

THIẾT KẾ TỐI ƯU GEN ĐỂ BIỂU HIỆN DỊ LOẠI SỬ DỤNG HỆ THỐNG THƯƠNG MẠI BẰNG MỘT SỐ PHẦN MỀM TỐI ƯU MIỄN PHÍ

Hà Thúc Đức Tùng, Đoàn Trọng Bích, Nguyễn Ngọc Lương*

Khoa Sinh học, Trường Đại học Khoa học, Đại học Huế

*Email: daigoro77@gmail.com

Ngày nhận bài: 21/12/2018; ngày hoàn thành phản biện: 21/12/2018; ngày duyệt đăng: 21/12/2018

TÓM TẮT

Tối ưu gen để biểu hiện khác loài là một bước quan trọng trong công nghệ DNA tái tổ. Đã có nhiều công trình nghiên cứu về các tiêu chí cần đạt được trong quá trình tối ưu gen. Một số tiêu chí phổ quát gồm thiên vị mã bộ ba, thiên vị cặp mã bộ ba và tối ưu cấu trúc thứ cấp của mRNA. Để có thể đồng thời đáp ứng nhiều tiêu chí tối ưu, các nhà Tin sinh học phát triển các phần mềm tối ưu gen nhằm tự động hóa quá trình này. Tuy nhiên để có được gen theo ý muốn, cần có sự can thiệp của con người. Ở đây chúng tôi điếm qua một số phần mềm tối ưu gen phổ biến và miễn phí cho mục đích tối ưu biểu hiện gen ở một số vật chủ phổ biến. Chúng tôi cũng mô tả quy trình sử dụng các phần mềm này và một số kinh nghiệm rút ra trong quá trình sử dụng phần mềm khi tối ưu gen để biểu hiện một số protein có nhiều ứng dụng trong Y dược học ở *Escherichia coli* và nấm men.

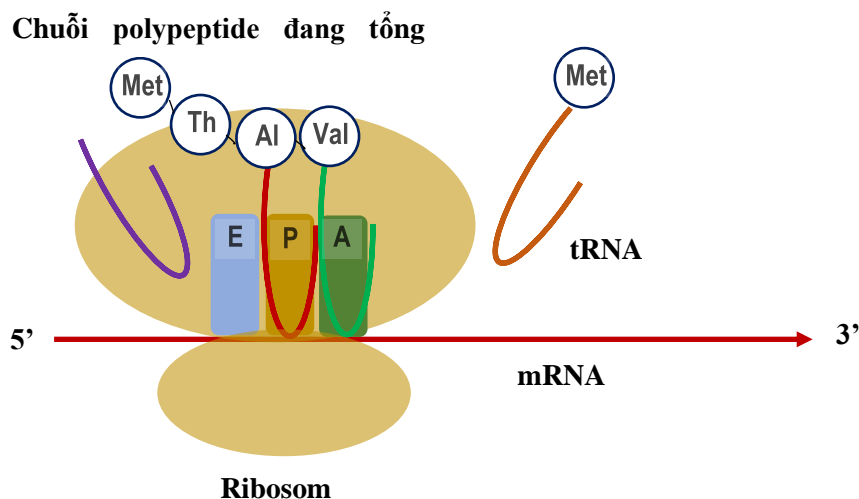
Từ khóa: Tối ưu gen, biểu hiện dị loài, *E. coli*, nấm men, tiêu chí tối ưu.

1. MỞ ĐẦU

Biểu hiện gen dị loài, trong đó gen được lấy từ loài này được đem đi biểu hiện ở một sinh vật thường dùng làm vật chủ, ví dụ *Escherichia coli* (*E. coli*), nấm men bánh mì *Saccharomyces cerevisiae* (*S. cerevisiae*), tế bào thực vật và tế bào động vật, đặt nền móng cho ngành công nghiệp công nghệ sinh học có trị giá hàng trăm tỉ USD [1], [2], [3]. Gen biểu hiện dị loài sẽ chịu sự điều hòa biểu hiện gen của vật chủ, tức gồm các bước điều hòa ngoại di truyền (epigenetics), điều hòa phiên mã, điều hòa hậu phiên mã thông qua miRNA và điều hòa sau dịch mã [4], [5]. Trong thực tế, người dùng chỉ có thể kiểm soát quá trình dịch mã của gen quan tâm do những yếu tố điều hòa ngoại di truyền, điều hòa phiên mã và hậu dịch mã đều đã được tối ưu hóa thông qua các hệ thống biểu hiện thương mại (vector và tế bào vật chủ thương mại) [1].

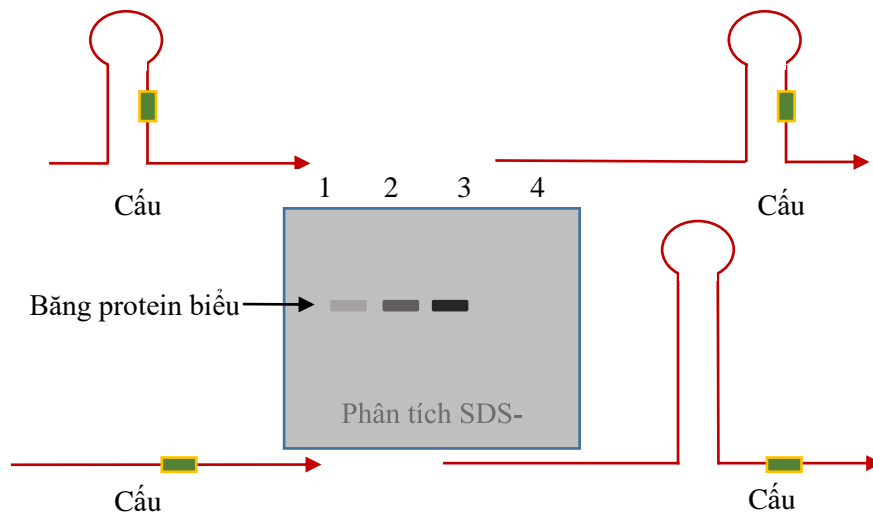
Hai phương pháp phổ biến để thu nhận gen cần biểu hiện là tạo dòng gen từ nguồn tự nhiên và tổng hợp gen. Tạo dòng gen từ nguồn tự nhiên sử dụng các kỹ thuật tạo dòng phổ biến như PCR, hoặc kết hợp các kỹ thuật tạo dòng phức tạp hơn như PCR môi trường hóa, sàng lọc thư viện hệ gen hoặc thư viện cDNA để thu nhận gen quan tâm. Kỹ thuật này thường đi kèm một số khó khăn như khó thu nhận nguồn sinh vật hiến (ví dụ sinh vật hiến là virus gây bệnh) và phải biết trước trình tự DNA của gen cần tạo dòng. Đặc biệt việc sử dụng gen tự nhiên để biểu hiện dị loài thường không đảm bảo mức độ biểu hiện tốt để có thể thương mại hóa sản phẩm [6], [7]. Phương pháp tối ưu và tổng hợp gen ngày nay được đón nhận và áp dụng rộng rãi bởi một số ưu việt của nó như dễ thu nhận từ trình tự trên các cơ sở dữ liệu sinh học, tính ổn định cao khi biểu hiện và giá thành tổng hợp đã giảm mạnh trong thời gian qua [1], [3].

