UNIVERSIDADE DE SÂO PAULO
INSTITUTO DE BIOCIÊNCIAS
GRADUAÇÃO EM CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE GENÉTICA E BIOLOGIA EVOLUTIVA

BIO0307
BIOLOGIA MOLECULAR PARA BACHARELADO

ROTEIRO DE AULAS PRÁTICAS

ANÁLISES COMPUTACIONAIS DE SEQUÊNCIAS DE ÁCIDOS NUCLÉICOS

4ª Edição

PROFESSORES RESPONSÁVEIS:

Profa. Dra. Luciana Amaral Haddad

Prof. Dr. Eduardo Gorab

Profa. Dra. Regina Célia Mingroni Netto

SÃO PAULO
2017

Aula Prática 1 – Análise computacional de sequências de ácidos nucléicos

Objetivos: Revisão sobre a reação em cadeia da polimerase (PCR). Desenho de iniciadores no NCBI para reação em cadeia da polimerase (PCR) e conhecimento dos bancos de dados Herança Mendeliana no Homem *Online* (OMIM) e *Medical Genetics* (*MedGen*).

Estratégia: Desenhar iniciadores para identificação pela PCR e sequenciamento da mutação ΔF508 no gene *CFTR* (do inglês, *Cystic Fibrosis Transmembrane conductance Regulator*).

(1) A fibrose cística é uma doença de herança autossômica recessiva, causada por mutações no gene *CFTR*.

(2) A forma mais simples e imediata para obter informações concisas sobre uma doença genética é hoje no banco de dados *Medical Genetics* (MedGen) do NCBI (*National Center for Biotechnology Information*, Bethesda, MD). O NCBI fornece acesso a informações genômicas e biomédicas.

(3) Entrando no *site* MedGen (**http://www.ncbi.nlm.nih.gov/medgen**), uma busca por 'cystic fibrosis' recupera no primeiro resultado um resumo sobre a doença. Clique neste resultado. Informações mais aprofundadas podem ser encontradas nos bancos Gene Reviews e OMIM.

(4) No NCBI, o *site* OMIM (do inglês, *Online Mendelian Inheritance in Man*) cita brevemente algumas informações disponíveis sobre todas as doenças genéticas humanas conhecidas, recuperadas da literatura científica no *pubmed*. Uma vez que o OMIM foi construído a partir do *pubmed* de forma manual e não compreensiva, o seu texto não tem o objetivo de ser completo embora tenda a conter as primeiras informações de associação de um gene a um fenótipo.

(5) Verifique algumas informações sobre a fibrose cística no OMIM, para o qual há um *link* nesta mesma página: OMIM 219700.

(6) Cerca de 70% dos pacientes com fibrose cística, em populações Caucasianas, tem a mutação ΔF508, uma deleção de três bases levando à deleção de um resíduo de fenilalanina na posição 508 da proteína CFTR. Esses dados podem ser observados se voltarmos a página e mudarmos para a *database* 'Clinical Variants – ClinVar' no mesmo menu. **ClinVar** é o banco de dados do NCBI que agrega informações sobre variantes de sequência do genoma humano e sua relação com a saúde humana. Faça uma busca por 'cystic fribrosis deltaF508'. Clique em 'Search'.

(7) Clique sobre a mutação p.Phe508DelPhe e observe na nova página alguns aspectos importantes:

   a. 'Clinical Significance': pathogenic

   b. (Clinical) 'Condition': cystic fibrosis

   c. 'Variant type': deletion

   d. 'Genomic location': 7q31.2

   e. Protein change: F508DelF

f. *Links* para OMIM (descrito acima), dbSNP (banco de dados de polimorfismos de base única), 1000genomes (aula P6).

(8) Tendo conhecido algumas portas de entrada no NCBI para obter informação sobre a doença, iremos agora desenhar um par de iniciadores para amplificar a sequência de DNA do **éxon 10 do *CFTR*** que contém a mutação que estudaremos.

(9) A descrição **p.Phe508delPhe** é a denominação correta desta mutação em relação à **sequência proteica (p.)** da CFTR, definindo a **deleção da fenilalanina em sua posição 508.** Em relação à **sequência de código (c.) do gene**, define-se corretamente a mutação como **c.1521-1523delCTT**, i.e, a **deleção de três bases, CTT**, nas posições **1521 a 1523** da sequência de código do gene *CFTR*.

(10) Para amplificação pela reação em cadeia da polimerase (PCR), os iniciadores deverão hibridar com as sequências intrônicas adjacentes ao éxon 10, isto é, os íntrons 9 e 10.

(11) Abra o arquivo *Word* enviado por email, contendo a sequência do éxon 10 do *CFTR* e regiões intrônicas adjacentes. Copie a sequência incluindo seu cabeçalho.

(12) Para o desenho dos iniciadores, utilizaremos o *primer3*, um programa de computador que busca identificar em uma sequência de DNA curtos trechos de fita simples com características que os definam como bons iniciadores para a PCR. Assim, os resultados vêm como pares de iniciadores em ordem decrescente de acordo com a adequação aos critérios adotados. Estes critérios compreendem:

a. temperatura de fusão (do inglês, *melting*) em torno de 60ºC (tamanho do oligonucleotídeo entre 18 e 25 bases);
b. produto de PCR entre 200 e 1000 pb;
c. inexistência de auto-complementariedade;
d. inexistência de complementariedade com seu par;
e. presença preferencial de C ou G nas duas últimas bases a 3´;
f. 40 a 60% da sequência de cada iniciador com conteúdo de [C+G]; e
g. sequência a ser amplificada (amplicom) com 40 a 60% de [C+G].

(13) O *site* do NCBI usa o *Primer3* para o desenho de iniciadores e então os submete a análise pelo BLAST com um banco de sequências definido pelo usuário. O objetivo do BLAST após o desenho dos iniciadores é verificar se estes podem identificar por complementariedade mais de um alvo no genoma do organismo selecionado, o que não é desejável, pois queremos especificidade à PCR.

(14) Para acessar o *site* Primer-blast, entre em http://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi

(15) O *site* aconselha usar uma RefSeq (aula P2) como molde para a PCR sempre que possível, porque auxilia o programa a identificar sequências inespecíficas, tendo-a como padrão. No entanto, hoje usaremos a nossa sequência do arquivo *Word*. Cole-a no local indicado.

(16) Determine o tamanho desejado para sequência a ser amplificada (amplicom, '*PCR product size'*) que, para nós, será entre 300 e 700 pb.

(17) A temperatura de fusão (Tm) será 60ºC (opt), variando entre 57ºC e 63ºC. O máximo de diferença de Tm entre os dois iniciadores será de 3ºC.

(18) Escolha um banco de sequências não-redundantes, no nosso caso, '*Genome (reference)*' e '*Homo sapiens*' como organismo.

(19) Clique em '*Get primers*'. Aguarde pelos resultados.

(20) Uma observação importante é que há **486 pb** de sequência do **íntron 9** e **1.323 pb do íntron 10** no arquivo. **O tamanho do éxon 10 é de 192 pb**.

(21) Observe os resultados. O que você identifica? Escolha um par de iniciadores.

(22) Com a sua escolha, uma vez sintetizados os oligonucleotídeos e padronizada a reação da PCR, como você poderá observar se há a mutação **c.1521-1523delCTT** (ΔF508)?

(23) Para restringir as alternativas, você pode definir que o primeiro iniciador deverá estar entre a base 1 e a 486 (íntron 9). O segundo iniciador deverá ser no íntron 10 (da base 679 à 2001). Mantenha os parâmetros anteriores e clique em 'Get primers'. Aguarde.

(24) Um banco de dados de mutações do gene *CFTR*, relacionadas à fibrose cística, está no *site* do *Clinical and Functional Translation of CFTR* (CFTR2) e tem o objetivo de facilitar a comunicação entre pesquisadores, acelerar o processo de triagem de mutações no gene e uniformizar o formato para os relatos de mutações e sua localização precisa. Para acessá-lo, entre em http://www.cftr2.org. Você deverá concordar com os termos de uso do *site*.

(25) NO menu, haverá 'Variant List History'. Na página que se abriu, selecione 'CFTR2_8August2016.xlsx'. No arquivo excel que se abre, observe a diversidade de mutações causadoras de fibrose cística. Busque pela mutação do nosso estudo digitando F508. Compare sua frequência na população (coluna F) à de outras mutações.

Curiosidades

(1) Como você acha que a fibrose cística é triada pelo teste do pezinho no recém-nascido?

(2) Uma gestante deseja realizar diagnóstico pré-natal para fibrose cística. Ela e o marido são primos e têm, em sua família, vários afetados pela doença e alta taxa de consanguinidade. Em testes genéticos realizados, não se encontrou mutação no gene *CFTR* dos familiares afetados, mas o diagnóstico clínico é evidente. Como você poderia fazer um diagnóstico pré-natal e aconselhamento genético para o casal?

Referências bibliográficas

Steve Rozen and Helen J. Skaletsky (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) *Bioinformatics Methods and Protocols: Methods in Molecular Biology.*Humana Press, Totowa, NJ, pp 365-386

Aula Prática 2 – Análise computacional de sequências de ácidos nucléicos

Objetivos: Revisão sobre bibliotecas genômicas de DNA. Recuperação de sequência genômica do gene *piruvato carboxilase* de *Caenorhabditis elegans* no banco de dados.

Estratégia: No *site* do NCBI, identificar o genoma de *Caenorhabditis elegans* e, neste, a sequência genômica do gene *piruvato carboxilase*.

Leituras para esta aula: Genbank, RefSeq & Entrez (páginas 24-29 deste Roteiro).

(1) Nossa porta de entrada será o Mapview do *site* do NCBI (*National Center for Biotechnology Information*). Para isso, você deverá entrar no *site* http://www.ncbi.nlm.nih.gov/mapview/.

(2) A página apresenta uma lista de organismos para os quais dados de sequenciamento e mapeamento genômicos estão disponíveis nos bancos de dados do NCBI. As linhas azuis podem expandir e listar os organismos dentro daquele grupo. Isto é feito clicando na ponta de seta no início da linha. Clique na ponta de seta à frente de '**Invertebrates**'.

(3) Surgem cinco novos níveis. Hoje vamos trabalhar com um nematodo. Clique na ponta de seta em frente a '**Nematodes**'.

(4) Há dois nematodos cujos genomas se encontram neste banco de dados. Vamos nos ater ao organismo-modelo ***Caenorhabditis elegans***.

(5) Na segunda coluna à direita do nome ***Caenorhabditis elegans***, encontra-se o nome/número da versão mais atualizada deste banco de dados. Denomina-se '**Build**' às diversas versões ou *runs* do banco de dados. Selecione o 'Build' mais recente para *Caenorhabditis elegans*, que é o que se apresenta: '**WS195**'. Clique sobre este.

(6) Na nova página, estão representados em figuras os cinco autossomos, um cromossomo sexual (X) e o genoma mitocondrial de *C. elegans*. Observe. Logo acima, há as estatísticas do Build WS195. Clique sobre '**WS195 statistics**'.

