Tạp chí Phát triển Khoa học và Công nghệ – Khoa học Tự nhiên, 5(2):1068-1077

Open Access Full Text Article

Bài Nghiên cứu

Nghiên cứu dự đoán gene biểu hiện cao cho *Escherichia coli* dựa trên dữ liệu mRNA microarray

Võ Trí Nam[1,2,3], Phạm Trung Nghĩa[1,3], Trương Hà Minh Nhật[1,3], Trần Linh Thước[3,4], Nguyễn Đức Hoàng[1,3,4,*]

Use your smartphone to scan this QR code and download this article

TÓM TẮT

Các gene biểu hiện cao (Highly expressed genes – HEG) là những gene có sẵn trong sinh vật, mang những codon ưa thích đối với hệ thống biểu hiện. Việc xác định được các gene biểu hiện cao giúp tìm ra các codon ưa thích và sử dụng trong tối ưu hóa gene nhằm biểu hiện protein mục tiêu với mức độ mong muốn. Hiện nay, HEG-DB là cơ sở dữ liệu (CSDL) duy nhất lưu trữ dữ liệu gene biểu hiện cao của nhiều chủng vi sinh vật, tuy nhiên dữ liệu hiện không còn được cập nhật và duy trì. Vì vậy chúng tôi tiến hành dự đoán các gene biểu hiện cao ở chủng *E. coli* K-12 MG1655 dựa trên các bộ tham chiếu là gene mã hóa protein ribosome được sử dụng phổ biến hiện nay và những gene có độ phiên mã cao từ dữ liệu microarray do chúng tôi đề xuất. Kết quả dự đoán được phân tích bằng cách so sánh giữa các bộ tham chiếu trên cũng như so sánh với gene biểu hiện cao thu nhận từ CSDL HEG-DB. Kết quả cho thấy bộ tham chiếu gồm 69 gene mã hóa protein ribosome và 100-mRNA cho kết quả hoàn toàn trùng khớp và dự đoán được gene biểu hiện cao nhiều hơn và có độ tin cậy cao hơn so với dữ liệu từ CSDL HEG-DB thể hiện qua các gene dự đoán được có giá trị CAI cao hơn và số lượng gene tham gia vào các con đường chuyển hóa trong tế bào, đặc biệt là các con đường chuyển hóa quan trọng đều cao hơn. Nghiên cứu này đề xuất có thể sử dụng bộ tham chiếu từ dữ liệu microarray của *E. coli* thay cho bộ tham chiếu protein ribosome.

Từ khoá: gene biểu hiện cao, Escherichia coli, protein ribosome, mRNA microarray, CAI

[1]*Trung tâm Khoa học và Công nghệ sinh học, Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM, Việt Nam*

[2]*Phòng Thí nghiệm Công nghệ sinh học phân tử, Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM, Việt Nam*

[3]*Đại học quốc gia thành phố Hồ Chí Minh, Việt Nam*

[4]*Khoa Sinh học – Công nghệ Sinh học, Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM, Việt Nam*

Liên hệ

Nguyễn Đức Hoàng, Trung tâm Khoa học và Công nghệ sinh học, Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM, Việt Nam

Đại học quốc gia thành phố Hồ Chí Minh, Việt Nam

Khoa Sinh học – Công nghệ Sinh học, Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM, Việt Nam

Email: ndhoang@hcmus.edu.vn

Lịch sử

- Ngày nhận: 25-8-2020
- Ngày chấp nhận: 22-3-2021
- Ngày đăng: 30-4-2021

DOI : 10.32508/stdjns.v5i2.945

Check for updates

MỞ ĐẦU

Trong những năm gần đây, kỹ thuật sản xuất protein tái tổ hợp ngày càng phát triển và ứng dụng rộng rãi trong nhiều lĩnh vực như y tế, công nghiệp, nông nghiệp và các nghiên cứu khoa học khác. Các protein tái tổ hợp được dùng để tổng hợp vắc xin, hỗ trợ điều trị bệnh (như insulin), sản xuất các enzyme công nghiệp. Bên cạnh đó, protein tái tổ hợp còn được dùng để cải thiện giống cây trồng, tạo ra các loài thực vật và động vật chuyển gene… Tuy nhiên, các gene thu nhận từ sinh vật ban đầu khi đưa vào hệ thống biểu hiện của vật chủ thường không tương thích, dẫn đến giảm khả năng biểu hiện protein mục tiêu. Thấy được hạn chế trên trong việc sản xuất protein tái tổ hợp, nhiều nghiên cứu về thiết kế gene đã được đưa ra nhằm tăng mức độ biểu hiện của các protein tái tổ hợp và thay thế cho các gene tự nhiên biểu hiện thấp ở sinh vật[1–3]. Một trong những nguyên lý để thiết kế lại gene là thay đổi một số codon trên trình tự gene sẵn có bằng các codon đồng nghĩa để cải thiện các đặc trưng về trình tự như chỉ số thích nghi codon, thành phần GC, trình tự lặp lại, khả năng hình thành cấu trúc bậc 2 của mRNA. Các đặc trưng này đã được chứng minh có ảnh hưởng đến độ biểu hiện của protein mục tiêu[3–5].

Những gene biểu hiện cao (HEG- Highly expressed genes) là những gene có sẵn, mang các đặc trưng trình tự phù hợp để có thể biểu hiện ở mức cao trong sinh vật. Vì vậy việc dự đoán gene biểu hiện cao là một bước quan trọng để từ đó, tìm ra các đặc trưng phù hợp cho từng hệ thống biểu hiện như bộ các codon ưa thích, chỉ số phần trăm GC, độ dài trình tự lặp lại, năng lượng tự do của các cấu trúc bậc hai của mRNA hay các chỉ số khác ảnh hưởng đến độ biểu hiện của protein mục tiêu. Các kết quả này sẽ được ứng dụng vào quá trình thiết kế gene để gia tăng độ biểu hiện trong từng hệ thống biểu hiện cụ thể[6,7].

