Análise Final: Características incompatíveis do corona vírus brasileiro

A. Pereira do Lago[1]

4 de agosto de 2020

Depois de realizadas as duas primeiras fases do trabalho de mac0465[1], Biologia Computacional, apresentamos as respostas a cinco perguntas que visaram fomentar uma análise final da comparação e inferência da filogenia das diferentes cepas do corona vírus causador da *covid 19*. Abaixo colocamos nossas respostas, que servem de gabarito estendido para o trabalho apresentado pelos alunos. Elas foram estendidas com considerações que abordam questões que apareceram nos diversos trabalhos e levam a uma análise mais aprofundada sobre as características incompatíveis da primeira cepa do corona vírus sequenciada no Brasil, e que tem espalhado a covid 19 pelo país.

1) Quantas e quais são as posições em que há mutação (substituição no alinhamento ótimo com a sequência de referência[2]) na cepa brasileira? (# 56/108, id MT126808)

O arquivo `mutacoes108.txt` fornecido lista as posições em que há mutação entre a sequência de referência e cada uma das 107 sequências completas não ambíguas publicadas no Genbank até 27 de março de 2020, mais a sequência italiana, que apresenta apenas uma ambiguidade. Das 29903 posições da sequência de referência, são apenas quatro as que apresentam substituição no alinhamento ótimo exibido contra a sequência brasileira:

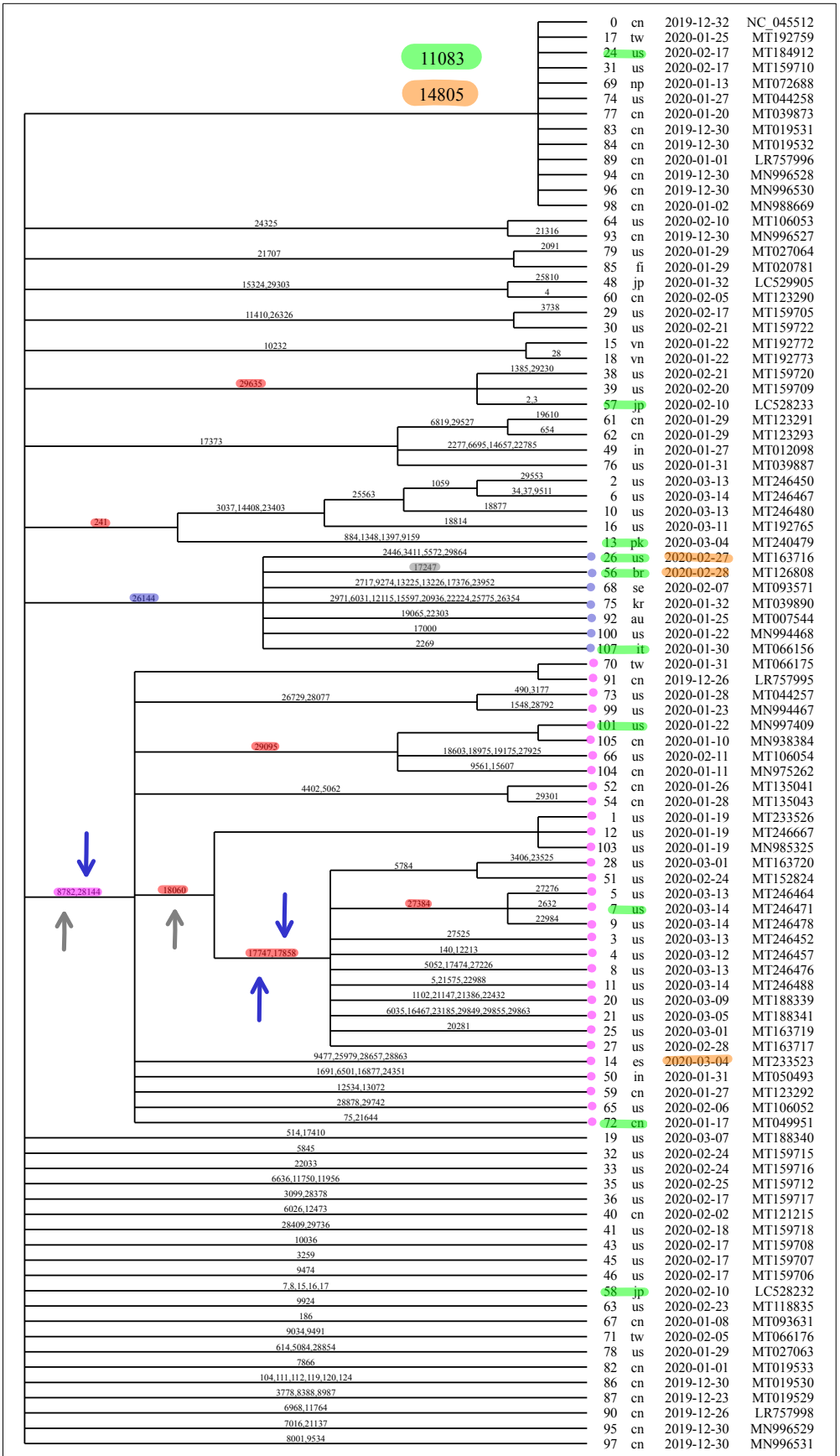
```
  0   NC_045512 29903 2019-12-32 China
      NC_045512 |Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1| complete genome
 56    MT126808 29876 2020-02-28 Brazil
      MT126808 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/SP02/human/2020/BRA| complete genome
Parâmetros do alinhamento: 2 0 -1 -1
Distância:  31
Primeiro ATG:    106          106     Delta:    0
 NC_045512   0 29903   --   29876  56    MT126808

identidades:   29872
remoções:         27
               29871 A -> - 29871
               29872 A -> - 29871
               29873 A -> - 29871
               29874 A -> - 29871
               29875 A -> - 29871
               29876 A -> - 29871
               29877 A -> - 29871
               29878 A -> - 29871
               29879 A -> - 29871
               29880 A -> - 29871
               29881 A -> - 29871
               29882 A -> - 29871
               29883 A -> - 29871
               29884 A -> - 29871
               29885 A -> - 29871
               29886 A -> - 29871
               29887 A -> - 29871
               29888 A -> - 29871
               29889 A -> - 29871
               29890 A -> - 29871
               29891 A -> - 29871
               29892 A -> - 29871
               29893 A -> - 29871
               29894 A -> - 29871
               29895 A -> - 29871
               29896 A -> - 29871
               29897 A -> - 29871
inserções:         0
substituições:     4
               11083 G -> T 11083
               14805 C -> T 14805
               17247 T -> C 17247
               26144 G -> T 26144
```

as posições 11083, 14805, 17247 e 26144. Observe que quatro posições representam apenas *0,013%* em relação às 29903 posições da sequência de referência.

1 Dep. de Ciência da Comp., Inst. de Mat. e Estat. da Univ. de São Paulo. Cf. <http://www.ime.usp.br/~alair/mac0465>.

2 Sequência do Genbank id NC_045512, Wuhan, China, 2019-12, https://www.ncbi.nlm.nih.gov/nuccore/NC_045512

2) Quantas delas estão presentes no grupo de características compatíveis selecionadas na segunda fase do projeto e quantas não?



Filogenia das 107 sequências completas sem ambiguidade de vírus da SARS2 mais a sequência italiana com uma ambiguidade, publicadas no Genbank a 27 de março de 2020. Há repetições, mas nunca de países distintos. A raiz é a barra vertical à esquerda e são 94 as sequências distintas listadas à direita nesta filogenia perfeita que ignora seis características incompatíveis. Os números nas arestas são de posições que sofreram um evento de mutação.

Ao lado vemos a filogenia obtida do algoritmo descrito na seção 6.1, que acrescenta à árvore filogenética arestas com características, grupos de posições, das relativas à maior quantidade de cepas para as relativas à menor quantidade.

A cepa 56 brasileira tem duas mutações características que formam a filogenia: nas posições 26144 e 17247, destacadas em azul e cinza.

Das 182 posições com substituição, resultam apenas 94 grupos de posições (que formam 94 características) por terem substituições no mesmo conjunto de cepas. Destas 94 características, 19 são incompatíveis com alguma outra e na segunda fase recomendamos "o critério de descartar características [que são] incompatíveis com outras de maior cardinalidade que [já] não tenham sido descartadas".

