DOI: 10.15625/vap.2016.00017

DỰ ĐOÁN GEN BIỂU HIỆN CAO CHO THIẾT KẾ GEN DÙNG TRONG TÁI TỔ HỢP

Dương Thị Kim Chi[1], Trần Văn Lăng[2], Huỳnh Xuân Hiệp[3]

[1] Khoa Công nghệ Thông tin, Trường Đại học Thủ Dầu Một

[2] Viện Cơ học và Tin học ứng dụng, Viện Hàn lâm Khoa học và Công nghệ Việt Nam

[3]Khoa Công nghệ Thông tin và Truyền thông, Trường Đại học Cần Thơ

chidtk@tdmu.edu.vn, langtv@vast.vn, hxhiep@ctu.edu.vn

TÓM TẮT*—Dự đoán gen biểu hiện cao HEG (Highly Expressed Gene) là một công đoạn quan trọng trong việc tìm gen tối ưu cho quá trình tái tổ hợp. Các gen biểu hiện cao trong tế bào thường có xu hướng có các đặc trưng tương tự nhau, chủ yếu là đặc trưng về xu hướng sử dụng codon. Bài viết này đề xuất một hướng tiếp cận mới để phân cụm dữ liệu ứng dụng để xác định nhóm các gen có đặc trưng giống nhau về xu hướng sử dụng codon để dự đoán HEG. Các thực nghiệm được triển khai trên hai thuật toán PAM (Partitioning Around Medoids), CLARA (Clustering for Large Applications) cho việc phân cụm dự đoán HEG. Các kết quả thu được cho thấy CLARA vượt trội hơn PAM về thời gian, chất lượng phân cụm.*

Từ khóa*— DNA tái tổ hợp, gen B.subtilis, PAM, và CLARA, HEG, HSCU (Relative Synonymous Codon Usage).*

I. GIỚI THIỆU

Dự đoán gen, phân loại gen để hiểu rõ hơn về cấu trúc và chức năng của gen phục vụ cho các mục đích nghiên cứu cơ bản về sinh học phân tử, chẩn đoán bệnh, sản xuất dược phẩm, cải tạo môi trường, cải tạo giống cây trồng. Một ứng dụng khác của phân loại gen đang được quan tâm hiện nay là chọn lựa được gen tốt nhất cho công nghệ tái tổ hợp. Việc sản xuất protein tái tổ hợp thường được bắt đầu bằng việc lựa chọn một gen mong muốn, tiếp theo là phân lập gen và cắt gen bằng các enzyme hạn chế. Gen tách được gắn vào một *véctơ tạo dòng* (plasmid) và đưa vào một vật chủ, ở đó đoạn gen này sẽ được dịch mã thành một protein đặc biệt được gọi là protein tái tổ hợp. Để có thể chọn được một đoạn gen mong muốn, gen này phải được dự đoán là có khả năng nâng cao biểu hiện gen mục tiêu. Gen với đặc tính như vậy được gọi là gen biểu hiện cao HEG. Có hai phương án dự đoán HEG được sử dụng:

Phương án 1: dựa vào chỉ số thích nghi codon CAI (Codon Adaptation Index) và dùng thống kê để xác định HEG, phương pháp này được đề xuất bởi Pere Puigbò và cộng sự năm 2007 [4]. Có thể tổng quan phương pháp này như sau:

(1) Tính giá trị CAI của các gen trong nhóm gen biểu hiện cao thu nhận được từ cơ sở dữ liệu HEG-DB.

(2) Dùng biểu đồ Boxplot thống kê khoảng tập trung giá trị CAI nhằm loại bỏ các giá trị cá biệt.

(3) Thực hiện dự đoán gen biểu hiện cao với lần lượt các giá trị ngưỡng CAI trong khoảng tập trung giá trị CAI từ bước 2, khoảng cách giữa các giá trị khảo sát là 0,05.