Tối ưu gen để biểu hiện dị loài đã đạt được một số khám phá quan trọng, giúp các nhà khoa học thiết kế tối ưu gen với kết quả biểu hiện ổn định và dễ dự đoán hơn. Trong thời kỳ đầu tối ưu gen được hiểu đơn giản là thay thế các mã bộ ba đồng nghĩa hiếm trong gen cần biểu hiện bằng các mã bộ ba phổ biến hơn ở vật chủ biểu hiện. Tuy nhiên cách tiếp cận đơn giản này thường cho kết quả không ổn định do khi tất cả mã bộ ba trong gen đều là bộ ba phổ biến, tRNA sẽ nhanh chóng bị cạn kiệt dẫn đến tốc độ tăng trưởng bị ảnh hưởng. Bằng cách “bắt chước” các gen biểu hiện mạnh ở vật chủ, gen cần biểu hiện được tối ưu sao cho mã bộ ba được phân bố theo một số đặc trưng nội tại của các gen ở vật chủ. Cách tiếp cận này không chỉ đảm bảo gen cần tổng hợp được dịch mã tốt mà các yếu tố ảnh hưởng đến mức độ dịch mã khác như độ bền mRNA, các trình tự lặp, mã kết thúc ẩn cũng được tối ưu hóa [6].



Hình 1. Tại mỗi thời điểm bất kỳ luôn có hai phân tử tRNA gắn với tiểu phần lớn của ribosome tại vị trí P và A. Hai phân tử tRNA này phải tương thích về cấu trúc để quá trình dịch mã không bị gián đoạn.

Tại thời điểm này một số tiêu chí quan trọng trong thiết kế tối ưu gen để tổng hợp dị loại đã được xác lập. Tiêu chí quan trọng nhất chính là tối ưu mã bộ ba sao cho một số mã bộ ba hiếm bị thay thế để quá trình dịch mã gen này tối ưu, nhưng không ảnh hưởng đến tốc độ tăng trưởng. Tiêu chí này được đánh giá bằng nhiều phương pháp khác nhau nhưng phổ biến nhất là chỉ số thích ứng mã bộ ba (Codon Adaptation Index, viết tắt CAI) [1], [2], [3], [6]. Một tiêu chí không kém phần quan trọng là sự hài hòa về các cặp mã bộ ba nằm cạnh nhau. Điều kiện này xuất phát từ thực tế rằng tại bất kỳ thời điểm nào trong bước kéo dài của quá trình dịch mã, ribosome luôn luôn gắn với hai mã bộ ba [1], [3], [8] (xem Hình 1). Tiêu chí thứ ba là cấu trúc thứ cấp của mRNA, được phân thành hai tiêu chí phụ gồm cấu trúc thứ cấp tổng thể của mRNA và cấu trúc thứ cấp của mRNA ở vùng 5' không dịch mã (UTR 5'). mRNA phải có cấu trúc thứ cấp đủ bền để đảm bảo không dễ dàng bị các exonuclease phân hủy, nhưng phải có đầu 5' có cấu trúc thứ cấp yếu để tạo điều kiện cho ribosome gắn vào, đặc biệt là ở sinh vật nhân chuẩn [1], [3], [9], [10], [11]. Thực tế, đã có thí nghiệm đẹp cho thấy khi cấu trúc thứ cấp của mRNA ở vùng 5' quá chặt, gen hầu như không được dịch mã cho dù có rất nhiều mRNA [12] (xem Hình 2). Ba tiêu chí trên trở thành những tiêu chí quan trọng nhất trong thiết kế gen để biểu hiện dị loại. Ngoài những tiêu chí quan trọng nói trên một số tiêu chí khác cũng thường được đưa vào quá trình tối ưu gen như tỉ lệ GC trong gen [13], các trình tự làm suy yếu mRNA [14], [15], mã kết thúc ẩn và các vị trí cắt hạn chế không mong muốn [1].



Hình 2. Cấu trúc thứ cấp của mRNA ở vùng 5' không dịch mã ảnh hưởng rất lớn đến hiệu quả dịch mã, qua đó ảnh hưởng đến mức độ biểu hiện. Mã mở đầu được biểu diễn bằng hộp màu xanh lục trong hình. Hiệu quả dịch mã lớn nhất khi không có cấu trúc thứ cấp ở vùng UTR 5' (cấu trúc 3). Ở cấu trúc 1, do vị trí để ribosome gắn vào mRNA quá gần nên hiệu quả dịch mã giảm. Ở cấu trúc 2, tuy ribosome có thể gắn vào mRNA nhưng sự hiện diện của cấu trúc bậc 2 ở vùng chứa mã mở đầu khiến hiệu quả dịch mã kém hơn so với cấu trúc 3. Ở cấu trúc 4, do cấu

Thiết kế tối ưu gen để biểu hiện dị loài sử dụng hệ thống thương mại bằng một số phần mềm tối ưu miễn phí

trúc thứ cấp có năng lượng quá lớn, ribosome không thể tiếp cận mã mở đầu dẫn đến protein không được tổng hợp (Kozak, 1989).

Tuy có sự đồng thuận về các tiêu chí quan trọng trong thiết kế và tối ưu gen để biểu hiện dị loài ở mức đại thể, một số chi tiết về các tiêu chí này đã được chứng minh đóng vai trò rất quan trọng đối với hiệu quả dịch mã. Mark Welch và cs phát hiện rằng một số mã bộ mã mã hóa cho một số axit amin, cụ thể là serine (mã hóa bởi AGC), threonine (ACG) và leucine (AAG) ảnh hưởng rất lớn đến mức độ dịch mã. Những mã bộ ba này giúp tăng hiệu quả dịch mã vì các tRNA mang axit amin tương ứng với các mã bộ ba này được nạp axit amin mới nhanh hơn các tRNA mang mã bộ ba đồng nghĩa cho dù chúng phong phú hơn. Điều này đặc biệt đúng với biểu hiện gen ở *E. coli* do mức độ phiên mã ở *E. coli* rất mạnh [16]. Thomas E. Goro-chowski và cs nghiên cứu kỹ hơn về cấu trúc của mRNA ở các gen tự nhiên và sự phong phú của các tRNA có khả năng nạp nhanh cho thấy có những áp lực chọn lọc để hai yếu tố này bù trừ nhau để quá trình dịch mã diễn ra thuận lợi hơn cho tất cả các gen. Nghiên cứu này cũng cho thấy cần có những tiêu chí cụ thể hơn nữa trong quá trình thiết kế tối ưu gen [17].

Để thiết kế tối ưu gen thỏa mãn đồng thời các tiêu chí nói trên là một nhiệm vụ bất khả thi nếu thực hiện bằng tay. Vì vậy đã có nhiều phần mềm được xây dựng để hỗ trợ cho người dùng thiết kế tối ưu gen. Một số phần mềm phổ biến gồm DNAworks, Jcat, Synthetic gene designer, GeneDesign, OPTIMZER, Visual Gene Developer, Eugene, COOL, D-Tailor và Costar. Các phần mềm này cho phép người dùng tùy biến các tiêu chí phổ biến nói trên, ưu tiên một số tiêu chí so với các tiêu chí còn lại tùy theo kinh nghiệm và quan điểm riêng. Tất cả các phần mềm này đều miễn phí và chạy trên nền tảng Windows/OS hoặc trực tiếp chạy trên web [1], [3].