(7) Algumas informações interessantes sobre o genoma de *C elegans* estão listadas na nova página. Explore estas informações. Por exemplo, na tabela no pé da página, listam-se quantos *contigs* para cada cromossomo? O que isso representa? Qual o tamanho do genoma de *C. elegans*?

(8) (Pausa para discussão)

(9) Volte à página anterior e realize a seguinte busca no genoma de *C elegans* – lembre que todas as buscas têm que ser feitas em inglês!

(10) Escreva no *menu*, no alto à esquerda, na lacuna em frente a 'Search for', o elemento de sua busca que, hoje, será o gene piruvato carboxilase de *C elegans*. Para isso, podemos escrever simplesmente: **elegans pyruvate carboxylase**. Clique em '**Find**'.

Recentemente, a busca por termos no mapa do genoma de *C elegans* esteve fora do ar. Se isso ocorrer, vamos recuperar a sequência genômica do gene pela busca direta pelo gene, ao invés do mapa. Isso pode ser feito como instruído na página 7 deste Roteiro [Item 6].

(11) Cinco resultados (*hits*) são encontrados no cromossomo V. Mais abaixo, nesta página, os resultados estão listados. Clique sobre o código de acesso do quarto

resultado: '**pyc-1'**, que se encontra na coluna de 'Map element', com a descrição de 'gene'.

(12) Você chegou à parte do mapa do cromossomo V de *C elegans* onde se localiza o gene da piruvato carboxilase, indicado com **sombreado rosa**. No mapa, em vertical, a linha vermelha indica a sua extensão. Na vertical, à esquerda, em paralelo à linha vermelha, a organização genômica é representada por blocos escuros, verticais, indicando éxons unidos por linhas representando os íntrons.

(13) Uma seta ao lado do nome do gene mostra sua orientação; neste caso na fita *minus*, à esquerda da linha cinza. *Links* para outros bancos de dados são mostrados, se existentes. São eles: *Worm Genes* (WG), *Worm base* (WB), *Sequence Viewer* (sv), *Protein Sequence* (pr), *Sequence Download* (dl), Homology (hm) e *Sequence tagged sites* (STS).

(14) Na parte vertical à esquerda da página, encontra-se o *zoom* do mapa. Clicando algumas vezes no *zoom out*, você pode verificar outros genes que mapeiam na mesma região, como para a membro da família do colágeno e semelhante a aciltransferase.

(15) Neste mapa, a maneira mais direta para recuperação da sequência genômica do gene Pyc-1 é pelo link *dl*. Clicando nele, a próxima página mostrará os parâmetros para a sequência e a definição de um intervalo de cerca de 5 kilobases (kb). Clicando em 'display', você obterá a sequência no formato FASTA. Esta é uma RefSeq, cujo número de acesso é NC_003283.9. O formato **FASTA** é utilizado em vários programas de análise de sequência: i**nicia-se pelo sinal > seguido por um nome (identificação). A sequência começa na linha seguinte.**

(16) Selecione toda a sequência de DNA no formato FASTA (incluindo a identificação na linha >). Copie-a e cole-a em um novo arquivo *word*. No *word*, você pode simplificar a identificação por um nome mais curto, por exemplo, >Celegans_Pyc1.

(17) Selecione tudo e formate a fonte para '**Courier New**', um tipo de caracter que ocupa sempre o mesmo espaço para qualquer letra e, por isso, é apropriado a análises de sequências, inclusive alinhamentos.

(18) Salve este arquivo (nuvem, email, dropbox, etc...). Ele será utilizado em outra aula.

(19) Você pode converter a sequência NC_003283.9 do formato FASTA ao formato GenBank. Para isso, você deve ir ao topo da página da guia do formato FASTA e, à esquerda, clicar sobre '**GenBank**'. Você observará que a sequência contínua do formato FASTA tornou-se numerada e dividia em blocos espaçados de dez bases cada, no pé da página. Antes, há algumas informações sobre esta sequência. Este é o formato **GenBank**. Este formato é informativo, mas não é utilizado em programas de análise de sequências.

(20) (pausa para discussão) – **formatos Genbank e FASTA**

(21) Se você clicar sobre 'See current version', você encontrará os números de acesso de sequências que compuseram o contig correspondente à sequência completa (> 20 Mb, GenBank) do cromossomo V de *C elegans*, incluindo o gene Pyc-1.

(22) Em uma busca nesta página (Ctrl+F) por pyc-1, você localiza este gene. Esta forma não é prática, mas uma curiosidade sobre um genoma inteiramente sequenciado.

Observações para antes da aula P3:

(1) Voltando ao loco '**pyc-1**' na área sombreada rosa [**item (12) desta aula**], clique sobre '**pyc-1**'.

(2) Na página que se abre, você encontrará várias informações disponíveis sobre o loco pyc-1. Esta é a página deste gene e espécie, no NCBI. Há links para sequências, referências, etc...

(3) Analisando em detalhe, no *site*, a figura da organização genômica, pode-se observar o **número de éxons (blocos verdes) e íntrons (linhas ligando esses blocos)** do gene Pyc-1 (Figura 1A da página 9 deste roteiro). Se clicar sobre essa figura do site, um novo diagrama se apresentará (Figura 1B da página 9 deste roteiro), com linhas de blocos azuis representando o RNAm e linhas de blocos vermelhos representando a sequência proteica. Os números correspondentes são *links* às respectivas sequências de RNAm e proteica.

(4) Se você mantiver o cursor sobre o diagrama de blocos verdes, sem clicar, você abrirá uma caixa de informações sobre a sequência. Ao abrir a caixa de informações, sob 'Links and Tools', você pode clicar com o botão direito do *mouse* sobre Genbank view: NC_003283.11 para abrir o *link* em uma nova guia ou sobre FASTA view: NC_003283.11.

Outras formas:

(5) Explore a página e *links* para contexto genômico (sequências vizinhas), artigos científicos relacionados (*pubmed*), proteínas que interagem com e domínios descritos para a piruvato carboxilase. Mais abaixo nesta página, sob 'NCBI reference sequences (RefSeq)', pode-se clicar diretamente sobre FASTA ou GenBank da sequência genômica ('Reference Assembly > Genomic').

(6) Como alternativa para recuperar sequência do Genbank, ao invés de entrar no 'Mapview' como no início desta aula, você poderá entrar diretamente na página do gene pyc1 de *C elegans*, pelo site **http://www.ncbi.nlm.nih.gov/gene**, escrevendo 'pyc1 elegans' ou 'elegans pyruvate carboxylase' no menu. Aparecerá a página que utilizamos acima.

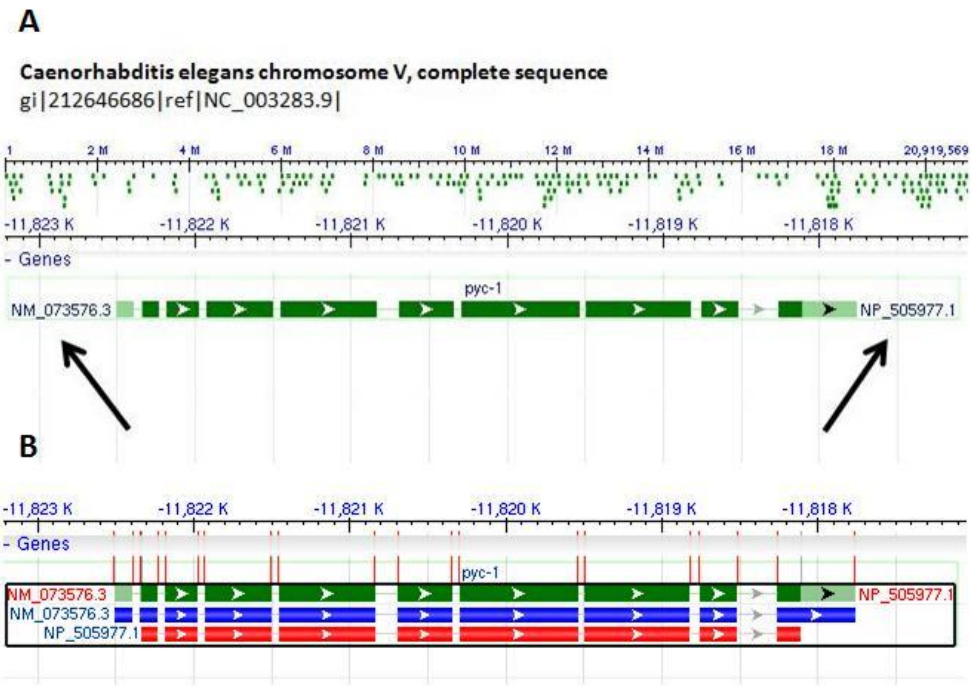

Figura 1: Mapa do gene *Pyc-1* de *C elegans* no NCBI. (A): Organização genômica, fora de escala, mostrando éxons como blocos verdes, separados por íntrons (linhas). Pontas de seta brancas indicam éxons com sequência codificadora para proteínas. As setas pretas indicam os *links* para a sequência de cDNA (seta à esquerda, iniciada por NM) e proteica (seta à direita, iniciada por NP). **(B)**: Relação fora de escala entre a estrutura gênica (linha superior, verde), transcrita (linha intermediária, azul) e proteica (linha inferior, vermelha).

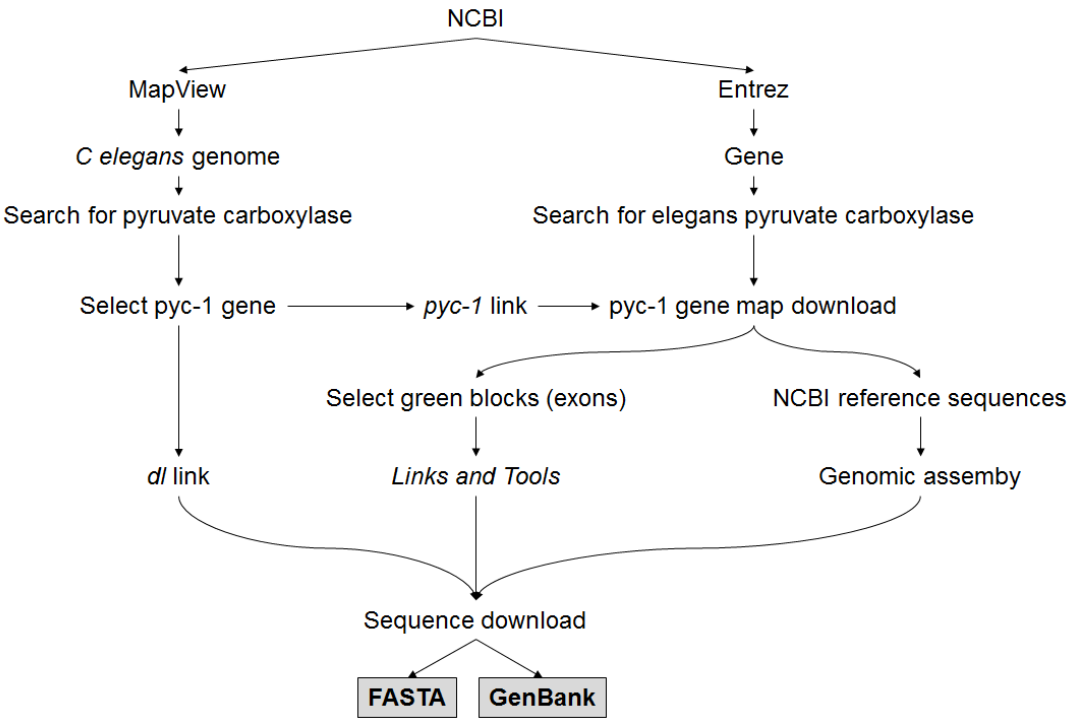

Figura 2: Resumo da estratégia adotada na aula P2 para recuperação de sequência gênica a partir de genomas sequenciados.

