DESTAQUES EVOLUÇÃO

Os 5 passos essenciais para construção de árvores filogenéticas

By **Filipe Zimmer**

27 de março de 2021

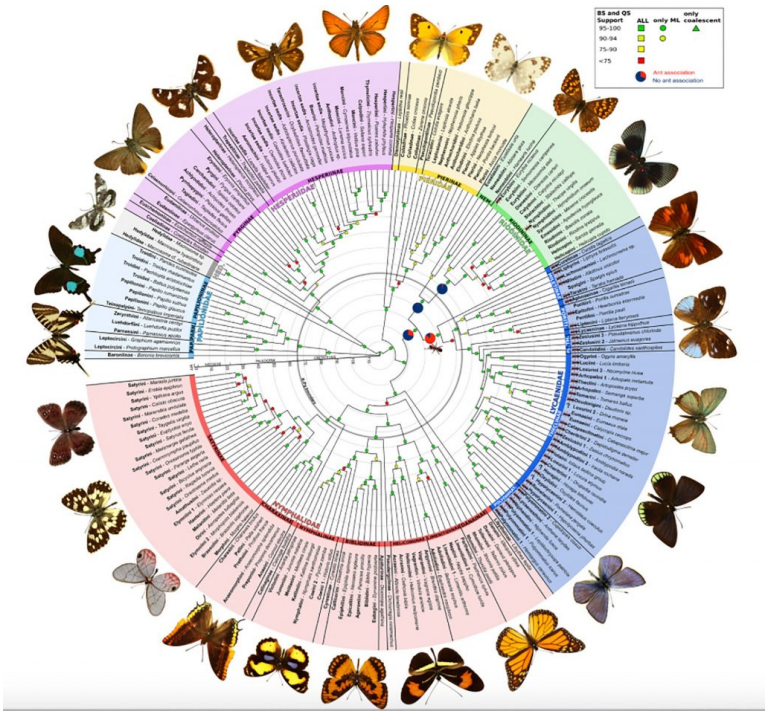
Os 5 passos essenciais para construção de árvores filogenéticas

Filipe Zimmer Dezordi

Revisão: Diego Mariano

BIOINFO – Revista Brasileira de Bioinformática. Edição #01. Julho, 2021.

DOI: 10.51780/978-6-599-275326-20



At last, butterflies get a bigger, better evolutionary tree.
Fonte: https://www.floridamuseum.ufl.edu/science/at-last-butterflies-get-a-bigger-better-evolutionary-tree/

Caro(a) leitor(a), esse será o primeiro de uma série de pequenos artigos com dicas em bioinformática. A iniciativa vêm da produção de conteúdos na minha página do Instagram, e a ideia é reunir uma coletânea de dicas voltadas para um determinado assunto, neste primeiro texto, falarei sobre construção de árvores filogenéticas.

Eu sigo a filosofia do "*Antes feito do que perfeito*", pois é a única forma que eu tenho de conciliar um doutorado, meus desenhos e a produção de conteúdos para uma página, então, se você está atrás de conteúdos super explicados, diferenças entre filogenética e filogenômica; filograma ou dendograma;

inferência bayesiana ou de máxima verossimilhança, eu recomendo a você procurar algum livro de bioinformática ou artigos científicos, pois o conteúdo desses artigos será extremamente direto e sem referências (que feio para um cientista né? mas vamos lá, quantas vezes você rodou uma ferramenta porque seu orientador mandou, sem nem se perguntar o porque? haha, antes feito do que perfeito!). Então vamos lá!

Passo 0: Comece pelo suplementar

Como todo bom programador, sabemos que a contagem na computação começa sempre pelo caracter 0. Fora isso, eu resolvi não inserir esse passo como algo relacionado especificamente a construção de árvores filogenéticas, mas é uma etapa que você sempre deve tentar realizar no seu estudo, seja em bioinformática ou em bancada: Comece pelo suplementar!