Tuy thiết kế tối ưu gen đã được ứng dụng rộng rãi trong công nghiệp Công nghệ sinh học, ở quy mô phòng thí nghiệm, đặc biệt ở Việt Nam, quá trình biểu hiện gen dị loài vẫn được thực hiện theo phương pháp tạo dòng và biểu hiện các gen tự nhiên. Chính vì vậy trong nghiên cứu này chúng tôi mô tả một số thí nghiệm tối ưu gen và phương pháp để có được gen tối ưu tổng hợp với giá thành rẻ nhất để tiến hành các thí nghiệm tiếp theo. Ba gen được chọn cho nghiên cứu là gen mã hóa miền III của protein vỏ virus Dengue tuýp 1 (EDIII) gây bệnh sốt xuất huyết [18], gen mã hóa kháng nguyên vỏ của virus Coxsackievirus A16 gây bệnh chân tay miệng ở trẻ em (Polyprotein) [19], và gen mã hóa protein độc tố nhạy nhiệt của *E. coli* (LTB) gây bệnh tiêu chảy ở người [20]. EDIII được tối ưu để biểu hiện ở *E. coli*, trong khi LTB và Polyprotein được tối ưu để biểu hiện ở nấm men *Pichia pastoris* và *Saccharomyces cerevisiae*. Ở đây chúng tôi cũng sẽ bàn một số ưu và nhược điểm của những phần mềm chúng tôi đã chọn để tối ưu các gen này, ở đây cụ thể là Visual Gene Developer và COOL.

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Vật liệu

Trình tự DNA hoặc protein của các gen nói trên được thu nhận từ cơ sở dữ liệu (CSDL) Genbank hoặc Genpept của NCBI thông qua tìm kiếm bằng từ khóa trên Entrez. Trường hợp gen virus, có thể thấy có sự biến dị lớn ở trình tự của các chủng khác nhau, do đó cần phải chọn trình tự đại diện nhất bằng cách tìm kiếm tất cả các trình tự polypeptide của protein cần tìm kiếm trên CSDL, tiến hành sắp giống cột nhiều trình tự để tìm trình tự đại diện nhất (consensus).

2.2. Phương pháp nghiên cứu

2.2.1 Thu nhận trình tự

Để tìm kiếm các trình tự của EDIII, Polyprotein và LTB, trước tiên ta tìm kiếm các trình tự protein trên CSDL genpept bằng các từ khóa liên quan. Ví dụ để tìm kiếm trình tự polypeptide của EDIII của virus Dengue tuýp 1 ta có thể đánh câu lệnh:

```
Dengue virus 1[organism] AND Envelope protein AND 495[sequence length]
```

Theo câu lệnh tìm kiếm này kết quả trả lại sẽ là các trình tự protein envelope của virus Dengue tuýp 1 có chiều dài chính xác là 495 axit amin, là chiều dài đầy đủ của protein này. Tương tự ta cũng có thể tìm kiếm trình tự của polyprotein của Coxsackievirus A16 bằng câu lệnh:

```
Coxsackievirus A16[organism] AND polyprotein AND 2193[sequence length]
```

Sau đó ta tải tất cả các trình tự tìm thấy và sắp giống cột bằng các phần mềm sắp giống cột nhiều trình tự phổ biến như MEGA [21] hoặc các phần mềm sắp giống cột mạng có khả năng xử lý dữ liệu lớn. Kết quả sắp giống sẽ được dùng làm dữ liệu đầu vào cho phần mềm Weblogo [22] để tạo trình tự đại diện nhất. Trình tự đại diện này sẽ được dùng để tìm kiếm BLAST xem liệu có trình tự polypeptide giống như vậy hay không trong tự nhiên. Tiếp đến trình tự DNA mã hóa cho các protein này sẽ được thu nhận để dùng làm đối chứng (gen tự nhiên).

2.2.2. Tối ưu trình tự

Để tối ưu EDIII biểu hiện ở *E. coli*, chúng tôi vận dụng kết quả của Mark Welch và cs làm tiêu chí tối ưu chính, có nghĩa là ràng buộc sao cho các mã bộ ba mã hóa serine, threonine và leucine sẽ lần lượt là AGC, ACG, và AAG. Để làm điều này chúng tôi sử dụng phần mềm Visual Gene Developer [23] và thay đổi tỉ lệ các mã bộ ba đồng nghĩa sao cho đối với serine, threonine và leucine, các mã bộ ba AGC, ACG và AAG sẽ chiếm tỉ lệ cao nhất và tất cả các mã bộ ba đồng nghĩa còn lại chiếm tỉ lệ thấp như nhau.

Thiết kế tối ưu gen để biểu hiện dị loài sử dụng hệ thống thương mại bằng một số phần mềm tối ưu miễn phí

Để tối ưu EDIII và polyprotein Coxsackievirus A16 chúng tôi sử dụng phần mềm COOL [24]. Tuy nhiên do quá trình dịch mã các gen biểu hiện cao khác với quá trình dịch mã các gen biểu hiện bình thường hoặc thấp, do đó không thể dùng dữ liệu chứa các gen biểu hiện thấp ở vật chủ cần biểu hiện mà nên chọn các gen biểu hiện cao. Để xác định các gen biểu hiện cao ở một số loài, chúng tôi sử dụng cơ sở dữ liệu về độ phong phú của protein ở sinh giới (PaxDb) để xác định các protein có hàm lượng cao ở *E. coli* và nấm men, và sau đó sử dụng các gen biểu hiện cao này làm mô hình để phần mềm tái thiết kế gen quan tâm [25].

Do polyprotein sẽ tự phân cắt thành 3CD là enzyme proteinase sẽ xử lý P1 thành các protein VP1, VP2, VP3 và VP4, chúng tôi tiến hành tối ưu protein này dưới dạng hai protein riêng rẽ là 3CD và P1.

Các phần mềm thường sẽ cho một số trình tự ứng viên để người dùng lựa chọn. Trong trường hợp này chúng ta có thể chọn ngẫu nhiên ba trình tự (ví dụ trường hợp EDIII) hoặc chọn một trình tự có tiêu chí nào đó trong các tiêu chí được chúng ta quan tâm hơn (ví dụ có chỉ số thích ứng cặp mã bộ ba cao), đồng thời đảm bảo các chỉ số của các tiêu chí khác nằm trong khoảng chấp nhận được.

Các trình tự tối ưu được chọn sẽ được gửi đi tổng hợp tại công ty Phusa Biochem (<http://www.phusabiochem.com/vi/.html>) dưới dạng geneblock, với giá thành khoảng 1,4 triệu cho một gen dài dưới 500 bp. Đối với những gen dài trên 500 bp, chúng sẽ được tổng hợp dưới dạng các trình tự geneblock dưới 500 bp và sau đó nối lại với nhau bằng phản ứng overlap extension PCR. Để tạo ra các trình tự geneblock này, phần mềm DNAworks 2.0 được sử dụng [26].