Aula Prática 3 – Análise computacional de sequências de ácidos nucléicos

Objetivos: Revisão sobre bibliotecas de cDNA. Organização de éxons e íntrons na sequência genômica do gene. Recuperação de sequência transcrita do gene *piruvato carboxilase* (*pyc1*) de *Caenorhabditis elegans* no banco de dados.

Estratégia: Introdução ao *site UCSC Genome Browser*. Nele, identificar o genoma de *Caenorhabditis elegans* e, neste, a organização genômica do gene *pyc1*. Identificar a sequência transcrita de *pyc-1*.

Leituras para esta aula: BLAST (páginas 30-32 deste Roteiro).

(1) Nossa porta de entrada será o 'browser' de genomas da Universidade da Califórnia em Santa Cruz (Santa Cruz, CA; 'UCSC Genome Browser'). Trata-se de uma ferramenta para visualização gráfica, análise e 'download' de sequências de genomas variados. Foi lançado em 2000, logo após a finalização do sequenciamento do genoma humano. Entre no *site*: **https://genome.ucsc.edu**

(2) Na página que se abre clique sobre 'Genome browser'.

(3) Ao chegar na página do genoma de *C elegans*, o 'assembly' de Feb. 2013 deve estar indicado. Redija **pyc-1** como termo de busca e clique sobre 'Go'.

(4) Como é grande a variedade de buscas para o 'Genome Browser' na UCSC, é possível que a nova página com os resultados da busca mostre-se com excesso de informações. Assim, vamos restringir os resultados aos objetivos desta aula. Para isso, vamos cancelar algumas alternativas de busca, clicando sobre 'Hide'.

(5) **Clique sobre 'Hide'** para as seguintes alternativas:

   a. Sob 'Mapping and Sequencing': todas as alternativas;
   b. Sob 'Genes and Gene Predictions': todas as alternativas, exceto RefSeq (*pack*)
   c. Sob 'mRNA and ESTs': todas as alternativas, exceto *C elegans* mRNAs (*pack*);
   d. Sob 'Expression and Regulation': todas as alternativas;
   e. Sob 'Comparative Genomics': todas as alternativas;
   f. Sob 'Variation and Repeats': todas as alternativas.

(6) Clique sobre 'Refresh' ao pé da página.

(7) No painel gráfico dos resultados, observe a distribuição dos éxons (blocos em cor azul) e íntrons (linhas entre blocos).

(8) Pausa para discussão.

(9) Para obter a sequência de cDNA, vamos selecionar o RNAm (blocos pretos). Clique sobre seu diagrama com o botão direito do *mouse* e selecione 'Show Details for AF237467'.

Uma forma de localizar éxons e íntrons quando se dispõe de ambas as sequências, genômica e cDNA, de um gene é através do alinhamento das duas sequências. O algoritmo mais usado para este fim é o BLAST (*Basic Local Alignment Search Tool*) que compara sequências de nucleotídeos ou proteínas a sequências de banco de dados de sequências e calcula a significância de cada região de similaridade encontrada. A alternativa 'Global Align', no NCBI permite empregar o BLAST para alinhar duas sequências.

(10) O alinhamento BLAST entre as sequências genômica e transcrita de *pyc-1* já se encontra pronto no 'genome browser' da UCSC. Na janela de detalhes sobre o gene *pyc1*, os alinhamentos entre sequências de cDNA e genômica podem ser acessados clicando sobre o *link* abaixo de 'mRNA/Genomic alignments'.

(11) A partir daí, clique sobre '*C elegans* chr V'. Isso permitirá observar o alinhamento entre a sequência transcrita que você havia selecionado e a sequência genômica mapeada no cromossomo V de *C elegans*.

(12) Examine blocos individuais de alinhamentos ou o alinhamento total. O que é evidenciado em cada alinhamento?

(13) Quais elementos da sequência genômica você reconhece no alinhamento?

(14) No final do alinhamento, a que corresponde a sequência vermelha após o códon de parada TGA?

(15) Conclusões.

(16) Clicando sobre o número de acesso para a sequência de RNAm de *pyc-1* ('Information on mRNA AF237467'), você será direcionado ao site do NCBI diretamente ao acesso ao formato GenBank da sequência.

(17) Você poderá observar o formato e verificar que se encontra na mesma página a tradução desse RNAm, i.e., a sequência proteica, abaixo de CDS.

Cds é a sequência codificadora ('coding sequence'). Este subtítulo identifica a *cds* na sequência do RNAm/cDNA, com início na base 36, onde você deve esperar encontrar o A do ATG (codón de início da tradução); e fim na base 3563, onde você espera encontrar a última base do códon de parada da tradução, a qual é, neste caso, o A de TGA.

(18) Embora a sequência proteica já esteja disponível, convém, neste ponto, aprender a obtê-la com ferramentas *in silico*, quando se dispõe somente da sequência do cDNA. Assim, inicialmente você deve salvar o formato FASTA desta sequência do RNAm.

(19) Selecione-a e salve em arquivo *word*.

(20) Na próxima aula, avaliaremos a fase de leitura desse RNAm.

Aula Prática 4 – Análise computacional de sequências de ácidos nucléicos

Objetivo: identificação de famílias gênicas.

Estratégia: identificar ORFs, buscar por similaridade entre sequências proteicas usando o algoritmo BLAST (e Smart BLAST), analisar alinhamento de grupo de sequências ortólogas utilizando os servidores Cobalt e Clustal.

(1) Para encontrar a fase de leitura e o produto proteico do RNAm de *pyc-1*, dirija-se ao seguinte site: **https://www.ncbi.nlm.nih.gov/orffinder/.**

(2) A partir do NCBI, vamos buscar pela maior moldura aberta de leitura (ORF, do inglês, *open reading frame*) para a sequência de cDNA (cds) do gene *pyc-1* de *C elegans*.

(3) Na nova página, há um **quadro em branco**, no qual você deverá colar a sequência **cds** completa do gene *pyc-1* de *C elegans*.

(4) Apesar de a piruvato carboxilase ser uma enzima de ação mitocondrial, seu gene é nuclear e, por isso, utilizaremos o código genético padrão ['(1)Standard']. Selecione o tamanho mínimo de ORF como 150 nucleotídeos. Clique sobre '**Submit**'.

(5) Na página que se abre, você encontrará um diagrama de retângulos alaranjados, longos e curtos, finos, horizontais, dispostos de modo não relacionado entre si. Cada um deles representa uma moldura de leitura. Cada retângulo corresponde a uma ORF numerada. As mesmas ORFs estão apresentadas na tabela abaixo do diagrama, à direita da página.

(6) Na tabela, a segunda coluna identifica a fita de DNA em que a ORF foi encontrada: fita (+)/senso ou fita (-)/antissenso.

(7) A terceira coluna identifica a 'frame' ou fase de leitura para a tradução. A fase de leitura também é denominada, em língua portuguesa, quadro ou moldura de leitura.

> Entende-se por ORF a tradução dos códons compreendidos entre um ATG (códon de iniciação da tradução) e um códon de parada da tradução.
>
> O *site* não considera a possibilidade de o primeiro ATG na ORF não ser o verdadeiro códon de iniciação da tradução.

(8) Pausa para discussão

(9) Na tabela, cada ORF está listada em ordem decrescente de tamanho – veja o comprimento de cada ORF listado abaixo de '**length**'. Em geral, as ORFs serão curtas, exceto pela ORF cuja tradução comumente gerará a proteína, produto daquele gene.

(10) Observe que a ORF mais longa do gene *Pyc-1* (ORF4) está na fita (+), 'frame' 3. Ela tem 3.528 pb de extensão que inclui o códon de parada da tradução. Isto significa 3.525 nt ou 1.175 códons para a síntese de uma proteína com 1.175 aminoácidos.

(11) A sequência proteica, produto da ORF4, encontra-se ao lado esquerdo da página.

(12) É possível apresentar os códons e a sequência protéica, clicando sobre 'Display ORF as...' e escolhendo 'cds translation'.

(13) A partir desta tela, há a possibilidade de realizar o BLAST para sequência proteica (blastp) com sequências proteicas de um banco de dados de RefSeq. Escolha este banco de dados, abaixo de 'BLAST Database'.

(14) Clique sobre '**BLAST ORF4**'.

(15) A nova página representa a página de busca pelo BLAST no NCBI.

Uma busca, em outra ocasião deve ser iniciada pelo endereço https://ncbi.nlm.nih.gov/blast

Observe que, no alto do cabeçalho da página de busca do BLAST, encontram-se cinco abas. Estas permitem realizar a busca por similaridade das maneiras identificadas no quadro abaixo.

Ferramenta	**Sequência ('query')**	**Banco de dados de sequências**
blastn	nucleotídeos	nucleotídeos
blastp	proteica	proteicas
blastx	de nucleotídeos (serão traduzidos)	proteicas
tblastn	proteica	de nucleotídeos traduzidos
tblastx	de nucleotídeos (serão traduzidos)	de nucleotídeos traduzidos

Explore essas alternativas em outra ocasião.

(16) Nosso interesse hoje é o BLASTp e, da maneira como iniciamos, os dados já estão completos nesta página. Eles devem ser preenchidos a cada nova busca e consistem de:

a. Sequência da busca ('query') no quadro maior;
b. 'Database';
c. 'Organism' (hoje deixaremos em branco);
d. 'Algorythm'.

(17) Os parâmetros do algoritmo podem também ser modificados, mas não os alteraremos hoje.

(18) Clique sobre 'BLAST' no pé da página.

(19) Na página que se abre, há um cabeçalho de fundo azul claro, seguido logo abaixo pelo '**Graphic summary**'. Abaixo, no diagrama horizontal '**Show conserved domais**', há uma reta numerada, que corresponde à sequência da proteína Pyc-1 de *C elegans*. Abaixo desta, encontra-se a localização dos domínios conservados da piruvato carboxilase. Há seis domínios conservados mostrados como uma sequência horizontal de caixas de cores diferentes, de acordo com sua posição em relação à sequência proteica numerada.

(20) Uma alternativa, que **NÂO** será utilizada hoje, é clicar sobre os domínios, por exemplo, a caixa marrom denominada '**Biotin carboxylase**'. Teste-a em outra ocasião.