Normalmente, no material suplementar de artigos nós colocamos informações adicionais que servem para reforçar as informações descritas no artigo, mas que não são informações essenciais para apresentação do texto. Uma das boas práticas de pesquisa é documentar tudo que está sendo feito, na bancada por exemplo, podemos documentar linhagens celulares ou cepas que estamos usando, os kits utilizados nas análises e até os parâmetros setados nos equipamentos. Em bioinformática, podemos documentar praticamente tudo: Origem das sequências, versões das ferramentas, etapas realizadas, linhas de comando utilizadas, modificações realizadas entre um arquivo e outro.

Quando estamos falando de análises evolutivas, um suplementar muito importante é informações sobre as sequências utilizadas, aqui você já começa a fazer uma análise dos seus dados, levantando informações secundárias (metadados) que ajudarão você a anotar a sua árvore filogenética ao final das análises, veja o exemplo:

Espécie	Strain	Genus	Sub-Family	Access	Host
Microhyla letovirus 1	MLeV	Alphaletovirus	Letovirinae	GECV01031551.1	Frog
Bat coronavirus CDPHE15	BtCoV-CDPHE15 *	Alphacoronavirus	Orthocoronavirinae	NC_022103.1	Bat
Bat coronavirus HKU10	BatCoV-HKU10	Alphacoronavirus	Orthocoronavirinae	NC_018871.1	Bat
Rhinolophus ferrumequinum alphacoronavirus I	BtRf-AlphaCoV-HuB2013	Alphacoronavirus	Orthocoronavirinae	NC_028814	Bat
Human coronavirus 229E	HCoV-229E	Alphacoronavirus	Orthocoronavirinae	AF304460.1	Human
Lucheng Rn rat coronavirus	LRNV	Alphacoronavirus	Orthocoronavirinae	NC_032730.1	Rodent
Mink coronavirus 1	MCoV-WD1127	Alphacoronavirus	Orthocoronavirinae	NC_023760.1	Mink
Miniopterus bat coronavirus 1	BtCoV-1A	Alphacoronavirus	Orthocoronavirinae	NC_010437.1	Bat
Miniopterus bat coronavirus HKU8	BtCoV-HKU8	Alphacoronavirus	Orthocoronavirinae	NC_010438.1	Bat
Myotis ricketti alphacoronavirus Sax-2011	BtMr-AlphaCoV-SAX2011	Alphacoronavirus	Orthocoronavirinae	NC_028811	Bat
Nyctalus velutinus alphacoronavirus SC-2013	BtNv-AlphaCoV-SC2013	Alphacoronavirus	Orthocoronavirinae	NC_028833.1	Bat
Pipistrellus kuhlii coronavirus 3398	BatCoV-Ita4	Alphacoronavirus	Orthocoronavirinae	MH938449.1	Bat
Porcine epidemic diarrhea virus	PEDV	Alphacoronavirus	Orthocoronavirinae	NC_003436.1	Pig
Scotophilus bat coronavirus 512	BtCoV-512	Alphacoronavirus	Orthocoronavirinae	NC_009657.1	Bat
Rhinolophus bat coronavirus HKU2	BtCoV-HKU2	Alphacoronavirus	Orthocoronavirinae	NC_009988	Bat
Human coronavirus NL63	HCoV-NL63	Alphacoronavirus	Orthocoronavirinae	NC_005831.2	Human
NL63-related bat coronavirus strain BtKYNL63-	BtKYNL63-9b	Alphacoronavirus	Orthocoronavirinae	NC_032107.1	Bat

Exemplo de material suplementar para vírus da família Coronaviridae

Este é um exemplo de suplementar que pode ser criado para a família *Coronaviridae*, onde na primeira coluna temos o nome da "espécie" viral, presente no ICTV, na segunda temos o código da cepa (que podemos usar como nome no arquivo fasta, o que reduz a poluição textual nos *tips* da filogenia), nas colunas adicionais temos os metadados, onde nas colunas 2 e

3 temos informações da taxonomia desses vírus, que podemos usar para colorir os clados (agrupamentos) nas árvores filogenéticas, na coluna 5 temos o código de acesso à sequência do vírus (o que permite a reprodutibilidade das suas análises) e na última coluna temos os hospedeiros, que podemos utilizar para entender a possível co-evolução vírus hospedeiro ou a transmissão entre hospedeiros após a construção e anotação das árvores.

