PHÂN TÍCH CẤU TRÚC VACCINE BNT162b2 DỰA TRÊN mRNA TRONG VIỆC PHÒNG DỊCH BỆNH COVID-19

Trần Thụy Ánh Quỳnh[1], Trần Văn Lăng[2*]

[1] Faculty of Science Life, HAN University, Neitherland
[2] Tạp chí Khoa học HUFLIT, Trường Đại học Ngoại ngữ -Tin học TP.HCM
tt.anhquynh@gmail.com, langtv@huflit.edu.vn

TÓM TẮT— Bài báo trình bày những kiến thức mang tính cơ sở của vaccine dựa trên mRNA,; đồng thời phân tích cấu trúc dựa trên góc nhìn Tin học về trình tự sinh học của của vaccine BNT162B2 ngừa virus SARS-CoV-2 đã gây ra dịch COVID-19. Đây là vaccine đã được Công ty Pfizer và BioNTech phát triển. Từ đó đưa ra thuật toán và chương trình viết bằng Python để chọn lựa trình tự mRNA trên cơ sở tối ưu tỷ lệ GC.

Từ khóa— Sinh tin học, vaccine, COVID-19

I. GIỚI THIỆU

Virus là một ký sinh trùng nhỏ như là một tác nhân truyền nhiễm mà không tự sinh sản được. Chúng dùng DNA rồi phiên mã thành RNA như là vật liệu di truyền sau đó dùng bộ máy của tế bào sống làm cơ chế sinh sản để lan truyền. Nên vì vậy người ta thường nói virus không phải là một cơ thể sống mà nó vật ký sinh dùng cơ thể sống để tồn tại và phát triển. Những nhà khoa học đã nghiên cứu từ đó dùng vaccine để ngăn ngừa virus, hiện nay có các loại vaccine khác nhau theo cách tiếp cận khác nhau để tạo ra; có loại vaccine tìm cách triệt phá năng lực hoạt động của virus rồi đưa vào cơ thể sống, có loại nhờ virus khác không nguy hại để đưa hình hài của virus vào cơ thể sống; có loại đưa vào cơ thể sống những nguyên vật liệu để từ đó sử dụng cơ chế tạo sinh tế bào để có được hình hài virus trong cơ thể sống.

Để trở thành vaccine đưa vào sử dụng, thì phải thoả mãn điều kiện quan trọng là số trường hợp tử vong trong số những người bị nhiễm bệnh sau khi họ được tiêm vaccine là rất nhỏ, thậm chí phải bằng không. Vì vậy, việc nghiên cứu vaccine cần phải huy động cả một nguồn lực lớn từ Hoá học, Sinh học cho đến Tin học và cả Toán học.

Hiện nay đã có nhiều loại vaccine liên quan đến phòng dịch bệnh COVID-19, trong số đó vaccine sử dụng các nguyên vật liệu đưa vào cơ thể sống rồi nhờ bộ máy sinh sản tế bào tạo ra hình dạng virus có ý nghĩa khoa học cao, là một cách tiếp cận mới trong việc tạo ra vaccine. Vacine thuộc loại này có thể là vaccine dựa trên DNA, hoặc vaccine dựa trên RNA.

Bài viết trình bày một số nội dung liên quan đến vaccine dựa trên RNA thể hiện qua các phần như sau trong các phần tiếp theo của bài viết. Phần thứ hai tiếp theo đây trình bày những khái niệm của Sinh học phân tử để làm nền tảng kiến thức cho những hiểu biết tiếp theo, đồng thời đưa ra những công trình liên quan gần đây về nghiên cứu về vaccine phòng virus SARS-COV-2. Phần thứ ba phân tích cấu trúc của vaccine dựa trên RNA; phần thứ tư là những tính toán thực nghiệm để đưa ra giải pháp tìm vaccine thoả mãn một tiêu chí nào đó; phần cuối cùng còn lại là một số kết luận sau quá trình khảo sát và thực nghiệm tính toán.

II. MỘT SỐ KHÁI NIỆM VÀ CÔNG TRÌNH LIÊN QUAN

A. MỘT VÀI KHÁI NIỆM SINH HỌC LIÊN QUAN

1. DNA, RNA, PROTEIN, RIBOSOME

Học thuyết trung tâm được F.Crick đưa ra từ nằm 1956 cho đến nay vẫn còn tính đúng đắn đó là thông tin di truyền từ DNA được phiên mã (*transcipt*) thành RNA sau đó dịch mã (*translate*) thành protein. Trong đó **DNA** (*DeoxyriboNucleic Acid*) chứa trong nhân tế bào, đây là một chuỗi xoắn kép gồm 2 mạch đơn, mỗi mạch đơn là một chuỗi các nucleotide. Chính vì vậy có thể nói DNA là một đại phân tử sinh học mà mỗi phân tử của nó là một nucleotide. Mỗi nuclootide ở đây bao gồm phophate, đường deoxyribose và một trong 4 base hữu cơ là Adenine, Cytosine, Guanine, Thymine. Các nucleotide này liên kết với nhau theo liên kết cộng hoá trị giữa đường deoxyribose của nucleotide này với phophate của nucleotide kế tiếp. Một đặc điểm quan trọng là các nucleotide này có phophate và đường doxyribose là giống nhau; chỉ các nhau các base hữu cơ; chính vì vậy, người ta dùng một trong 4 base hữu cơ để đặc trưng cho một nucleotide [1].

* Coressponding Author

Trong sinh học phân tử, 4 base hữu cơ được ký hiệu bởi các chữ cái ban đầu là A, C, G, T; nên một mạch đơn trong chuỗi DNA được biểu diễn bởi một chuỗi gồm các ký tự chữ A, C, G, T. Chẳng hạn, một đoạn của virus SARS-CoV-2 được cung cấp trên ngân hàng dữ liệu của NCBI† như sau:

```
ATTAAAGGTTTATACCTTCCCAGGCAAACCAACCAACTTTCGATCTCTTGTAGATCT
```

Ngoài ra, do DNA là chuỗi xoắn kép mà 2 mạch đơn này liên kết với nhau theo liên kết tĩnh điện yếu giữa hydro mang điện tích dương và hydro mang điện tích âm của 2 base hữu cơ (gọi là liên kết hydro), trong đó Adenine liên kết với Cytosine, còn Guanine liên kết với Thymine. Từ đó với một mạch đơn này của chuỗi xoắn kép có thể suy ra được mạch đơn thứ hai. Chính vì vậy, một DNA được biểu diễn hình thức bởi chỉ một chuỗi các ký tự A, C, G, T như trên.

RNA (*RiboNucleic Acid*) được phiên mã từ DNA; RNA có đặc điểm là chỉ có một mạch đơn, trong đó Thymine của DNA được thay bởi Uracine (ký hiệu là U). Chính vì vậy một trình tự RNA được hình thức hoá bởi 4 ký tự A, C, G, U. Với cấu trúc đơn giản như vậy nên trong lưu trữ các trình tự sinh học này người ta không cần lưu trữ trình tự RNA.