(21) Desça um pouco a página e, sob '**Descriptions**', observe as sequências que se alinharam à da proteína piruvato carboxilase de *C elegans*. São sequências ortólogas de diferentes espécies. Note que para todas elas há um **valor 'E'** (valor esperado), apresentado em ordem decrescente. O valor 'E' é o principal parâmetro na análise de resultados pelo algoritmo BLAST. O valor 'E' representa o número de

resultados que se pode 'esperar' observar ao acaso ao realizar uma busca com aquela sequência, em um banco de dados de determinado tamanho. Quanto menor o valor 'E', maior a significância daquele resultado.

(22) Neste caso, como se observa mais abaixo na página, a elevada significância dos resultados significa que, em longos trechos de sequência proteica, há alta identidade entre os resíduos de aminoácidos, refletindo extensa similaridade entre cada par de sequências alinhadas.

(23) Você pode verificar o alinhamento de alguns desses resultados clicando sobre a descrição da sequência, sob '**description**' ou indo diretamente à seção '**Alignments**', mais abaixo nesta página.

(24) Observe os alinhamentos indo do alto ao pé da página. Além do valor 'E', os resultados do BLAST incluem valores como '**gaps**', '**identities**', '**positives**'. O que eles identificam?

(25) Uma forma adaptada do BLASTp, SmartBLAST, apresenta para uma dada sequência proteica um resumo da similaridade entre a sequência 'query' e três sequências com mais alta similaridade do banco de sequências não redundantes, além de duas outras sequências de espécies bem estudadas. Estas espécies, cujos proteomas são conhecidos, compõem um banco de dados usado pela ferramenta Smart BLAST, denominada 'Landmark Database'. Hoje, este banco de dados contém proteomas de 27 espécies.

(26) No alto desta página dos resultados do BLASTp, há um link para os resultados do Smart BLAST. Clique sobre 'Analize your query with SmartBLAST'.

Se você desejar acessar o SmartBLAST em outra ocasião, o site é encontrado pelo endereço http://blast.ncbi.nlm.nih.gov/smartblast/smartBlast.cgi?CMD=Web

(27) Além das descrições e alinhamentos de pares como apresentado no resultado do BLAST, o SmartBLAST traz o cladograma da sequência 'query' e das cinco citadas acima.

(28) Clicando sobre 'See full multiple alignment', obtém-se o alinhamento de sequências proteicas da piruvato carboxilase de *C. elegans*, *Drosophila melanogaster*, *Danio rerio*, *Homo sapiens*, *Mus musculus* e *Saccharomyces cerevisae*, o qual é realizado pela ferramenta COBALT (*Constraint-based multiple alignment tool*).

(29) Analise os dados.

(30) Explore a alternativa 'Conservation setting: identity'. Observe os resíduos idênticos em vermelho.

(31) Vamos comparar esses resultados aos de alinhamento múltiplo das mesmas sequências em outro servidor.

(32) Salve as sequências proteicas em formato FASTA em arquivo *word*. Para isso, clique sobre os números de acesso iniciados por NP_ e obtenha o formato FASTA.

(33) No arquivo, convém simplificar o nome que identifica cada sequência por > seguido do nome da espécie. Isso facilitará a comparação ao final do exercício.

(34) Em uma nova aba, vamos ao *site* do Clustal Omega: **http://www.ebi.ac.uk/Tools/msa/clustalo/**

(35) Clustal Omega é uma ferramenta para alinhamento múltiplo para sequências proteicas, buscando obter alinhamento de espécies divergentes com significado biológico.

(36) Cole as sequências no espaço indicado.

(37) Observe o alinhamento. Clique sobre 'Show colors'.

(38) Discuta.

(39) Clique sobre 'Phylogenetic tree'.

(40) Observe.

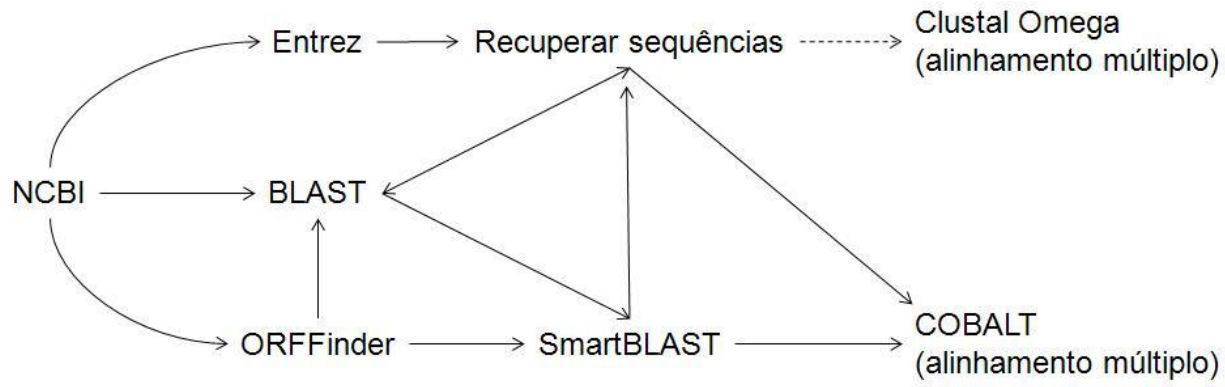

Figura 3: Resumo da estratégia adotada na aula P4 para recuperação de sequência proteica a partir da tradução do RNAm *in silico* (ORFFinder) ou pelo resultado do BLAST. Alinhamentos múltiplos de sequências proteicas, realizados pelo Clustal (UK) ou observado no COBALT. A linha tracejada indica mudança de *site* (fora do NCBI).

Curiosidades:

(1) Um alinhamento de sequências de cDNA de genes do Piruvato Carboxilase com mais alta similaridade agrupa somente algumas sequências de cDNA de invertebrados. A similaridade não é alta (dados não mostrados). Como você compara e explica esses dados em relação ao observado hoje com o alinhamento das seis sequências proteicas?

Aula Prática 5 – Conhecendo um pouco sobre Ensembl

Autor: Leandro Ucela (2015). Aluno de Doutorado da Universidade de São

Objetivo: Apresentar aos estudantes algumas das funcionalidades do portal Ensembl e utilizar uma de suas ferramentas para fazer a previsão de patogenicidade de uma sequência nucleotídica humana com alteração.

Parte 1 - Conhecendo o Ensembl

O Ensembl é um projeto conjunto do EMBL-EBI (*European Molecular Biology Laboratory - European Bioinformatics Institute*) e do Instituto Wellcome Trust Sanger que tem como principal meta o desenvolvimento de um sistema de produção e manutenção de bancos de dados genômicos de vertebrados e alguns outros eucariotos. Desde que o *website* foi criado no ano de 2000, muitos genomas têm sido adicionados ao Ensembl e a gama de dados disponíveis tem expandido para incluir dados de genômica comparativa, variação gênica dentro das espécies e informações sobre elementos regulatórios.

O grupo Ensembl consiste de uma equipe de 50 pessoas em diversos grupos de trabalho, os quais são responsáveis por: (i) criar um "*set*" de genes para diferentes espécies, colocar esses resultados em bancos de dados; (ii) desenvolver e manter os dados de ferramentas de busca; (iii) manter e atualizar os dados de variações de DNA e regiões regulatórias, gênicas e genômicas; (iv) manter esses dados disponíveis e de fácil acesso; e (v) organizar "workshops" sobre o uso do Ensembl.

(1) Entre no *site* do Ensembl - **http://www.ensembl.org**. Na interface inicial do Ensembl é possível ter uma visão geral das ferramentas disponíveis. No quadro à esquerda "*Browse a genome*" estão disponíveis acessos para bancos de dados genômicos de diversos vertebrados e outros eucariotos. Na área central é possível ver alguns acessos como, por exemplo, "*ENCODE data in Ensembl* " e "*Variant Effect Predictor*". Todos esses acessos direcionam a páginas iniciais, explicativas de cada uma das funções.

(2) Note que na caixa em frente a "*Search*" é possível escolher bancos de dados de diferentes espécies. Vamos escolher "**Human**". Abaixo do quadro de preenchimento observe os exemplos de como fazer uma pesquisa. É possível buscar por informações inserindo o número correspondente a um endereço cromossômico, o nome de um gene ou mesmo o nome de alguma condição clínica. Vamos digitar o nome do gene "***TP63***" (importante fator de transcrição expresso durante o desenvolvimento, principalmente em membros) e clicar em "***Go***". Em seguida, clique no primeiro resultado da lista de resultados da busca em que aparece "***TP63 (Human gene)"***

(3) Na tela que se abre é apresentada a descrição do gene, seus sinônimos ("aliases"), a sua localização e os transcritos alternativos já identificados para esse gene. Clique em "***show transcript table***". Observe na tabela que nem todos os transcritos codificam para proteínas. Alguns deles não estão anotados no banco de dados do "RefSeq". Logo abaixo, em "*Summary,"* é possível ver um resumo do gene *TP63*, seguido da representação gráfica de seus diferentes transcritos. Tanto pela tabela quanto pela figura, é possível notar que os transcritos se dividem em dois grupos em relação ao seu comprimento, mais longos ou mais curtos.

(4) Clique em "***Phenotype***", no menu do lado esquerdo. Observe a primeira tabela intitulada "*Phenotype*". Nesta tabela estão descritas diversas doenças que já foram associadas a alterações no gene *TP63*, seis delas descritas no OMIMGENE. O acesso ao OMIMGENE está mostrado para cada uma das doenças caso o pesquisador queira saber mais sobre o quadro clínico associado.

(5) Vamos observar um transcrito específico em seus detalhes. Clique em "Show transcript table". Escolheremos o maior transcrito do gene *TP63*, clicando em "**ENST00000264731**" na coluna "transcript ID", desça a barra de rolagem e observe a representação do gene e as informações que estão na linha "***Statistics***".

(6) Suba a barra de rolagem; no menu à esquerda clique em "**exons**" e desça novamente a barra de rolagem. É possível observar a sequência de todos os 14 éxons, e parte dos íntrons deste transcrito. No menu à esquerda, clique em "***Configure this Page***". Na tela que se abre, coloque "**Yes**" em frente a "***show variations***" e confirme clicando no ícone superior à direita da página. Agora serão visíveis todas as variantes de DNA já descritas para esse gene em bancos de dados. Observe que, no cabeçalho da sequência, há a legenda explicando o significado das cores representando os tipos de variantes (substituições sinônimas, não sinônimas, mutações intrônicas e mutações que criam códon de parada, entre outras).

Parte 2 – Prevendo o efeito de variantes

Suponha que você estudou a amostra de DNA de um paciente clinicamente diagnosticado com a síndrome EEC (*ectodermal, ectrodactyly and cleft-lip*; OMIM #604292). Você realizou o sequenciamento de Sanger dos éxons do gene *TP63*, no qual já foram identificadas mutações que causam a síndrome. Você encontrou no éxon 8 uma variante [C/G] em heterozigose.

CTTTGAGGCCCGGATCTGTG[**C/G**]TTGCCCAGGAAGAGACAGG

Como você faria para descobrir se esta variante é patogênica ou não?