Enfim, além de facilitar a sua vida na hora de responder aos revisores do artigo, você pode gerar *insights* sobre seus resultados além de permitir a reprodutibilidade da sua pesquisa, então, comece pelo suplementar! (vai dar muito menos trabalho do que fazer o suplementar depois de meses de análises).

Dicas do Passo 0:

- Use sed, awk, grep e loop em bash (for) para automatizar a edição e estruturação desses materiais;
- Use o E-utilities do NCBI (em python as funções estão na biblioteca biopython), para recuperar automaticamente as sequências por código ou por informação taxonômica, preparei um script bem basicão pra isso, disponível nesse link;
- Limite o nome das suas sequências ao código de acesso mais alguma informação curta, nesse exemplo do texto, o código da cepa viral;
- Sempre tente trabalhar com padrões e com análises reprodutíveis, por menor que seja a tarefa, evite fazer de forma manual, acredite em mim, você vai precisar refazer a mesma tarefa pelo menos 3 vezes até a publicação do artigo (valor diretamente da minha imaginação).

Passo 1: Construa um bom conjunto de sequências

Assim como em todo bom experimento de bancada, para boas análises de bioinfo precisamos de amostras de qualidade. Se em bancada precisamos de material genético ou proteínas em amostras de alta qualidade (material intacto e sem contaminação), em bioinformática precisamos de sequências confiáveis, mas o que isso significa?

É normal que ocorram erros durante o sequenciamento, ou até sequenciamentos incompletos, o que resulta, ao final das etapas de montagem, em sequências repletas de NNNNs (quando você não estiver trabalhando com sequências *hard-masked*), o que concomitantemente resulta em sequências proteicas repletas de XXX após as análises de obtenção de ORFs. Caso você tenha sequências com esses caracteres, saiba que você estará perdendo informação genética nas estimativas evolutivas, dependendo do foco do seu estudo, e se essas regiões não resolvidas estiverem em regiões chaves para as análises (como domínios ou motivos

proteicos, por exemplo), pode ser que você tenha dificuldade em gerar árvores bem suportadas.

Para resolver este pequeno problema, primeiro você deve ter um bom conhecimento teórico da molécula que você está estudando, se as regiões não resolvidas estiverem fora das regiões chave para as análises evolutivas, talvez o impacto na topologia final da árvore seja pequeno, mas é sempre bom ter noção da qualidade dos dados ao início de qualquer análise.

Outro problema que pode aparecer, é a criação de *datasets* gigantescos, cheios de sequência que não trarão nenhuma informação valiosa para seus resultados, explico. Uma das análises normalmente utilizadas em análises evolutivas é recuperar as sequências por duas estratégias:

- Pegar sua/suas sequência(s) alvo e realizar uma análise de BLAST para recuperar sequências similares (e fique de olho no BLAST que você vai rodar, blastn padrão roda o megablast e você só recupera sequências altamente similares), geralmente, o pesquisador que vai por esse caminho tende ou a recuperar o arquivo fasta com os *matches*, ou selecionar no olho alguns *matches*;
- Pegar sequências de referência diretamente em bancos de dados personalizados, por exemplo: BOLD systems, FlyBase, Vectorbase.