3. KẾT QUẢ VÀ THẢO LUẬN

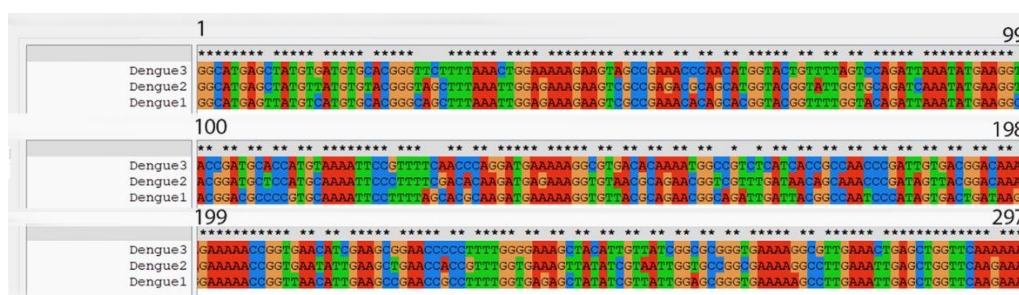
3.1. Tối ưu gen EDIII của virus Dengue tuýp 1

Dựa trên cách làm mô tả ở phần PHƯƠNG PHÁP NGHIÊN CỨU chúng tôi đã thu nhận được hơn 200 gen tối ưu mã hóa cho EDIII. Do phần mềm Visual Gene Developer thiếu chức năng tính chỉ số thích hợp cặp mã bộ ba, chúng tôi buộc phải sử dụng phần mềm Anaconda 2.0 để tính toán độ thích nghi cặp mã bộ ba của các gen tối ưu [27]. So sánh độ thích nghi cặp mã bộ ba của các gen tối ưu này với một số gen tự nhiên có mức độ biểu hiện mạnh ở *E. coli* giúp chọn ra ít nhất 18 trình tự để có thể sử dụng cho thí nghiệm tổng hợp gen. 3 gen được chọn ngẫu nhiên cho các thí nghiệm tiếp theo được trình bày ở đây dưới dạng kết quả giống cột trình tự, chỉ số CAI, năng lượng gập cuộn cấu trúc bậc hai mRNA và chỉ số thích nghi cặp mã bộ ba. Dựa trên kết quả này, 3 gen được chọn đều có CAI thấp hơn gen tự nhiên (0.72) nhưng các chỉ số khác đều tốt hơn. Theo nhiều nghiên cứu CAI hiện tại không còn là một chỉ số đáng tin

cây để dự đoán mức độ biểu hiện [3] do đó chúng tôi giả định rằng mức độ biểu hiện của các gen được chọn sẽ xấp xỉ hoặc hơn gen tự nhiên.

Bảng 1. Gen mã hóa EDIII của virus Dengue tuýp 1 đã được tối ưu để biểu hiện ở *E. coli*

Tên trình tự	Hàm lượng GC3	CAI	Chỉ số thích nghi cặp mã bộ ba	Năng lượng cấu trúc bậc hai mRNA
Dengue 1	50,5	0,69	-0,12086	-77,5
Dengue 2	45,4	0,69	-0,12042	-81,4
Dengue 3	47,4	0,67	-0,11857	-83,5



Hình 3. Các trình tự EDIII của virus Dengue tuýp 1 tối ưu để biểu hiện ở *E. coli*. Kết quả sắp giống cột được thực hiện bằng phần mềm ClustalX.

Để ràng buộc điều kiện gen tối ưu phải có các mã bộ ba nhất định cho một số axit amin (serine mã hóa bởi AGC, threonine mã hóa bởi ACG và leucine mã hóa bởi AAG), chúng tôi thay đổi tần số của các mã bộ ba mã hóa cho các axit amin này sao cho AGC, ACG và AAG có tần số cao nhất có thể có. Tần số này được tính bằng cách tìm tần số thấp nhất của nhóm mã bộ ba đồng nghĩa trong bảng mã chuẩn của *E. coli* (ở đây là AGG mã hóa cho Arg, tần số 0.02), sau đó áp dụng tần số này cho tất cả mã bộ ba đồng nghĩa còn lại. Ví dụ đối với Thr, bốn mã bộ ba đồng nghĩa là ACC (0,440) ACG (0,270), ACA (0,130), ACT (0,160). Bằng cách ràng buộc điều kiện ACG sẽ mã hóa cho Thr khi có thể, ba mã bộ ba ACC, ACA và ACT sẽ được gán tần số thấp nhất là 0,02 và tần số của ACG mã hóa cho Thr sẽ là $1 - 3 \times 0,02 = 0,94$. Tương tự như vậy tần số của AGC mã hóa cho serine sẽ là $1 - 5 \times 0,02 = 0,9$ và tần số của AAG mã hóa cho leucine sẽ là $1 - 5 \times 0,02 = 0,9$. Bằng cách gán tần số như vậy ta đảm bảo rằng AGC, ACG và AAG sẽ xuất hiện nhiều nhất có thể trong trình tự gen tối ưu, nhưng vẫn bị ràng buộc theo các tiêu chí khác như độ thích nghi cặp mã bộ ba, tỉ lệ GC và năng lượng gập cuộn của mRNA.

3.2. Tối ưu gen LTB của *E. coli*

Trình tự của LTB được nhập vào phần mềm COOL và tối ưu bằng cách tối ưu chỉ số CAI, tối ưu độ thích nghi cặp bộ ba và tối ưu tỉ lệ GC. Tại thời điểm chúng tôi tối ưu trình tự LTB phần mềm COOL chưa có tính năng tối ưu cấu trúc thứ cấp mRNA nên chúng tôi sử dụng phần mềm Visual Gene Developer để tính toán năng lượng gập cuộn cấu trúc thứ cấp mRNA để chọn ra gen tối ưu. Kết quả chúng tôi đã chọn một

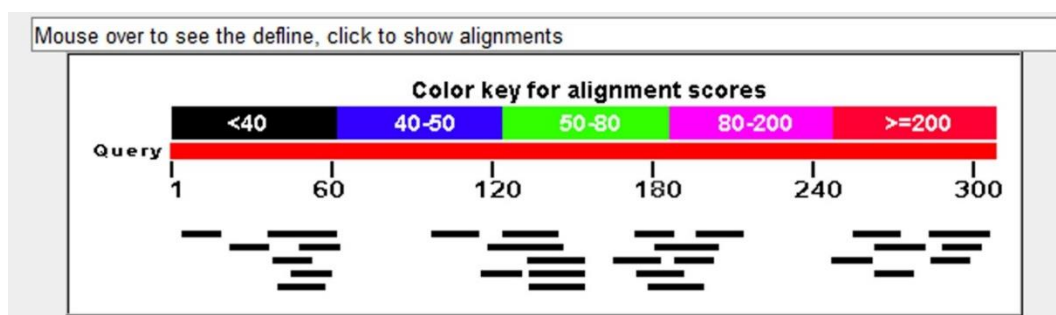
Thiết kế tối ưu gen để biểu hiện dị loài sử dụng hệ thống thương mại bằng một số phần mềm tối ưu miễn phí

gen để tối ưu. Trình tự gen này được cung cấp ở phần phụ chú. Kết quả tìm kiếm gen này trên CSDL mRNA của *S. cerevisiae* được chọn là tiêu chí phụ để chọn gen từ tập hợp các gen được tối ưu, và kết quả này được trình bày ở Hình 4.