Năm 2007, Pere Puigbò và cộng sự đã thiết lập CSDL HEG-DB, chứa thông tin về các gene biểu hiện cao của gần 200 chủng vi sinh vật[6]. Nhưng đến thời điểm hiện tại, CSDL HEG-DB vẫn chưa được cập nhập tính từ lúc thành lập. Các thông tin trong dữ liệu đã cũ, không còn chính xác để làm cơ sở cho các nghiên cứu khác. Hiện tại dữ liệu gene biểu hiện cao của các chủng vi sinh vật trên CSDL HEG-DB không được cập nhật và duy trì tốt, dữ liệu gene biểu hiện cao của chủng *Bacillus subtilis* 168 đã không truy cập được, các thông tin hiển thị đã không còn đầy đủ. Bên cạnh đó, phương pháp dự đoán gene biểu hiện cao của CSDL HEG-DB vẫn chưa được nêu rõ, đặc biệt là các thông số sử dụng trong quá trình dự đoán. Một

Trích dẫn bài báo này: Nam V T, Nghĩa P T, Nhật T H M, Thước T L, Hoàng N D. **Nghiên cứu dự đoán gene biểu hiện cao cho *Escherichia coli* dựa trên dữ liệu mRNA microarray**. *Sci. Tech. Dev. J. - Nat. Sci.*; 5(2):1068-1077.

nghiên cứu khác được đưa ra bởi nhóm tác giả Kim Chi và cộng sự vào năm 2016 sử dụng thuật toán phân cụm PAM và CLARA để dự đoán gene biểu hiện cao cho *B. subtilis*[8]. Tuy nhiên, kết quả dự đoán chỉ được đánh giá thông qua các chỉ số về thuật toán chứ chưa được đánh giá về mặt sinh học nên cần thêm các minh chứng và phản hồi từ các nghiên cứu khác. Từ thực tế đó, chúng tôi đặt ra mục tiêu nghiên cứu phương pháp dự đoán gene biểu hiện cao được đề xuất bởi Puigbò và cộng sự[9] một cách rõ hơn về các thông số cũng như đưa ra các chỉ tiêu đánh giá cho kết quả dự đoán để có thể chủ động trong việc dự đoán, từ đó xây dựng một CSDL mới để lưu trữ các thông tin về gene biểu hiện cao ở vi sinh vật.

VẬT LIỆU VÀ PHƯƠNG PHÁP

Thu nhận và xử lý dữ liệu

Dữ liệu trình tự bộ gene hoàn chỉnh của chủng vi sinh vật *E. coli* K-12 MG1655 được thu nhận từ NCBI với mã số [GeneBank: U00096.3]. Dữ liệu sau đó được lọc bỏ các gene giả (pseudogenes), các gene mã hóa RNA mà không được dịch mã (ncRNA), các gene mã hóa tRNA và các gene mã hóa rRNA. Tiếp theo tiến hành lọc bỏ các gene có trình tự không chẵn bộ ba và các gene không có mã mở đầu hoặc mã kết thúc.

Dữ liệu độ phiên mã gene của *E. coli* - K-12 được thu nhận từ CSDL E COLI EXPRESSION2[10]. Dữ liệu sau khi thu nhận về được tiến hành loại bỏ các dữ liệu của các chủng đột biến. Đối với các dữ liệu của các lần lặp lại trong cùng một thí nghiệm với cùng một điều kiện môi trường, tiến hành tính giá trị trung bình để đại diện cho các dữ liệu này. Sau đó, sử dụng ngôn ngữ R và gói Package preprocessCore để áp dụng phương pháp chuẩn hóa "Quantile normalization" cho tất cả các giá trị biểu hiện trung bình ở mỗi điều kiện vừa tính được và cuối cùng là tính trung bình cho tất cả các giá trị biểu hiện cho từng gene.

Dữ liệu gene biểu hiện cao của *E. coli* - K-12 được thu nhận từ CSDL HEG-DB. Dữ liệu này được dùng cho bước đánh giá kết quả dự đoán gene biểu hiện cao phía sau.

DỰ ĐOÁN GENE BIỂU HIỆN CAO

Quy trình dự đoán

Quá trình dự đoán gene biểu hiện cao được thực hiện bằng phương pháp được nêu trong nghiên cứu của Pere Puigbò và cộng sự năm 2007[9] có thay đổi về bộ tham chiếu và tiêu chí chọn kết quả, cụ thể theo quy trình sau: (i) Bước 1. Chọn các bộ tham chiếu độc lập nhau: (i.1) Gene mã hóa ribosomal protein dựa theo phương pháp được nêu ở bài báo của Puigbò và cộng sự, dựa trên thông tin trong bộ gene thu nhận từ NCBI tiến hành thu nhận các gene mã hóa cho protein ribosome để dùng làm bộ tham chiếu ban đầu; (i.2) Các gene có giá trị biểu hiện mRNA trung bình cao nhất (từ dữ liệu microarray). Dựa trên dữ liệu microarray thu nhận từ E COLI EXPRESION2, chọn ra lần lượt 100, 200 và 300 gene có giá trị biểu hiện trung bình cao nhất làm bộ tham chiếu ban đầu cho quá trình dự đoán; (ii) Bước 2. Lần lượt dựa trên các bộ tham chiếu ban đầu ở bước 1, tiến hành tính giá trị w_i và từ đó tính giá trị CAI cho từng gene của toàn bộ gene theo công thức:

$$w_i = \frac{f[i]}{max_{f[j]}}$$

$$CAI = \left[\prod_{i=1}^{L} w_i\right]^{\frac{1}{L}}$$

Trong đó: i, j là các codon đồng nghĩa, cùng mã hóa cho một amino acid, f[i] là tần số của codon i, f[j] là tần số của codon có tần số cao nhất, L là chiều dài của gene (đơn vị là codon); (iii) Bước 3. So sánh các ngưỡng giá trị CAI, chọn ra những gene có giá trị CAI cao hơn ngưỡng làm bộ tham chiếu mới và quay lại bước 2. Giá trị ngưỡng CAI khảo sát được chọn dựa trên phân tích giá trị CAI của các gene HEG thu nhận từ CSDL HEG-DB. Quá trình lặp lại được thực hiện bằng một thuật toán viết bằng ngôn ngữ lập trình Python cho đến khi bộ tham chiếu thu được ở lần cuối cùng giống với lần ngay trước đó. Bộ tham chiếu ở lần cuối này cũng là các gene biểu hiện cao đã dự đoán được.

Đánh giá bộ tham chiếu

Kết quả từ 4 bộ tham chiếu, gene mã hóa cho protein ribosome và các gene có mức biểu hiện cao nhất từ dữ liệu microarray [100 gene, 200 gene và 300 gene] được so sánh với nhau dựa trên các tiêu chí bao gồm: Số lượng gene biểu hiện cao, số lượng gene mã hóa protein ribosome có trong các gene biểu hiện cao, tỉ lệ giữa gene mã hóa protein ribosome và các gene biểu hiện cao dự đoán được, khoảng giá trị CAI thấp nhất đến cao nhất. Đồng thời, việc so sánh đặc trưng codon của hai bộ tham chiếu protein ribosome và 100 gene từ dữ liệu microarray được tiến hành bằng cách sử dụng giá trị w_i của hai bộ tham chiếu để vẽ biểu đồ.