O Ensembl possui uma ferramenta própria, que acessa alguns dos programas de predição de patogenicidade de variantes, SIFT e POLYPHEN, para prever o efeito da alteração na proteína e se pode ser patogênica ou não. Essa ferramenta chama-se VEP (*variation effect predictor*). Para fazer isso, é preciso saber a posição cromossômica da variante. Se você não tiver esse dado, pode encontrar a posição cromossômica do nucleotídeo realizando o BLAST da sequência obtida (como, por exemplo, a fornecida acima).

(7) Entre na página inicial do Ensembl (http://www.ensembl.org/). Na barra superior, clique em "**BLAST/BLAT**". Digite os 20 nucleotídeos que antecedem a base da variante no espaço "***Sequence data***". Desça a barra de rolagem, clique em "***Run***" e aguarde o resultado ser gerado.

(8) Clique em "***View results***". A página que se abre mostrará a localização cromossômica da sequência. Para identificar a posição exata da variante de DNA no cromossomo clique em "***[sequence]***" na célula "*genomic location*". Observe que a sequência inserida no "BLAST" está em vermelho e que no início de cada linha está o número da posição cromossômica do primeiro nucleotídeo da linha. Para descobrir a posição cromossômica exata da variante basta usar o número à esquerda da linha em que aparece o resultado do BLAST e somar 20 (a quantidade de nucleotídeos inseridos no "BLAST"). Anote a posição cromossômica da variante de DNA.

(9) No menu lateral clique em "***Variant Effect Predictor***". No campo "***Either paste data***" digite o cromossomo (3), a posição cromossômica duas vezes e a alteração (C/G)

como exemplificada a seguir: 3 189868624 189868624 C/G (Observação: a posição é repetida porque se trata de substituição de nucleotídeo. Se fosse, por exemplo, uma deleção, o número do primeiro e último nucleotídeos deveriam ser inseridos).

(10) Desça a barra de rolagem e observe que há diversos parâmetros a serem escolhidos para que sejam gerados como resultado da análise. Vamos alterar o *default* colocando "**RefSeq *transcript***". No quadro "*identifiers*" adicione "***protein***", "**Uniprot**" e "**HGVS**". Em seguida, clique em "Run" no final da página e aguarde o resultado ser gerado.

(11) Após o resultado ser gerado clique em "***view results***". Note que é apresentada a predição da patogenicidade da variante para todos os produtos do gene *TP63.* Observe que há valores nas duas últimas colunas, referentes a resultados de SIFT e Polyphen. Esses dois programas realizam predições de patogenicidade de mutações de sentido trocado (*missense*). Um resultado com valor próximo a 0 no SIFT e próximo a 1 no Polyphen indica que a mutação provavelmente é patogênica. Observe que pode haver divergência entre os dois s*oftwares*, pois algumas mutações são preditas como patogênicas por um programa e neutras pelo outro. Observe que também há diferença nos resultados desses *softwares* entre os diferentes polipeptídeos previstos. O que isso pode significar?

A utilização das ferramentas "Polyphen" e "SIFT" não está restrita à entrada pelo VEP. Elas frequentemente são utilizadas diretamente a partir de seus respectivos portais: http://genetics.bwh.harvard.edu/pph2/ e http://sift.jcvi.org/

Além do "Polyphen" e do "SIFT", há várias outras ferramentas *online* que podemos utilizar na predição de patogenicidade de variantes. Uma delas é o "Mutation Taster". Uma vantagem deste pacote computacional sobre o "Polyphen" e o "SIFT" é que o "Mutation Taster" pode realizar também a predição de patogenicidade de variantes do tipo inserção e deleção. Para exemplificar seu uso, vamos fazer outra predição:

(12) Entre na página do "*Mutation Taster*" www.mutationtaster.org. Primeiramente, na primeira linha, digite o nome do gene no qual você encontrou a variante, no nosso caso, digite "TP63". Clique em "*show available transcripts*". Selecione a primeira opção "ENST00000264731". Desça a página e observe que há três quadros em "*Alteration*". Cada quadro corresponde a uma maneira diferente de inserir dados para realizar a predição do efeito da variante.

(13) No primeiro quadro "*all types by sequence*" é preciso inserir a variante entre colchetes (ex "[A/T]") e as sequências de nucleotídeos que flanqueiam a variante. Note que não é necessário digitar a posição da mutação. No segundo quadro "*single base exchange by position*", na primeira linha, é preciso inserir a posição da variante (contada a partir do ATG que corresponde ao códon de iniciação) e, abaixo, inserir apenas a base da variante. No terceiro quadro "*insertion or deletion by position*" pode-se descobrir o efeito de deleções e inserções inserindo a numeração da base (também contada a partir do ATG) que antecede a alteração (primeira linha), a que sucede a alteração (segunda linha) e na terceira linha, em caso de inserção, escrever a base inserida.

(14) Vamos escolher o primeiro quadro, "all types by sequence". Digite a sequência, com a variante, entre colchetes, como fornecida no início deste exercício. Marque a opção "*show nucleotide alignment*" no quadro ao lado e clique em "continue" no final da página.

(15) A página que se abre apresenta ao topo, em destaque, a predição da patogenicidade dessa variante (no caso dessa variante, "*disease causing*"). Abaixo há um resumo sobre o efeito dessa variante. Examine essas informações.

(16) Há duas linhas chamadas de "*conservation*", ambas mostram a similaridade dessa região entre diferentes espécies. A primeira mostra a conservação ao nível de aminoácidos e a segunda, ao nível de nucleotídeos. Na linha "*protein features*" é mostrado que a mutação reside em uma região da proteína de ligação ao DNA.

(17) Além dessas três ferramentas para predição de patogenicidade (Polyphen, SIFT e Mutation Taster) há também a ferramenta Provean que pode ser acessada no *site* http://provean.jcvi.org/index.php.Não será abordada nesta prática.

Todos os programas de predição de patogenicidade apresentados acima, Polyphen, SIFT e Mutation Taster, utilizam informações sobre a conservação de aminoácidos para calcular o índice relativo à patogenicidade de variantes. Além de cada um desses programas utilizarem algoritmos específicos para fazer tal inferência, o alinhamento de aminoácidos do Polyphen, SIFT e Mutation Taster é realizado por meio de diferentes *softwares* como PSIC, PSI-BLAST e JASPAR, respectivamente.

O algoritmo do Polyphen considera em suas estimativas se a alteração ocorreu em um sítio específico (por exemplo: de ligação ao DNA, transmembrânico, ligação a metais), além de, realizar uma simulação 3D da alteração estrutural da proteína e análise de conservação. O resultado qualitativo gerado após a análise é dado como *benign*, *possibly damaging* (predição menos confiante) ou *probably damaging* (predição mais confiante).

O SIFT avalia tanto a conservação do aminoácido entre diferentes espécies quanto à conservação das estruturas da proteína em virtude das propriedades químicas dos aminoácidos trocados. Os resultados qualitativos do SIFT são classificados em *damaging* ou *tolerated*.

O software Mutation Taster rastreia a variante em diversos bancos de dados de variantes (por exemplo: "1000 genomas"); será considerada "neutra" qualquer variante que aparecer mais de quatro vezes nesses bancos de dados, ou seja, sem efeito clínico. *Mutation Taster* prediz uma alteração como uma de quatro tipos: *disease causing* (provavelmente deletéria), *disease causing automatic* (relatada como deléteria), *polymorphism* (provavelmente inofensiva) ou *polymorphism automatic* (relatada como inofensiva).

Devido às diferenças entre os *softwares*, tanto entre os dados utilizados como base da estimativa quanto nos próprios algoritmos, os resultados da predição de patogenicidade algumas vezes não coincidem entre os diferentes programas e também podem não representar a realidade. Se a variante encontrada é nova, nunca descrita, uma forma final de validá-la como patogênica seria por meio de estudos funcionais, pois a predição *in silico* não substitui experimentos laboratoriais. Em casos familiais das doenças investigadas, é importante verificar se a variante segrega coerentemente com o fenótipo da doença.

Aula Prática 6: Frequências de variantes genéticas - banco de dados '1000 Genomas'

Autores: Lucas Alvizi & Vagner Carvalho (2014). Alunos de Doutorado da Universidade de São Paulo

(1) O Polimorfismo de Nucleotídeo Único (*Single-Nucleotide Polymorphism* - SNP) é uma variação na sequência de DNA que ocorre quando um nucleotídeo (A, T, C ou G), situado em determinada posição genômica, difere entre genomas de indivíduos da mesma espécie. Bancos de dados de bioinformática também existem para os SNPs. O '**dbSNP**' é um bando de dados público de variações genéticas em diferentes espécies desenvolvido pelo *National Center for Biotechnology Information* (NCBI). Embora o nome do banco de dados sugira uma coleção apenas de SNPs, na verdade, ele contém um leque de variações moleculares como: SNPs; polimorfismos de pequenas deleções e inserções (indels/DIPs); marcadores de microssatélites ou STRs e etc.

(2) No entanto, informações sobre variações genéticas humanas também podem ser encontradas em outros bancos, como os do **Projeto 1000 Genomas** e do projeto **HapMap**. O Projeto 1000 genomas é um consórcio internacional formado com o objetivo de sequenciar o genoma completo de aproximadamente 1000 indivíduos, permitindo o estudo populacional das variações moleculares úteis para a biomedicina. Atualmente, o Projeto 1000 Genomas expandiu-se incluindo o sequenciamento de mais de 2500 amostras. Já o Projeto HapMap objetiva o desenvolvimento de mapas de haplótipos (**Hap**lotype **Map**) do genoma humano por meio da genotipagem de SNPs em *chips* de *microarray*, com o objetivo de descrever os padrões comuns da variação genética humana como um recurso para encontrar variantes genéticas que afetam a saúde e respostas a drogas e a fatores ambientais.

(3) Considere que você investiga o gene *IRF6* (*interferon regulatory factor* 6), no cromossomo 1 humano, em busca de variantes polimórficas de base única (SNPs – *single nucleotide polimorfism*) para um estudo populacional. Algumas variantes neste gene foram associadas às fissuras lábio palatinas não sindrômicas (comumente chamadas de "lábio leporino"). Iremos aqui investigar a frequência dessas variantes em diferentes populações por meio dos dados do Projeto 1000 Genomas.

(4) Por meio da técnica de sequenciamento dos éxons do *IRF6*, foi identificada na amostra de um paciente a variante c.820G>A, cuja posição genômica equivale a 1: 209790735 (C). Agora iremos investigar essa posição genômica no banco de dados UCSC Genome Browser. Para isso entre em **http://genome.ucsc.edu/cgi-bin/hgGateway**.

(5) Será aberta a página do UCSC Genome Browser, em que podemos escolher o genoma do organismo a ser analisado, procurando-o por grupo (ex: Mammal), genoma (ex: Human) e *assembly* (ex: GRCh38/hg38), buscando-se pelo nome do gene ou pela região de interesse. Neste caso, procuraremos pela região de interesse. Para tal, digite chr1: 209790735-209790735 no campo "Search term" e clique em "Submit".