Ambas estratégias podem resultar em um conjunto diverso de sequências, mas que podem esconder no meio desse *balaio de gato* algumas sequências redundantes. Essas sequências redundantes, na minha visão, podem ser tanto sequências 100% idênticas, ou sequências com mais de 99% de identidade de uma mesma espécie, e o potencial dessas sequências redundantes em aumentar o tempo computacional ou gerar árvores com topologias com politomia é grande (sim, vou usar termos subjetivos aqui).

Então sempre é bom pensar na sua questão biológica (e eu sempre vou bater muito nessa tecla), se você vai fazer uma filogenia de genes/proteínas distribuídos ao longo de um grupo taxonômico, por exemplo, proteínas de envelope dos Flavivirus, é necessário você ter a sequência da proteína de 200 linhagens do vírus da Dengue? ou é melhor você construir um banco de proteínas com as proteínas Env de todas as “espécies” do gênero Flavivirus?

Mas se você estiver trabalhando com um gene de subpopulações da mesma espécie, por exemplo, um gene de resistência à inseticidas em *Aedes aegypti,* então você deve usar o maior número de sequências possíveis de diferentes populações. Esses questionamentos iniciais quase sempre vão iluminar a escolha da estratégia de recuperação de sequências.

Dicas do Passo 1:

- Tenha conhecimento dos resíduos não resolvidos (NNNNNs/XXXXXs) nas sua sequências, você pode contabilizar isso facilmente com alguns scripts básicos;
- Defina um padrão claro para recuperação das sequências, e documente isso no material suplementar (Passo 0);
- Defina para o seu estudo o que seriam sequências redundantes, e remova-as, cd-hit e cd-hit-est são boas ferramentas para isso;
- Faça uma última checagem dos seus dados: Nomes das sequências (fasta *headers*) formatados; Estratégia de recuperação das sequências bem clara; Remoção das sequências redundantes, se necessário.

Passo 2: Pense que você está alinhando dados biológicos, não apenas caracteres digitais

Praticamente as duas metodologias de análise evolutiva mais robustas utilizam a mesma lógica básica: A melhor árvore filogenética será a que explica melhor os dados de entrada, seja por probabilidade de máxima verossimilhança (*Maximum Likelihood*), ou pelas melhores topologias em populações imensas de árvores filogenéticas (inferência Bayesiana). E um dos dados de entrada mais importante que temos, é justamente o alinhamento das sequências. Um alinhamento incorreto, sempre resultará em uma árvore filogenética incorreta, dessa forma, vamos pensar em como realizar uma boa análise de alinhamento!

Antigamente basicamente cada uma das várias estratégias de alinhamento era implementada em uma ferramenta específica, até que um grupo maravilhoso de pessoas criou a ferramenta MAFFT, a qual eu uso desde a graduação, por dois principais motivos:

- Sua disponibilidade em plataforma web ou por linha de comando, com fácil implementação;
- A gama de estratégias de alinhamento bem como a gama de parâmetros que podem ser ajustados para otimização das análises.

Dessa forma, caso você venha a utilizar o MAFFT, ou qualquer outra ferramenta para alinhamento de sequências, é sempre bom ter em mente que você está trabalhando com informações biológicas, e alguns parâmetros podem/devem ser ajustados, para inserir um sentido biológico na análise computacional. Entre os parâmetros que devemos ficar de olho no MAFFT são:

- UPPERCASE / lowercase: Caso você esteja trabalhando com sequências nucleotídicas *soft-masked*, é melhor setar *same as input*, para evitar confusões entre nucleotídeos em letras maiúsculas/minúsculas.
- Direction of nucleotide sequences: Essa opção permite corrigir o sentido das sequências de nucleotídeos, é muito importante caso algum grupo tenha depositado a sequência invertida nos bancos de dados;
- Scoring matrix: Matriz Blosum 62 ou 80 para proteínas mais conservadas, 30 ou 45 para menos conservadas e Matriz PAM 1 ou 20 para sequências de nucleotídeos mais conservadas, e PAM 200 para menos conservadas.
- A estratégia de alinhamento pode ser ajustada se você tiver um conhecimento da estrutura das sequências a serem alinhadas (um ou vários domínios conservados por exemplo), geralmente a opção *default* **Auto** serve pra maioria dos casos, pois a ferramenta irá detectar qual será a melhor abordagem de alinhamento de acordo com o conteúdo do arquivo fasta.