Trình tự DNA ở trên có thể chuyển về trình tự RNA bằng các dòng lệnh Python với thư viện Biopython 1.78 như sau:

```
from Bio.Seq import Seq
DNA = Seq("ATTAAAGGTTTATACCTTCCCAGGCAAACCAACCAACTTTCGATCTCTTGTAGATCT")
RNA = DNA.transcribe()
```

Kết quả trình tự RNA là

```
AUUAAAGGUUUAUACCUUCCCAGGCAAACCAACCAACUUUCGAUCUCUUGUAGAUCU
```

Ngoài ra, RNA còn có nhiều loại khác nhau, nhưng dưới góc nhìn Tin học chúng ta chỉ quan tâm đến mRNA (*message RNA*) trong đó chứa đầy đủ thông tin để dịch mã sang protein. Đây chính là bản sao của các trình tự DNA, nhằm chuyển thông tin mã hóa trên DNA đến bộ máy giải mã protein tương ứng. Nếu so sánh với máy tính thì DNA được lưu trữ trên thiết bị lưu trữ (storage), nhưng khi đưa vào máy tính để xử lý thì DNA phải chuyển thành (phải phiên mã thành) mRNA để chứa trong bộ nhớ; từ đó máy tính dịch mã sang protein. Chính vì vậy nên DNA tồn tại như là dữ liệu gốc, còn mRNA hay protein là những biểu hiện trong quá trình xử lý và tồn tại mang tính nhất thời.

Phiên mã ở đây chính là bước đầu tiên của việc biểu hiện gen dựa trên DNA, trong đó một đoạn DNA cụ thể được sao chép thành RNA (đặc biệt là *mRNA*) bởi chất xúc tác RNA (*Enzyme* RNA *polymerase*) có trong cơ thể sống. Như vậy, trong một cơ thể sống có công cụ để chuyển vật liệu di truyền từ DNA sang RNA rồi thành protein.

Còn **Protein** cũng là một đại phân tử sinh học mà mỗi phân từ là một amino acid. Mỗi amino acid được cấu tạo từ 3 base hữu cơ liên tiếp nhau trên trình tự RNA. Do có 4 base hữu cơ nên về mặt lý thuyết có thể có $4^3 = 64$ amino acid; tuy nhiên cho đến nay người ta cũng chỉ giải mã được 20 amino acid; nhưng do có đến 64 bộ 3 base hữu cơ, nên người ta dùng thuật ngữ khác - mà gọi là *codon* để chỉ các bộ ba này. Từ đó suy ra là có những codon khác nhau nhưng có cùng tên gọi về amino acid. Cũng chính điều này khi giải mã ngược từ một trình tự protein ta lại có nhiều trình tự RNA khác nhau, trong khi đó từ một trình tự RNA thì có suy nhất một trình tự protein.

Trong một trình tự protein các amino acid liên kết với nhau theo liên kết peptide; đó là liên kết giữa đầu Carboxylic (COOH) của amino acid này với đầu Amin (NH_2) của amino acid khác và loại bỏ đi một phân tử nước (H_2O). Về hình thức, 20 amino acid này được ký hiệu bởi 20 ký tự A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y; nên một trình tự protein là một chuỗi gồm những ký tự trong 20 ký tự nêu trên.

Cũng với trình tự RNA ở trên, với dòng lệnh Python tiếp theo:

```
Protein = RNA.translate()
```

Đưa về trình tự protein gồm những chữ cái trong 20 chữ nêu trên là:

```
IKGLYLPRQTNQLSISCRS
```

Về mặt sinh học, protein còn được gọi là chất đạm, thời gian tồn tại của chúng cũng tuỳ theo từng loại protein khác nhau; có loại chỉ tồn tại trong cơ thể sống được vài phút rồi bị thoái hoá, nhưng có loại tồn tại đến hàng năm. Việc thiếu protein sẽ gây hậu quả như chậm lớn, bệnh tật, v.v... Protein là những thành phần cơ bản của sự sống, tuỳ theo chức năng mà có các loại protein khác nhau. Chẳng hạn protein mang tính cấu tạo rồi phát triển để duy trì sự sống; protein mang tính cảm biến để tiếp nhận thông tin bên ngoài như tín hiệu âm thanh, ánh sáng, từ

† https://www.ncbi.nlm.nih.gov/nuccore/NC_045512

trường, nồng độ glucoze, nồng độ pH, v.v…; có loại protein là men (*enzyme*) làm xúc tác cho các phản ứng hoá học từ đó tạo nên những phân tử mới khác. Cũng lưu ý rằng, trong số 20 amino acid mà loài người phát hiện ra được hiện nay lại có 9 amino acid thiết yếu mà một cơ thể sống không thể tự tạo ra, nên phải được đưa từ ngoài vào như là một nguồn thực phẩm.

Để chuyển từ RNA sang protein tế bào có một bộ phận được gọi là **Ribosome** đóng vai trò tổng hợp nên protein trong tế bào. Ribosome làm nhiệm vụ dịch chuỗi các phân tử tạo nên RNA thành ra chuỗi các amino acid tạo nên trình tự protein. Như vậy có thể nói Ribosome làm nhiệm vụ dịch mã RNA sang protein như những công đoạn của học thuyết trung tâm đề cập ở trên.

2. VIRUS, VACCINE

Virus là một tác nhân truyền nhiễm nên được coi như là một ký sinh trùng. Chúng có vật liệu di truyền là DNA và có cả emzyme để phiên mã thành RNA hoặc virus chỉ có RNA; nhưng tất cả đều không có Ribosome. Chính vì vậy virus không thể tự sinh sản theo nghĩa không thể tự nó tạo ra protein. Virus được coi là một vật ký sinh bởi chúng dùng bộ máy Ribosome của vật chủ (cơ thể sống mà nó ký sinh) để sinh sản. Từ đó tế bào vật chủ như là một nhà máy sản xuất virus mới, để rồi tiếp tục lây nhiễm vào các tế bào khác. Để làm được điều này, virus liên lạc và chèn RNA vào tế bào vật chủ thông qua gai bám *glycoprotein* của mình; từ đó Ribosome vật chủ đã vô tình sản sinh ra các trình tự protein là các amino acid trong đó có các amino acid tạo nên trình tự protein của virus.

Virus SARS-CoV-2 hay còn gọi là Wuhan Coronavirus trong giai đoạn đầu do xuất phát đầu tiên được phát hiện ở một chợ của thành phố Wuhan - Trung Hoa, sau đó lây lan hầu như cả thành phố này; đây chính là virus gây ra đại dịch CoViD-19. Virus này chứa một trình tự RNA với 29.903 nucleotide; đây là virus có số nucleotide lớn nhất trong các virus loại RNA [2]. Cấu tạo của virus này có nhiều protein, trong đó có một protein nằm trên bề mặt với nhiều gai (*spike*) tạo nên hình dáng giống vương miện (*corona*); vì vậy protein nảy được gọi là Spike Protein (*hay Protein S, hay PS2*).

Virus chính là một mầm bệnh (*pathogen*), mầm bệnh xâm nhập vào cơ thể sống rồi sinh sôi dẫn đến làm nhiễm trùng một hoặc nhiều bộ phận trong cơ thể sống. Để chống lại mầm bệnh khi bị nhiễm trung, cơ thể sống tìm đến cái để tiêu diệt đó là kháng nguyên (*antigen*), khi đó cơ thể sống sản sinh ra kháng thể (*antibody*) gắn vào kháng nguyên để tiêu diệt mầm bệnh. Cơ thể sống sản sinh ra rất nhiều kháng thể cho đến khi hết nhiễm trùng thì các kháng thể này sẽ tự bị loại bỏ. Bên cạnh đó, cơ thể sống cũng sản sinh ra các tế bào ghi nhớ (*memory cells*) để phòng khi kháng nguyên gây nhiễm trùng xuất hiện trở lại thì cơ thể sống nhanh chóng sản xuất kháng thể [3].