So sánh kết quả dự đoán với dữ liệu từ HEG-DB

So sánh kết quả các gene biểu hiện cao dự đoán được với các gene biểu hiện cao thu nhận từ CSDL HEG-DB. Nội dung so sánh bao gồm: (i) Các thông số trong quá trình dự đoán, bao gồm: khoảng giá trị CAI của nhóm gene biểu hiện cao, số lượng gene biểu hiện, số lượng gene mã hóa cho protein ribosome và tỉ lệ giữa gene mã hóa protein ribosome và gene biểu

hiện cao; (ii) Độ phiên mã mRNA: dùng Excel để vẽ biểu đồ Boxplot thể hiện độ phân bố độ biểu hiện mRNA của các nhóm, 4021 gene trùng khớp với dữ liệu microarray, gene biểu hiện cao dự đoán được, gene biểu hiện cao từ CSDL HEG-DB và gene mã hóa protein ribosome; (iii) Số lượng gene tham gia vào các con đường chuyển hóa: thông tin dữ liệu các con đường chuyển hóa được lấy từ CSDL DAVID [11,12]. Dựa vào tên của các gene biểu hiện cao để tìm các con đường chuyển hóa mà các gene này tham gia. Sau đó thu nhận dữ liệu về số lượng các con đường chuyển hóa mà nhóm gene biểu hiện cao tham gia và số lượng gene có ở mỗi con đường. Cuối cùng so sánh số lượng gene ở cả hai nhóm gene biểu hiện cao tham gia vào 08 con đường chuyển hóa quan trọng, gồm có RNA polymerase, Oxidative phosphorylation, Glycolysis, Pentose phosphate pathway, Pyrimidine metabolism, Purine metabolism, TCA cycle và Carbon metabolism.

KẾT QUẢ VÀ THẢO LUẬN

Thu nhận và xử lý dữ liệu

Dữ liệu bộ gene *E. coli* K-12 sau khi thu nhận từ NCBI được xử lý lọc bỏ các gene giả, gene mã hóa rRNA, tRNA và ncRNA do các gene này không được biểu hiện thành protein. Đồng thời các gene không có mã mở đầu hoặc mã kết thúc cũng được loại bỏ. Từ đó thu nhận được 4238 gene. Các gene này sẽ được dùng vào dự đoán gene biểu hiện cao.

Đối với dữ liệu độ phiên mã gene thu nhận từ CSDL E COLI EXPRESION2 bao gồm 213 bộ dữ liệu tương ứng với 71 môi trường biểu hiện (mỗi thí nghiệm được lặp lại 3 lần). Sau khi tính trung bình cho từng thí nghiệm cũng như lọc bỏ bộ dữ liệu của các chủng đột biến, các dữ liệu được chuẩn hóa bằng phương pháp Quantile để đảm bảo tính đồng nhất về khoảng giá trị giữa các thí nghiệm với nhau. Kết quả thu được 35 bộ dữ liệu biểu hiện với 7312 mẫu dò trong từng bộ. Trong số các mẫu dò, có các mẫu dò nằm trên cùng một gene cũng như các mẫu dò không nằm trên gene, từ đó tiến hành so khớp với dữ liệu 4238 tên gene của *E. coli* và kết quả thu được 4021 gene trùng nhau. Giá trị trung bình của 35 bộ dữ liệu của từng gene được sử dụng như độ biểu hiện đại diện của gene đó. Kết hợp lại chúng tôi thu nhận được 4021 gene cùng với độ biểu hiện của chúng. Dữ liệu này được sử dụng cho xác định bộ tham chiếu dựa trên độ biểu hiện mRNA và đánh giá kết quả dự đoán gene biểu hiện cao tiếp theo.

Bộ gene biểu hiện cao của *E. coli* K-12 được thu nhận từ CSDL HEG-DB bao gồm tên gene và giá trị CAI của chúng. Kết quả thu được 253 gene HEG của chủng *E. coli* K-12.

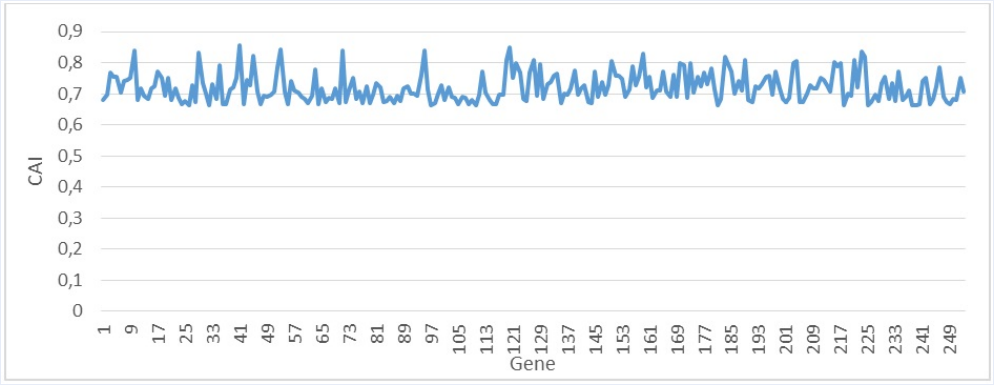
Dự đoán gene biểu hiện cao

Nghiên cứu sử dụng phương pháp của nhóm tác giả Puigbò và cộng sự công bố năm 2007 để dự đoán gene biểu hiện cao. Trong phương pháp của nhóm tác giả trên, các gene mã hóa cho protein ribosome được sử dụng làm bộ tham chiếu ban đầu để từ đó tìm kiếm các gene có xu hướng sử dụng codon tương tự thông qua tính toán giá trị CAI. Lý do tác giả lựa chọn gene mã hóa protein ribosome vì cho rằng đây là các gene được biểu hiện cao trong tế bào. Tuy nhiên, đôi khi việc xác định gene mã hóa protein ribosome gặp khó khăn do bộ gene chưa được chú thích hoặc chú thích chưa đầy đủ. Trong nghiên cứu này, bên cạnh bộ tham chiếu protein ribosome do tác giả đề xuất, chúng tôi sử dụng thêm các bộ tham chiếu dựa trên dữ liệu mRNA microarray thể hiện mức độ phiên mã của gene, một yếu tố góp một phần quan trọng vào lượng protein tạo ra. Cụ thể chúng tôi sử dụng bộ tham chiếu gồm 69 gene mã hóa protein ribosome từ thông tin của bộ gene *E. coli* đã thu nhận (gọi là bộ tham chiếu RP) cùng với 03 bộ tham chiếu chứa lần lượt 100, 200 và 300 gene có độ biểu hiện mRNA cao nhất (gọi là bộ tham chiếu 100-mRNA, 200-mRNA và 300-mRNA) cho quy trình dự đoán gene biểu hiện cao.