(4) Đánh giá kết quả dự đoán gen biểu hiện cao để chọn ngưỡng CAI thích hợp theo hai tiêu chí:

- Số lượng gen biểu hiện cao: khoảng 5% số gen trong bộ gen.

- Độ nhạy (sensitive): Tỉ lệ giữa số gen mã hóa cho Protein Ribosome trong tập gen biểu hiện cao dự đoán được và tổng số gen mã hóa cho Protein Ribosome.

Phương án 2: dựa vào chỉ số sử dụng codon đồng nghĩa RSCU (Relative Synonymous Codon Usage) [5] của từng gen và phân cụm các gen dựa trên tiêu chí này. Các gen biểu hiện cao trong tế bào thường có xu hướng có các đặc trưng tương tự nhau, chủ yếu là đặc trưng về xu hướng sử dụng codon. Phương pháp này dựa trên các gen vốn đã được biết là HEG, được đặt tên là "kernel", có thể khái quát phương pháp này như sau:

(1) Tính RSCU cho từng gen.

(2) Áp dụng các thuật toán phân cụm dữ liệu tìm ra ở bước (1), hình thành các cụm và tìm nhân "kernel mới".

(3) Đánh giá một nhóm được phân cụm càng có nhiều kernel càng chứng tỏ nhóm đó càng gần với kernel. Do đó, nhóm này có khả năng cao là HEG.

Bài viết này tiếp cận theo phương án 2 để tìm HEG, thuật toán được chọn để áp dụng là PAM và CLARA để phân cụm dữ liệu nhằm tìm HEG cho quá trình thiết kế gen cho tái tổ hợp. Các phần còn lại của bài viết bao gồm: phần 2 giới thiệu bài toán tìm HEG, cách tính các chỉ số RSCU và các độ đo được dùng trong các thuật toán phân cụm, phần 3 giới thiệu hai thuật toán PAM và CLARA trong thực nghiệm, phần 4 trình bày kết quả thực nghiệm trên bộ gen *B.subtilis* và cuối cùng là phần kết luận.

II. BÀI TOÁN TÌM HEG

Bài toán tìm HEG cho thiết kế gen tái tổ hợp

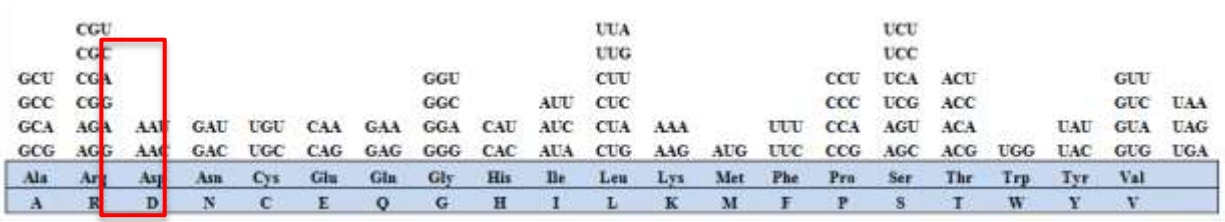
DNA tái tổ hợp:

DNA tái tổ hợp là phân tử DNA được tạo thành từ hai hay nhiều trình tự DNA của các loài sinh vật khác nhau. Trong kỹ thuật di truyền, DNA tái tổ hợp thường là được tạo thành từ việc gắn những đoạn DNA có nguồn gốc khác nhau vào trong véctơ tách dòng. Những véctơ tách dòng mang DNA tái tổ hợp này có thể biểu hiện hành các protein tái tổ hợp trong các sinh vật. [7].

Codon đồng nghĩa (Synonymous Condon)

Cơ sở khoa học của việc chọn lựa gen biểu hiện cao dựa trên hiện tượng codon đồng nghĩa. Một codon gồm ba nucleotide, nên có $4^3 = 64$ codon. Nhưng chỉ có 20 amino acid, nên một amino acid có thể được mã hóa bởi ít nhất hai codon khác nhau, các codon như vậy được gọi là các codon đồng nghĩa [9].