Por este critério, foram descartadas seis características incompatíveis com alguma outra dentre as 19 mencionadas. Quatro destas seis

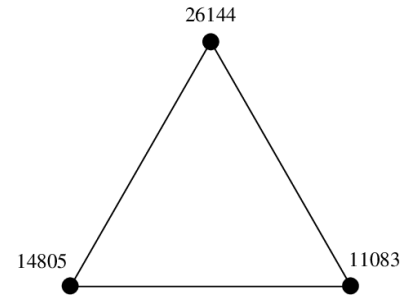
são incompatíveis com a característica formada pelas posições 8782, 28144 – que estão destacadas em roxo e que sofrem a mesma mutação em 35 das 108 cepas –, inclusive as duas das quatro posições onde a cepa brasileira apresenta mutação e que não rotulam arestas da árvore filogenética obtida: a posição 11083 e a posição 14805, destacadas em verde e amarelo-ouro.

A mutação na posição 14805 é uma mutação silenciosa (não produz alteração de aminoácidos) e ocorre apenas em três sequências, destacadas todas em amarelo-ouro. Já a mutação na posição 11083 produz uma substituição de Leucina por Fenilalanina, e ocorre em dez das 108 cepas, destacadas em verde na filogenia acima.

3) Há incompatibilidades entre as características/posições da cepa brasileira? Se houver, desenhe o grafo de incompatibilidades entre as características que apresentam alguma incompatibilidade. (Uma aresta ligando duas características indica que elas são incompatíveis entre si.)

É imediato verificar que mutações que ocorrem em apenas uma sequência não são incompatíveis com nenhuma outra; o que é o caso da mutação na posição 17247, cuja única sequência que apresenta esta mutação é a brasileira: *Seq(17247) = {56}*. Assim, resta examinar as características relativas às posições 11083, 14805 e 26144. Observando na filogenia acima também os destaques em amarelo-ouro e verde, notemos que

Seq(26144) = { 26, 56, 68, 75, 92, 100, 107 },
Seq(14805) = { 14, 26, 56 },
Seq(11083) = { 7, 13, 24, 26, 56, 57, 58, 72, 101, 107 }.



Grafo de incompatibilidades das mutações presentes na variante brasileira do vírus da SARS2. As características relativas às posições 11083, 14805 e 26144 são duas a duas incompatíveis.

Assim, as computações feitas à tabela seguinte são imediatas:

p	*q*	*Seq(p) ∩ Seq(q)*	*Seq(p) – Seq(q)*	*Seq(q) – Seq(p)*
26144	*14805*	*{ 26, 56 }*	*{ 68, 75, 92, 100 , 107}*	*{ 14 }*
26144	*11083*	*{ 26, 56 }*	*{ 68, 75, 92, 100, 107 }*	*{ 7, 13, 24, 57, 58, 72, 101, 107 }*
14805	*11083*	*{ 26, 56 }*	*{ 14 }*	*{ 7, 13, 24, 57, 58, 72, 101, 107 }*

Qualquer que seja a escolha de posições distintas *p* e *q* em *{ 11083, 14805, 26144 }*, tanto a intersecção de *Seq(p)* e *Seq(q)* quanto as diferenças recíprocas são não vazias, o que implica que

$$Seq(p) \not\subseteq Seq(q) \not\subseteq Seq(p).$$

Portanto, *p* e *q* são incompatíveis entre si. Isto completa a demonstração de que o grafo de incompatibilidades das mutações da cepa brasileira com alguma incompatibilidade é o da figura.

4) Para cada posição de uma característica incompatível da cepa brasileira, cuja informação foi ignorada na elaboração da filogenia obtida, listar que outras cepas de que outros países apresentam a mesma mutação. Existe algum padrão geográfico observável? Considere a representabilidade de cada país no conjunto das cepas observadas; compare as probabilidades de se ter o país em questão dado que sejam satisfeitas as seguintes condições: a cepa pertence ao conjunto de todas as cepas; a cepa pertence ao conjunto das cepas que apresentam a característica incompatível em questão.

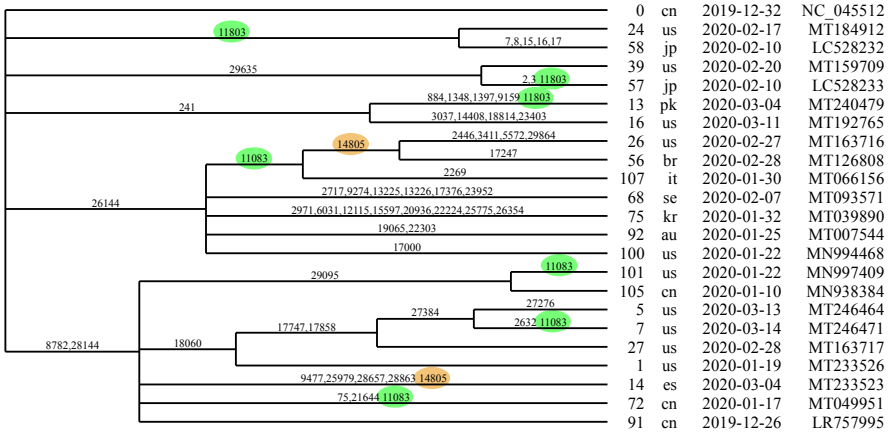
Como exposto na resposta à questão 2, são duas as posições da cepa brasileira que foram ignoradas na elaboração da filogenia obtida em nossa segunda etapa: 14805 e 11083. Elas foram descartadas por serem incompatíveis com uma característica cuja mutação se manifesta num número maior de

sequências: a característica associada às posições 8782 e 28144, cujas mutações são observadas nas mesmas 35 cepas destacadas em roxo. Observamos que a da posição 8782 é silenciosa (não altera o aminoácido codificado na proteína) e a da posição 28144 substitui um aminoácido não polar (a leucina L) por um polar (a serina S). Note que é esperado que uma mutação que altere a classe dos aminoácidos seja mais difícil de ser aceita pela Natureza e que seja pouco observada. É o caso desta substituição rara (pontuação -8 na matriz de substituições PAM-30). Ainda assim ocorre em quatro das 96 posições com aminoácidos distintos e só 4 substituições de aminoácidos são mais frequentes!

A reconsideração das posições omitidas no processo de inferência de uma árvore filogenética perfeita (sem características incompatíveis) requer eventos de mutações paralelas, repetidas: em princípio, uma mutação paralela para cada sequência que sofreu mutação para cada posição que foi descartada. Algumas destas mutações repetidas na mesma posição podem ainda ser fatoradas, reduzindo a quantidade de mutações paralelas, como se discutirá a seguir.

A mutação na posição 14805 é uma mutação silenciosa que ocorre em três sequências: na 56, brasileira; na 14, espanhola; na 26, americana. Coletadas entre 27 de fevereiro e 4 de março, as três provêm de regiões distantes entre si, ainda que todas ocidentais. Se observarmos que 63 (58.3%) das 108 sequências foram coletadas no Ocidente, 100% dos que manifestam esta mutação são de pacientes no Ocidente. As mutações desta posição nas sequências 26 e 56, americana e brasileira, podem ser fatoradas num único evento devido à sua proximidade na árvore filogenética da figura: depois da mutação na posição 26144, teria havido uma única mutação na posição 14805, antes de acontecerem as mutações nas posições 17247 e 2446, 3411, 5572, 2986 que teriam levado às sequências 56 e 26, respectivamente. Nenhuma delas sofre mutação nas posições 8782 e 28144, ao contrário da sequência espanhola 14. Por esta razão, a menos que se conceda uma nova mutação paralela em cada posição 8782 e 28144, não é possível unificar num único evento as mutações na posição 14805 nas três sequências.

Já a mutação na posição 11083 produz uma substituição de Leucina por Fenilalanina (tida como relativamente rara por possuir pontuação -3 na tabela PAM-30), mas ocorre nas dez cepas de *Seq(11083)* destacadas em verde na filogenia acima: 7, 13, 24, 26, 56, 57, 58, 72, 101, 107. Algumas destas sequências são de vértices relativamente próximos na árvore filogenética, de modo que podemos fatorar algumas mutações paralelas. Assim, podemos supor que a sequência 58, japonesa, tenha sido obtida depois de mutações nas posições 7,8,15,16,17 (de uma cópia anterior) da sequência 24, americana, não requerendo uma nova mutação paralela. Também as mutações da posição 11083 nas sequências 26, 56 e 107 (americana, brasileira e italiana) podem ser fatoradas num único evento devido à sua proximidade na árvore filogenética: depois da mutação na posição 26144, teria havido uma única mutação na posição 11803. A partir desta nova sequência obtida: uma mutação silenciosa na posição 2269 teria originado a sequência italiana 107; e uma única mutação na posição 14805 teria gerado uma cepa intermediária que, sofrendo mutações nas posições 17247 e 2446, 3411, 5572, 2986, teria respectivamente gerado as sequências 56 e 26. Assim, sete mutações paralelas na posição 11803 são suficientes para as dez cepas com esta mutação. Entretanto, mais fatoração pode ser feita, contanto que se adicione ao menos uma mutação paralela em outra posição.