Phân tích giá trị CAI của các gene HEG thu nhận từ CSDL HEG-DB, khoảng giá trị nhận được dao động từ 0,662 đến 0,857 (Hình 1). Do đó, nghiên cứu tiến hành dự đoán gene biểu hiện cao với các ngưỡng giá trị CAI từ 0,650 đến 0,860.

Đánh giá bộ tham chiếu

Kết quả dự đoán gene biểu hiện cao từ 04 bộ tham chiếu được thể hiện trên Bảng 1. Đối với kết quả của từng bộ tham chiếu, chúng tôi so sánh để chọn ra kết quả có độ tin cậy cao nhất. Tiêu chí đầu tiên được xem xét là số lượng gene HEG được dự đoán trong kết quả. Số lượng gene biểu hiện cao ở một chủng vi khuẩn là bao nhiêu tùy thuộc vào ngưỡng xét, mức độ nào là biểu hiện cao, mức độ nào là biểu hiện thấp. Do đó, không có một quy tắc chung nào đưa ra để xác định chính xác số lượng gene biểu hiện cao ở *E. coli*. Trong nghiên cứu này, để chọn lựa kết quả dự đoán dựa trên số lượng gene biểu hiện cao trả ra, chúng tôi dựa trên số lượng gene biểu hiện cao của *E. coli* trong hai công bố: công bố trên HEG-DB là 5% và trong nghiên cứu của Karlin và cộng sự năm 2000 là 8% [7]. Từ đó, chúng tôi chọn các kết quả cho số lượng gene biểu hiện cao trong khoảng từ 4% đến 9% tổng số gene của bộ gene, tương ứng từ 170 đến 381 gene. Như vậy, bộ tham chiếu RP và bộ tham chiếu 100-mRNA cho các kết quả chấp nhận được với ngưỡng giá trị CAI giao động từ 0,688 đến 0,692 trong khi hai bộ



Hình 1: Đồ thị biểu diễn giá trị CAI của 253 gene biểu hiện cao từ CSDL HEG-DB

tham chiếu 200-mRNA và 300-mRNA không có kết quả được chấp nhận. Điều này có thể lý giải do khi chọn bộ tham chiếu có số lượng gene quá lớn (200 và 300 gene) làm cho đặc trưng về codon của các gene này không tập trung làm ảnh hưởng đến quá trình dự đoán nên không cho kết quả tốt. Tiếp theo, các kết quả được chọn tiếp dựa trên số lượng gene mã hóa ribosome vì đây là các gene có khả năng biểu hiện cao trong tế bào. Kết quả tương ứng ngưỡng CAI 0,688 và 0,689 với số gene mã hóa protein ribosome nhiều nhất là 44 gene. Cuối cùng, chúng tôi chọn kết quả dự đoán HEG với ngưỡng CAI là 0,689 do có tỉ lệ gene mã hóa protein ribosome cao hơn (14,2%).

Khi so sánh các gene biểu hiện cao thu được từ bộ tham chiếu RP và bộ tham chiếu 100-mRNA, kết quả cho thấy sự trùng khớp 100% của hai kết quả này. Để lý giải sự trùng khớp này, chúng tôi tiến hành so sánh giá trị w_i là đại lượng đặc trưng cho tần suất sử dụng các codon trong nhóm gene của hai bộ tham chiếu này (Hình 2). Kết quả so sánh ở Hình 2 cho thấy tần suất sử dụng codon của hai bộ tham chiếu có sự tương đồng rất cao. Trong đó, 16 trong 20 amino acid có trật tự tần suất các codon hoàn toàn giống nhau giữa hai bộ tham chiếu. Đối với 04 amino acid còn lại, codon có tần suất cao nhất vẫn giống nhau giữa hai bộ tham chiếu, sự khác biệt chỉ xảy ra ở các codon còn lại và chênh lệch giữa tần suất của các codon này trong mỗi bộ tham chiếu là không nhiều (Bảng 2). Chính sự tương đồng rất cao về tần suất sử dụng codon giữa hai bộ tham chiếu này dẫn đến sự trùng khớp trên kết quả dự đoán gene biểu hiện cao. Việc tương tự về tần suất sử dụng codon giữa hai bộ tham chiếu cho thấy có mối liên hệ chặt chẽ giữa mức độ biểu hiện mRNA và độ biểu hiện protein cũng như cho thấy tính khả thi trong việc sử dụng dữ liệu mRNA microarray thay cho các gene mã hóa protein ribosome để làm bộ tham chiếu ban đầu trong dự đoán gene biểu hiện cao.

So sánh kết quả dự đoán với dữ liệu từ HEG-DB

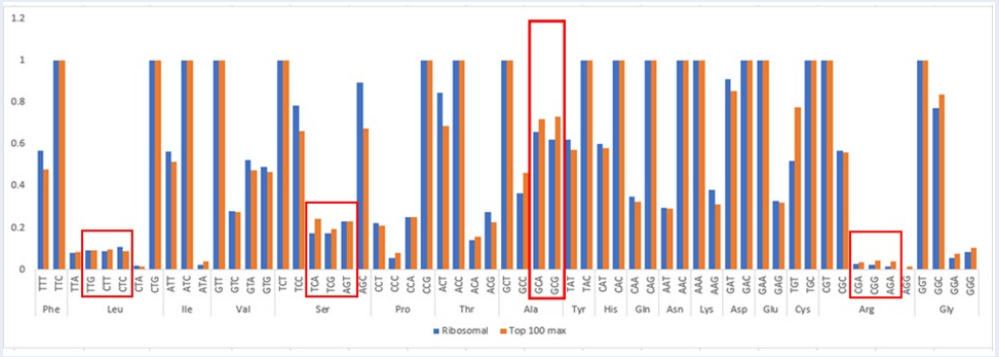
Kết quả dự đoán gene biểu hiện cao từ nghiên cứu dựa trên dữ liệu bộ gene *E. coli* được cập nhật vào năm 2018 với mã số U00096.3 [4238 gene với 69 gene mã hóa protein ribosome], khác với dữ liệu tác giả Puigbò sử dụng vào năm 2007 [4243 gene với 54 gene mã hóa protein ribosome]. Đồng thời ngưỡng giá trị CAI chúng tôi cũng khảo sát lại cùng với tiêu chí chọn kết quả cũng được đề xuất mới. Chính vì vậy, kết quả dự đoán thu được có sự khác biệt so với kết quả công bố trên CSDL HEG-DB. Do đó, chúng tôi tiến hành so sánh 310 gene biểu hiện cao dự đoán được với 253 gene biểu hiện cao thu nhận từ CSDL HEG-DB để đánh giá hiệu quả dự đoán trong nghiên cứu này. Các tiêu chí được dùng để so sánh bao gồm các thông số trong quá trình dự đoán gene biểu hiện cao, độ phiên mã của các gene biểu hiện cao [dựa vào dữ liệu microarray] cùng với số lượng gene tham gia vào các con đường chuyển hóa.