Con người đã và đang tìm cách để tìm ra phương thực diệt được virus gây hại trong cơ thể sống, đặc biệt là cơ thể người. Tuy nhiên, đây là một vấn đề khó vì virus sau khi xâm nhập có thể tồn tại rồi thích nghi để tiến hoá theo cùng. Chính vì vậy, việc tạo ra **vaccine** để đưa vào cơ thể sống, từ đó chống lại sự lây lan của virus là một giải pháp đã và đang tồn tại. Vaccine hoạt động bằng cách đưa kháng nguyên của mầm bệnh vào cơ thể để kích thích sản xuất kháng thể, đồng thời các tế bào ghi nhớ cũng được tạo ra.

Như vậy vaccine là chế phẩm được con người tạo ra với kháng nguyên từ virus gây bệnh, hoặc đôi khi được lấy từ vi sinh vật có cấu trúc kháng nguyên giống như của virus gậy bệnh. Ngoài ra vaccine cũng có thể là toàn bộ mầm bệnh (*là virus gây bệnh*) đã được thay đổi về mặt hóa học để không thể là mầm bệnh gây hại cho cơ thể sống. Dạng vaccine này được gọi là loại vaccine bất hoạt (*inactivated vaccine*). Lượng vaccine đưa vào cơ thể sống phải vừa đủ lượng (có dấu hiệu nhiễm trùng) sao cho đủ để kháng thể được tạo ra. Đây là cách tạo ra vaccine truyền thống.

Trong thời gian qua, hãng dược phẩm Pfizer của Mỹ, kết hợp với Công ty BioNTech của Đức; cũng như Công ty Moderna của Mỹ đã tạo ra vaccine khác với vaccine truyền thống, người ta gọi đó là vaccine dựa trên mRNA để ngăn ngừa virus SARS-CoV-2 gây ra đại dịch COVID-19 bắt đầu từ cuối năm 2019.

Spike protein của virus SARS-CoV-2 có thể coi đó là kháng nguyên của virus này, chính vì vậy vaccine mRNA là vaccine chứa các trình tự nucleotide để cơ thể sống tạo ra Spike protein. Từ đó cơ thể sống sẽ sản sinh ra kháng thể để chống lại, cũng như để tạo ra các tế bào ghi nhớ dùng cho việc tạo nhanh kháng thể vào những lần sau.

Vaccine mRNA có ưu điểm là được tạo ra nhanh chóng vì đó là các trình tự RNA, đồng thời nó chỉ chứa một protein (là Protein S) trong nhiều protein của virus SARS-CoV-2 nên mức độ nguy hiểm rất thấp khi đưa vào cơ thể sống; đồng thời là vaccine RNA nên không lưu trữ trong cơ thể sống để di truyền cho đời sau. Nhưng vaccine loại này có nhược điểm lớn đó là việc bảo quản trước khi đưa vào cơ thể sống vì các trình tự này chưa phải là một sinh vật; cần phải bảo quản trong một môi trường sao cho không bị phân huỷ trước khi nó sử dụng bộ máy Ribosome của cơ thể sống để hình thành và phát triển.

B. CÔNG TRÌNH LIÊN QUAN

Công ty Pfizer kết hợp với Công ty BioNTech cho ra đời vaccine mRNA để phòng ngừa virus SARS-CoV-2 có mã hiệu BTN162b2. Tên thương mại của vaccine này là Tozinameran hay là Comirnaty. Một tuần sau đó, ngày

18/12/2020 Công ty Moderna (được gộp từ cụm từ Modern RNA) cũng phát triển vaccine dựa trên nguyên lý mRNA, vacine nảy này có mã hiệu mRNA-1273. Theo [4], vaccine BTN162b2 có khả năng năng ngừa virus SARS-CoV-2 lên đến 95% cho những người trên 16 tuổi khi đã có tiêm vaccine. Vaccine này kích hoạt cơ thể sống sản xuất kháng thể một cách tự nhiên và kích thích các tế bào miễn dịch bảo vệ chống lại căn bệnh COVID-19.

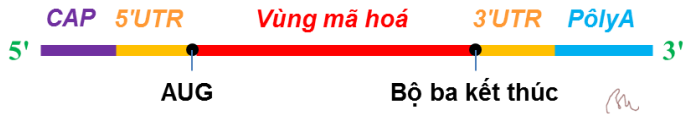
Vaccine BTN162b2 mã hoá protein PS2 nên được phát triển nhanh chóng, đặc biệt là tránh được nguy cơ tích hợp vào bộ gen cơ thể sống (tế bào vật chủ), đồng thời nhanh chóng tạo ra protein của virus (protein PS2) tinh khiết. Hơn nữa mRNA được biểu hiện mang tính nhất thời, nên cho phép tạo ra protein trong tế bào. Ngoài ra cũng do RNA được xác định về mặt phân tử, vì vậy được tổng hợp bằng quy trình phiên mã trong phòng thí nghiệm mà không cần lấy tế bào từ các khuôn mẫu DNA của virus. Điều đó cho thấy vaccine BTN162b2 không chứa các vật liệu có nguồn gốc từ virus SARS-CoV-2 [5].

Để giải quyết vấn đề cơ bản của vaccine dựa trên mRNA, đó là làm sao đưa một phần lạ vào cơ thể sống mà cơ thể sống không chống lại. Kết quả này là thành tựu chính từ những nghiên cứu của giáo sư *Katalin Karikó* người gốc Hungary; bà ta đã tạo ra một nucleoside (*là một nucleotide nhưng không có photphate*) biến đổi một chút để thay cho cho Uracine (U). Có nhiều nucleoside được tạo ra để sử dụng trong nghiên cứu, đặc biệt là trong điều trị thuốc; trong đó pseudouridine là chất có tên gọi *1-methyl-3'-pseudouridylyl* với ký hiệu là Ψ thay cho U được dùng để tạo ra vaccine BTN162b2 này [6]. Từ đó có thể tạo ra được mRNA lai có thể xâm nhập vào tế bào mà không làm báo động khả năng phòng vệ của cơ thể sống. Nói cách khác, với cách thay thế U bởi Ψ thì hệ thống Rebosome của cơ thể sống cũng sử dụng để tạo ra các amino acid để hình thành nên protein như bình thường với mRNA có U; sau đó mRNA kiểu mới này bị cơ thể sống phân huỷ để không còn tồn tại.

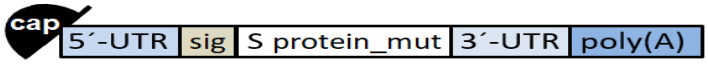
Những nghiên cứu xuất sắc này là xuất xứ để từ đó những nhà nghiên cứu về công nghệ sinh học góp phần thành lập nên Công ty Moderna với tên ban đầu là ModeRNA. Hiện nay Bà Katalin Karikó với vai trò là Phó chủ tịch cấp cao (*Senoir Vice President*) của BioNTech phụ trách những nghiên cứu về mRNA.

III. PHÂN TÍCH CẤU TRÚC CỦA VACCINE

Cấu trúc của một mRNA gồm các vùng như Hình 1, nên vaccine dựa trên mRNA cũng phải có tổ chức của các vùng tương tự. Cấu trúc mRNA đầy đủ về spike glycoprotein (protein S) của virus SARS-CoV-2 được WHO đưa ra trong thông báo 11889 vào tháng 9/2020 như Hình 2.