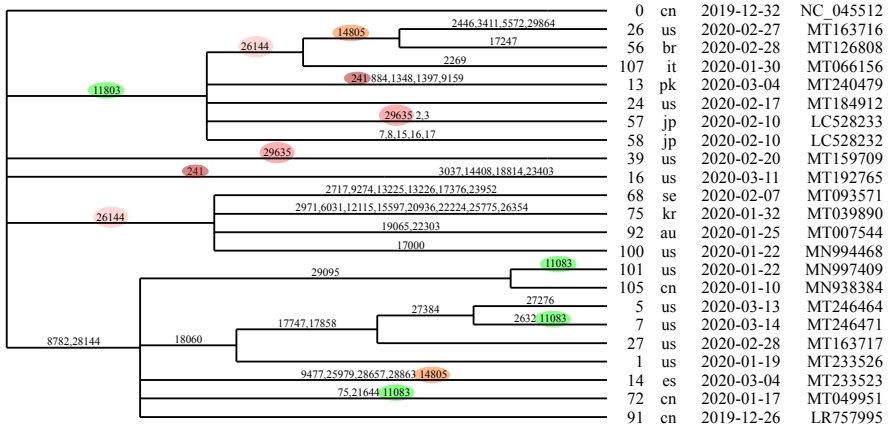
Filogenia restrita a um subconjunto das 108 sequências. Incluem-se as que sofreram mutação nas posições 11803 ou 14805, e próximas na Filogenia. Em amarelo destacam-se dois eventos de mutações paralelas na posição 14805, donde derivam as três sequências com esta mutação. Em verde destacam-se sete eventos de mutações paralelas na posição 11083, donde derivam as dez com esta mutação.

As 10 sequências que apresentam mutação na posição 11083 foram coletadas de pessoas dos: EUA (7, 24, 26, 101); China (72); Japão (57, 58); Brasil (56); Itália (107); Paquistão (13). Portanto, 40% das cepas que apresentam mutação nesta posição são dos EUA, 10% da China, 20% do Japão, 10% do Brasil, 10% da Itália e 10% do Paquistão. Cada um destes países possui probabilidade a priori respectivamente 53,7%, 28,7%, 2,8%, 0,9%, 0,9% e 0,9%. Se China e EUA representam 82% das 108 cepas, apenas 50% das cepas que sofreram mutação na posição 11083 infectaram pessoas que vivem nestes países: uma redução em 1,6 vezes. Os EUA reduziram sua participação de 53.7% para 40%, uma redução em 1.4 (1 / 0.74) vezes. Já a China reduziu sua participação em 2.9 (1 / 0.35) vezes, de 28.7% para 10%. Já os 50% restantes ficaram restritos aos países de fora deste eixo e que antes representavam apenas 18% das cepas, um aumento em 2,8 vezes.

Pais	N	%	nMut	Odds
Australia	1	0,9	0	0,0
Finland	1	0,9	0	0,0
Nepal	1	0,9	0	0,0
SouthKorea	1	0,9	0	0,0
Spain	1	0,9	0	0,0
Sweden	1	0,9	0	0,0
India	2	1,9	0	0,0
VietNam	2	1,9	0	0,0
Taiwan	3	2,8	0	0,0
China	31	28,7	1	0,35
USA	58	53,7	4	0,74
Japan	3	2,8	2	7,1
Brazil	1	0,9	1	11,1
Italy	1	0,9	1	11,1
Pakistan	1	0,9	1	11,1
	108	100	10	

Países das 108 sequências ordenados pela razão de chances da probabilidade de mutação na posição 11083 (nMut/10) e a probabilidade a priori deste país (N/108). Se agrupados, os países fora do eixo China-EUA aumentaram a participação em 2,8 vezes, de 18% para 50%.

Observe que não pode ser o caso de que alguma destas cepas derive da sequência 72 de Yunnan, província do sudoeste chinês, que além da posição 11083 apresenta mutações nas posições 8782, 28144, já citadas, bem como mutações exclusivas nas posições 75, 21644. Coletada a 17 de janeiro, ela é também a mais antiga do grupo, seguida pela sequência 101 coletada cinco dias depois nos EUA, com mutação na posição 29095, ao contrário das demais do grupo. A sequência americana 24, coletada a 17 de fevereiro, é a única do grupo que sofreu apenas a mutação de substituição da posição 11083 e, como visto antes, podemos supor que a sequência japonesa 58 seja uma descendente de uma cópia sua (a japonesa foi coletada sete dias antes). Poderia ocorrer o mesmo fenômeno com outras sequências do grupo? Não sem a introdução de novas mutações paralelas: as sequências 26, 56 e 107 requereriam uma mutação paralela na posição 26144; a sequência 13 requereria uma mutação paralela na posição 241; a sequência 57 requereria uma mutação paralela na posição 29635. A figura ao lado mostra uma filogenia alternativa para o mesmo subconjunto de sequências acima visto que incorpora todas estas três fatorações. Ao custo de três mutações paralelas em três posições, ela reduz o número de mutações paralelas na posição 11803 de sete para quatro. Conforme se incorpore cada uma das três fatorações ou não, pode-se oferecer oito filogenias alternativas, todas com a mesma quantidade de eventos de mutações. Outras fatorações mais complexas requereriam o acréscimo de mais de uma mutação paralela: a sequência 7 requereria mutações paralelas nas posições 8782, 28144, 18060, 17747, 17858, 27384; a sequência 101 requereria mutações paralelas nas posições 8782, 28144, 29095. A sequência 72 é um caso a parte: mesmo mutações paralelas nas posições 8782, 28144 não seriam suficientes pois as mutações na posição 11083 substituem uma guanina G da sequência 0 de Wuhan para: uma citosina C, no caso da sequência 72; uma timina T, no caso das demais sequências do grupo. Isto não altera a substituição da leucina L pela fenilalanina F na cadeia proteica (ao



Filogenia restrita a um subconjunto das 108 sequências. Incluem-se as que sofreram mutação nas posições 11803 ou 14805, e algumas próximas. Em amarelo destacam-se dois eventos de mutações paralelas na posição 14805, donde derivam as três sequências com esta mutação. Em verde destacam-se quatro eventos de mutações paralelas na posição 11083, donde derivam as dez com esta mutação. Este número reduzido de eventos paralelos nesta posição requer a introdução de mutações paralelas nas posições 241, 26144 e 29635.

contrário do que ocorreria se houvesse substituição para adenina A), mas garante que a sequência 72 não derive da sequência 24 sem que ocorra uma nova mutação na mesma posição 11083.

Todas as características das posições antes citadas onde eventualmente se requereriam mutações paralelas são de fato incompatíveis com a característica da posição 11083 e nas filogenias anteriores estão destacadas em rosa ou vermelho, mas também em roxo e azul.

O programa fornecido `MutacoesDeAminoacidos.py` *processa o arquivo de mutações* `mutacoes108.txt` *e verifica se cada mutação se encontra em região codificante ou não. As regiões codificantes são os ORFs (Open Reading Frames) estudados e anotados na sequência de referência. Caso pertença a uma região codificante, o programa também calcula o códon a que pertence a posição em que ocorre a mutação, bem como o aminoácido codificado. Isto é feito diretamente, no caso da sequência de referência, e requer que se calcule em que posição do códon ocorreu a mutação. As mutações silenciosas são mais fáceis de serem aceitas e costumam ocorrer na posição 2 do códon, enquanto que mutações ocorridas nas posições 0 ou 1 costumam produzir alteração no aminoácido. Como quase não há operações de inserção e remoção nas regiões codificantes, e quando há ocorrem em múltiplos de três, a mutação correspondente na outra sequência, também mencionada como sequência alvo (target), deve se dar na mesma posição no códon correspondente. Decodificados os aminoácidos associados aos referidos códons nas sequências de referência e alvo, o programa contabiliza a mutação nos contadores de frequências de mutações de aminoácidos associados ao par de aminoácidos em questão. Além de listar estas informações relativas a cada substituição, o programa lista ao final toda a tabela de contadores. No caso de sequências similares, a matriz PAM normalmente recomendada é a PAM-30, de modo que sua pontuação para o par de aminoácidos em questão também é exibida nesta listagem. Tal como observado na seção 3.5.1,*[3] *o cálculo das matrizes PAM faz uso de frequências* f_{ab} *em seus cálculos. No programa, as frequências relativas a todas as substituições presentes no arquivo* `mutacoes108.txt` *são armazenadas no dicionário* `FreqAbs`*. Já o dicionário* `FreqPos` *registra quais são as posições envolvidas, de modo que a cardinalidade do conjunto de posições funciona como um contador da frequência de eventos de mutações.*