Các thông số trong quá trình dự đoán gene biểu hiện cao

Bảng 3 thể hiện kết quả so sánh các thông số cho thấy chúng tôi dự đoán được nhiều hơn 57 gene so với gene biểu hiện cao từ CSDL HEG-DB. Kết quả này cho thấy đã dự đoán được thêm các gene biểu hiện cao khác mà CSDL HEG-DB không có. Trong đó, số gene mã hóa protein ribosome ở hai dữ liệu là ngang nhau [44 gene]cho thấy độ tin cậy ở tiêu chí này là ngang nhau. Tuy nhiên, khi tính tỉ lệ gene mã hóa protein ribosome của CSDL HEG-DB cao hơn 3,2% so với kết quả chúng tôi thu được, điều này là do số lượng gene dự đoán được nhiều hơn nhưng vẫn giữ nguyên số gene mã hóa protein ribosome, do đó tiêu chí này không ảnh hưởng đến độ tin cậy của kết quả. Ở một tiêu chí

Bảng 1: Kết quả dự đoán gene biểu hiện cao ở *E. coli*. Các kết quả cho số lượng gene dự đoán trong khoảng 4%–9% được gạch dưới. Kết quả được chọn cuối cùng được tô đậm

Bộ tham chiếu	Ngưỡng giá trị CAI	Gene biểu hiện cao [4% - 9% = 170 - 381]	Gene mã hóa protein ribosome	Tỉ lệ
RP	≤ 0,687	≥ 507	50	9,9%
	0,688	324	44	13,6%
	0,689	**310**	**44**	**14,2%**
	0,690	244	42	17,2%
	0,691	235	42	17,9%
	0,692	198	41	20,7%
	≥ 0,693	≤ 32	13	40,63%
200-mRNA	≤ 0,693	≥ 451	50	11,09%
	≥ 0,694	≤ 31	12	38,7%
300-mRNA	≤ 0,693	≥ 451	50	11,07%
	≥ 0,694	≤ 31	12	38,7%



Hình 2: Biểu đồ thể hiện giá trị wi ở hai bộ tham chiếu RP và 100-mRNA

khác quan trọng hơn, khoảng giá trị CAI ở các gene biểu hiện cao chúng tôi dự đoán được [0,689 – 0,879] cao hơn so với CSDL HEG-DB [0,662 – 0,848]. Giá trị CAI cao cho thấy các gene trong nhóm gene biểu hiện cao có chung xu hướng sử dụng codon, các gene này đang sử dụng những codon ưa thích trong hệ thống biểu hiện của tế bào, có lượng tRNA tương ứng dồi dào, hoạt động hiệu quả hơn. Vì vậy, kết quả dự đoán này có độ tin cậy cao hơn kết quả của CSDL HEG-DB.

Độ phiên mã của các gene biểu hiện cao dựa trên dữ liệu microarray

Tiếp theo, hai bộ gene biểu hiện cao được so sánh dựa trên độ biểu hiện ở mức phiên mã. Hình 3 cho thấy tổng số gene thu nhận từ dữ liệu microarray (4021 gene) có khoảng phân bố tập trung giá trị về độ phiên mã gene thấp hơn những nhóm dữ liệu còn lại. Điều này chứng minh được rằng, ở mỗi thời điểm của tế bào, các gene có mức biểu hiện trung bình hoặc thấp chiếm nhiều nhất, các gene biểu hiện cao sẽ có số lượng ít hơn. Đối với các gene mã hóa protein ribosome, khoảng tập trung giá trị về độ phiên mã gene cao hơn hẳn so với độ phiên mã của tổng số gene trong dữ liệu microarray [NCBI], nhóm gene biểu hiện cao dự đoán được và của CSDL HEG-DB. Phần lớn các gene mã hóa protein ribosome đều tập trung ở mức độ phiên mã cao, điều này phù hợp với vai trò của các gene mã hóa protein ribosome vì những gene này thường được biểu hiệu cao trong tế bào, do đó được sử dụng để làm bộ tham chiếu dự đoán gene biểu hiện cao.

So sánh giữa kết quả dự đoán trong nghiên này với dữ liệu microarray của tất cả các gene thu nhận từ NCBI cho thấy hầu hết các gene dự đoán được đều có độ

Bảng 2: So sánh giá trị wi ở 4 amino acid Leu, Ser, Ala và Arg ở hai bộ tham chiếu RP và 100-mRNA. Những codon có thứ tự khác nhau giữa 2 bộ tham chiếu được tô đậm

Amino acid	Codon	Protein ribosome	Microarray	Amino acid	Codon	Protein ribosome	Microarray
Leu	CTG	1	1	Ser	TCT	1	1
	CTC	0,1076	0,0885		AGC	0,8944	0,6727
	TTG	0,0934	0,0935		TCC	0,7826	0,6623
	CTT	0,087	0,0957		AGT	0,2298	0,2312
	TTA	0,0791	0,0827		TCA	0,1739	0,2416
	CTA	0,019	0,0151		TCG	0,1739	0,1922
Arg	CGT	1	1	Ala	GCT	1	1
	CGC	0,5696	0,5583		GCA	0,6585	0,7194
	CGA	0,0267	0,0367		GCG	0,6192	0,7326
	CGG	0,0229	0,0446		GCC	0,3636	0,4604
	AGA	0,0133	0,0393				
	AGG	0,0038	0,0131				

Bảng 3: Kết quả so sánh các thông số trong quá trình dự đoán gene biểu hiện cao

Tiêu chí	HEG dự đoán	HEG-DB
Khoảng giá trị CAI	0,689 – 0,879	0,662 – 0,848
Số gene biểu hiện cao	310	253
Số gene mã hóa protein ribosome	44	44
Tỉ lệ gene mã hóa protein ribosome	14,2%	17,4%

phiên mã cao hơn 75% các gene từ NCBI [giá trị Q1 của cột HEG cao hơn Q3 của cột NCBI]. Điều này cho thấy kết quả dự đoán được các gene hầu hết đều thuộc nhóm có độ phiên mã cao. Kết quả tương tự cũng thu được khi so sánh dữ liệu từ HEG-DB với dữ liệu các gene từ NCBI.