Hình 1. *Phân bổ các vùng chính của một mRNA trưởng thành [https://vi.m.wikipedia.org/wiki/Vùng_không_được_dịch_mã]*



Hình 2. *Sơ đồ cấu trúc (Schematic) mRNA về spike Glycoprotein của SARS-CoV-2*

Trên cơ sở đó, BioNTech và Pfizer đã tạo ra vaccine BNT162b2 là một trình tự RNA dài 4284 nucleotide bao gồm các phân tử A, C, G, Ψ như sau‡:

GAGAAΨAAAC ΨAGΨAΨΨCΨΨ CΨGGΨCCCCA CAGACΨCAGA GAGAACCCGC 50
CACCAΨGΨΨC GΨGΨΨCCΨGG ΨGCΨGCΨGCC ΨCΨGGΨGΨCC AGCCAGΨGΨG 100
ΨGAACCΨGAC CACCAGAACA CAGCΨGCCΨC CAGCCΨACAC CAACAGCΨΨΨ 150
ACCAGAGGCG ΨGΨACΨACCC CGACAAGGΨG ΨΨCAGAΨCCA GCGΨGCΨGCA 200
CΨCΨACCCAG GACCΨGΨΨCC ΨGCCΨΨΨCΨΨ CAGCAACGΨG ACCΨGGΨΨCC 250
ACGCCAΨCCA CGΨGΨCCGGC ACCAAΨGGCA CCAAGAGAΨΨ CGACAACCCC 300
GΨGCΨGCCCΨ ΨCAACGACGG GGΨGΨACΨΨΨ GCCAGCACCG AGAAGΨCCAA 350
CAΨCAΨCAGA GGCΨGGAΨCΨ ΨCGGCACCAC ACΨGGACAGC AAGACCCAGA 400
GCCΨGCΨGAΨ CGΨGAACAAC GCCACCAACG ΨGGΨCAΨCAA AGΨGΨGCGAG 450
ΨΨCCAGΨΨCΨ GCAACGACCC CΨΨCCΨGGGC GΨCΨACΨACC ACAAGAACAA 500
CAAGAGCΨGG AΨGGAAAGCG AGΨΨCCGGGΨ GΨACAGCAGC GCCAACAACΨ 550
GCACCΨΨCGA GΨACGΨGΨCC CAGCCΨΨΨCC ΨGAΨGGACCΨ GGAAGGCAAG 600
CAGGGCAACΨ ΨCAAGAACCΨ GCGCGAGΨΨC GΨGΨΨΨAAGA ACAΨCGACGG 650
CΨACΨΨCAAG AΨCΨACAGCA AGCACACCCC ΨAΨCAACCΨC GΨGCGGGAΨC 700
ΨGCCΨCAGGG CΨΨCΨCΨGCΨ CΨGGAACCCC ΨGGΨGGAΨCΨ GCCCAΨCGGC 750
AΨCAACAΨCA CCCGGΨΨΨCA GACACΨGCΨG GCCCΨGCACA GAAGCΨACCΨ 800
GACACCΨGGC GAΨAGCAGCA GCGGAΨGGAC AGCΨGGΨGCC GCCGCΨΨACΨ 850
AΨGΨGGGCΨA CCΨGCAGCCΨ AGAACCΨΨCC ΨGCΨGAAGΨA CAACGAGAAC 900
GGCACCAΨCA CCGACGCCGΨ GGAΨΨGΨGCΨ CΨGGAΨCCΨC ΨGAGCGAGAC 950
AAAGΨGCACC CΨGAAGΨCCΨ ΨCACCGΨGGA AAAGGGCAΨC ΨACCAGACCA 1000
GCAACΨΨCCG GGΨGCAGCCC ACCGAAΨCCA ΨCGΨGCGGΨΨ CCCCAAΨAΨC 1050

‡ https://web.archive.org/web/20210105162941/https://mednet-communities.net/inn/db/media/docs/11889.doc