Ví dụ: Như minh họa ở hình 1- được đóng khung đỏ, Amino acid *Ala (hay A)* có bốn codon đồng nghĩa là GCA, GCC, GCG, GCU.



Hình 1. Minh họa các codon đồng nghĩa

Codon hiếm:

Trong quá trình tiến hóa, các loài khác nhau có tính thiên vị với một loại mã bộ ba nhất định trong số các codon đồng nghĩa cùng mã hóa cho một amino acid. Do đó có thể codon được ưu tiên sử dụng ở loài này nhưng lại trở thành codon hiếm ở loài khác. Ví dụ tần suất sử dụng AGG, AGA mã hoá cho Arg ở người là 11,2%, trong khi đó, giá trị này ở E. coli là 2,1% và 2,4% [8].

Chỉ số sử dụng codon đồng nghĩa RSCU:

Chaperon là các protein giúp hỗ trợ các phân tử peptide gấp cuộn chính xác để tạo thành protein có hoạt tính. Các protein này thường tồn tại trong tế bào với số lượng lớn. Cả ribosomal và chaperone protein gen đều là HEG và có xu hướng sử dụng codon cao [5,11]. Ý tưởng chính của việc dự đoán HEG dựa trên các HEG có quan hệ gần gũi nhau, trong khi những gen còn lại thì không. Đặc trưng cho các gen được sử dụng là chỉ số RSCU của các codon trong gen [4]:

$$r_{ac}(g) = \frac{O_{ac}}{\frac{1}{k_a}\sum_{c \in C_a} O_{ac}} = \frac{O_{ac}}{\overline{O_{ac}}} \quad (1)$$

Trong đó:

r_{ac}: RSCU của codon *c* mã hoá cho amino acid *a.*

O_{ac}: tần suất xuất hiện của codon *c* trong trình tự gen;

C_a: tập hợp các codon cùng mã hoá cho amino acid *a;*

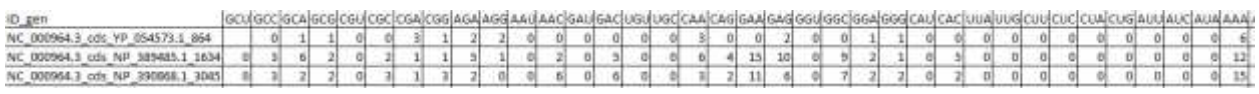
K_a: số lượng các loại codon *c* mã hóa cho amino acid *a.*

Các gen biểu hiện cao trong tế bào thường có xu hướng có các đặc trưng tương tự nhau, chủ yếu là đặc trưng về xu hướng sử dụng codon. Do đó, phương pháp phân cụm dữ liệu được ứng dụng để xác định nhóm các gen có đặc trưng giống nhau về xu hướng sử dụng codon nhằm dự đoán HEG.

Bên cạnh đó, việc tính toán RSCU cho các codon kết thúc (UAA,UGA,UGA) và các codon không có codon đồng nghĩa (AUG – mã hóa cho Methionine, UGG – mã hóa cho Trytophan) là không cần thiết. Do đó, gen được biểu diễn lại bởi một véctơ RSCU 59 chiều có dạng như sau:

$$r(g)=\{ r_1, r_2, r_3 \ldots r_{59}\}^T \quad (2)$$

Hình 2 là mô tả cho việc biểu diễn tập gen bằng tập các RSCU.