Após a etapa de alinhamento, você pode utilizar uma ferramenta para visualizar o resultado, eu recomendo a ferramenta Aliview por vários motivos (**veja esse post**). Nessa primeira visualização com Aliview, você já terá noção da qualidade do alinhamento (checando os sítios conservados e não conservados pelas opções de visualização, sério mesmo, veja esse post), sequências muito divergentes no alinhamento podem:

- Ter sido depositadas incorretamente no banco de dados;
- Ser de baixa qualidade (gerada por um sequenciamento de baixa qualidade);
- Estar invertida;
- Simplesmente ser uma sequência evolutivamente divergente, e cabe a você se vale a pena investir energia em tempo para solucionar o problema dessa sequência específica, ou simplesmente removê-la do alinhamento (*spoiler* do Passo 3).

Então sempre devemos ter em mente nossa pergunta, e quais os dados necessários para respondê-la, e por último, mas não menos importante, o quanto de tempo/recursos temos para investir em determinado problema.

Dicas do Passo 2:

- Tomar cuidado com o possível sentido inverso de algumas sequências no seu dataset;
- Ajustar as matrizes de distância;
- Crie um pequeno alinhamento de referência, e use o MAFFT -add (spoiler passo 3);

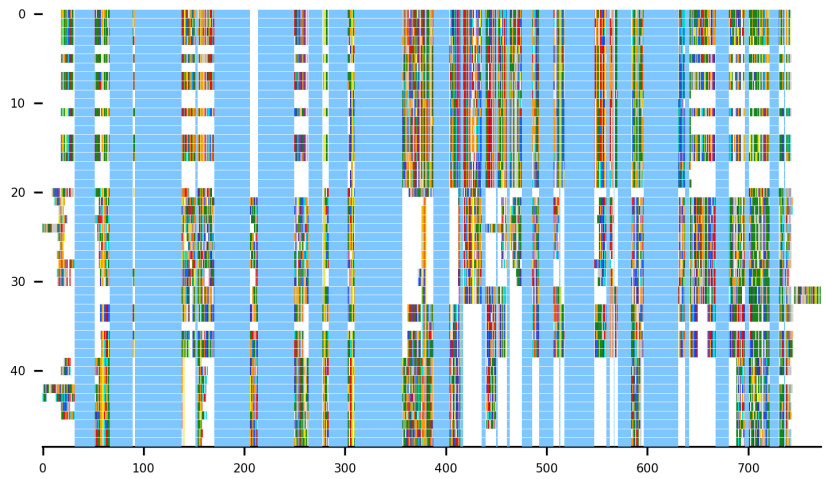
Passo 3: Nem sempre mais é melhor, edite seu alinhamento, mas tente evitar os vieses do operador

Uma prática comum antes de partir para as análises filogenéticas é realizar a edição do alinhamento. Essa edição serve geralmente para remover sequências, ou regiões específicas do alinhamento, que resultam em algum ruído no alinhamento, o que potencialmente modificará a topologia final da árvore. Normalmente essa edição é realizada de forma a remover sequências "pobremente" alinhadas, ou regiões cheias de SNPs ou *Indels*, que não trarão informações para resolução dos agrupamentos da filogenia.

Nos primórdios da bioinformática, essa edição era feita na mão, o que criava um viés gigantesco entre os estudos, pois a forma que eu editaria um alinhamento na mão, você não editaria da mesma forma, então foram surgindo algumas ferramentas que automatizam algumas dessas etapas, deixando apenas poucos detalhes para serem ajustados na mão.