5) Seja A o conjunto dos pares de aminoácidos distintos com pontuação estritamente positiva na matriz PAM-30, e seja m o número destes pares. A grosso modo, estes pares de aminoácidos participam mais frequentemente de substituições que os outros pares. Considere o conjunto C formado pelos m pares de aminoácidos mais frequentes na lista de posições com mutações (segundo contadores dados pela cardinalidade de `FreqPos[par]`*). Liste os conjuntos A e C, compare-os, e comente. (Além da própria função* `ListaConjunto` *fornecida, ajuda nesta comparação o cálculo do índice de* `Jaccard` *entre A e C.) Busque explicação a qualquer discrepância observada.*

```
mac0465:
mac0465: ./MutacoesDeAminoacidos.py | grep -- '->' | wc -l
415
mac0465: ./MutacoesDeAminoacidos.py | grep codificante| wc -l
95
mac0465: ./MutacoesDeAminoacidos.py | grep pam30 | wc -l
320
mac0465: ./MutacoesDeAminoacidos.py | grep pam30 | sort -k 2 | awk 'last!=$2{ print; last =$2;}' | wc -l
148
mac0465: ./MutacoesDeAminoacidos.py | grep pam30 | sort -k 2 | awk 'last!=$2{ print; last =$2;}' | awk '$10==$13' | wc -l
52
mac0465: ./MutacoesDeAminoacidos.py | grep pam30 | sort -k 2 | awk 'last!=$2{ print; last =$2;}' | awk '$10!=$13' | wc -l
96
mac0465: ./MutacoesDeAminoacidos.py | awk '/^[A-Z] [A-Z]/ && $7 > 0' | wc -l
39
mac0465:
```

Estatísticas gerais tomadas da listagem provida por `MutacoesDeAminoacidos.py` em sistema linux

Temos ao todo, 415 mutações (substituições) analisadas: 95 mutações que se dão em posições não codificantes; 320 mutações numa posição codificante de nossa sequência de referência (0) contra outra das 108. Contudo, as posições internas a regiões codificantes envolvidas nestas 320 mutações são apenas 148, das quais: 52 posições onde as mutações são silenciosas, não alterando o aminoácido gerado; 96 posições cujas mutações envolvem dois aminoácidos distintos. Ademais, dos 190 possíveis pares de aminoácidos distintos, apenas 39 pares foram observados nestas 96 posições.

3 Meidanis & Setúbal, Introduction to Computational Molecular Biology, 1997, PWS Publishing.

A execução de `MutacoesDeAminoacidos.py` de fato imprime três tabelas: na primeira delas cada mutação listada em `mutacoes108.txt` é complementada com informações adicionais, como os aminoácidos eventualmente envolvidos; na segunda (`tabela`), uma contabilidade das mutações é feita segundo cada um dos 190 pares de possíveis aminoácidos distintos; na terceira, cálculos diversos evolvendo razões de chance e conjuntos de pares de aminoácidos. A segunda tabela também informa na terceira coluna (`col_pam30`) quais os valores da pontuação segundo a PAM-30 (`ProteinSubsMatrix[par]`) relativos ao par dos aminoácidos das duas primeiras colunas. Desta forma é imediato computar $m = 7$ contando as linhas que têm valor estritamente positivo na terceira coluna, como nas computações ao lado. Elas mostram em seguida que uma ordenação pela terceira coluna nesta segunda tabela permite selecionar quais são os pares de aminoácidos nas duas primeiras colunas das linhas de maior pontuação segundo a tabela PAM-30. Desta forma é imediato verificarmos que o conjunto $A = \{ (E,Q), (L,M), (H,Q), (D,N), (F,Y), (D,E), (I,V) \}$. Ademais, para os aminoácidos a e b das duas primeiras colunas, a quarta coluna (`col_pam1`) contém $P(b|a)P(a) + P(a|b)P(b)$,[4] ou seja, a probabilidade de que a mude para b mais a probabilidade de que b mude para a. Como se pode observar nas computações acima, somando-se a quarta coluna vemos que o conjunto A oferece uma cobertura de 20% dos pares associados aos 1572 eventos de mutação contabilizados nos cálculos originais das tabelas PAM (Margaret Dayhoff, A model of Evolutionary Change in Proteins, 1978) resumidos na seção 3.5.1 do livro de Meidanis e Setúbal.

```
mac0465:
mac0465: ./MutacoesDeAminoacidos.py | awk '/^[A-Z] [A-Z]/ && $3 > 0' | LANG=C sort -n -k 3 -k 7 | wc -l
7
mac0465: ./MutacoesDeAminoacidos.py | awk '/^[A-Z] [A-Z]/ && $3 > 0' | LANG=C sort -n -k 3 -k 7
E Q   1 0.02680 0.00404  6.628
L M   1 0.01355 0.00271  4.994
H Q   1 0.01542 0.00275  5.608   2 0.02083  7.577   133.3
D N   2 0.03372 0.00400  8.428   1 0.01042  2.604    30.8
F Y   2 0.01680 0.00255  6.578   1 0.01042  4.079    61.6
D E   2 0.05282 0.00500 10.561   2 0.02083  4.166    39.3
I V   2 0.04254 0.00512  8.311   3 0.03125  6.106    72.9
mac0465: ./MutacoesDeAminoacidos.py | awk '/^[A-Z] [A-Z]/ && $3 > 0' | LANG=C sort -n -k 3 -k 7 | sumcol.awk 4
0.20165
mac0465: ./MutacoesDeAminoacidos.py | awk '/^[A-Z] [A-Z]/ ' | LANG=C sort -n -k 7 -k 3 | tail -7
H Y  -3 0.00256 0.00217  1.179   4 0.04167 19.194  1399.8
G S  -2 0.02894 0.01326  2.183   4 0.04167  3.143   141.9
P S  -2 0.01707 0.00760  2.247   4 0.04167  5.484   238.3
A V  -2 0.02301 0.01203  1.912   5 0.05208  4.328   221.3
L P  -7 0.00323 0.00923  0.350   7 0.07292  7.904  1841.7
F L  -3 0.01030 0.00724  1.423   7 0.07292 10.077   661.1
I T  -2 0.00813 0.00457  1.780  14 0.14583 31.931  1520.9
mac0465: ./MutacoesDeAminoacidos.py | awk '/^[A-Z] [A-Z]/ ' | LANG=C sort -n -k 7 -k 3 | tail -7 | sumcol.awk 7
45
mac0465:
```

Processamento da execução de `MutacoesDeAminoacidos.py` em sistema linux que oferece amparo ao cálculo de m = 7 e dos conjuntos A = { (E,Q), (L,M), (H,Q), (D,N), (F,Y), (D,E), (I,V) } e C = { (H,Y), (G,S), (P,S), (A,V), (L,P), (F,L), (I,T) }, cuja intersecção é vazia. Da terceira coluna vê-se que as pontuações da PAM-30 são positivas para os pares de A e negativas para os de C. Ao somar a coluna 7, vemos que os pares de C cobrem 47% (45) das 96 posições codificantes associadas a uma mutação de aminoácidos. Somando a coluna 4, vemos que o conjunto A cobre 20% dos pares associados a eventos de mutação contabilizados nos cálculos originais das tabelas PAM.

A sétima coluna desta mesma segunda tabela contém a quantidade de posições cuja substituição de aminoácidos listada na primeira tabela envolve os aminoácidos das duas primeiras colunas, de modo que uma ordenação pela sétima coluna permite selecionar os sete pares mais envolvidos nas 96 substituições contabilizadas. Isto é feito nas execuções descritas na figura acima, resultando que $C = \{ (H,Y), (G,S), (P,S), (A,V), (L,P), (F,L), (I,T) \}$.[5] Assim, é imediato verificar que os dois conjuntos A e C são completamente disjuntos, até porque as pontuações da PAM-30 presentes na terceira coluna são estritamente positivas para os pares de A e negativas (≤ -2) para os de C. Isto apesar do conjunto C cobrir 47% (45) das 96 posições com mutações que substituem aminoácidos!

Quão discrepante do esperado é isto? Não se deveria esperar uma grande intersecção em função das tabelas PAM refletirem as probabilidades das mutações de aminoácidos observadas na Natureza conhecida?