3.3. Tối ưu gen CD3 và P1 của Cocksackievirus A16

Để tối ưu các gen này chúng tôi chọn ra 14 gen mã hóa cho các protein có độ phong phú cao nhất ở chủng nấm men *Pichia pastoris* (tên mới *Komagataella phaffii* GS115) dựa trên thông tin tìm thấy trên CSDL PaxDb [25] và danh mục các protein của *Pichia pastoris* sắp theo thứ tự hàm lượng cao đến thấp [28] (Bảng 2). Các gen này được sắp giống cột để tìm trình tự Kozak (Hình 5) và dùng để làm gen mô hình để phần mềm mô phỏng theo trong quá trình tối ưu gen. Kết quả chúng tôi chọn được một gen 3CD và một gen P1 để tổng hợp cho các thí nghiệm tiếp theo. Các trình tự này được cung cấp ở phần phụ lục.

Do các gen này khá lớn (1961 bp đối với 3CD và 2609 bp đối với P1), chúng được chia nhỏ ra thành các gen có kích thước khoảng 350 bp để đặt hàng tổng hợp dưới dạng geneblock. Trình tự các gen này được trình bày ở phần phụ lục.



Hình 4. Kết quả BLAST gen LTB tối ưu trên CSDL mRNA của *S. cerevisiae*. Gen có nhiều vùng giống nhau với các trình tự trên CSDL RNA của *S. cerevisiae* nhất được kiểm tra bằng mắt và chọn để tổng hợp.



Hình 5. Trình tự Kozak của gen mã hóa protein ở nấm men *P. pastoris* xây dựng từ các gen tự nhiên có mức độ biểu hiện cao ở *P. pastoris* bằng phần mềm Weblogo.

4. KẾT LUẬN

Tối ưu và tổng hợp gen để biểu hiện dị loài là xu hướng chủ đạo trong công nghệ DNA tái tổ hợp. Hiện nay các tiêu chí tối ưu và sự hiện diện của nhiều phần mềm miễn phí có khả năng tối ưu gen theo các tiêu chí của người dùng để biểu hiện ở nhiều vật chủ phổ biến. Tối ưu gen được chứng minh là giúp cải thiện kết quả biểu hiện, đồng thời tối ưu gen cũng cho phép thu nhận các gen có nguồn gốc từ các vi sinh vật gây bệnh.

Bảng 2. Danh sách các gen được dùng để tối ưu 3CD và P1 của Cocksackievirus A16 [28]

Xếp hạng độ phong phú	Mã truy cập	Tên protein
1	254568470	Glyceraldehyde-3-phosphate dehydrogenase, isozyme 3
2	254567507	Nhân tố kéo dài phiên mã EF-1 alpha
3	254570367	Enolase I
4	254568544	Alcohol dehydrogenase isozyme III ti thể
5	254573764	ATPase tham gia gấp cuộn protein và vận chuyển nhân
6	254566601	3-phosphoglycerate kinase
7	254566257	Protein sốc nhiệt Hsp90
8	254567798	Protein giả định
9	254570575	Pyruvate decarboxylase chính trong số các isozyme
10	254571387	Tiểu phần alpha của ATP synthase F1F0 ti thể
11	254571425	Thành phần protein của tiểu phần nhỏ của ribosome
12	254568572	Protein dung hợp, giống hệt với Rpl40Bp
13	254571679	Protein dung hợp được cắt thành protein thành phần tiểu phần nhỏ của ribosome và ubiquitin
14	254569858	Tiểu phần beta của ATP synthase F1F0 ti thể

Trong nghiên cứu này chúng tôi trình bày một số kết quả tối ưu một số gen kháng nguyên từ virus (Dengue tuýp 1, Cocksackievirus A16 và *E. coli*) bằng hai phần mềm tối ưu gen phổ biến là Visual Gene Developer và COOL. Do các phần mềm sẽ cho ra nhiều gen tối ưu ứng viên, việc lựa chọn một số gen để kiểm tra đòi hỏi người dùng phải có một số kinh nghiệm và tri thức riêng. Ở đây chúng tôi trình bày một số kinh nghiệm như vậy trong việc lựa chọn một số gen vi sinh vật tối ưu. Các tiêu chí lựa chọn gồm: mức độ giống với các gen tự nhiên trên toàn bộ chiều dài, hoặc chọn ngẫu nhiên ba gen để kiểm tra, hoặc trong trường hợp bất khả kháng (khi chiều dài gen quá lớn dẫn đến giá thành tổng hợp gen quá lớn) chọn một gen có các thông số tốt nhất có thể.

LỜI CẢM ƠN

Nghiên cứu này được thực hiện dưới sự tài trợ của Đề tài ĐHH mã số ĐHH2016-01-87. Tác giả tuyên bố không có mâu thuẫn gì về quyền lợi.

TÀI LIỆU THAM KHẢO

- [1]. Gould N., Hendy O., và Papamichail D. (2014). Computational Tools and Algorithms for Designing Customized Synthetic Genes. *Frontiers in Bioengineering and Biotechnology*, 2.
- [2]. Parret A.H., Besir H., và Meijers R. (2016). Critical reflections on synthetic gene design for recombinant protein expression. *Current Opinion in Structural Biology*, 38, 155–162.
- [3]. Webster G.R., Teh A.Y.-H., và Ma J.K.-C. (2017). Synthetic gene design-The rationale for codon optimization and implications for molecular pharming in plants: Synthetic Gene Design. *Biotechnology and Bioengineering*, 114(3), 492–502.
- [4]. Lodish H.F., Berk A., A. Kaiser C. và cộng sự. (2013). Transcriptional control of gene expression. *Molecular Cell Biology*. 7th, W. H. Freeman, 279–388.
- [5]. Lodish H.F., Berk A., A. Kaiser C. và cộng sự. (2013). Post-transcriptional Gene Control. *Molecular Cell Biology*. 7th, W. H. Freeman, 345–391.
- [6]. Elena C., Ravasi P., Castelli M.E. và cộng sự. (2014). Expression of codon optimized genes in microbial systems: current industrial applications and perspectives. *Frontiers in Microbiology*, 5.
- [7]. Gustafsson C., Minshull J., Govindarajan S. và cộng sự. (2012). Engineering genes for predictable protein expression. *Protein Expression and Purification*, 83(1), 37–46.
- [8]. Papamichail D., Liu H., Machado V. và cộng sự. (2018). Codon Context Optimization in Synthetic Gene Design. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(2), 452–459.
- [9]. Bai C., Wang X., Zhang J. và cộng sự. (2014). Optimisation of the mRNA secondary structure to improve the expression of interleukin-24 (IL-24) in *Escherichia coli*. *Biotechnology Letters*, 36(8), 1711–1716.
- [10]. Griswold K.E., Mahmood N.A., Iverson B.L. và cộng sự. (2003). Effects of codon usage versus putative 5'-mRNA structure on the expression of *Fusarium solani* cutinase in the *Escherichia coli* cytoplasm. *Protein Expression and Purification*, 27(1), 134–142.
- [11]. Gaspar P., Moura G., Santos M.A.S. và cộng sự. (2013). mRNA secondary structure optimization using a correlated stem-loop prediction. *Nucleic Acids Research*, 41(6), e73–e73.
- [12]. Kozak M. (1989). Circumstances and mechanisms of inhibition of translation by secondary structure in eucaryotic mRNAs. *Molecular and Cellular Biology*, 9(11), 5134–5142.
- [13]. Kudla G., Lipinski L., Caffin F. và cộng sự. (2006). High Guanine and Cytosine Content Increases mRNA Levels in Mammalian Cells. *PLoS Biology*, 4(6), e180.
- [14]. Brown C.Y., Lagnado C.A., và Goodall G.J. (1996). A cytokine mRNA-destabilizing element that is structurally and functionally distinct from A+U-rich elements. *Proceedings of the National Academy of Sciences*, 93(24), 13721–13725.
- [15]. Putland R.A., Sassinis T.A., Harvey J.S. và cộng sự. (2002). RNA Destabilization by the Granulocyte Colony-Stimulating Factor Stem-Loop Destabilizing Element Involves a