Uma ferramenta publicada recentemente e com um funcionamento excelente é a ferramenta CIAlign, com essa ferramenta você pode realizar inúmeros tipos de análises, desde cálculo de matriz de distância, o que permite você estimar a identidade média entre as sequências no seu alinhamento (e você pode calcular essa identidade das sequências de referência, e das sequências que você está estudando para estimar se está tudo certinho). Você pode remover sequências que não estão bem alinhadas (inclusive usando o limiar de distância estipulado na análise anterior); remover os gaps, remover sequências pequenas (estipulando um tamanho mínimo específico), e remover as regiões do início e do fim de cada sequência que podem apresentar ruídos.

Além de todas as possibilidades de análise, a cada análise o CIAlign gera figuras (desde que isso seja solicitado) do esquema de edição do alinhamento, como na figura abaixo:



output gráfico da ferramenta CIAlign

Nesse caso, temos um alinhamento de cerca de 50 sequências (eixo Y) de quase 800 aminoácidos (eixo X), onde a região em azul (detectadas como inserções ou regiões não conservadas) foram removidas.

Existem outras ferramentas para edição, mas o ClAlign além de implementar praticamente todas as funcionalidades das demais ferramentas, implementou a geração dessas figuras de input e disponibilizou o código de fácil implementação, basta instalar (ou baixar o script e rodar com python) e testar as possibilidades (farei uma postagem futura só com o ClAlign, pois tem estratégias bem interessantes de uso dessa ferramenta).

Dicas do Passo 3:

- Instale o ClAlign;
- Faça uma análise de matriz de distância do seu alinhamento;
- Utilize a média da matriz de distância (lembre de remover os *outliers*) como um valor de *threshold* para remoção das sequências divergentes;
- Remova as pontas com ruídos e os gaps;
- Faça uma última visualização com o Aliview para ver se está tudo certo com o alinhamento;
- Se você criar um alinhamento de referência e usar o mafft -add, a etapa de edição praticamente não é necessária!
- Documente tudo que seja importante para a reprodutibilidade do seu estudo, sim vou voltar no Passo 0 toda hora!

Passo 4: Várias estratégias e várias ferramentas, como rodar a análise filogenética correta?

Uns anos atrás, eu escreveria um passo só falando sobre a estimativa dos modelos evolutivos, mas praticamente as ferramentas mais utilizadas hoje já calculam o modelo evolutivo e o aplicam na análise filogenética.

Um modelo evolutivo é basicamente uma fórmula estatística que explica como os nucleotídeos ou aminoácidos vão mudar no seu alinhamento, e os parâmetros de otimização (variações gamma e similares) explicam com que frequência essas mudanças acontecem. Se você pretende utilizar uma ferramenta de análise filogenética que não estima o modelo automaticamente, indico o ModelFinder como ferramenta para realizar esta estimativa e depois aplicar o modelo na sua análise evolutiva.

Feitas as considerações sobre os modelos, atualmente existem 2 métodos largamente utilizados para reconstruções evolutivas, o método de máxima verossimilhança e a inferência bayesiana, quando utilizar cada um deles?

Métodos de máxima verossimilhança geralmente são empregados em análises iniciais (uma análise rápida para verificar se o alinhamento do jeito que foi construído é suficiente para gerar uma árvore com topologia que faz sentido). Essas análises geralmente utilizam um método de cálculo rápido para o suporte de ramo (aLRT ou *ultrafast-bootstrap*), pois nessa análise

inicial precisamos de um resultado rápido e que consuma baixo poder computacional. Após essa análise inicial (que fica indicado aqui ser feita com o PhyML online, ou com a ferramenta fasttree), você pode realizar as análises pra valer, eu tenho utilizado muito a ferramenta IQ-TREE, devido sua fácil implementação, sua velocidade, estimativa automática de modelo e uma série de opções de análise que essa ferramenta fornece.