Ocorre que no cálculo das matrizes PAM computa-se uma matriz de razões de chances tal que "Os pares de aminoácidos com valores acima de 1 substituem-se mutuamente com mais frequência do que em sequências aleatórias de mesma composição,[6] enquanto que os de valores abaixo de 1 substituem-se com menos frequência." Cada elemento da matriz PAM propriamente dita é dez vezes o logaritmo da razão de chances[7], de forma que a pontuação neutra é zero e uma "pontuação

4 Vide fator `M[(a,b)]*ProbPriori[a]+M[(b,a)]*ProbPriori[b]` atribuído a `abpam1`, coluna `col_pam1`.
5 Há ambiguidade na definição pois (L,S) é tão frequente quanto os três menos frequentes de C ... e não pertence a A.
6 Vide fator `ProbPriori[a]*ProbPriori[b]` atribuído a `abrand`, coluna `col_rand`, ou 5ª na figura.
7 Vide fator `abpam1/abrand` atribuído a `abodds`, coluna `col_pamodds`, ou 6ª na figura.

de −10 significa que se espera que o par ocorra em sequências relacionadas apenas um décimo das vezes do quanto seria previsto ao acaso, e uma pontuação de +2 significa que se espera que o par ocorra 1,6 [= $10^{0,2}$] vezes mais" (Dayhoff, 1978, p.351-352) A maximização da soma destas pontuações nos algoritmos de alinhamento equivale à solução de um problema de máxima verosimilhança envolvendo produto de razões de chances envolvendo probabilidades condicionais.

Ademais, as 1572 mutações estudadas em transições de 71 árvores filogenéticas cuidadosamente inferidas a partir de grupos de proteínas similares com no máximo 15% de diferenças entre as sequências mais distantes, levantados a partir do Protein Data Bank (PDB) na década de 70, constitui o primeiro grande trabalho de modelagem estatística das diferentes probabilidades de substituição entre aminoácidos. Esta modelagem reflete as diferentes abundâncias dos aminoácidos e uma comparação entre os extremos relata que a glicina G ocorre 8,9 vezes mais que o triptofano W no amplo conjunto de dados então estudado (tabela 22). Ainda que a glicina seja 2,7 vezes mais propensa a sofrer mutações que o triptofano, o mais estável dos aminoácidos, ela é também 2,7 vezes menos propensa a sofrer mutações que a asparagina, o mais instável (tabela 21). A glicina é o 14º aminoácido de maior mutabilidade e "a baixa mutabilidade da glicina deve ser devido à sua pequenez única que é vantajosa em muitos lugares" (Dayhoff, 1978, p.347). Além desta modelagem estatística que leva em consideração a abundância de cada aminoácido e sua mutabilidade, este trabalho se propõe a reconhecer padrões de conhecimento que melhor expliquem a distribuição de pares de aminoácidos observados nas mutações aceitas pela Natureza. "Os padrões têm sido visíveis nas mutações pontuais aceitas desde o início do trabalho de sequenciamento de proteínas. Isoleucina-valina e serina-treonina foram alternativas frequentemente observadas. Era óbvio que essa intercambiabilidade tinha algo a ver com suas semelhanças químicas. Na grande quantidade de informações que agora existe, correlações muito mais detalhadas são visíveis e muitas inferências funcionais podem ser feitas. […] os grupos de aminoácidos quimicamente semelhantes que tendem a substituir um ao outro: o grupo hidrofóbico; o grupo aromático; o grupo básico; o ácido, grupo ácido-amida; cisteína; e os outros resíduos hidrofílicos. […] Esses padrões são impostos principalmente pela seleção natural e apenas secundariamente pelas restrições do código genético: refletem a semelhança das funções dos resíduos de aminoácidos em suas fracas interações entre si na conformação tridimensional das proteínas." (Dayhoff, 1978, p.351) As tabelas PAM refletem este conhecimento inferido do mais tradicional banco de dados de proteínas que a ciência humana produziu e as matrizes de substituição são críticas nos algoritmos usados nas etapas mais fundamentais do trabalho de um biólogo que se põe a sequenciar uma nova proteína descoberta: o alinhamento da proteína obtida contra outras já previamente estudadas.

Uma questão que se coloca é qual das diferentes tabelas PAM se recomenda usar, já que elas levam em consideração a distância evolutiva das sequências estudadas. Poderia ser o caso de que o uso da tabela PAM não ideal seja a causa da discrepância observada? Diversas tabelas são calculadas e as tabelas PAM mais recomendadas para as sequências mais conservadas são as de menor valor. A tabela PAM-30, por exemplo, é recomendada para mudanças observadas em 25% das posições das sequências comparadas e a tabela PAM-1 para mutações em 1% das posições (tabela 23, p.351). Lembremos que as quatro mutações da variante brasileira contra o vírus de Wuhan representa apenas *0,013%* das 29903 posições da sequência de referência e não há tabela PAM publicada para uma taxa de mutação tão pequena. As tabelas PAM disponíveis num ajuste paramétrico feito no alinhador BLASTP disponível no NCBI são apenas `PAM30`, `PAM70` e `PAM250` e no `Biopython` a oferta é maior mas ainda limitada: `pam30`, `pam60`, `pam90`, `pam120`, `pam180`, `pam250`, `pam300`. A verdade é que se pode assumir que não haja sobreposição de mutações para taxas de mutação mais baixas, o que ao contrário é a própria razão dos cálculos que levam à PAM250, onde se supõem haver uma média de 2,5 mutações por posição. Mesmo o artigo citado de Dayhoff não publica a tabela com logaritmos de razões de chances PAM-1, mas a matriz de probabilidades condicionais (`M[(a,b)]`, ou $P(b|a)$) que elevada a uma potência k permite calcular PAM-k. Ainda que sem os coeficientes de normalização de Dayhoff, espera-se que a razão de chances `abpam1 / abrand`

que compõe a variável `abodds` (coluna `col_pamodds`) produza a mesma ordenação durante a execução de `MutacoesDeAminoacidos.py`.

Assim, estamos aptos a verificar se a discrepância entre *A* e *C* permaneceria ao trocarmos a PAM-30 pela PAM-1. Sendo *B* o conjunto dos *m* pares de aminoácidos com a maior PAM-1 (ou maior razão de chances na sexta coluna), as execuções ao lado mostram justamente que *B* possui seis dos sete pares de A, trocando *(L,M)* por *(S,T),* que *C* é disjunto não só de *A* como também de *B*, e que a cobertura de *B* aumenta de 20% para 23%. O par (L,S) não está em B.

```
mac0465: ./MutacoesDeAminoacidos.py | awk '/^[A-Z] [A-Z]/' | LANG=C sort -n -k 6 -k 7 | tail -7
S T   0 0.04444 0.00864  5.143
H Q   1 0.01542 0.00275  5.608   2 0.02083  7.577   133.3
F Y   2 0.01680 0.00255  6.578   1 0.01042  4.079    61.6
E Q   1 0.02680 0.00404  6.628
I V   2 0.04254 0.00512  8.311   3 0.03125  6.106    72.9
D N   2 0.03372 0.00400  8.428   1 0.01042  2.604    30.8
D E   2 0.05282 0.00500 10.561   2 0.02083  4.166    39.3
mac0465: ./MutacoesDeAminoacidos.py | awk '/^[A-Z] [A-Z]/' | LANG=C sort -n -k 6 -k 7 | tail -7 | sumcol.awk 4
0.23254
mac0465: ./MutacoesDeAminoacidos.py | awk '/^[A-Z] [A-Z]/' | LANG=C sort -n -k 6 -k 7 | tail -190 | sumcol.awk 4
1.00428
mac0465: ./MutacoesDeAminoacidos.py | awk '/^[A-Z] [A-Z]/ ' | LANG=C sort -n -k 7 -k 6 | tail -7
H Y  -3 0.00256 0.00217  1.179   4 0.04167 19.194  1399.8
G S  -2 0.02894 0.01326  2.183   4 0.04167  3.143   141.9
P S  -2 0.01707 0.00760  2.247   4 0.04167  5.484   238.3
A V  -2 0.02301 0.01203  1.912   5 0.05208  4.328   221.3
L P  -7 0.00323 0.00923  0.350   7 0.07292  7.904  1841.7
F L  -3 0.01030 0.00724  1.423   7 0.07292 10.077   661.1
I T  -2 0.00813 0.00457  1.780  14 0.14583 31.931  1520.9
mac0465: ./MutacoesDeAminoacidos.py | awk '/^[A-Z] [A-Z]/ ' | LANG=C sort -n -k 7 -k 6 | tail -7 | sumcol.awk 7
45
mac0465: ./MutacoesDeAminoacidos.py | awk '/^[A-Z] [A-Z]/ ' | LANG=C sort -n -k 7 -k 6 | tail -190 | sumcol.awk 7
96
```

Processamento da execução de `MutacoesDeAminoacidos.py` em sistema linux que obtém, para m = 7, os conjuntos B = { (E,Q), (S,T), (H,Q), (D,N), (F,Y), (D,E), (I,V) } e C = { (H,Y), (G,S), (P,S), (A,V), (L,P), (F,L), (I,T) }. O conjunto B está para a PAM-1 assim como A para a PAM-30. Da terceira coluna vê-se que só (S,T) é par novo e que a intersecção com C permanece vazia. Ao somar a coluna 7, vemos C cobre 47% (45) das 96 posições codificantes associadas a uma mutação de aminoácidos entre as variantes do vírus da SARS2 e, somando a coluna 4, vemos que o conjunto B cobre 23% dos pares associados aos 1572 eventos de mutação de aminoácidos contabilizados nos cálculos das tabelas PAM.

