacterium through inoculation, ingestions, or inhalations. The mortality rates are very high, though. In Thailand, about 40% of cases died. And that's even with treatment. With untreated patients, we estimate that the mortality rate could reach 90%. And currently, there's no vaccines available. And what is ***Burkholderia pseudomallei***. That's quite a wordy species name.

1:51

This is a Gram-negative bacterium. So this bacteria occupy wide ranges of host niches. It's live in soil and contaminated water. And it can survive in plants, in mammals, and in human host. **And how has genome sequencing improved our understanding about this bacterium?** Genome sequencing helps massively. For this particular bacterium, ***Burkholderia pseudomallei***, it has a huge genome, two chromosomes at least and a combined size of roughly **seven to eight megabase**. So, you can imagine that there's a lot of genes and a lot of inflammations in there.

2:32

This combined with the fact that the bacteria occupy wide ranges of host niche, so we hypothesized that some of the subsets of genes might allow the bacterium to adapt and to colonize in particular niche, including human host. **And how did you identify the genes that influence disease potential**? That's a very important question. So we thought that some of the genes or genetic variations within this pathogen might allow part of them to be better adapted to human infection. So, we used a **technique called genome by association study**. That is to mine the genomes for the prevalence of particular genes or genetic variations that are more prevalent in clinical isolates versus environmental isolates.

3:25

And we are very fortunate because there's a lot of genomes already available in the public database. So that's power our studies. So, we managed to identify some potentials virulence candidate. However, computational analysis alone wouldn't be adequate. We need some **experimental validation** to back up our hypothesis. However, it is quite tricky because ***Burkholderia pseudomallei*** **is highly virulent**. And it's categorized under **category three**. So, we need spatial lab containment to perform experiments involving these organisms. And such facilities **are not available everywhere**. So, to avoid that issue, we use another species, which is closely related, but less virulent called ***Burkholderia thailandensis*** as a proxy. And in order to do so, we mine the genomes of ***Burkholderia thailandensis*** for homologous genes.

4:30

So that's why blast search through the **NCBI database**. And once we identified that, we can knock out and manipulate the genes in ***Burkholderia thailandensis*** to investigate what we found. **How variable are these genes?** Very, very variable. And one of the interesting features that we saw is that there seems to be a strong geographic signals. That is, genes, virulence candidate (genes) detected in Thailand **are** quite different from what we've seen in Australia. So, I think we need to mine the data further, so in this sense, we need more information. **And how does your research translate to public health?** I think a good understanding of basic biology is key to inform a better choice of treatment.

5:24

And from what we found so far in the genomes; we saw a lot of diversities between different geography. And that suggests that there's not going to be any one rule that fit all, in terms of choice of **treatment** and also **vaccine**. So, **vaccine** design needs to take into account of bacterial diversity, geographically, and also, in some cases, even locally. **Thanks again, Claire. This was really fascinating.** My pleasure. Thank you so much.