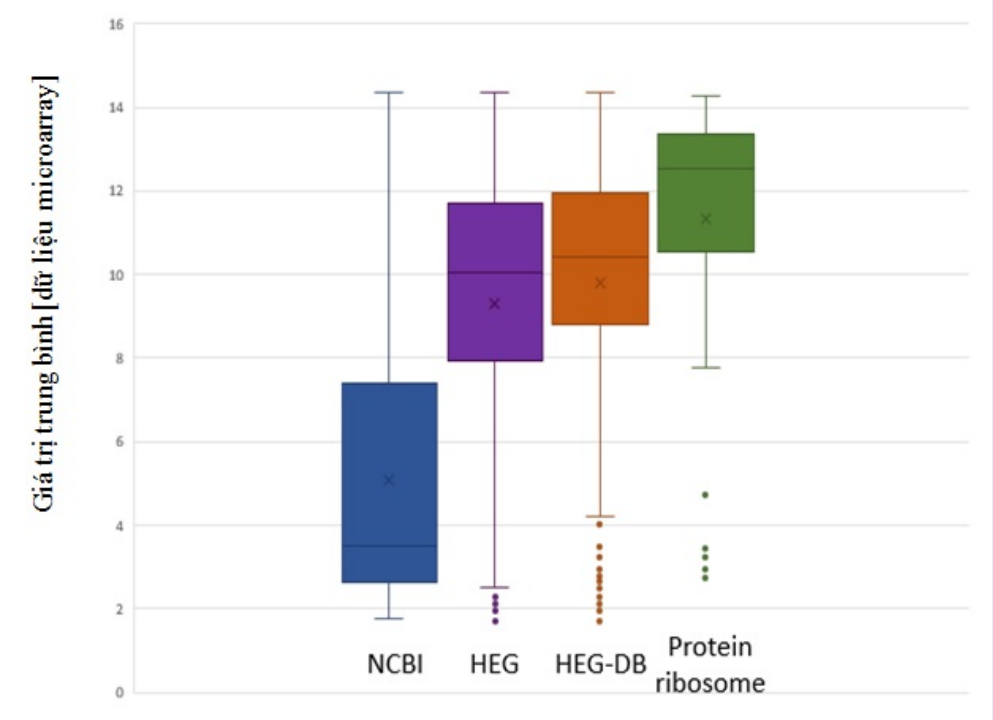
So sánh giữa nhóm gene biểu hiện cao dự đoán được và nhóm gene biểu hiện cao từ CSDL HEG-DB, mặc dù khoảng tập trung giá trị độ phiên mã của các gene biểu hiện cao từ CSDL HEG-DB nhỏ hơn, cũng như giá trị trung bình và trung vị về mức độ phiên mã của các gene từ CSDL HEG-DB cũng cao hơn so với kết quả của nghiên cứu, nhưng điều đó là do số lượng gene dự đoán được trong nghiên cứu này nhiều hơn số gene HEG trên cơ sở dữ liệu HEG-DB. Bằng chứng là khi lấy giá trị độ phiên mã của các gene từ CSDL HEG-DB so với số lượng tương ứng (253 gene) từ kết quả dự đoán cho thấy khoảng tập trung giá trị của kết quả từ nghiên cứu (14,373 – 5,469) tập trung hơn so với kết quả từ HEG-DB (14,373 – 1,692). Tương tự, giá trị trung bình và giá trị trung vị từ kết quả dự đoán (10,584 và 10,468) cũng cao hơn so với kết quả từ HEG-DB (9,712 và 10,418). Các số liệu độ phiên mã của các nhóm gene được thể hiện trong phụ lục.

So sánh số gene tham gia vào các con đường chuyển hóa

Để thu nhận thông tin về các con đường chuyển hóa mà các gene biểu hiện cao dự đoán được và của CSDL HEG-DB tham gia vào, chúng tôi sử dụng CSDL DAVID để tìm kiếm và thu nhận những gene tham gia vào con đường chuyển hóa của tế bào.

Từ kết quả được biểu thị ở Bảng 4, nhận thấy số gene tham gia vào con đường chuyển hóa ở nhóm gene biểu hiện cao dự đoán được là 166 gene tham gia vào 22 con đường chuyển hóa, nhiều hơn so với nhóm gene biểu hiện cao trên CSDL HEG-DB là 122 gene tham gia vào 16 con đường chuyển hóa. Vậy nghiên cứu đã dự đoán được các gene biểu hiện cao có nhiều gene tham gia vào các con đường chuyển hóa hơn các gene biểu hiện cao trên CSDL HEG-DB.

Tiếp theo chúng tôi tiến hành so sánh cụ thể 8 con đường chuyển hóa quan trọng đối với tế bào (Hình 4). Đây là những con đường tham gia vào chu trình