Há que se observar que o valor de *m* é definido em função da PAM-30, e que este limiar pode ser redefinido ao se usar a razão de chances da PAM-1. Permaneceriam ainda as discrepâncias observadas se permitirmos a inclusão de pares com PAM-30 nulos? Neste caso obtemos *m = 12,* e novas versões *A' , B'* e *C'* para os conjuntos *A, B* e *C*. A terceira tabela listada pelo programa `MutacoesDeAminoacidos.py` faz este tipo de exploração e as versões dos conjuntos *A, B* e *C* que dependem do parâmetro *m* correspondem às listas `TopPAM30`, `TopPAM1_Rand` e `TopSARS2`, respectivamente. Para *m = 12* em particular, podemos verificar que: as novas versões de *A* e *B* são iguais e cobrem 36,7% das 1572 mutações que deram suporte à modelagem feita com as tabelas PAM; ao passo que a nova versão de *C* passa a cobrir quase dois terços das mutações das variantes do vírus de Wuhan, inclusive 6 das 96 mutações que são associadas aos dois pares em comum entre as novas versões de A e de C.[8] A intersecção não é vazia, mas permanece a discrepância de que esta intersecção seja tão pequena e cubra apenas 6 das 96 mutações não silenciosas.

Talvez por um preciosismo se possa considerar uma comparação em que ao invés do conjunto C obtido por `TopSARS2` se venha a optar por um conjunto D obtido como `TopSARS2_Rand`,[9] ordenando não segundo as probabilidades inferidas pelas frequências `freq=len(FreqPos[par])` mas segundo a razão de chances com o que se esperaria ao acaso: `(freq/nposmult)/abrand`. Assim, para m = 12, a cobertura diminui de 61 / 96 em C para 50/96 = 25/48 (52%) em D ao passo que permanece a discrepância de que sua intersecção com A seja tão pequena quanto a de C (dois entre doze elementos) e cubra apenas 5 das 96 mutações não silenciosas.

Não poderia ser que as discrepâncias observadas entre os pares de aminoácidos que mais se destacam em relação ao que se esperaria ao acaso sejam porque as mutações observadas nas primeiras variantes do vírus de Wuhan possam não ser aceitas pela Natureza? É difícil de dizer que um vírus contagioso que prolifere dentro do organismo a ponto de provocar diversos pontos de embolia e um estado inflamatório em diversos órgãos não tenha já em certo grau se adaptado ao ambiente. De toda maneira, é certo que a posição mais frequente com mutação não silenciosa associada é a mais forte candidata a ser considerada aceita pela Natureza e, como já visto, a posição

8 Na terceira tabela, na linha em que *m = 12*, observa-se que `Jaccard(TopPAM30,TopPAM1_Rand)` é 1, `IntegraColuna(tabela,col_pam1,TopPAM30)` é 0,367, `Jaccard(TopPAM30,TopSARS2)` é 1/11 = 2/22 e `IntegraColuna(tabela,col_sars2,TopSARS2)` é 61/96.

9 Definidos respec. por `SelecionaMParesTopo(OrdenaTabelaPorColuna(tabela,col_sars2),m)` e por `SelecionaMParesTopo(OrdenaTabelaPorColuna(tabela,col_sars2odds),m)`.

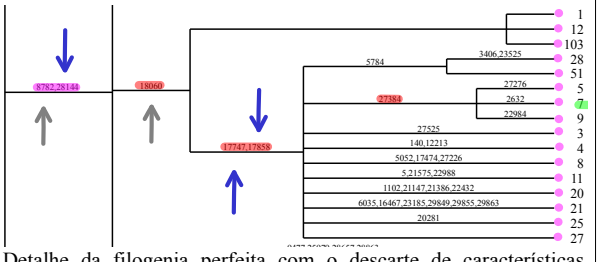
28144 apresenta mutação em cerca de um terço das sequências publicadas até 27 de março, todas destacadas em roxo na filogenia da figura. Ademais, as 35 linhas da primeira tabela impressa por `MutacoesDeAminoacidos.py` relatam sempre a mesma substituição de nucleotídeos (de T para C) e a mesma substituição de aminoácidos (de L para S). Pois este par de aminoácidos aparece também noutras três posições (28854, 09561, 28863) e somente outros quatro pares de aminoácidos são mais frequentemente associados às posições com mutação entre as variantes estudadas dos virus da covid 19. Em contraste, a mutação entre a leucina e a serina possui pontuação -8 na tabela PAM30 (há 121 pares de aminoácidos com pontuação superior de modo que (L,S) não pertence a A mesmo para m=121) e apenas o 97º posto entre os pares de aminoácidos ordenados segundo a probabilidade decrescentes de serem associados a uma mutação aceita entre as 1572 extraídas do PDB quando as tabelas PAM foram computadas. A raridade desta mutação na natureza explica-se pelo fato de que a leucina é apolar e hidrofóbica enquanto a serina é polar e seu substituinte tende a formar pontes de hidrogênio. De fato, a mutação de leucina para serina deve tornar a proteína mais reativa e potente, de forma que isto levanta a questão de como foi que esta variante foi gerada.

Teria havido alguma forma de fator externo à Natureza que teria gerado esta mutação, que segundo a Química deveria ser extremamente rara? Afinal, quando foi que a mutação na posição 28144 se deu? De fato, como se verifica na filogenia da figura, duas sequências apresentam mutação apenas nesta posição: a sequência 78, de Taiwan, coletada a 31 de janeiro de 2020; a sequência 91, da China, coletada a 26 de dezembro de 2019. São apenas 13 as sequências coletadas em 2019 e somente a sequência 91 tem mutação nesta posição. A proporção aumentou de 1/13 para 35/108 em três meses, de forma que esta mutação é definitivamente uma mutação aceita.

Se para um leigo pode parecer à primeira vista que não há por que se deva esperar que a substituição envolvendo leucina e serina deva ser esperada de ocorrer muito raramente, tal como se observa nos dados dos estudos das mutações pontuais aceitas (PAM), a verdade que a nós emerge desta análise é que a introdução destas mutações não obedecem às leis da natureza inferidas por uma aprendizagem computacional calcada numa modelagem estatística bastante sofisticada.

O por quê das discrepâncias observadas requer análises futuras, que podem ser facilitadas pelo estudo da região onde se dão as mutações observadas na sequência 86, coletada ainda em 2019.

A título de curiosidade, quisemos pegar a próxima mutação não silenciosa mais frequente, manifesta em 13 sequências. Em todas elas, a mutação observada à posição 17747 sempre muda o nucletídeo C para T e um aminoácido P para L. Com valor -7 na tabela PAM-30, esta substituição de aminoácidos não é presente em A (para m = 7,12,...,103) e constitui o 63º par de aminoácidos mais frequentemente observado entre os 1572 eventos de mutação inferidos nas 71 árvores filogenéticas reconstituídas nos estudos da tabelas PAM. Ordenados os pares segundo seu correspondente valor na matriz da razão de chances da PAM-1, seu posto cai de 63 para 90, de modo que a substituição também não pertence a B, para m < 90. Em contraste, esta mutação envolvendo a prolina e a leucina é a terceira mais frequente e pertence a C, manifestando-se em sete posições: 3177, 3738, 5052, 6501, 9159, 14408 e 17747. Como a razão de chances entre a probabilidade associada a esta quantidade de posições e o que se esperaria do acaso a coloca no 6º posto, esta substituição de aminoácidos também pertence a D para m ≥ 6. Tanto a prolina quanto a leucina estão entre os aminoácidos classificados como apolares. Contudo, o índice de hidrofobia de ambas as coloca em subclasses diferentes, o que talvez ajude a explicar por que razão a figura 84 do trabalho de Dayhoff que publicando a tabela PAM250 coloque a prolina em subgrupo distante dos hidrofóbicos M, I, L, V, reunindo-a com as polares treonina T e Serina S.