(6) É possível ver a posição nucleotídica em questão sobre uma caixa azul, correspondente a um dos exons do *IRF6*. Iremos resgatar agora as informações sobre esta variante, a partir dos dados do Projeto 1000 Genomas. Podemos acessar estes dados por meio do banco de dados "Ensembl", o qual está também conectado ao

"UCSC Genome Browser". Na aba superior da página, clique em "View" e em seguida em "Ensembl".

(7) Será aberta a página do banco de dados "Emsembl Genome Browser" para a região do gene *IRF6*, em que o gene se encontra na posição da faixa vermelha vertical. Caso o *IRF6* não esteja visível, arraste a barra branca e preta inferior para baixo. As barras horizontais amarelas representam genes conhecidamente codificadores para proteínas, enquanto as azuis representam genes recentemente descritos na fita anti-sentido e as barras rosa referem-se a genes de RNAs não codificadores. Clique sobre *IRF6*, na figura. Uma caixa com uma breve descrição do gene se abrirá. Então, clique sobre o código de identificação, ENSG00000117595, logo após 'Gene'.

(8) A página que se abriu apresenta a descrição do gene *IRF6*, com informações sobre seus transcritos e produtos protéicos. Podemos acessar todas as variantes catalogadas pelo projeto 1000 genomas e outros projetos através do *link* 'Variation Table', na coluna à esquerda. Clique sobre ele.

(9) A página será atualizada e veremos uma lista de classes de variantes já descritas e o número de variantes por classes, além de outras informações. Ao final da página, clique sobre 'Show' logo à frente de 'ALL', para que todas as variantes sejam listadas. Espere a página carregar. Após carregada, por meio da busca de CTRL+F, busque pela posição da sua variante, no nosso caso, **209790735**.

(10) Veja que a variante já foi descrita e é possível notar ao menos oito entradas para esta posição, referentes aos transcritos alternativos do *IRF6* e aos diferentes bancos de dados em que a variante foi anotada.

(11) Clique sobre o primeiro SNP (rs2235371). Na página que se abre, vemos a descrição da variante e os possíveis alelos encontrados nesse lócus.

(12) Agora acessaremos os dados de frequência destes alelos em diferentes populações, depositados pelo Projeto 1000 Genomas. Para isso, clique na aba à esquerda em 'Population Genetics'.

(13) Já podemos visualizar os dados fornecidos pelo *1000 Genomes*, como os gráficos de frequências dos alelos nas diferentes populações. Notamos um gráfico para a frequência geral dos alelos, um para a população africana (AFR), um para a população americana (AMR), um para a asiática (ASN) e um para a população europeia (EUR).

(14) Com base nestes gráficos:

   a. A variante encontrada pelo sequenciamento da amostra do paciente equivale ao alelo de maior ou menor frequência?
   b. Qual população apresenta a maior frequência da variante encontrada no sequenciamento? E a com menor frequência?
   c. Levando-se em conta a sua frequência nas populações, você consideraria essa variante como patogênica, responsável por fissura lábio-palatina com herança dominante?
   d. Como você faria para testar a hipótese de que, em um sistema multifatorial, essa variante pode relacionar-se a aumento de suscetibilidade a fissura lábio-palatina?

Textos de apoio (P1-P4)

Glossário (P1-P4)

(Extraído de: National Center for Biotechnology Information (www.ncbi.nlm.nih.gov)

Accession number: An Accession number is a unique identifier given to a sequence when it is submitted to one of the DNA repositories (GenBank, EMBL, DDBJ). The initial deposition of a sequence record is referred to as version 1. If the sequence is updated, the version number is incremented, but the Accession number will remain constant.

BLAST: Basic Local Alignment Search Tool (Altschul *et al.*, J Mol Biol 215:403-410; 1990). A sequence comparison algorithm that is optimized for speed and used to search sequence databases for optimal local alignments to a query.

Blastn: nucleotide–nucleotide BLAST. blastn takes nucleotide sequences in FASTA format, GenBank Accession numbers, or GI numbers, and compares them against the NCBI Nucleotide databases.

Blastp: protein–protein BLAST. blastp takes protein sequences in FASTA format, GenBank Accession numbers, or GI numbers and compares them against the NCBI Protein databases.

BLAT: A DNA/protein sequence analysis program to quickly find sequences of 95% and greater similarity of length 40 bases or more. It may miss more divergent or shorter sequence alignments. BLAT on proteins finds sequences of 80% and greater similarity of length 20 amino acids or more. BLAT is not BLAST.

Build: A run of the genome assembly and annotation process of the set of products generated by that run.

CDS: coding region, coding sequence. CDS refers to the portion of a genomic DNA sequence that is translated, from the start codon to the stop codon, inclusively, if complete. A partial CDS lacks part of the complete CDS (it may lack either or both the start and stop codons). Successful translation of a CDS results in the synthesis of a protein.

Contig: A contiguous segment of the genome made by joining overlapping clones or sequences. A clone contig consists of a group of cloned (copied) pieces of DNA representing overlapping regions of a particular chromosome. A sequence contig is an extended sequence created by merging primary sequences that overlap. A contig map shows the regions of a chromosome where contiguous DNA segments overlap. Contig maps provide the ability to study a complete and often large segment of the genome by examining a series of overlapping clones, which then provide an unbroken succession of information about that region.

Draft sequence: Draft sequence refers to DNA sequence that is not yet finished but is generally of high quality (i.e., an accuracy of greater than 90%). Draft sequence data are mostly in the form of 10,000 base pair-sized fragments, the approximate chromosomal locations of which are known. The following keywords are associated with draft sequence: **phase 0**, light-pass coverage of a clone, generally only 1× coverage; **phase 1**, 4–10× coverage of a **BAC** clone (order and orientation of the fragments are unknown); and **phase 2**, 4–10× coverage of a BAC clone (order and orientation of the fragments are known). **Phase 3** refers to the completely **finished sequence**.

E-value: Expect value. The E-value is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size. It decreases exponentially with the score (S) that is assigned to a match between two sequences. Essentially, the E-value describes the random background noise that exists for matches between sequences. For example, an E-value of 1 assigned to a hit can be interpreted as meaning that in a database of the current size, one might expect to see one match with a similar score simply by chance. This means that the lower the E-value, or the closer it is to "0", the higher is the "significance" of the match. However, it is important to note that searches with short sequences can be virtually identical and have relatively high E-value. This is because the calculation of the E-value also takes into account the length of the query sequence. This is because shorter sequences have a high probability of occurring in the database purely by chance.

Entrez: a retrieval system for searching several linked databases. It provides access to the following NCBI databases: PubMed, GenBank, Protein, Structure, Genome, PopSet, OMIM, Taxonomy, Books, ProbeSet, 3D Domains, UniSTS, SNP, and CDD.

FASTA: The first widely used algorithm for similarity searching of protein and DNA sequence databases. The program looks for optimal local alignments by scanning the sequence for small matches called "words". Initially, the scores of segments in which there are multiple word hits are calculated ("init1"). Later, the scores of several segments may be summed to generate an "initn" score. An optimized alignment that includes gaps is shown in the output as "opt". The sensitivity and speed of the search are inversely related and controlled by the "k-tup" variable, which specifies the size of a "word". **Also refers to a format for a nucleic acid or protein sequence.**

Finished sequence: High-quality, low-error DNA sequence that is free of gaps. To qualify as a finished sequence, only a single error out of every 10,000 bases (i.e., an accuracy of 99.999%) is allowed.

Gap: A gap is a space introduced into an alignment to compensate for insertions and deletions in one sequence relative to another. To prevent the accumulation of too many gaps in an alignment, introduction of a gap causes the deduction of a fixed amount (the gap score) from the alignment score. Extension of the gap to encompass additional nucleotides or amino acid is also penalized in the scoring of an alignment.

GenBank: a database of nucleotide sequences from more than 100,000 organisms. Records that are annotated with coding region features also include amino acid translations. GenBank belongs to an international collaboration of sequence databases that also includes **EMBL** (The European Molecular Biology Laboratory) and **DDBJ** (DNA Data Bank of Japan).

GeneID: a unique identifier that is assigned to a gene record in **Entrez Gene**. It is an integer and is species specific. In other words, the integer assigned to dystrophin in human is different from that in any other species. For genomes that had been represented in LocusLink, the GeneID is the same as the LocusID. The GeneID is reported in RefSeq records as a 'db_xref' (e.g. /db_xref="GeneID:856646", in GenBank format).

HTGS: High-Throughput Genomic Sequences. The source of HTGS are large-scale genome sequencing centers; **unfinished sequence**s are in phases 0, 1, and 2, and **finished sequence**s are in phase 3.

IMAGE Consortium: Integrated Molecular Analysis of Genomes and their Expression. A consortium of academic groups that share high-quality, arrayed cDNA libraries and place sequence, map, and expression data of the clones in these arrays into the public domain. With the use of this information, unique clones can be rearrayed to form a "master array",

with the aim of ultimately having a representative cDNA from every gene in the genome under study. To date, human, mouse, rat, zebrafish, and *Xenopus laevis* genomes have been studied.

OMIM: Online Mendelian Inheritance in Man. OMIM is a directory of human genes and genetic disorders, with links to literature references, sequence records, maps, and related databases.

RefSeq: RefSeq is the NCBI database of reference sequences; a curated, non-redundant set including genomic DNA contigs, mRNAs and proteins for known genes, and entire chromosomes.

WGS sequence: Whole Genome Shotgun sequence. In this semi-automated sequencing technique, high-molecular-weight DNA is sheared into random fragments, size selected (usually 2, 10, 50, and 150 kb), and cloned into an appropriate vector. The clones are then sequenced from both ends. The two ends of the same clone are referred to as mate pairs. The distance between two mate pairs can be inferred if the library size is known and has a narrow window of deviation. The sequences are aligned using sequence assembly software. Proponents of this approach argue that it is possible to sequence the whole genome at once using large arrays of sequencers, which makes the whole process much more efficient than the traditional approaches.

Textos de apoio – P2

GenBank: The Nucleotide Sequence Database

Ilene Mizrachi (http://www.ncbi.nlm.nih.gov)

The GenBank sequence database is an annotated collection of all publicly available nucleotide sequences and their protein translations. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the ftp site. The release notes for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. GenBank growth statistics for both the traditional GenBank divisions and the (Whole Genome Shotgun) WGS division are available from each release. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release. See the growth from the last ten years in Figure 4.