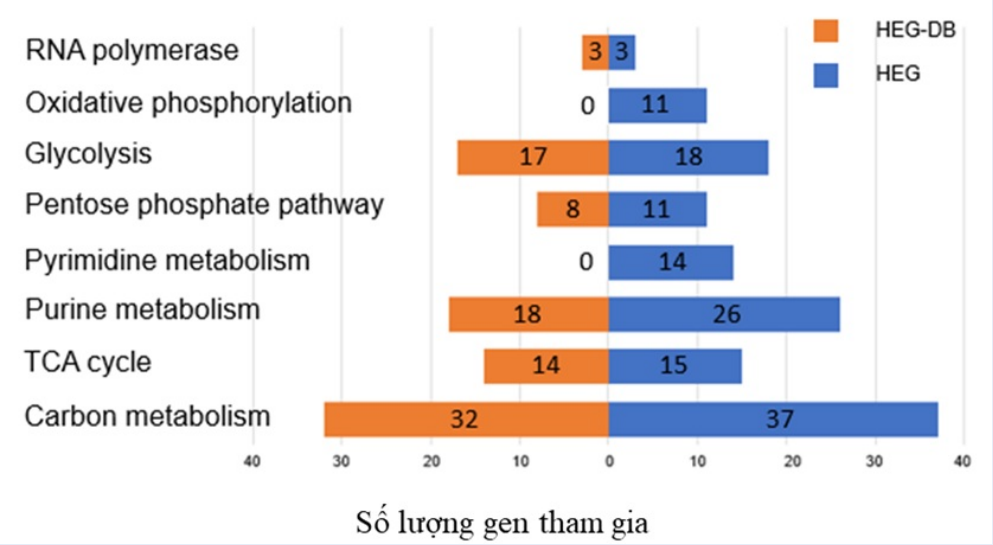
Hình 3: Biểu đồ Boxplot biểu diễn mức độ dao động về độ phiên mã của các nhóm gene ở *E. coli*. NCBI là tổng 4021 gene của *E. coli* thu nhận từ NCBI có dữ liệu microarray. HEG là 309/310 gene từ kết quả dự đoán trong nghiên cứu (gene ypdK không có dữ liệu microarray). HEG-DB là 253 gene HEG thu nhận từ CSDL HEG-DB. Protein ribosome là 68/69 gene mã hóa protein ribosome dùng trong bộ tham chiếu RP (gene ykgO không có dữ liệu microarray).

Bảng 4: Kết quả phân tích các gene tham gia vào các con đường chuyển hóa

Tiêu chí	Gene biểu hiện cao dự đoán được	Gene biểu hiện cao từ HEG-DB
Tổng số lượng gene	310 gene	253 gene
Có dữ liệu trong CSDL DAVID	306 gene	238 gene
Số gene tham gia vào các con đường chuyển hóa	166 gene	122 gene
Số lượng con đường chuyển hóa	22	16

chuyển hóa carbon, chuyển đổi năng lượng và cung cấp nguyên liệu tổng hợp DNA, RNA trong tế bào. Vì thế nó cần thiết cho sự sinh trưởng và phân chia của tế bào vi khuẩn. Tám con đường chuyển hóa này bao gồm: RNA polymerase, Oxidative phosphorylation, Glycolysis, Pentose phosphate pathway, Pyrimidine metabolism, Purine metabolism, TCA cycle (Citrat cycle) và Carbon metabolism.

Kết quả thể hiện trên biểu đồ cho thấy gene biểu hiện cao dự đoán được có tham gia vào cả 8 con đường chuyển hóa quan trọng. Trong khi đó, các gene biểu hiện cao thu nhận từ CSDL HEG-DB chỉ tham gia vào 6/8 con đường chuyển hóa quan trọng, 2 con đường chuyển hóa không tham gia là: oxidative phosphorylation và pyrimidine metabolism. Cả hai nhóm gene đều có 3 gene tham gia vào con đường chuyển hóa RNA polymerase trong tổng số 4 gene trong con đường chuyển hóa này trong CSDL DAVID. Đối với 5 con đường chuyển hóa còn lại (glycolysis, pentose phosphate pathway, purine metabolism, TCA cycle và carbon metabolism), số gene thuộc nhóm gene biểu hiện cao do chúng tôi dự đoán được tham gia ở mỗi con đường là nhiều hơn so với nhóm gene biểu hiện cao thu từ CSDL HEG-DB. Tóm lại, nhóm gene biểu hiện cao do nghiên cứu này dự đoán được tham gia vào các con đường chuyển hóa quan trọng nhiều hơn so với kết quả từ CSDL HEG-DB. Từ đó chứng minh được kết quả dự đoán gene biểu hiện cao của nghiên



Hình 4: Kết quả so sánh số lượng gene tham gia các con đường chuyển hóa quan trọng

cứu có độ tin cậy cao và chính xác hơn so với CSDL HEG-DB.

KẾT LUẬN

Nghiên cứu dự đoán gene biểu hiện cao cho chủng *E. coli* K-12 MG1655 bằng phương pháp đề xuất bởi tác giả Puigbò với ngưỡng giá trị CAI tự khảo sát và các tiêu chí chọn kết quả tự đề xuất cùng với sử dụng bộ tham chiếu ban đầu dựa trên dữ liệu độ biểu hiện mRNA bên cạnh bộ tham chiếu đề xuất bởi tác giả dựa trên các gene mã hóa protein ribosome. Kết quả dự đoán với ngưỡng giá trị CAI = 0,689 cho kết quả thống nhất hoàn toàn giữa hai bộ tham chiếu RP và 100-mRNA cho thấy khả năng ứng dụng của bộ tham chiếu dựa trên dữ liệu mRNA microarray vào dự đoán gene biểu hiện cao. Nghiên cứu cũng đã so sánh kết quả dự đoán được với dữ liệu thu nhận từ CSDL HEG-DB và cho thấy kết quả dự đoán có số lượng gene biểu hiện cao nhiều hơn, giá trị CAI cao hơn và số lượng gene tham gia vào các con đường chuyển hóa tổng và chuyển hóa quan trọng đều cao hơn so với dữ liệu từ HEG-DB. Mặc dù chỉ mới tiến hành trên chủng *E. coli* K-12 MG1655 nên sẽ cần đánh giá sâu hơn trên các chủng loài khác, nhưng các kết quả của nghiên cứu đã giúp chi tiết hóa quy trình dự đoán gene biểu hiện cao từ xác định ngưỡng giá trị CAI đến các tiêu chí chọn lọc kết quả cũng như đưa ra một lựa chọn mới trong việc chọn bộ tham chiếu ban đầu. Bộ gene biểu hiện cao dự đoán được có độ tin cậy cao và có thể ứng dụng trong các phân tích và nghiên cứu sâu hơn về ảnh hưởng của đặc trưng gene lên độ biểu hiện cũng như thiết kế gene.

LỜI CẢM ƠN

Nghiên cứu được tài trợ bởi Đại học Quốc gia Thành phố Hồ Chí Minh [ĐHQG-HCM] trong khuôn khổ đề tài mã số C2017-18-17.

DANH MỤC TỪ VIẾT TẮT

HEG: Highly expressed genes [gene biểu hiện cao]
CSDL: Cơ sở dữ liệu
CAI: Codon adaptation index [chỉ số thích nghi codon]

CAM KẾT XUNG ĐỘT LỢI ÍCH

Các tác giả tuyên bố không có xung đột lợi ích liên quan đến việc xuất bản của bài viết này.