ACCAAΨCΨGΨ GCCCCΨΨCGG CGAGGΨGΨΨC AAΨGCCACCA GAΨΨCGCCΨC 1100
ΨGΨGΨACGCC ΨGGAACCGGA AGCGGAΨCAG CAAΨΨGCGΨG GCCGACΨACΨ 1150
CCGΨGCΨGΨA CAACΨCCGCC AGCΨΨCAGCA CCΨΨCAAGΨG CΨACGGCGΨG 1200
ΨCCCCΨACCA AGCΨGAACGA CCΨGΨGCΨΨC ACAAACGΨGΨ ACGCCGACAG 1250
CΨΨCGΨGAΨC CGGGGAGAΨG AAGΨGCGGCA GAΨΨGCCCCΨ GGACAGACAG 1300
GCAAGAΨCGC CGACΨACAAC ΨACAAGCΨGC CCGACGACΨΨ CACCGGCΨGΨ 1350
GΨGAΨΨGCCΨ GGAACAGCAA CAACCΨGGAC ΨCCAAAGΨCG GCGGCAACΨA 1400
CAAΨΨACCΨG ΨACCGGCΨGΨ ΨCCGGAAGΨC CAAΨCΨGAAG CCCΨΨCGAGC 1450
GGGACAΨCΨC CACCGAGAΨC ΨAΨCAGGCCG GCAGCACCCC ΨΨGΨAACGGC 1500
GΨGGAAGGCΨ ΨCAACΨGCΨA CΨΨCCCACΨG CAGΨCCΨACG GCΨΨΨCAGCC 1550
CACAAAΨGGC GΨGGGCΨAΨC AGCCCΨACAG AGΨGGΨGGΨG CΨGAGCΨΨCG 1600
AACΨGCΨGCA ΨGCCCCΨGCC ACAGΨGΨGCG GCCCΨAAGAA AAGCACCAAΨ 1650
CΨCGΨGAAGA ACAAAΨGCGΨ GAACΨΨCAAC ΨΨCAACGGCC ΨGACCGGCAC 1700
CGGCGΨGCΨG ACAGAGAGCA ACAAGAAGΨΨ CCΨGCCAΨΨC CAGCAGΨΨΨG 1750
GCCGGGAΨAΨ CGCCGAΨACC ACAGACGCCG ΨΨAGAGAΨCC CCAGACACΨG 1800
GAAAΨCCΨGG ACAΨCACCCC ΨΨGCAGCΨΨC GGCGGAGΨGΨ CΨGΨGAΨCAC 1850
CCCΨGGCACC AACACCAGCA AΨCAGGΨGGC AGΨGCΨGΨAC CAGGACGΨGA 1900
ACΨGΨACCGA AGΨGCCCGΨG GCCAΨΨCACG CCGAΨCAGCΨ GACACCΨACA 1950
ΨGGCGGGΨGΨ ACΨCCACCGG CAGCAAΨGΨG ΨΨΨCAGACCA GAGCCGGCΨG 2000
ΨCΨGAΨCGGA GCCGAGCACG ΨGAACAAΨAG CΨACGAGΨGC GACAΨCCCCA 2050
ΨCGGCGCΨGG AAΨCΨGCGCC AGCΨACCAGA CACAGACAAA CAGCCCΨCGG 2100
AGAGCCAGAA GCGΨGGCCAG CCAGAGCAΨC AΨΨGCCΨACA CAAΨGΨCΨCΨ 2150
GGGCGCCGAG AACAGCGΨGG CCΨACΨCCAA CAACΨCΨAΨC GCΨAΨCCCCA 2200
CCAACΨΨCAC CAΨCAGCGΨG ACCACAGAGA ΨCCΨGCCΨGΨ GΨCCAΨGACC 2250
AAGACCAGCG ΨGGACΨGCAC CAΨGΨACAΨC ΨGCGGCGAΨΨ CCACCGAGΨG 2300
CΨCCAACCΨG CΨGCΨGCAGΨ ACGGCAGCΨΨ CΨGCACCCAG CΨGAAΨAGAG 2350
CCCΨGACAGG GAΨCGCCGΨG GAACAGGACA AGAACACCCA AGAGGΨGΨΨC 2400
GCCCAAGΨGA AGCAGAΨCΨA CAAGACCCCΨ CCΨAΨCAAGG ACΨΨCGGCGG 2450
CΨΨCAAΨΨΨC AGCCAGAΨΨC ΨGCCCGAΨCC ΨAGCAAGCCC AGCAAGCGGA 2500
GCΨΨCAΨCGA GGACCΨGCΨG ΨΨCAACAAAG ΨGACACΨGGC CGACGCCGGC 2550
ΨΨCAΨCAAGC AGΨAΨGGCGA ΨΨGΨCΨGGGC GACAΨΨGCCG CCAGGGAΨCΨ 2600
GAΨΨΨGCGCC CAGAAGΨΨΨA ACGGACΨGAC AGΨGCΨGCCΨ CCΨCΨGCΨGA 2650
CCGAΨGAGAΨ GAΨCGCCCAG ΨACACAΨCΨG CCCΨGCΨGGC CGGCACAAΨC 2700
ACAAGCGGCΨ GGACAΨΨΨGG AGCAGGCGCC GCΨCΨGCAGA ΨCCCCΨΨΨGC 2750
ΨAΨGCAGAΨG GCCΨACCGGΨ ΨCAACGGCAΨ CGGAGΨGACC CAGAAΨGΨGC 2800
ΨGΨACGAGAA CCAGAAGCΨG AΨCGCCAACC AGΨΨCAACAG CGCCAΨCGGC 2850
AAGAΨCCAGG ACAGCCΨGAG CAGCACAGCA AGCGCCCΨGG GAAAGCΨGCA 2900
GGACGΨGGΨC AACCAGAAΨG CCCAGGCACΨ GAACACCCΨG GΨCAAGCAGC 2950
ΨGΨCCΨCCAA CΨΨCGGCGCC AΨCAGCΨCΨG ΨGCΨGAACGA ΨAΨCCΨGAGC 3000
AGACΨGGACC CΨCCΨGAGGC CGAGGΨGCAG AΨCGACAGAC ΨGAΨCACAGG 3050
CAGACΨGCAG AGCCΨCCAGA CAΨACGΨGAC CCAGCAGCΨG AΨCAGAGCCG 3100
CCGAGAΨΨAG AGCCΨCΨGCC AAΨCΨGGCCG CCACCAAGAΨ GΨCΨGAGΨGΨ 3150
GΨGCΨGGGCC AGAGCAAGAG AGΨGGACΨΨΨ ΨGCGGCAAGG GCΨACCACCΨ 3200
GAΨGAGCΨΨC CCΨCAGΨCΨG CCCCΨCACGG CGΨGGΨGΨΨΨ CΨGCACGΨGA 3250
CAΨAΨGΨGCC CGCΨCAAGAG AAGAAΨΨΨCA CCACCGCΨCC AGCCAΨCΨGC 3300
CACGACGGCA AAGCCCACΨΨ ΨCCΨAGAGAA GGCGΨGΨΨCG ΨGΨCCAACGG 3350
CACCCAΨΨGG ΨΨCGΨGACAC AGCGGAACΨΨ CΨACGAGCCC CAGAΨCAΨCA 3400
CCACCGACAA CACCΨΨCGΨG ΨCΨGGCAACΨ GCGACGΨCGΨ GAΨCGGCAΨΨ 3450
GΨGAACAAΨA CCGΨGΨACGA CCCΨCΨGCAG CCCGAGCΨGG ACAGCΨΨCAA 3500
AGAGGAACΨG GACAAGΨACΨ ΨΨAAGAACCA CACAAGCCCC GACGΨGGACC 3550
ΨGGGCGAΨAΨ CAGCGGAAΨC AAΨGCCAGCG ΨCGΨGAACAΨ CCAGAAAGAG 3600
AΨCGACCGGC ΨGAACGAGGΨ GGCCAAGAAΨ CΨGAACGAGA GCCΨGAΨCGA 3650
CCΨGCAAGAA CΨGGGGAAGΨ ACGAGCAGΨA CAΨCAAGΨGG CCCΨGGΨACA 3700
ΨCΨGGCΨGGG CΨΨΨAΨCGCC GGACΨGAΨΨG CCAΨCGΨGAΨ GGΨCACAAΨC 3750
AΨGCΨGΨGΨΨ GCAΨGACCAG CΨGCΨGΨAGC ΨGCCΨGAAGG GCΨGΨΨGΨAG 3800
CΨGΨGGCAGC ΨGCΨGCAAGΨ ΨCGACGAGGA CGAΨΨCΨGAG CCCGΨGCΨGA 3850
AGGGCGΨGAA ACΨGCACΨAC ACAΨGAΨGAC ΨCGAGCΨGGΨ ACΨGCAΨGCA 3900
CGCAAΨGCΨA GCΨGCCCCΨΨ ΨCCCGΨCCΨG GGΨACCCCGA GΨCΨCCCCCG 3950
ACCΨCGGGΨC CCAGGΨAΨGC ΨCCCACCΨCC ACCΨGCCCCA CΨCACCACCΨ 4000
CΨGCΨAGΨΨC CAGACACCΨC CCAAGCACGC AGCAAΨGCAG CΨCAAAACGC 4050
ΨΨAGCCΨAGC CACACCCCCA CGGGAAACAG CAGΨGAΨΨAA CCΨΨΨAGCAA 4100
ΨAAACGAAAG ΨΨΨAACΨAAG CΨAΨACΨAAC CCCAGGGΨΨG GΨCAAΨΨΨCG 4150
ΨGCCAGCCAC ACCCΨGGAGC ΨAGCAAAAAA AAAAAAAAAA AAAAAAAAAA 4200
AAAAGCAΨAΨ GACΨAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA 4250
AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAA 4284

Trong đó,

- **Cap**: mã hoá bởi 2 nucleotide "GA" chỉ định các phần tiếp theo đến từ nhân tế bào để có thể sử dụng đươc
- **UTR**: vùng không được dịch mã (*untranslated region*), khi đó những nucleotite của vùng này bộ máy Risbosome không dùng để dịch mã sang protein. Có 2 vùng: 5'-UTR và 3'-UTR, với 5'UTR chứa 52 nucleotide kể từ vị trí thứ 3 đến vị trí 54; còn 3'-UTR gồm 295 nucleotide kể từ vị trí 3880 đến 4174

- **Sig**: peptide về tín hiệu gồm 48 nucleotide từ vị trí 55 đến vị trí 102, nhằm định vị để chuyển glycoprotein S đến vị trí phù hợp trên tế bào, nhằm sinh ra protein hiệu quả nhất.
- **S protein_mut**: trình tự glycoprotein S chứa đột biến K986P and V987P, đây là trình tự codon tối ưu gồm 3777 nucletotide từ vị trí 103 đến 3879, được dịch mã thành 1259 codon. Đây chính là thành phần được sử dụng để Ribosome tạo ra Protein S
- **poly(A)**: đuôi tín hiệu polyadenylate gồm 110 nucleotide, trong đó gồm 30 nucleotide A, 10 nucldeotide mang tính liên kết GCAΨAΨGACΨ tiếp theo là 70 nucleotide A.