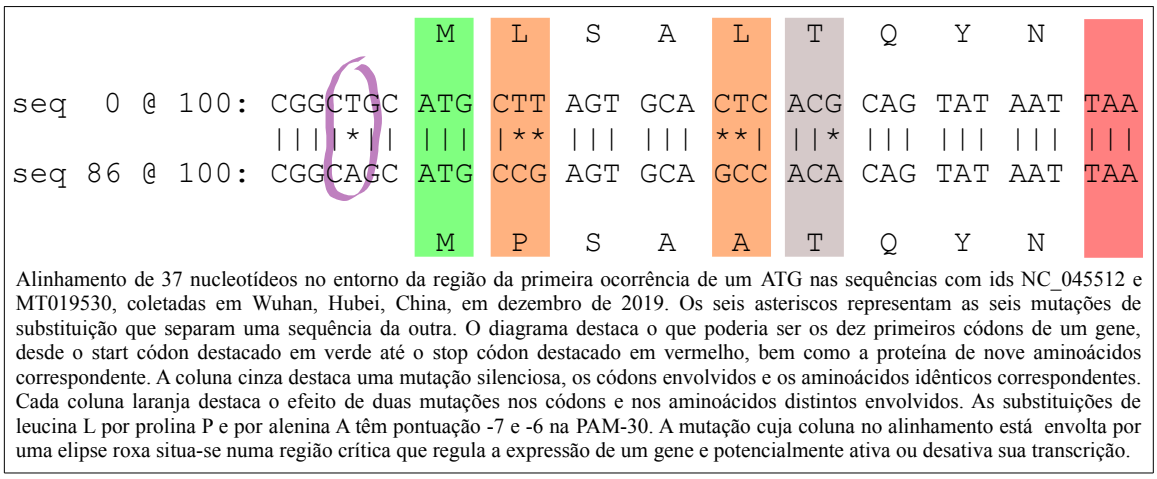
Detalhe da filogenia perfeita com o descarte de características incompatíveis. As cinco flechas apontam para as cinco posições cujas mutações mais se manifestam nas 108 sequências. As duas flechas cinza apontam para mutações silenciosas e as três flechas azuis apontam para mutações entre aminoácidos distintos.

As treze sequências com mutação à posição 17747 também possuem uma segunda mutação não

silenciosa: a da posição 17858, que substitui a trianina Y pela cistina C no único sítio que envolve este par de aminoácidos polares. Este par não pertence a nenhum dos conjuntos A, B, C ou D, nem para m = 7, nem para m = 12.

***

A ciência, fruto do maravilhamento humano diante daquilo que existe, ultimamente recorre à própria Ontologia na busca do conhecimento daquilo que existe. O vírus de referência cuja sequência 0 foi extraída é um ser, e podemos chamar de s_1 o primeiro ser da mesma espécie cuja sequência não registra qualquer mutação de diferença em relação à sequência 0. Analogamente, podemos chamar de s_2 o primeiro ser da mesma espécie cuja sequência não registra qualquer mutação de diferença em relação à sequência 86. Há um movimento que levou de s_1 a s_2. Este próprio movimento é. Ele é um ser, a relação entre s_1 e s_2, e que chamamos s_3. Está na genética de s_3 o conjunto de seis mutações que transforma a sequência 0 na sequência 86. Suas posições são {104, 111, 112, 119, 120, 124}, todas próximas entre si e à posição 107, onde se encontra a primeira ocorrência de um ATG na sequência de referência. As anotações do Genbank não relatam que esta ocorrência de um ATG abra uma janela de leitura para formar o que seria o primeiro gene deste vírus, e que deveria codificar uma proteína de apenas nove aminoácidos.



Alinhamento de 37 nucleotídeos no entorno da região da primeira ocorrência de um ATG nas sequências com ids NC_045512 e MT019530, coletadas em Wuhan, Hubei, China, em dezembro de 2019. Os seis asteriscos representam as seis mutações de substituição que separam uma sequência da outra. O diagrama destaca o que poderia ser os dez primeiros códons de um gene, desde o start códon destacado em verde até o stop códon destacado em vermelho, bem como a proteína de nove aminoácidos correspondente. A coluna cinza destaca uma mutação silenciosa, os códons envolvidos e os aminoácidos idênticos correspondentes. Cada coluna laranja destaca o efeito de duas mutações nos códons e nos aminoácidos distintos envolvidos. As substituições de leucina L por prolina P e por alenina A têm pontuação -7 e -6 na PAM-30. A mutação cuja coluna no alinhamento está envolta por uma elipse roxa situa-se numa região crítica que regula a expressão de um gene e potencialmente ativa ou desativa sua transcrição.

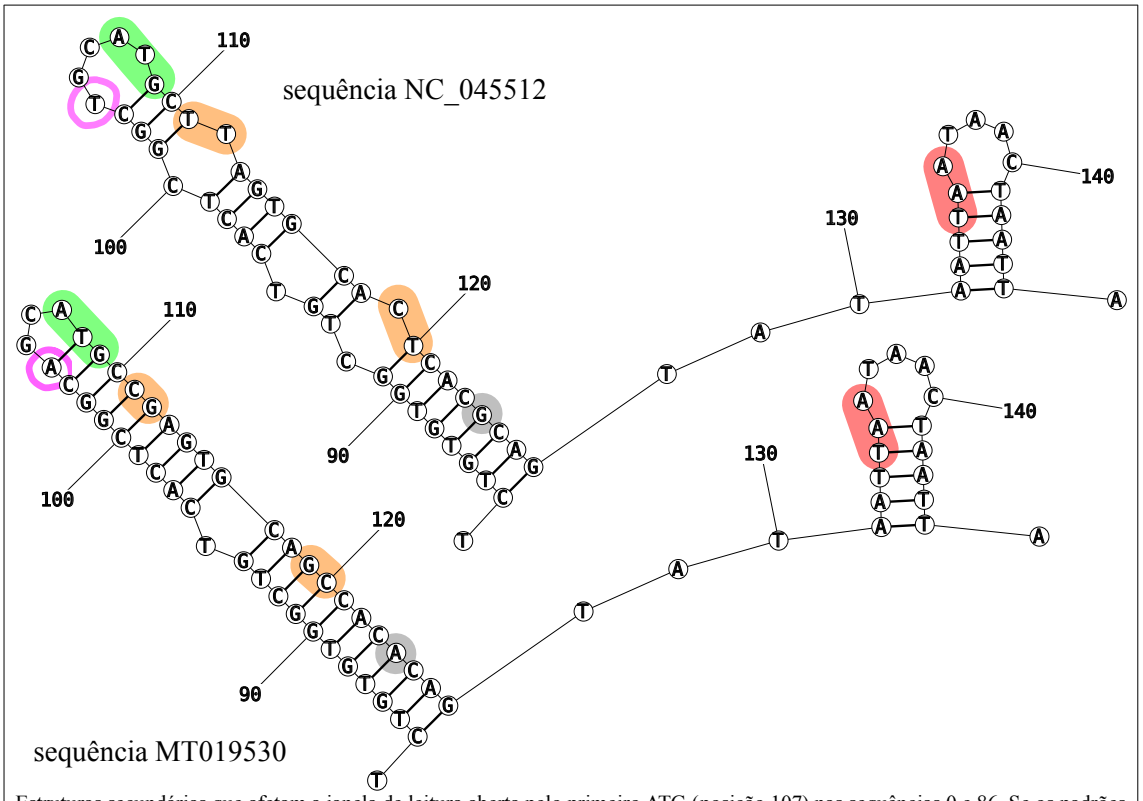
É por conta de padrões curtos e apropriados que imediatamente precedem a ocorrência de um ATG que um potencial gene é promovido a um gene real, codificante de uma proteína de verdade; e este ser s_3 coincidentemente opera uma mutação na posição 104, que pertence à região promotora deste gene e onde potencialmente se ligaria algum tipo de fator de transcrição. Teria esta mutação na posição 104 a capacidade de ativar ou desativar um potencial gene na posição 107? Teria o ser s_3 esta finalidade? Não se pode fazer um estudo apropriado sobre um processo evolutivo quem não usar os conceitos empregados por Aristóteles em Metafísica... Levantar as perguntas certas a partir da realidade é condição necessária a uma boa ciência. Assim, qual a causa final desta relação s_3?

Sendo esta mutação bem sucedida em efetivamente promover a ativação deste gene, as demais mutações deste grupo produziriam duas mutações de aminoácidos: uma envolvendo leucina L e prolina P, um par de aminoácidos de pontuação −7 na PAM-30 presente na posição 17747 examinada acima; outra que produziria a primeira mutação entre as 108 sequências analisadas envolvendo leucina L e alanina A, par de aminoácidos com pontuação −6 na PAM-30 que requer mutação nos dois primeiros nucleotídeos do códon por conta do código genético.