Normalmente nessas análises de máxima verossimilhança usamos o valor de suporte de ramo para estimar a confiabilidade das nossas análises, o valor padrão é o bootstrap (cada valor de bootstrap implica uma réplica de análise onde o alinhamento foi embaralhado e foi gerada uma árvore específica, então se foram feitas 100 réplicas, e um clado apresenta 80 no valor de bootstrap, significa que em 80 árvores aquele clado foi reconstruído daquela forma), e podemos ter alguns valores que se equivalem ao bootstrap, mas executam uma análise mais rápida, como o aLRT (equivalência à 80% do bootstrap tradicional) e o ultrafast-bootstrap (ficarei devendo a equivalência).

E quando usar a inferência bayesiana no lugar da máxima verossimilhança? Na verdade, poucas questões na biologia terão uma resposta padrão, ou correta em 100% dos casos. No caso da escolha do método, não existe uma explicação do porquê usar a bayesiana no lugar da máxima verossimilhança, mas...

Quando não conseguimos clados bem suportados por análise de máxima verossimilhança, normalmente rodamos análises bayesianas, eu uso o MrBayes para análises bayesianas mais simples, e colegas usam o BEAST para inferência bayesiana com datação, onde você pode inserir informação de datações de sequências ancestrais, e a análise estipulará a datação da formação dos clados na árvore evolutiva (isso também está aplicado no IQ-TREE para máxima verossimilhança, mas ainda está um tanto limitado).

Outro ponto que temos que ter em mente é a disponibilidade de recursos computacionais, quanto mais complexa a análise, mais poder de processamento é necessário, então em uma regra extremamente geral, em ordem crescente de custo computacional teríamos: ML com aLRT -> ML com boostrap -> Bayesiana. Fique atento(a) na literatura, quais métodos vem sendo desenvolvido e quais as vantagens e desvantagens de cada método para cada tipo de problema biológico (e não acredite em respostas definitivas para tudo, geralmente quem oferta essas respostas “não sabe que não sabe”).

Dicas do Passo 4:

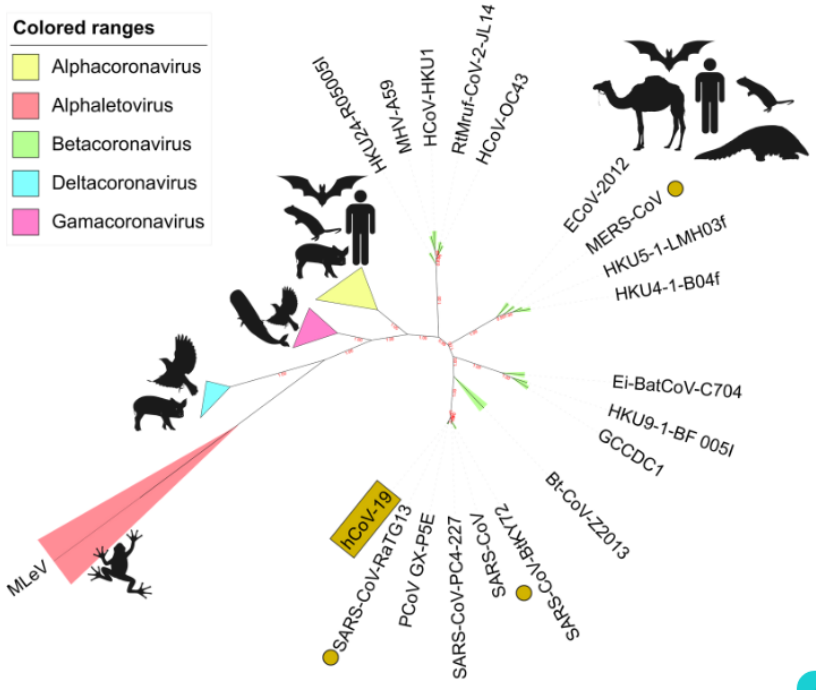
- Tenha um bom alinhamento (Passos anteriores);
- Faça uma análise de teste (PhyML aLRT ou fasttree);
- Faça uma análise de ML (500 bootstrap geralmente são o suficiente);
- Veja se a árvore consenso gerada faz sentido, e se está bem suportada;
- Se não estiver bem suportada, tente uma Bayesiana (normalmente começando com 3 árvores iniciais e parando quando as árvores geradas tiverem um desvio padrão menor de 0.05, mas isso não é uma regra);
- Se precisar de datação, utilize o BEAST (normalmente avaliando as métricas com o TRACER, ajustando os pesos entre as métricas até todas estiverem num valor limiar ideal).