Để kiểm tra, trình tự mRNA được lưu vào tập tin định dạng FASTA có tên BNT162b2.fasta với nội dung như sau:

```
>gi|BNT162b2
GAGAAΨAAACΨAGΨAΨΨCΨΨCΨGGΨCCCCACAGACΨCAGAGAGAACCCGC
CACCAΨGΨΨCGΨGΨΨCCΨGGΨGCΨGCΨGCCΨCΨGGΨGΨCCAGCCAGΨGΨG
ΨGAACCΨGACCACCAGAACACAGCΨGCCΨCCAGCCΨACACCAACAGCΨΨΨ
ACCAGAGGCGΨGΨACΨACCCCGACAAGGΨGΨΨCAGAΨCCAGCGΨGCΨGCA
CΨCΨACCCAGGACCΨGΨΨCCΨGCCΨΨΨCΨΨCAGCAACGΨGACCΨGGΨΨCC
…
```

Từ đây có thể rút trích các vùng 5'-UTR, Sig, 3'-UTR, S Protein_mut, Poly(A)

```
from Bio.Seq import Seq
from Bio import SeqIO
from Bio.SeqUtils import GC
for record in SeqIO.parse("BNT162b2.fasta", "fasta"):
   seq = record.seq
UTR5  = seq[2:54]
sig   = seq[54:102]
PS2   = seq[102:3879]
UTR3  = seq[3879:4174]
polyA = seq[4174:len(seq)]

print( "Tổng số nucleotide: ", len(seq) )
print( "5'UTR: ", UTR5, len(UTR5) )
print( "sig:       ", sig, len(sig) )
print( "Poly(A): ", polyA, len(polyA) )
print( "Số nucleotide của 3'UTR: ", len(UTR3) )
print( "Số nucleotide của Protein S đột biến: ", len(PS2) )
```

Để phân tích trình tự của vaccine so với trình tự mRNA của Protein S, ta chuyển Ψ trở thành U, sau đó so sánh tỷ lệ GC của virus SARS-CoV-2 với vaccine BNT162b2.

Ta biết vùng **sig** của virus là

```
AUGUUUGUUUUUCUUGUUUUAUUGCCACUAGUCUCUAGUCAGUGUGUU
```

Nên tỷ lệ GC trong trình tự của vaccine lớn hơn

```
sig_virus = Seq("AUGUUUGUUUUUCUUGUUUUAUUGCCACUAGUCUCUAGUCAGUGUGUU ")
sig_U = Seq( str(sig).replace("Ψ","U") )
print( "Tỷ lệ GC của virus  : %5.2f%%" % GC(sig_virus) )
print( "Tỷ lệ GC của vaccine: %5.2f%%" % GC(sig_U) )
```

Kết quả

```
Tỷ lệ GC của virus  : 32.65%
Tỷ lệ GC của vaccine: 58.33%
```

Từ đây cho thấy tỷ lệ GC của vaccine lớn hơn nên ổn định hơn. Ngoài ra, khi dịch mã sang protein, cả 2 phần Sig này đều giống nhau là MFVFLVLLPLVSSQCV. Điều này có được là do mặc dù có 64 codon nhưng chỉ có 20 amino acid, nên có những condon khác nhau nhưng lại cùng một amino acid:

```
print( sig_virus.translate() )
print( sig_U.translate() )
```

Trên cơ sở đó, thuật toán để tìm tỷ lệ GC tốt nhất được viết như sau với số lần tính thử là 500.000:

Thuật toán: Tìm GC tối ưu nhất
Nhập Trình tự protein P
Xuất GCmax, Trình tự mRNA R
Bắt đầu
    GCmax = GC(reverse_translate(P).transcribe())
    Lặp lại 500000 lần

DNA = reverse_translate(P)
R = DNA.transcribe()
GC = GC(R)
Nếu GC > GCmax
GCmax = GC
Kết thúc Nếu
Kết thúc lặp
Kết thúc

IV. THỰC NGHIỆM

Trên cơ sở trình tự mRNA của vaccine BNT162b2, chúng ta trích xuất phần Sig và phần Glycoprotein như đề cập ở trên, từ đó sử dụng thư viện Biopython (https://biopython.org) và thư viện DNA Chisel (https://pypi.org/project/dnachisel/) để hiện thực thuật toán tìm tỷ lệ GC tốt nhất.

Dữ liệu về trÌnh tự của vaccine BNT162b2 được lưu trữ trong tập tin dạng FASTA có tên BNT162b2.fasta; từ đó đưa vào trình tự để trích xuất.

```
from Bio import SeqIO
for record in SeqIO.parse("BNT162b2.fasta", "fasta"):
   seq = record.seq
print(seq)
```

Sa đó trích xuất các vùng **Sig** và **S Protein_mut** như sau:

```
sig     = seq[54:102]
PS2    = seq[102:3879] # S Protein_mut
```

Để coi biểu hiện protein của các đoạn trình tự này cần phải chuyển Nucleoside Ψ thành U:

```
from Bio.Seq import Seq
sig = Seq(str(sig).replace("Ψ","U"))
sig = sig.translate()
print("Sig: ", sig)

PS2 = Seq(str(PS2).replace("Ψ","U"))
PS2 = PS2.translate()
print("Glycoprotein S: ", PS2)
```

Từ đây ta có biểu hiện protein của 2 thành phần này là:

```
Sig:  MFVFLVLLPLVSSQCV
Glycoprotein S:
NLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFG
TTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNI
DGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAV
DCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCY
GVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDIST
EIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKF
LPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGC
LIGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYIC
GDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIK
QYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQ
FNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDPPEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIR
ASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGTHWFVTQ
RNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLID
LQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT**
```

Sau khi đã có trình tự protein, để tìm tỷ lệ GC tốt nhất, chúng ta hiện thực thuật toán ở trên bằng ngôn ngữ Python với các thư viện đã nêu như sau. Ở đây, dữ liệu trình tự protein được lưu vào file với tên là Sig.fa và PS2.fa tương ứng.

```
from Bio.Seq import Seq
from Bio.SeqUtils import GC
import dnachisel
from dnachisel.biotools import reverse_translate

record = dnachisel.load_record("PS2.fa")
max = GC((record.seq).transcribe())
print( "GC%% của vaccine: %5.2f" % max )
for index in range(1000):
```

```
        dna = Seq(reverse_translate(str(record.seq),randomize_codons=True))
        rna = dna.transcribe()
        gc  = GC(rna)
        if gc > max:
                max = gc
print( "GC%% tối ưu là %5.2f \nVới mRNA là %s" %(max,rna) )
```

Kết quả tính toán sau 500.000 lần lặp có tỷ lệ GC của vaccine và tỷ lệ tính toán, từ đó đề xuất trình tự mRNA tương ứng cho vaccine.