- Single Stem-Loop That Promotes Deadenylation. *Molecular and Cellular Biology*, 22(6), 1664–1673.
- [16]. Welch M., Govindarajan S., Ness J.E. và cộng sự. (2009). Design Parameters to Control Synthetic Gene Expression in *Escherichia coli*. *PLoS ONE*, 4(9), e7002.
- [17]. Gorochofski T.E., Ignatova Z., Bovenberg R.A.L. và cộng sự. (2015). Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate. *Nucleic Acids Research*, 43(6), 3022–3032.
- [18]. Chen Y., Maguire T., Hileman R.E. và cộng sự. (1997). Dengue virus infectivity depends on envelope protein binding to target cell heparan sulfate. *Nature Medicine*, 3(8), 866–871.
- [19]. Gong M., Zhu H., Zhou J. và cộng sự. (2014). Cryo-Electron Microscopy Study of Insect Cell-Expressed Enterovirus 71 and Coxsackievirus A16 Virus-Like Particles Provides a Structural Basis for Vaccine Development. *Journal of Virology*, 88(11), 6444–6452.
- [20]. Sixma T.K., Pronk S.E., Kalk K.H. và cộng sự. (1991). Crystal structure of a cholera toxin-related heat-labile enterotoxin from *E. coli*. *Nature*, 351(6325), 371–377.
- [21]. Kumar S., Stecher G., và Tamura K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*, 33(7), 1870–1874.
- [22]. Crooks G.E. (2004). WebLogo: A Sequence Logo Generator. *Genome Research*, 14(6), 1188–1190.
- [23]. Jung S.-K. và McDonald K. (2011). Visual gene developer: a fully programmable bioinformatics software for synthetic gene optimization. *BMC Bioinformatics*, 12(1), 340.
- [24]. Chin J.X., Chung B.K.-S., và Lee D.-Y. (2014). Codon Optimization OnLine (COOL): a web-based multi-objective optimization platform for synthetic gene design. *Bioinformatics*, 30(15), 2210–2212.
- [25]. Wang M., Weiss M., Simonovic M. và cộng sự. (2012). PaxDb, a Database of Protein Abundance Averages Across All Three Domains of Life. *Molecular & Cellular Proteomics*, 11(8), 492–500.
- [26]. Hoover D.M. (2002). DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Research*, 30(10), 43e – 43.
- [27]. Moura G., Pinheiro M., Arrais J. và cộng sự. (2007). Large Scale Comparative Codon-Pair Context Analysis Unveils General Rules that Fine-Tune Evolution of mRNA Primary Structure. *PLoS ONE*, 2(9), e847.
- [28]. Renuse S., Madugundu A.K., Kumar P. và cộng sự. (2014). Proteomic analysis and genome annotation of *Pichia pastoris*, a recombinant protein expression host. *PROTEOMICS*, 14(23–24), 2769–2779.

GENE REDESIGN FOR HETEROLOGOUS EXPRESSION IN COMMERCIAL SYSTEMS USING FREE GENE OPTIMIZATION PROGRAMS

Ha Thuc Duc Tung, Doan Trong Bich, Nguyen Ngoc Luong*

Faculty of Biology, University of Sciences, Hue University

* Email: daigoro77@gmail.com

ABSTRACT

Gene optimization for heterologous expression is the foundation in modern recombinant DNA technology for biopharmaceutical productions. Several studies have examined and established optimization criteria. The most popular criteria include codon bias, codon context bias and mRNA secondary structure. To simultaneously optimize several criteria, software has been developed to automate this process. However, to obtain genes for subsequent work, human intervention is necessary, particularly in selecting a few among several candidates, based on users' experience, insights or own criteria. In this study, we overviewd some free gene optimization programs for gene optimization. We described in details procedures for using these programs and our own experience in using these programs to optimize some proteins that have applications in biopharmaceutical biotechnology industries, particularly in common expression hosts such as Escherichia coli and yeasts.

Keywords: E. coli, Gene optimization, heterologous expression,, optimization criteria, yeasts.



Hà Thúc Đức Tùng sinh ngày 26/05/1993 tại Thừa Thiên Huế. Hiện ông là sinh viên ngành Công nghệ Sinh học tại Trường Đại học Khoa học, Đại học Huế.

Đoàn Trọng Bích sinh ngày 24/02/1992 tại Thừa Thiên Huế. Năm 2015, ông tốt nghiệp kỹ sư ngành Công nghệ Sinh học tại Trường Đại học Khoa học, Đại học Huế. Hiện ông là nhân viên kiểm định chất lượng tại Công ty Carlsberg Việt Nam, chi nhánh nhà máy bia Huda Thừa Thiên Huế.



Nguyễn Ngọc Lương sinh ngày 28/03/1977 tại Hà Tĩnh. Năm 2002 ông tốt nghiệp đại học ngành Giải phẫu người tại Đại học Melbourne, năm 2007 tốt nghiệp Thạc sĩ ngành Sinh học tại Đại học Huế và năm 2014 tốt nghiệp Tiến sĩ tại Đại học Quốc gia Chonbuk. Hiện ông là giảng viên Khoa Sinh học, Trường Đại học Khoa học, Đại học Huế.

Lĩnh vực nghiên cứu: biểu hiện gen dị loài.