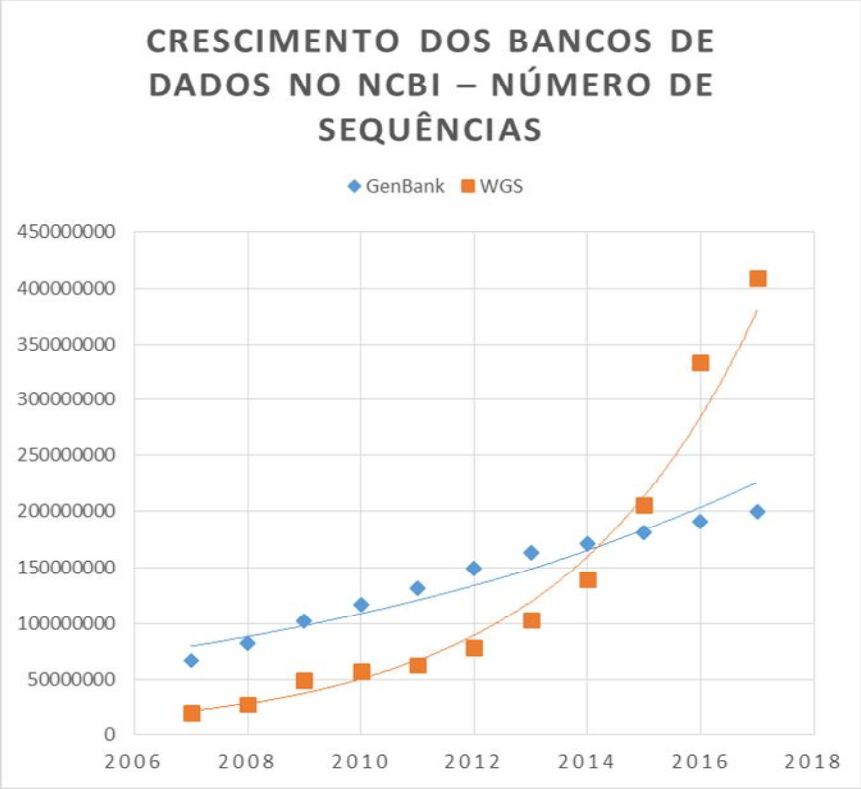

Figura 4: Crescimento do número de sequências depositadas nos bancos de dados GenBank e WGS, no NCBI, registradas nas releases de fevereiro de cada ano, entre 2007 e 2017.

GenBank and its collaborators receive sequences produced in laboratories throughout the world from more than 100,000 distinct organisms. Upon receipt of a sequence submission, the GenBank staff assigns an Accession number to the sequence and performs quality assurance checks. The submissions are then released to the public database, where the entries are retrievable by Entrez or downloadable by FTP.

Initially, GenBank was built and maintained at Los Alamos National Laboratory (LANL). In the early 1990s, this responsibility was awarded to NCBI through congressional mandate. NCBI undertook the task of scanning the literature for sequences and manually typing the sequences into the database. Staff then added annotation to these records, based upon information in the published article. All sequences are now deposited directly by the labs that generate the sequences. This is attributable to, in part, a requirement by most journal publishers that nucleotide sequences are first deposited into publicly available databases (DDBJ/EMBL/GenBank) so that the Accession number can be cited and the sequence can be retrieved when the article is published. NCBI began accepting direct submissions to GenBank in 1993 and received data from LANL until 1996.

In the mid-1990s, the GenBank database became part of the International Nucleotide Sequence Database Collaboration with the EMBL database (European Bioinformatics Institute, Hinxton, United Kingdom) and the Genome Sequence Database (GSDB; LANL, Los Alamos, NM). Subsequently, the GSDB was removed from the Collaboration (by the National Center for Genome Resources, Santa Fe, NM), and DDBJ (Mishima, Japan) joined the group. Each database has its own set of submission and retrieval tools, but the three databases exchange data daily so that all three databases should contain the same set of sequences. Members of the DDBJ, EMBL, and GenBank staff meet annually to discuss technical issues, and an international advisory board meets with the database staff to provide additional guidance. An entry can only be updated by the database that initially prepared it to avoid conflicting data at the three sites.

Because GenBank is an archival database and includes all sequence data submitted, there are multiple entries for some loci. Just as the primary literature includes similar experiments conducted under slightly different conditions, GenBank may include many sequencing results for the same loci. These different sequencing submissions can reflect genetic variations between individuals or organisms, and analyzing these differences is one way of identifying single nucleotide polymorphisms.

The Reference Sequence (RefSeq) Project

Kim Pruitt, Garth Brown, Tatiana Tatusova, and Donna Maglott. (http://www.ncbi.nih.gov)
Document created: October 09, 2002 / Updated: April 6, 2012

NCBI's Reference Sequence (RefSeq) database is a collection of taxonomically diverse, non-redundant and richly annotated sequences representing naturally occurring molecules of DNA, RNA, and protein. Included are sequences from plasmids, organelles, viruses, archaea, bacteria, and eukaryotes. Each RefSeq is constructed wholly from sequence data submitted to the International Nucleotide Sequence Database Collaboration (INSDC). Similar to a review article, a RefSeq is a synthesis of information integrated across multiple sources at a given time. RefSeqs provide a foundation for uniting sequence data with genetic and functional information. They are generated to provide reference standards for multiple purposes ranging from genome annotation to reporting locations of sequence variation in medical records. The RefSeq collection is available without restriction and can be retrieved in several different ways, such as by searching or by available links in NCBI resources, including PubMed, Nucleotide, Protein, Gene, and Map Viewer, searching with a sequence via BLAST, and downloading from the RefSeq FTP site..

NCBI's Reference Sequence (RefSeq) collection is a freely accessible database of naturally occurring DNA, RNA, and protein sequences. It is a unique resource because it provides a large, multi-species, curated sequence database representing separate but explicitly linked records from genomes to transcripts and translation products, as appropriate. Unlike the sequence redundancy found in the public sequence repositories that comprise the INSDC, (i.e., NCBI's GenBank, the European Nucleotide Archive [ENA], and the DNA Data Bank of Japan [DDBJ]), the RefSeq collection aims to provide, for each included species, a complete set of non-redundant, extensively cross-linked, and richly annotated nucleic acid and protein records. It is recognized, however, that the coverage and finishing of public sequence data varies from organism to organism so intermediate genomic records are provided in some circumstances.

The non-redundant nature of the RefSeq collection facilitates database inquiries based on genomic location, sequence, or text annotation. Be aware, however, that the RefSeq collection does include alternatively spliced transcripts encoding the same protein or distinct protein isoforms, in addition to orthologs, paralogs, and alternative haplotypes for some organisms, which will affect the outcome of a database query.

RefSeq records are based on sequence records submitted to the INSDC. However, the RefSeq collection is a distinct database. The public archival databases house sequences and annotations supplied by original authors and cannot be altered by others. The RefSeq collection differs from the archival databases in the same way that a review article differs from a related collection of primary research articles on the same subject. Each RefSeq record represents a synthesis, by a person or group, of the primary information that was generated and submitted by others. Other organizing principles or standards of judgment are possible, which is why the work is attributed to the synthesizing "editors". The RefSeq dataset is curated on an ongoing basis by collaborating groups and by NCBI staff. Sequence records are presented in a standard format and subjected to computational validation. The INSDC source of the RefSeq record, the curation status, and attribution to the curation group are also indicated.

The RefSeq collection establishes a useful baseline for integrating diverse data types, including sequence, genetic, expression, and functional information, into one consistent framework with a uniform set of conventions and standards. The RefSeq collection supports the following activities:

a) genome annotation
b) gene characterization
c) comparative genomics
d) reporting sequence variation, and
e) expression studies

The RefSeq collection is the result of data extraction from INSDC submissions, curation, and computation, combined with extensive collaboration with authoritative groups. Each molecule is annotated as accurately as possible with the organism name, strain (or breed, ecotype, cultivar, or isolate), gene symbol for that organism, and informative protein name. Collaborations with authoritative groups outside of NCBI provide a variety of information, including curated sequence data, nomenclature, feature annotations, and links to external organism-specific resources. When no collaboration has been established, NCBI staff assembles the data from the INSDC submission. Each record has a COMMENT, indicating the level of curation that it has received, and attribution of the collaborating group. Thus, a RefSeq record may be an essentially unchanged, validated copy of the original INSDC submission, or include updated or additional information supplied by collaborators or NCBI staff.

If multiple INSDC submissions represent the same molecule for an organism, the "best" sequence is chosen to represent as the RefSeq record. Known mutations, sequencing errors, cloning artifacts and erroneous annotation are avoided. Sequences are validated to confirm that the genomic sequence corresponding to an annotated mRNA feature matches the mRNA sequence record, and that coding region features translate into the corresponding protein sequence.

Working groups using distinct process pipelines compile the RefSeq collection for different organisms (Figure 5). RefSeq records are provided via several distinct approaches including:

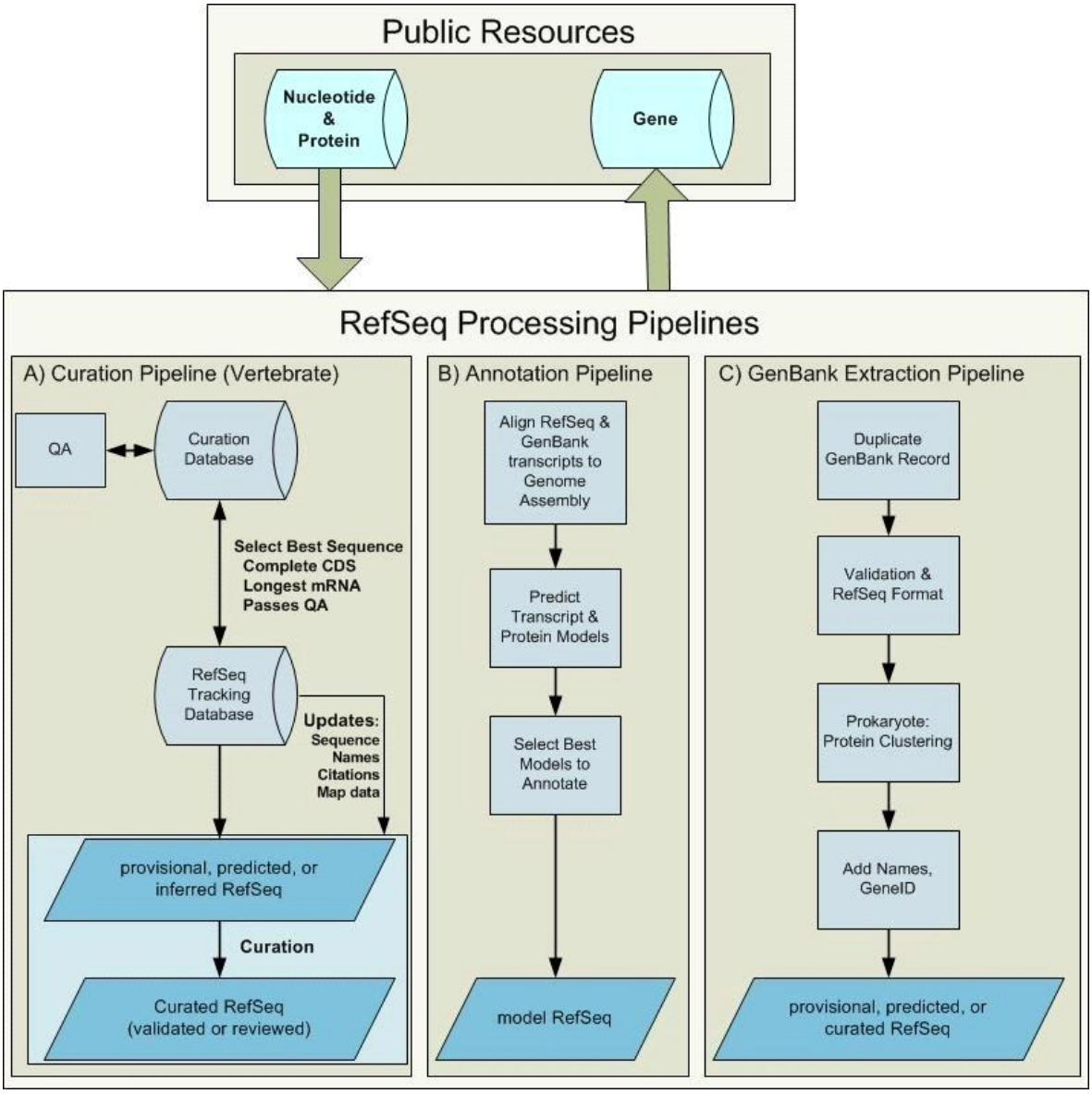

Figure 5. RefSeq Processing Pipelines. Sequence data deposited in the public archival databases is available for RefSeq processing. Processing pipelines include the vertebrate curation pipeline, the computational genome annotation pipeline, and extraction>