GC% của vaccine: 17.32
GC% tối ưu là 48.40
Với mRNA là
AACΨΨAACΨACΨCGCACACAAΨΨACCCCCGGCGΨAΨACCAAΨΨCCΨΨΨACGCGGGGΨGΨAΨACΨACCCAGAΨAAAGΨGΨΨ
ΨCGAΨCGΨCGGΨCCΨΨCAΨΨCCACCCAGGAΨΨΨAΨΨΨCΨCCCΨΨΨCΨΨΨAGCAACGΨGACCΨGGΨΨCCACGCAAΨΨCA
CGΨΨΨCAGGAACCAAΨGGAACGAAACGΨΨΨΨGAΨAACCCGGΨΨΨΨACCGΨΨCAACGACGGAGΨGΨACΨΨCGCCΨCCACΨ
GAGAAGAGΨAAΨAΨAAΨACGΨGGGΨGGAΨΨΨΨΨGGΨACGACΨΨΨGGACAGCAAGACΨCAAΨCΨCΨCCΨGAΨΨGΨΨAACA
ACGCCACAAAΨGΨΨGΨAAΨAAAAGΨΨΨGCGAGΨΨΨCAAΨΨCΨGCAACGAΨCCΨΨΨCCΨΨGGAGΨAΨAΨΨACCAΨAAGAA
ΨAACAAGAGΨΨGGAΨGGAAAGCGAAΨΨΨCGΨGΨΨΨACΨCCΨCCGCCAACAACΨGΨACGΨΨΨGAGΨAΨGΨCΨCΨCAACCG
ΨΨΨCΨΨAΨGGACCΨAGAGGGAAAGCAAGGAAACΨΨΨAAGAAΨΨΨGAGGGAAΨΨΨGΨGΨΨΨAAGAACAΨAGACGGAΨAΨΨ
ΨCAAGAΨΨΨAΨΨCΨAAACACACGCCAAΨAAACΨΨAGΨACGAGAΨΨΨGCCΨCAGGGGΨΨΨΨCAGCACΨGGAACCAΨΨAGΨG
GAΨCΨGCCGAΨAGGΨAΨΨAAΨAΨAACACGGΨΨΨCAAACGΨΨACΨGGCGΨΨACACAGAΨCAΨAΨCΨCACACCCGGCGAΨAG
ΨAGΨAGΨGGΨΨGGACGGCAGGCGCCGCCGCCΨACΨACGΨAGGAΨACΨΨACAACCCCGCACCΨΨCCΨCΨΨAAAAΨACAACG
AGAAΨGGΨACCAΨCACΨGAΨGCAGΨAGACΨGΨGCCCΨCGACCCΨΨΨAΨCΨGAGACGAAGΨGCACΨΨΨAAAGAGΨΨΨΨAC
GGΨCGAAAAGGGCAΨCΨACCAAACΨAGΨAACΨΨΨCGAGΨGCAGCCCACAGAAAGΨAΨCGΨACGCΨΨCCCGAACAΨCACGAA
CCΨΨΨGΨCCΨΨΨCGGGGAGGΨCΨΨCAACGCGACACGCΨΨCGCGΨCGGΨΨΨACGCAΨGGAACAGGAAACGGAΨAΨCCAAC
ΨGCGΨAGCΨGACΨAΨΨCGGΨACΨAΨACAAΨΨCCGCΨΨCΨΨΨΨAGΨACΨΨΨΨAAGΨGCΨACGGΨGΨCΨCACCΨACCAAG
CΨGAACGACΨΨAΨGΨΨΨCACGAAΨGΨΨΨAΨGCCGACΨCGΨΨΨGΨAAΨΨAGAGGCGAΨGAAGΨGCGGCAAAΨΨGCΨCCAG
GGCAAACAGGCAAAAΨCGCGGAΨΨACAACΨAΨAAACΨΨCCAGACGAΨΨΨCACΨGGGΨGΨGΨCAΨΨGCΨΨGGAAΨΨCCAAC
AACCΨAGAΨΨCGAAGGΨGGGAGGGAAΨΨACAAΨΨACCΨAΨACAGACΨAΨΨCCGCAAAΨCGAAΨCΨGAAACCGΨΨΨGAGCG
GGAΨAΨCAGΨACAGAGAΨΨΨACCAAGCAGGΨΨCAACACCΨΨGCAAΨGGΨGΨΨGAGGGCΨΨΨAACΨGCΨACΨΨCCCGΨΨA
CAAΨCGΨACGGΨΨΨCCAACCCACCAACGGCGΨAGGAΨACCAGCCGΨACCGAGΨAGΨGGΨACΨGΨCΨΨΨCGAAΨΨACΨCCA
CGCΨCCCGCΨACΨGΨΨΨGCGGACCGAAAAAGAGΨACΨAACCΨCGΨΨAAGAACAAAΨGCGΨΨAACΨΨΨAAΨΨΨΨAAΨGGGC
ΨGACΨGGAACCGGAGΨGCΨΨACAGAGAGΨAACAAAAAAΨΨΨCΨCCCGΨΨCCAACAAΨΨΨGGACGCGAΨAΨAGCAGAΨACAA
CCGAΨGCΨGΨΨCGCGAΨCCΨCAAACAΨΨGGAAAΨΨCΨGGAΨAΨAACGCCCΨGΨAGCΨΨCGGCGGAGΨAΨCΨGΨAAΨCAC
GCCAGGΨACGAACACAΨCΨAAΨCAAGΨAGCAGΨGCΨCΨAΨCAAGAΨGΨΨAAΨΨGCACAGAAGΨCCCΨGΨCGCΨAΨACAΨG
CGGACCAGCΨCACGCCAACGΨGGAGGGΨGΨAΨΨCΨACCGGΨΨCGAAΨGΨAΨΨCCAAACGCGΨGCGGGGΨGCCΨGAΨΨGG
ΨGCAGAGCAΨGΨΨAACAACΨCGΨAΨGAAΨGCGAΨAΨCCCCAΨΨGGAGCCGGΨAΨΨΨGCGCAΨCCΨACCAGACCCAGACGA
ACΨCACCCCGAAGGGCCAGGΨCΨGΨGGCCΨCCCAGΨCGAΨAAΨCGCΨΨAΨACAAΨGΨCAΨΨAGGCGCCGAAAAΨΨCΨGΨC
GCΨΨACΨCCAAΨAACΨCGAΨAGCΨAΨACCAACGAACΨΨΨACGAΨΨΨCCGΨAACGACCGAAAΨAΨΨACCΨGΨCΨCGAΨGAC
ΨAAAACAΨCΨGΨCGAΨΨGΨACΨAΨGΨACAΨAΨGΨGGGGAΨΨCΨACAGAGΨGΨAGΨAACCΨACΨΨCΨGCAGΨAΨGGCAGC
ΨΨΨΨGCACACAACΨAAACAGAGCGΨΨAACGGGGAΨAGCCGΨAGAACAAGACAAGAAΨACACAGGAGGΨΨΨΨCGCGCAGGΨ
ΨAAGCAGAΨCΨACAAGACCCCCCCAAΨΨAAAGAΨΨΨCGGΨGGAΨΨΨAACΨΨΨAGΨCAAAΨΨCΨΨCCAGACCCGAGΨAAAC
CGΨCAAAGCGGAGΨΨΨΨAΨCGAAGAΨCΨCCΨCΨΨCAAΨAAAGΨGACΨCΨGGCΨGACGCCGGΨΨΨCAΨCAAGCAGΨAΨGG
GGAΨΨGCΨΨAGGGGACAΨAGCAGCGCGΨGAΨCΨΨAΨAΨGCGCGCAAAAGΨΨΨAAΨGGACΨΨACGGΨACΨACCCCCΨΨΨG
CΨCACAGACGAAAΨGAΨCGCGCAAΨAΨACCΨCGGCACΨACΨΨGCΨGGCACGAΨAACCΨCAGGGΨGGACAΨΨΨGGGGCAGG
GGCAGCGCΨACAGAΨΨCCCΨΨCGCAAΨGCAAAΨGGCCΨAΨCGGΨΨCAAΨGGCAΨAGGGGΨGACACAGAACGΨCΨΨAΨAΨ
GAAAACCAAAAGCΨCAΨΨGCCAAΨCAAΨΨCAACΨCΨGCΨAΨAGGCAAAAΨCCAGGACAGCCΨCΨCGΨCCACGGCΨΨCΨGCA
CΨΨGGGAAACΨACAGGAΨGΨΨGΨGAACCAAAACGCGCAGGCACΨGAAΨACΨΨΨAGΨΨAAACAGCΨAAGΨΨCCAAΨΨΨΨGG
ΨGCΨAΨAAGCΨCAGΨGCΨGAACGAΨAΨACΨΨΨCGCGΨΨΨAGAΨCCGCCCGAAGCGGAAGΨGCAAAΨCGACAGGΨΨGAΨΨ
ACAGGGCGGΨΨGCAAΨCGΨΨGCAGACΨΨACGΨAACCCAGCAGCΨGAΨAAGAGCΨGCCGAAAΨΨCGGGCAΨCΨGCΨAAΨCΨ
CGCAGCAACGAAGAΨGΨCAGAGΨGCGΨGΨΨGGGΨCAGΨCAAAGCGΨGΨGGACΨΨCΨGCGGGAAAGGAΨAΨCACCΨΨAΨG
AGCΨΨΨCCCCAGΨCGGCCCCΨCACGGGGΨAGΨCΨΨΨCΨΨCACGΨAACGΨAΨGΨΨCCAGCΨCAGGAAAAGAAΨΨΨΨACGA
CCGCCCCGGCGAΨAΨGΨCACGAΨGGAAAGGCGCACΨΨCCCACGΨGAAGGAGΨCΨΨΨGΨGΨCΨAAΨGGCACCCACΨGGΨΨ
ΨGΨAACCCAGCGGAAΨΨΨCΨAΨGAGCCΨCAGAΨCAΨΨACΨACGGACAACACΨΨΨCGΨAAGCGGΨAACΨGCGACGΨGGΨGA
ΨΨGGAAΨCGΨCAACAAΨACGGΨCΨACGACCCACΨCCAGCCCGAGCΨAGACAGCΨΨΨAAAGAAGAGΨΨAGAΨAAAΨAΨΨΨΨ
AAAAAΨCACACΨΨCCCCCGAΨGΨCGACCΨAGGΨGACAΨΨAGΨGGΨAΨCAACGCAΨCGGΨΨGΨGAAΨAΨCCAAAAAGAGAΨ
AGAΨCGCCΨGAACGAGGΨAGCCAAGAACΨΨGAACGAAΨCGCΨCAΨAGAΨΨΨGCAGGAGCΨGGGCAAAΨAΨGAGCAAΨAΨA
ΨAAAAΨGGCCAΨGGΨAΨAΨAΨGGCΨAGGAΨΨΨAΨCGCAGGAΨΨAAΨAGCΨAΨAGΨΨAΨGGΨΨACGAΨCAΨGCΨAΨGCΨ
GΨAΨGACCAGCΨGCΨGΨAGCΨGCΨΨGAAGGGGΨGΨΨGΨAGCΨGCGGCAGCΨGΨΨGΨAAAΨΨΨGAΨGAGGAΨGAΨΨCAG
AACCAGΨCCΨCAAGGGCGΨΨAAGΨΨACAΨΨAΨACCΨGAΨAG