Thiết kế tối ưu gen để biểu hiện dị loài sử dụng hệ thống thương mại bằng một số phần mềm tối ưu miễn phí

PHỤ CHÚ

Trình tự CD3 tối ưu:

CCAAAATGGCTGGACCATCACTTGACTTTGCACTTTCCTTTCGTCGAAACATTCGTCAAGTCCAAACGGACCAGG
GTCAC TTCACCATGTTGGGTGTCCGGGACAGGCTTGCTATTCTCCAAGACTCCCAACCAGGCAAAAACATTTGGG
TGGAGCACAAGCTCATCAACGTTCTGGATGCCGTCGAGCTGGTGGACGAACAAGGGGTCAACCTGGAGCTTACACTT
GTCACGTTGGACACCAACGAGAAGTTCAGAGATGTCACCAAGTTTATCCAGAAACCATCACTGGTGCTCAGATGC
TACGCTCATTATCAACACCGAACACATGCCCTCCATGTTTCGTCGCCAGTTGGGGACGTCGTTCAATATGGGTTCCCTCAA
TCTCTTGAAAGCCAACCCACCGTACCATGATGTACAACCTCCCGACCAAGGCCGACAATGTGGTGGTGTGTTAG
CTCTGTCGCAAGATTATTGGTATTCACATTGGTGGCAACGGTCGTCAGGGTTTCTGTGCTGGTCTCAAGAGAGGTTA
CTTCGCTCAGAGCAAGGCGAAATACAATGGATGAAGTCCAACAAGGAAACCGGGAGACTCAACATCAACGGTCCA
ACCAGAACCAAATGGAGCCAAGCGCTTTCTACGATGTTTCGAGGGTTCCAAGGAACCAGCTGTTCTCACTTCCAAG
GACCCACGCCTGGAAGTCGACTTCGAACAGGCCCTATTCTCCAATACGTTGGGAACACCTTGCACGAGCCGGACGA
GTATGTCAC TCAAGCTGCATTGCACTACGCAAACTCAAGCAATTGGACATCAACATCAACAAGATGTCATGG
AGGAGGCTTGTATGGTACGGAGTACTTGAAGCTATCAATCTTACACCTCAGCTGGTTACCCATACTCTGCTTTGG
GAGTCAAGAAGTCAGATATCTTGATCCAACCACCAGAGACTACCAAGATGAAGTTCTACATGGACAAGTACGGC
TTGGACCTGCCCTACTCTACCTACGTC AAGGACGAGTTGCGATCGCTGGACAAGATCAAGAAAGGAAAATCCCGCT
CATCGAAGCCTCTTCACTCAACGACTCCGTTTACCTACGAATGACCTTTGGACACTTGTACGAGACTTTCCACGCAAA
CCCTGGGACGGTTACAGGATCTGCCGTGGGGTGAATCCCGATGCTTCTGGTGAAGTTGCCAATTCTGCTTCCAGG
TTCCTTATTTGCCCTCGACTACTCGGGTTACGACGCCTCCTTGTCTCCAGTCTGGTTCCGAGCCTTGGAGGTTGTATTA
CGGGAGATCGGCTACTCCGAAGAAGCCGTTTCTCTTATCGAAGGTATCAACCACACTACCACGTGTACCGAAACAG
AACTTACTGTGTTTTGGGAGGTATGCCATCCGGTTGTCTGGGACCTCCATCTTCAACTCTATGATCAACAACATTATC
ATCAGAACTTTGCTTATCAAGACTTTCAAGGGTATCGATTTGGACGAGCTCAACATGGTTGCTTACGGCGACGACGTT
TTGGCCTCTACCCATTTCCAATCGACTGTTCCGAGCTTGCCAAGACCGGAAAGGAATACGGTCTTACCATGACTCCA
GCGGACAAGTCCCATGTTTCAACGGGGTTACATGGGAGAACGCTACCTTTCTCAAGAGAGGATTCTCCAGACCA
CCAATTTCCATTCTTATTCACCAACCATGCCAATGCGAGAGATCCACGAGTCTATCAGATGGACCAAGGATGCCA
GAAACACTCAGGATCAGTCTCAGATCTCTGTGTTTGTGGCCTGGCACAACGGAAAGGAGGAATACGAGAAATTCGTC
TCCACCATCAGATCTGTCCAGTTGAAAGGCTCTGGCTATTCCAAACTTCGAGAACCTCAGAAGAACTGGTTGGA
GTTGTTCTTAA

Trình tự các geneblock cần tổng hợp để tổng hợp gen 3CD vật lý:

Gene 1:

GAATTCACACAATGGCTGGACCATCACTTGACTTTGCACTTTCCTTTCGTCGAAACATTCGTCAAGTCCAAACGG
ACCAGGGTCACTTCACCATGTTGGGTGTCCGGGACAGGCTTGCTATTCTCCAAGACTCCCAACCAGGCAAAAACC
ATTTGGGTGGAGCACAAGCTCATCAACGTTCTGGATGCCGTCGAGCTGGTGGACGAACAAGGGGTCAACCTGGAGCT
TACACTGTACGTTGGACACCAACGAGAAGTTCAGAGATGTCACCAAGTTTATCCAGAAACCATCACTGGTGCTTC
AGATGCTACGCTCATTATCAACACCGAACACATG

Gene 2:

TCATTATCAACACCGAACACATGCCCTCCATGTTTCGTCCCAGTTGGGGACGTCGTTCAATATGGGTTCTCAATCTCTC
 TGGAAAGCCAACCCACCGTACCATGATGTACAACCTCCCAGCAAGGCCGACAATGTGGTGGTGTGTTACCTCTGT
 CGGCAAGATTATTGGTATTCACATTGGTGGCAACGGTCGTCAGGGTTTCTGTGCTGGTCTCAAGAGAGTTACTTCGC
 TTCAGAGCAAGGCGAAATACAATGGATGAAGTCCAACAAGGAAACCGGGAGACTCAACATCAACGGTCCAACCAGA
 ACCAAATTGGAGCCAAGCGCTTCTACGATGTTTCGAG

Gene 3:

CGCTTCTACGATGTTTTGAGGGTTCCAAGGAACCAGCTGTTCTCACTTCCAAGGACCCACGCCTGGAAGTCGACTT
 CGAACAGGCCCTATTCTCCAAATACGTTGGGAACACCTTGCACGAGCCGGACGAGTATGTCCTCAAGCTGCATTGC
 ACTACGCAAACCAACTCAAGCAATTGGACATCAACATCAACAAGATGTCCATGGAGGAGGCTTGTATGGTACGGAG
 TACTTGGAAAGCTATCAATCTTACACCTCAGCTGGTTACCCATACTCTGCTTTGGGAGTCAAGAAGTCAGATATCTTG
 GATCCAACCACAGAGACACTACCAAGATGAAGTTCTACA

Gene 4:

TGTGATGGTCTACTTCAAGATGTACCTGTTTCATGCCGAACCTGGACGGGATGAGATGGATGCAGTTCCTGCTCAACG
 CTAGCGACCTGTTCTAGTCTTTTCTTTTAGGGCGAGTAGCTTCGGAGAAGTGAGTTGTGAGGCAAATGGATGCTT
 ACTGGAAACCTGTGAACATGCTCTGAAAGGTGCGTTTGGGACCCTGCCAATGTCCTAGACGGCACCCCA
 CGTTAGGGCTACAGAAGACCAGCTTCAACGGTTAAGACGAAGGTCCAAGGAATAAACGGAAGCTGATGAGCCCAAT
 GCTGCGGAGGAACAGAGGTCAGACCAAGGCTCGGAACCTCCAACATAAT

Gene 5:

GCTCGGAACCTCCAACATAATGCCCTTAGCCGATGAGGCTTCTTCGGCAAAGAGAATAGCTTCCATAGTTGGTGTGA
 GTGGTGCACATGGCTTTGTCTTGAATGACACAAAACCTCCATACGGTAGGCCAACAAAGACCCTGGAGGTAGAAGTT
 GAGATACTAGTTGTTGTAATAGTAGTCTTGAAACGAATAGTTCTGAAAGTTCCCATAGCTAAACCTGCTCGAGTTGTA
 CCAACGAATGCCGCTGCTGCAAACCGGAGGATGGGTAAAGGGTTAGCTGACAAGCCTCGAACGGTCTGGCCTTTCC
 TTATGCCAGAATGGTACTGAGGTCGCCTGTTACGGGGTAC

Gene 6:

TCGCCTGTTACGGGTACAAAGTTGCCCAATGTACCCTCTTGCATGGAAAGAGTTCTCTCCTAAGGAGGGTCTGGT
 GGTTAAGGGTAAGGAATAAGTGGGTTGGTACGGTTACGCTCTCTAGGTGCTCAGATAGTCTACCTGGTTCTACGGTC
 TTTGTGAGTCTAGTGCAGTCTAGAGACACAAACAACCGGACCGTGTTCCTTTCTCTTATGCTCTTTAAGCAGAG
 GTGGTAGTCTAGACAAGGTCAACCTTTCCGAGACCGATAAGGTTTGAAGCTCTTGGAGTCTTCTTTGACCAACCTCAA
 CAAGATTCCATGG

Trình tự gen P1 tối ưu:

CCAAAATGGCTGGTCTCAAGTCTCTACGCAACGCTCTGGTCTCACGAGAAGTCCAATAGTGCATCAGAAGGTTCA
 ACGATCAACTACACAACCATCAATTAATAAGGATGCTACGCGCCAGCGCTGGACGTCAAGACATGTGCGCAAGA
 CCCAAAGAAATTCACGGACCCTGTCATGGATGTTATCCACGAGATGGCTCCGCCCTCAAGAGTCCAAGCGCGGAAG
 CCTGTGGTTATTACAGACAGAGTTGCACAACCTACCATCGGCAATTTACCATCACAACACAAGAGGCGGCAATATT

Thiết kế tối ưu gen để biểu hiện dị loài sử dụng hệ thống thương mại bằng một số phần mềm tối ưu miễn phí

GTTATCGCATATGGCGAATGGCCCGAGTACTGTCCAGACACGGATGCTACCGCAGTCGACAAAACCAACCAGACCGGA
TGTGTCTGTCAACCGTTCTTTACTTTGGACACCAAGTCTTGGGCGAAAGACTCGAAGGGATGGTATTGGAAGTTTCC
CGACGTTCTTACAGAGGTTGGTGTCTTCGGACAGAATGCCAATTCCACTATCTTTATCGCTCGGGATTTTGC GTTCAT
GTGCAGTGCATGCCTCGAAGTCCATCAGGGTGCTTTGCTGTGTGCGGTTTGGCAGAATATGTACTCGGGACCATT
GCCGGGGGTACGGGGAACGAAAACAGCCATCCCCCTTACGCCACGACCCAACCTGGTCAAGTTGGTGCCGTGCTCAC
GCATCCATATGTGCTGGATGCGGGTATCCCACTGTCACAGCTACCGTCTGTCCACATCAATGGATCAATCTCAGGAC
GAACAACTGCGCTACGATCATTGTACCCTACATGAATACCGTTCATTTCGATAGCGCTTTGAACCACTGTAACCTCGG
CCTGTTGGTTATCCCAGTTGTTCTTCTGACTTCAACACTGGAGCTACTTCTGAAATTCCTATCACTGTGACCATTGCT
CCAATGTGTGCTGAGTTTGCCGGTTTGCCTCAGGCTGTCAAGCAAGGTATTCCAAGTGAAGCTGAAAGCTGTAACAAAC
CAATTCCTTACCACCGATGACGGTGTTCGCCCCAATTTGCGCAGGTTTCCACCCAACCCACCAATCCACATCCCT
GGTGAGGTTCAACAACTGTTAGAGATCTGTAGAGTTGAGACCATTCTAGAAGTTAACAACCTGAAAACAAACGAAAC
TACTCCAATGCAAAGACTGTGTTTCCCAGTTTCTGTCCAATCCAAGACTGGCGAGCTTTGTGCCGCTTTCAGAGCTGA
CCCAGGTAGAGATGGTCCATGGCAGTCCACCATTTTGGGTCAGCTGTGTAGATACTATACTCAGTGGTCTGGTTCTCT
AGAGGTTACTTTCATGTTTCGCTGGTCTTTCATGGCCACTGGAAAAATGCTGATCGCTTACACTCCACCAGGTGGTAA
CGTCCCAGCTGACAGAATTACCCTATGTTGGGAACACACGTTATCTGGGACTTCGGATTGCAGTCTCCGTCACCCT
GGTTGTTCCATGGATTCTAATACCCACTACAGAGCTCACGCTAGAGCTGGTACTTCGACTACTATACTACTGGTATT
ATCACCATCTGGTATCAAACCAACTACGTTGTCCCTATTGGTGCCCCAACCCGCTTACATTGTAGCCTTGGCTGCT
GCTCAAGATAACTTCACTATGAAGTTGTGTAAGGATACTGAAGACATCGAACAAACTGCTGACATTCAAGGTGATCC
AATTGCCGACATGATTGACCAAACTGTCAACAACCAAGTCAACAGATCCTTGACGGCTTTGCAAGTCTTGCCACTGCTC
TGCTAACACCGAGGCTTCGTTCCACAGATTGGGTAAGTGGAGTTGCCAGCTTTGCAAGCTGCTGAAAACCGGTGCTTC
CTCCAACGCCTCCGATAAGAACCTGATTGAGACTCGTTGTGTTTTGAACCACCACTCAACTCAAGAGACAGCTATTGG
TAACTTCTTCTCCAGAGCTGGACTTGTCTTCTATTATCACCATGCCAACCATGGGAACTCAGAACACTGATGGCTATGC
AAACTGGGACATTGACTTGATGGGATACGCTCAGCTGAGAAGAAAGTGTGAGCTGTTCACTTACATGAGGTTTGATG
CCGAGTTCACCTTCGTTGTTGCCAAACCTAACGGTGAGTTAGTTCCACAATTGCTGCAATACATGTACGTCCACCTG
GTGCTCCTAAGCAACCTTAGAGACTCTTTCGCTGGCAAACCGCTACTAATCCATCTGTCTTCGTCGAAGATGACCG
ACCCACCAGCTCAGGTCTCAGTCCATTCATGTCCCAGCCTCTGCTTATCAATGGTCTACGATGGTTACCTACCTT
CGGGGAGCACTTGCAAGCCGACGATTGGACTACGGTCAATGTCCAAACACATGATGGGCACTTCTCCATTAGAA
CCGTGCGTACTAAGAAATCCCCACACTCTATCACCTTGAGAGTCTACATGAGAATCAAGCACGTTAGAGCCTGGATC
CCAAGACCATTGAGAAACAGCCATACTTGTCAAGACCAACCCAACTACAAGGGTAACGACATCAAGTGTACTTC
TACTTCCAGAGACAAGATCACACCTTGTAA

Mã mở đầu (ATG) và mã kết thúc của các gen được bôi đậm và gạch dưới (ghi chú của tác giả).