ĐÓNG GÓP TỪNG TÁC GIẢ

Tác giả Phạm Trung Nghĩa và Trương Hà Minh Nhật tiến hành thu nhận và xử lý dữ liệu; tác giả Võ Trí Nam tiến hành các tất cả các phân tích còn lại trong bài báo; tác giả Trần Linh Thước và Nguyễn Đức Hoàng định hướng, góp ý và nhận xét cho nghiên cứu; tất cả các tác giả đã xem xét và đồng ý với bản thảo bài báo.

TÀI LIỆU THAM KHẢO

1. Yu K, Ang KS, Lee D-Y. Synthetic genee design using codon optimization on-line [COOL]. Methods Mol Biol. 2017;1472:13-34;PMID: 27671929. Available from: https://doi.org/10.1007/978-1-4939-6343-0_2.
2. Grote A, Hiller K, Scheer M, Münch R, Nörtemann B, Hempel DC, et al. JCat: a novel tool to adapt codon usage of a target genee to its potential expression host. Nucleic Acids Res. 2005 1;33[Web Server issue]:W526-531;PMID: 15980527. Available from: https://doi.org/10.1093/nar/gki376.

3. Raab D, Graf M, Notka F, Schödl T, Wagner R. The GeneeOptimizer Algorithm: using a sliding window approach to cope with the vast sequence space in multiparameter DNA sequence optimization. Syst Synth Biol. 2010 ;4[3]:215-25;PMID: 21189842. Available from: https://doi.org/10.1007/s11693-010-9062-3.
4. Sharp PM, Li WH. The codon adaptation index–a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 1987 11;15[3]:1281-95;PMID: 3547335. Available from: https://doi.org/10.1093/nar/15.3.1281.
5. Gaspar P, Moura G, Santos MAS, Oliveira JL. mRNA secondary structure optimization using a correlated stem-loop prediction. Nucleic Acids Res. 2013;41[6]:e73;PMID: 23325845. Available from: https://doi.org/10.1093/nar/gks1473.
6. Puigbò P, Romeu A, Garcia-Vallvé S. HEG-DB: a database of predicted highly expressed genees in prokaryotic complete geneomes under translational selection. Nucleic Acids Res. 2008;36[Database issue]:D524-527;PMID: 17933767. Available from: https://doi.org/10.1093/nar/gkm831.
7. Karlin S, Mrázek J. Predicted highly expressed genees of diverse prokaryotic geneomes. J Bacteriol. 2000;182[18]:5238-50;PMID: 10960111. Available from: https://doi.org/10.1128/JB.182.18.5238-5250.2000.
8. Chi DTK, Lang TV, Hiep HX. Dự đoán gene biểu hiện cao cho thiết kế gene dùng trong tái tổ hợp. Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX -Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin [FAIR'9]. 2016;.
9. Puigbò P, Guzmán E, Romeu A, Garcia-Vallvé S. OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. Nucleic Acids Res. 2007;35[Web Server issue]:W126-131;PMID: 17439967. Available from: https://doi.org/10.1093/nar/gkm219.
10. Lewis NE, Cho B-K, Knight EM, Palsson BO. Genee Expression Profiling and the Use of Geneome-Scale In Silico Models of Escherichia coli for Analysis: Providing Context for Content. Journal of Bacteriology. 2009;191(11):3437. PMID: 19363119. Available from: https://doi.org/10.1128/JB.00034-09.
11. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large genee lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57. PMID: 19131956. Available from: https://doi.org/10.1038/nprot.2008.211.
12. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large genee lists. Nucleic Acids Res. 2009;37[1]:1-13;PMID: 19033363. Available from: https://doi.org/10.1093/nar/gkn923.

Science & Technology Development Journal – Natural Sciences, 5(2):1068-1077

Open Access Full Text Article

Research Article

Study on predicting highy expressed genes for *Escherichia coli* based on mRNA microarray data

Nam Tri Vo[1,2,3], Trung-Nghia Pham[1,3], Minh-Nhat Truong-Ha[1], Thuoc Linh Tran[3,4], Hoang Duc Nguyen[1,3,4,*]

Use your smartphone to scan this QR code and download this article

ABSTRACT

Highly expressed genes [HEG] are genes available in the organism, which carry the preferred codons for the expression system. Identifying HEG helps to find preferred codons and use them in the gene optimization to express target proteins. Currently, HEG-DB is the only database storing HEG data of many strains of microorganisms, but the data is not updated and maintained. Therefore, our research is carried out to predict HEG in the *E. coli* K-12 MG1655 strain based on reference sets that are the mostly used ribosomal protein coding genes and genes with high transcription levels from microarray data proposed by the research. Next, the results of HEG from the two above reference sets, HEG-RP and HEG-mRNA, were compared. Finally, we analyzed and compared the HEG that the project predicted with HEG from HEG-DB database. The results from RP and 100-mRNA reference sets were completely identical and were better than data from HEG-DB in the number of HEGs, CAI values and the number of genes contributing to important metabolic pathways. The results showed that it was possible to use reference sets from mRNA microarray data instead of ribosomal protein reference sets in HEG prediction.

Key words: highly expressed genes, Escherichia coli, ribosomal protein, mRNA microarray, CAI

[1]*Center for Bioscience and Biotechnology, University of Science, VNU-HCMC, Vietnam*

[2]*Laboratory of Molecular Biotechnology, University of Science, VNU-HCMC, Vietnam*

[3]*Vietnam National University, Ho Chi Minh City, VNU-HCM, Vietnam*

[4]*Faculty of Biology -Biotechnology, University of Science, VNU-HCMC, Vietnam*

Correspondence

Hoang Duc Nguyen, Center for Bioscience and Biotechnology, University of Science, VNU-HCMC, Vietnam

Vietnam National University, Ho Chi Minh City, VNU-HCM, Vietnam

Faculty of Biology -Biotechnology, University of Science, VNU-HCMC, Vietnam

Email: ndhoang@hcmus.edu.vn

History

- Received: 25-8-2020
- Accepted: 22-3-2021
- Published: 30-4-2021

DOI : 10.32508/stdjns.v5i2.945

Cite this article : Vo N T, Pham T, Truong-Ha M, Tran T L, Nguyen H D. **Study on predicting highy ex-pressed genes for *Escherichia coli* based on mRNA microarray data**. *Sci. Tech. Dev. J. - Nat. Sci.;* 5(2):1068-1077.