Tương tự như vậy, có thể tính toán cho trình tự Sig.

V. KẾT LUẬN

Máy tính cũng như các phương tiện của Tin học có vai trò rất lớn trong việc tạo ra vaccine dựa trên mRNA. Để thực hiện được điều này, những nghiên cứu cơ bản trong lĩnh vực Sinh hoá là vô cùng quan trọng, nhưng nếu thiếu những công cụ của Tin học thì việc tạo ra vaccine không thể dễ dàng và nhanh chóng được. Riêng đối với Tin học, việc lập trình để hỗ trợ những công đoạn tính toán cũng là vấn đề. Ngày nay với sự phát triển của thư viện ngôn ngữ lập trình, chúng ta cũng nhanh chóng tạo ra sản phẩm. Trong số các ngôn ngữ lập trình thông dụng ngày nay, Python là ngôn ngữ có nhiều thư viện hỗ trợ nhất cho lĩnh vực Sinh học phân tử. Những kết quả trong phần thực nghiệm là một minh chứng về điều này: với kiến thức về chuyên môn trong lĩnh vực Sinh học, thông qua các thư viện lập trình sẵn có sẽ đưa ra được giải pháp tính toán phù hợp. Bài viết này thử nghiệm tính

toán tỷ lệ thành phần G và C trong một trình tự để tìm tỷ lệ cao nhất qua đó tạo ra các các trình tự nucleotide có liên kết bền vững nhất bằng phương tiện Tin học.

VI. TÀI LIỆU THAM KHẢO

[1] Trần Văn Lăng (2008). *Ứng dụng Tin học trong việc giải một số bài toán của Sinh học phân tử*, NXB. Giáo Dục, 230tr.

[2] Bernd Sebastian Camps, Christian Hoffmann (04/92020). *Covid Reference*, Steinhauser Verlag, 290p (https:// amedeo.com/CovidReference04_vn.pdf)

[3] Ngọc Bích (12/2020). *Cơ chế hoạt động của vắc xin*, https://thetinypharmacist.org/2020/12/24, Truy cập: 05/4/2021.

[4] Polack FP, Thomas SJ, Kitchin N, Absalon J, Gurtman A, Lockhart S, Perez JL, Pérez Marc G, Moreira ED, Zerbini C, Bailey R, Swanson KA, Roychoudhury S, Koury K, Li P, Kalina WV, Cooper D, Frenck RW Jr, Hammitt LL, Türeci Ö, Nell H, Schaefer A, Ünal S, Tresnan DB, Mather S, Dormitzer PR, Şahin U, Jansen KU, Gruber WC; C4591001 Clinical Trial Group (2020). *Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine*. N Engl J Med. 2020 Dec 31;383(27):2603-2615. doi: 10.1056/NEJMoa2034577. Epub 2020 Dec 10. PMID: 33301246; PMCID: PMC7745181.

[5] WHO (01/2021). *Background document on mRNA vaccine BNT162b2 (Pfizer-BioNTech) against COVID-19*. https://www.who.int/publications/i/item/background-document-on-mrna-vaccine-bnt162b2-(pfizer-biontech)-against-covid-19. Truy cập : 23/4/2021.

[6] Karikó K, Buckstein M, Ni H, Weissman D (2005). *Suppression of RNA recognition by Toll-like receptors: the impact of nucleoside modification and the evolutionary origin of RNA*. Immunity. 2005 Aug;23(2):165-75. doi: 10.1016/j.immuni.2005.06.008. PMID: 16111635.

AN ANALYSIS OF mRNA-BASED STRUCTURE OF BNT162b2 COVID-19 VACCINE

Tran Thuy Anh Quynh, Tran Van Lang

ABSTRACT— In this paper, we describe the submission guidelines for preparing papers for the HJS (Style Abstract). The paper presents basic knowledge of mRNA-Based vaccine, and analyzes the structure of the biological sequence of BNT162B2 vaccine which prevents SARS-CoV-2 virus that caused COVID-19 epidemic. This is a vaccine developed by Pfizer and BioNTech Company. From there, building algorithms and programs written in Python to select mRNA sequence on the basis of optimal GC ratio.