Passo 5: Apresente seus dados de forma a responder seu problema biológico

Ao final da análise, você terá o arquivo com a topologia da árvore consenso (árvore gerada representando o conjunto de árvores criadas durante as análises), mas o trabalho não acaba por aí! Apresentar o arquivo como ele é gerado implica que o leitor do seu estudo terá que procurar por conta as informações que estão presentes na árvore, e isso diminui muito o interesse no seus resultados. Existem várias ferramentas para anotar árvores filogenéticas, eu indico três:

- iTOL: Para quem não tem familiaridade com linguagens de programação, permite realizar diversos tipos de anotação, e inserção de gráficos na filogenia, mas costuma travar com grandes quantidades de dados (mais de 5 mil sequências, por exemplo);
- toytree: Para quem gosta de brincar com python, possui uma limitação na questão de cores e possibilidades de anotação quando comparada com o iTOL, mas permite a anotação de árvores realmente gigantes (anotei árvores de SARS-CoV-2 com mais de 7 mil genomas, sem problemas);
- ggtree: Para quem tem familiaridade com R, eu realmente testei poucas funções dessa biblioteca, pois não tenho muita familiaridade com R, e na minha opinião a comunidade de R não se esforça muito para produzir manuais ou tutoriais de fácil entendimento, mas caso você seja um mago do R, está aí a dica!

Esse último passo vai de encontro ao Passo 0, então se você preparou o arquivo suplementar, basta puxar os metadados para anotar sua árvore, nesse exemplo temos uma árvore de máxima verossimilhança da polimerase da família *Coronaviridae*, onde usei as informações de código de *Strain* no nome das sequências, de gênero para coloração dos clados (Podemos ver que temos os gêneros bem definidos), e por fim fiz a adição de algumas figuras dos hospedeiros pelo Inkscape, marcado em amarelo escuro os vírus que já foram identificados em humanos.



Então se a nossa pergunta inicial fosse, qual a origem do SARS-CoV-2 (que cometi o erro de deixar como hCoV-19 na filogenia), podemos ver que está no mesmo clado do SARS-CoV e muito próximo ao SARS-CoV-RaTG13, que é um coronavírus encontrado em roedores e morcegos e ao PCoV GX-P5E, que é um coronavírus encontrado normalmente em pangolins, tudo casando com as teorias mais aceitas atualmente para a origem do SARS-CoV-2, certo?

Dicas do Passo 5:

- Tenha metadados para a anotação da árvore (sempre voltando ao Passo 0);
- Escolha a ferramenta de acordo com os dados que você quer anotar (iTOL permite anotar até piechart e *barplots* ao lado das filogenias);
- Padronize o nome das sequências, caso você pense em mostrar eles nas filogenias (filogenias com centenas de sequências ficam muito poluídas com os nomes nos *tips*);
- Pense em quais informações são importantes para serem mostradas na sua figura final;

Por esse artigo era isso pessoal, algumas informações podem ter sido passadas de forma muito direta, mas era essa a ideia mesmo, faça um check-list desses passos e se aprofunde em cada um deles na hora de realizar suas análises! E lembre-se estou longe de ser um especialista em filogenia, pense se esses passos fazem sentido para suas análises, e sempre busque a literatura científica na hora de tomar qualquer decisão para suas análises!!!