A probabilidade de que uma mutação se dê numa janela de cerca de 30 posições próximas à primeira ocorrência de um ATG dentro de uma sequência de 30000 posições é de cerca de 1/1000, a probabilidade de que um grupo de ao menos seis mutações se dê exatamente nesta região é da ordem de 10^{-18}. Seria simplesmente obra do acaso este conjunto de mutações? Quais são suas causas? Qual a causa final de s_3? Das treze sequências de vírus coletadas em 2019, a sequência 86 é aquela com mais mutações, seguida por uma com três. As demais têm no máximo duas …

***

Como já visto, a leucina L e a prolina P formam o terceiro par de aminoácidos mais envolvido em mutações do vírus de Wuhan publicados no Genbank até 27 de março. O segundo par é formado pela leucina e fenilanina F. Cada par participa de mutações em 7 posições na sequência de referência, mas é necessário acrescentar à contabilidade do par (L,F) o que é devido às mutações paralelas da posição 11083, como cuidadosamente exposto na resposta à pergunta 4, sem falar na possibilidade de que na posição ambígua da sequência italiana o nucleotídeo seja outro que não a timina presente na sequência brasileira ou a guanina presente na sequência de referência. Ainda que o posto do par se altere de 2º para 5º se ranqueado pela razão de chances, o par pertence a C e a D tanto para m = 7 quanto para m = 12. Com pontuação -3 na tabela PAM-30, ele não pertence nem a A nem a B para m ≤ 28. Além de incompatível, a mutação característica 11803 é testemunha eloquente da discrepância entre a realidade observada dos eventos de mutações inferidos entre as variantes do vírus de Wuhan completamente sequenciados e a modelagem da evolução das mutações aceitas obtida pela ciência humana na década de 70.

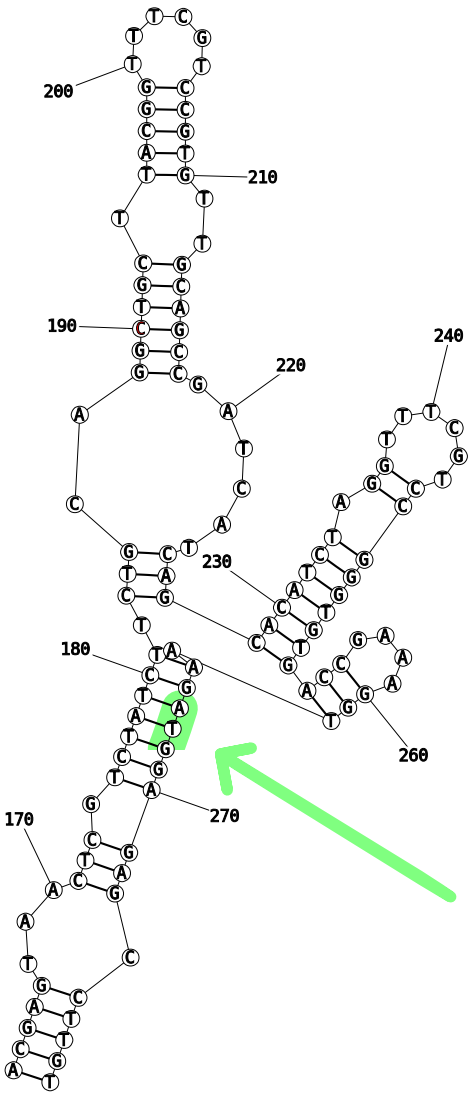


Estruturas secundárias que afetam a janela de leitura aberta pelo primeiro ATG (posição 107) nas sequências 0 e 86. Se os padrões que se repetem no comportamento dos assassinos ajudam na investigação de organizações criminosas, a conformação das repetições presentes nas sequências dos coronas vírus quando se alinha com seu reverso complementar revela estruturas ocultas aos olhos de quem examina a sequência de nucleotídeos apenas na ordem direta. A estrutura secundária tem papel primordial na ativação de um vírus de RNA e a ocorrência na extremidade do laço formado entre as posições 84 e 127 do sítio onde potencialmente se atracará um fator de transcrição para o ATG (AUG) da posição 107 destacado em verde deve dar-lhe a potência de ser ativável ao extremo. Nas posições destacadas em laranja encontram-se as mutações não silenciosas das posições 111, 112, 119, 120. A mutação silenciosa da posição 124 e o códon de parada da posição 134 recebem destaques cinza e vermelho, respec.

De fato, temos sempre suposto que o nucleotídeo da posição 11083 da sequência italiana 107 seja a timina, até porque esta seria uma condição mais favorável ao que se vem afirmando de que a sequência brasileira 56 tenha sido contraída no norte italiano. A sequência 107 é apontada como sendo a primeira sequenciada por pesquisadores do norte italiano e foi coletada a 30 de janeiro de um turista da província de Hubei, onde fica Wuhan, que teria levado a doença para a Itália. Corrobora com a tese da ancestralidade italiana da cepa brasileira o fato de que a sequência 56 e a sequência italiana 107 manifestarem a mutação da posição 26144. Mesmo que obscurecido pela ambiguidade, o estado da posição 11083 na sequência italiana também corrobora. Se daí se pergunta se a sequência brasileira 56 descende da única sequência italiana completa publicada no Genbank

até 27 de março, deve-se dizer que não, a menos que se suponha ter havido um fenômeno de reversão da mutação silenciosa que a sequência italiana sofreu na posição 2269, mas que a sequência brasileira não possui.

Se nossa sequência 56 não descende do primeiro ser com sequência genética idêntica à sequência 107, de quem então descenderia? Quem seria o pai da criança? Dentre as demais 107 sequências, de qual delas o primeiro vírus com a mesma sequência é o mais próximo ancestral de nossa sequência brasileira? Seria a sequência 0, a sequência de referência? A sequência 31, das sequências sem nenhuma mutação (de substituição) que as separe da sequência 0, aquela contra a qual nossa sequência brasileira possui a menor distância de edição? A sequência 24, das 107 demais sequências, aquela contra a qual a sequência brasileira possui a menor distância de edição? Dentro do teatro das 108 sequências, saindo de cena a sequência 107, o palco do teatro italiano permanece vazio e a criança brasileira está órfã. A menos que apareça um novo ator neste cenário, a peça que se encerra no teatro italiano é um autêntico suspense.



Se das sequências publicadas a 27 de março permanece o suspense sobre um eventual pai italiano da criança brasileira com suas incompatíveis características dotadas de posições discrepantes, num cenário tão distante quanto a China de 2019 projeta-se a luz da conjunção astral que produziu em ser tão improvável e de causa final misteriosa como s3 e sua relação entre s1 e s2. Tanto na sequência 0 quanto na 86, a segunda ocorrência de um ATG situa-se à posição 266 e inicia um gene com um raro frameshift -1 à posição 13468 e que termina apenas à posição 21555. O mal causado pela covid 19 deve-se principalmente a este gene, que se inicia numa substrutura secundária assemelhada a um tridente, e que evoca a memória do mal representado por Goethe em Mefisto, uma obra apreciada por Marx em sua juventude. A temática é dominante no livro de poemas que o jovem Karl Marx ofereceu a seu pai como presente de aniversário em 1837: em *Orgulho Humano,* o jovem Marx escreveu: “Com desdém lançarei meu desafio / Em cheio no rosto do Mundo / Verei o colapso deste gigante pigmeu / Cuja queda não poderá sufocar meu ardor. / Então eu passearei divino e vitorioso / Através das ruínas do mundo / Darei às minhas palavras uma força ativa, / Me sentirei igual ao Criador.”; já em *O violinista (The Fiddler)*, o jovem Marx escreveu: “Com Satanás eu fiz o meu pacto. / Ele escreve as partituras e marca o compasso, / Eu toco e canto a marcha da morte com rapidez e desembaraço.” (Karl Marx & Frederick Engels, Collected Works, International Publishers, New York, 1975, v.1, p. 586 e 22) A marcha da morte tocada pelo vírus SARS-CoV-2 ameaça dar corpo à pretensão do jovem Marx de sentir-se igual ao Criador vendo o colapso do Mundo. Que culpa tem o Criador se a humanidade tem dado ouvidos ao *Manifesto Comunista* de Marx e Engels que prega um comunismo que “abole as verdades eternas, abole toda a religião e toda a moralidade”? Será que o espírito que inspirou Marx inspirou os homens que deixaram a marca do tridente nestas estruturas virais? Ou será que se trata de uma imagem deformada da cruz? Carregar a cruz destes dias é disseminar a verdade. E a verdade é que também eu sou pecador, esperando ser açoitado com falácias *ad hominem* por quem tem se condenado a viver no inferno da desinformação e da mentira ainda que acredite ser ele mesmo o messias prometido. O que me provoca a meditar sobre o capítulo 53 de Isaías, e a esperar cheio de certeza, esperar que a marcha da morte tocada pelo violinista será vencida pela epidemia da verdade! Viralize.