The Entrez Search and Retrieval System

Jim Ostell (http://www.ncbi.nih.gov)
Document created: October 9, 2002 / Updated: August 13, 2003

Entrez is the text-based search and retrieval system used at NCBI for all of the major databases, including PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, OMIM, and many others. Entrez is at once an indexing and retrieval system, a collection of data from many sources, and an organizing principle for biomedical information. The first version of Entrez was distributed by NCBI in 1991 on CD-ROM. At that time, it consisted of nucleotide sequences from GenBank and PDB; protein sequences from translated GenBank, PIR, SWISS-PROT, PDB, and PRF; and associated citations and abstracts from MEDLINE (now PubMed).

An Entrez “node” is a collection of data that is grouped together and indexed together. It is usually referred to as an Entrez database. In the first version of Entrez, there were three nodes: published articles, nucleotide sequences, and protein sequences (see figure, central nodes). Each node represents specific data objects of the same type, e.g., protein sequences, which are each given a unique ID (UID) within that logical Entrez Proteins node. Records in a node may come from a single source (e.g., all published articles are from PubMed) or many sources (e.g., proteins are from translated GenBank sequences, SWISS-PROT, or PIR) (Figure 4).

A data-retrieval system succeeds when you can retrieve the same data you put in. A discovery system is intended to let you find more information than appears in the original data. By making links between selected nodes and making computed associations within the same node, Entrez is designed to infer relationships between different data that may suggest future experiments or assist in interpretation of the available information, although it may come from different sources.

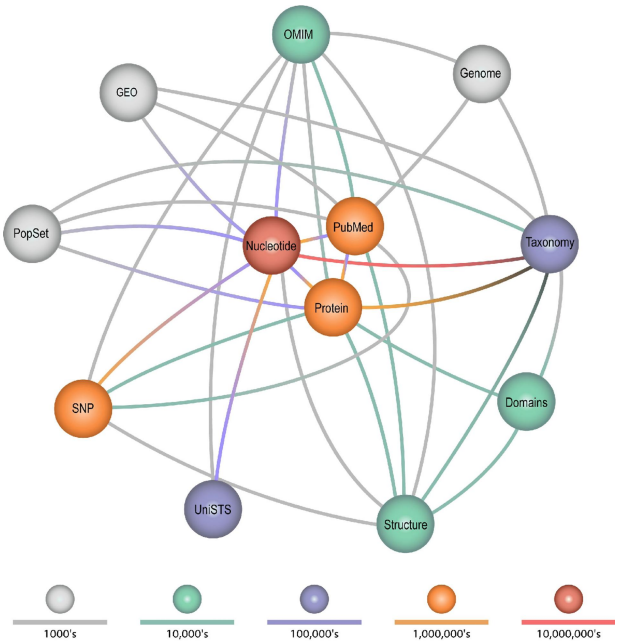

Figure 6: The original version of Entrez had just three nodes: nucleotides, proteins, and PubMed abstracts (center of the figure). In 2001, Entrez had nearly 20 nodes, as illustrated. Today, it contains approximately 50 nodes.

Texto de apoio – P4

The BLAST Sequence Analysis Tool

Tom Madden (http://www.ncbi.nih.gov)
Document created: October 9, 2002 / Updated: August 13, 2003

The comparison of nucleotide or protein sequences from the same or different organisms is a very powerful tool in molecular biology. By finding similarities between sequences, scientists can infer the function of newly sequenced genes, predict new members of gene families, and explore evolutionary relationships. Now that whole genomes are being sequenced, sequence similarity searching can be used to predict the location and function of protein-coding and transcription regulation regions in genomic DNA.

Basic Local Alignment Search Tool (BLAST) is the tool most frequently used for calculating sequence similarity. BLAST comes in variations for use with different query sequences against different databases. All BLAST applications, as well as information on which BLAST program to use and other help documentation, are listed on the BLAST homepage [http://www.ncbi.nlm.nih.gov/BLAST/].

The way most people use BLAST is to input a nucleotide or protein sequence as a query against all (or a subset of) the public sequence databases, pasting the sequence into the textbox on one of the BLAST Web pages [http://www.ncbi.nlm.nih.gov/BLAST/]. This sends the query over the Internet, the search is performed on the NCBI databases and servers, and the results are posted back to the person's browser in the chosen display format. The best known output is the default display from BLAST Web pages, the so-called “traditional report”. As well as obtaining BLAST results in the traditional report, results can also be delivered in structured output, such as a hit table, XML, or ASN.1. The optimal choice of output format depends upon the application.

How BLAST Works: The Basics

The BLAST algorithm is a heuristic program, which means that it relies on some smart shortcuts to perform the search faster. BLAST performs "local" alignments. Most proteins are modular in nature, with functional domains often being repeated within the same protein as well as across different proteins from different species. The BLAST algorithm is tuned to find these domains or shorter stretches of sequence similarity. The local alignment approach also means that a mRNA can be aligned with a piece of genomic DNA, as is frequently required in genome assembly and analysis. If instead BLAST started out by attempting to align two sequences over their entire lengths (known as a global alignment), fewer similarities would be detected, especially with respect to domains and motifs.

When a query is submitted via one of the BLAST Web pages, the sequence, plus any other input information such as the database to be searched, word size, expect value, and so on, are fed to the algorithm [http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/BLAST_algorithm.html] on the BLAST server. BLAST works by first making a look-up table of all the “words” (short subsequences, which for proteins the default is three letters) and “neighboring words”, i.e., similar words in

the query sequence. The sequence database is then scanned for these “hot spots”. When a match is identified, it is used to initiate gap-free and gapped extensions of the “word”.

BLAST does not search GenBank flatfiles (or any subset of GenBank flatfiles) directly. Rather, sequences are made into BLAST databases. Each entry is split, and two files are formed, one containing just the header information and one containing just the sequence information. These are the data that the algorithm uses. If BLAST is to be run in “stand-alone” mode, the data file could consist of local, private data, downloaded NCBI BLAST databases, or a combination of the two.

After the algorithm has looked up all possible "words" from the query sequence and extended them maximally, it assembles the best alignment for each query–sequence pair and writes this information to an SeqAlign data structure (in ASN.1 ; also used by Sequin). The SeqAlign structure in itself does not contain the sequence information; rather, it refers to the sequences in the BLAST database.

The BLAST Formatter, which sits on the BLAST server, can use the information in the SeqAlign to retrieve the similar sequences found and display them in a variety of ways. Thus, once a query has been completed, the results can be reformatted without having to re-execute the search. This is possible because of the QBLAST [http://www.ncbi.nlm.nih.gov/BLAST/blast_overview.html#blastq] system.

BLAST Scores and Statistics

Once BLAST has found a similar sequence to the query in the database, it is helpful to have some idea of whether the alignment is “good” and whether it portrays a possible biological relationship, or whether the similarity observed is attributable to chance alone. BLAST uses statistical theory [http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html] to produce a bit score and expect value (E-value) for each alignment pair (query to hit).

The bit score gives an indication of how good the alignment is; the higher the score, the better the alignment. In general terms, this score is calculated from a formula that takes into account the alignment of similar or identical residues, as well as any gaps introduced to align the sequences. A key element in this calculation is the “substitution matrix ”, which assigns a score for aligning any possible pair of residues. The BLOSUM62 matrix is the default for most BLAST programs, the exceptions being blastn and MegaBLAST (programs that perform nucleotide– nucleotide comparisons and hence do not use protein-specific matrices). Bit scores are normalized, which means that the bit scores from different alignments can be compared, even if different scoring matrices have been used.

The E-value gives an indication of the statistical significance of a given pairwise alignment and reflects the size of the database and the scoring system used. The lower the E-value, the more significant the hit. A sequence alignment that has an E-value of 0.05 means that this similarity has a 5 in 100 (1 in 20) chance of occurring by chance alone. Although a statistician might consider this to be significant, it still may not represent a biologically meaningful result, and analysis of the alignments is required to determine “biological” significance.

Nucleotide Sequence Databases for BLAST

- **nr/nt (Nucleotide collection)**
- **nr/nt Non**-redundant nucleotide collection
  Non-redundant sequences merged into one from GenBank + RefSeq Nucleotides + EMBL + DDBJ + PDB sequences (excluding HTGS0,1,2, EST, GSS, STS, PAT, WGS). ".
- **refseq_rna**
  RNA entries from NCBI's Reference Sequence project
- **refseq_genomic**
  Genomic entries from NCBI's Reference Sequence project
- **refseq representative genomes**
-
- **est**
  Database of GenBank + EMBL + DDBJ sequences from EST Divisions
- **est_human**
  Human subset of est.
- **est_mouse**
  Mouse subset.
- **est_others**
  Non-Mouse, non-Human subset of est.
- **gss**
  Genome Survey Sequence, includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences.
- **htgs**
  Unfinished High Throughput Genomic Sequences: phases 0, 1 and 2 (finished, phase 3 HTG sequences are in nr)
- **pat**
  Nucleotides from the Patent division of GenBank.
- **pdb**
  Sequences derived from the 3-dimensional structure from Brookhaven Protein Data Bank
- **month**
  All new or revised GenBank + EMBL + DDBJ + PDB sequences released in the last 30 days.
- **dbsts**
  Database of GenBank+EMBL+DDBJ sequences from STS Divisions .
- **NCBI genomes (chromosome)**
- A database with complete genomes and chromosomes from the NCBI Reference Sequence project.
- **wgs**
  A database for whole genome shotgun sequence entries.
- **SRA (Sequence read archive)**
  Raw sequence data from "next-generation" sequencing technologies including Illumina, 454, IonTorrent, Complete Genomics, PacBio and OxfordNanopores. In addition to raw sequence data, SRA now stores alignment information in the form of read placements on a reference sequence.