Hình 2. Minh họa cho biểu diễn gen với RSCU

Bài toán tìm HEG

Trong học máy, phép phân tích cụm thường dựa trên học không giám sát. Không giống như phân loại, phân cụm không dựa trên các lớp đã định nghĩa trước và các mẫu dữ liệu huấn luyện đã gắn nhãn lớp [6]. Nguyên tắc chính

của phân cụm vẫn là làm sao cho độ giống nhau trong cùng một cụm là cao và độ giống nhau giữa các cụm là thấp [6]. Do vậy trong bài viết này chọn phương pháp phân hoạch để phân cụm với mục đích chọn ra các nhóm gen có cùng xu hướng sử dụng codon vào một nhóm và có thể chọn HEG có khuynh hướng gần nhân của nhóm nhất.

Độ đo khoảng cách

Ma trận dữ liệu:

Tập dữ liệu được phân cụm là *n* gen được biểu diễn bằng *p* biến chính là RSCU của từng gen và tập dữ liệu này là cấu trúc có dạng bảng quan hệ, hay ma trận *n* x *p* với *n* gen; *p* biến còn được gọi là các phép đo hay các thuộc tính. Tập dữ liệu gen được biểu diễn lại bằng ma trận dữ liệu có dạng:

$$\begin{bmatrix} r_{11} & r_{12} & \ldots & r_{59} \\ \ldots & \ldots & \ldots & \ldots \\ r_{i1} & r_{i2} & \ldots & r_{i59} \\ \vdots & \vdots & \vdots & \vdots \\ r_{n1} & dr_{n2} & \ldots & r_{n59} \end{bmatrix} \quad (3)$$

Ma trận phân biệt:

Để biểu diễn khoảng cách giữa hai gen trong không gian dữ liệu gồm n gen theo thuộc tính RSCU ta dùng ma trận phân biệt:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \ldots & \ldots & 0 \end{bmatrix} \quad (4)$$

$d(i, j)$ là khoảng cách giữa gen *i* và gen *j*; khoảng cách này được tính theo công thức (5).

Độ đo khoảng cách Euclidean [11]**:**

$$\boldsymbol{d(i,j) = \sqrt{\left(|r_{i1} - r_{j1}|^2 + |r_{i2} - r_{j2}|^2 + \ldots + |r_{ip} - r_{jp}|^2\right)}} \quad (5)$$

Chỉ số Silhouette

Giả sử các HEG được chia thành *k* cụm. Với mỗi cụm gen *i*, đặt:

- $a(i)$ là khoảng cách trung bình từ *i* tới tất cả các gen trong cùng cụm với *i*.
- $b(i)$ là khoảng cách trung bình ngắn nhất từ *i* tới bất kỳ cụm nào không chứa *i*.

Cụm tương ứng với $b(i)$ này được gọi là cụm hàng xóm của *i*. Khi đó chỉ số Silhouette $s(i)$ được định nghĩ như sau:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (6)$$

Với chỉ số $s(i)$ nằm trong đoạn [-1,1] cho thấy $s(i)$ càng gần 1 thì gen *I* càng phù hợp với cụm mà nó được phân vào, $s(i) = 0$ thì không thể xác định được gen *i* nên thuộc về cụm nào giữa cụm hiện tại và cụm hàng xóm của nó, $s(i)$ càng gần -1 chứng tỏ *i* bị phân sai cụm, nó nên thuộc về cụm hàng xóm chứ không phải cụm hiện tại.

III. THUẬT TOÁN PAM VÀ CLARA TRONG THỰC NGHIỆM

Các phương pháp phân chia nổi tiếng và thường được dùng nhất là k-means (MacQueen 1967), k-medoids (Kaufman và Rousseew 1987) và các dạng biến đổi của chúng [6]. Đối với phương pháp k-means thường được áp dụng khi trung bình của một cụm được xác định. Đối với một số gen có loại codon hiếm thì có thể có những codon không được sử dụng trong tất cả tập gen nên có thể gây khó khăn cho việc tính trung bình cụm. Hơn nữa, nó nhạy cảm với các điểm dữ liệu nhiễu vào outlier, một số lượng nhỏ dữ liệu như vậy về căn bản có ảnh hưởng tới giá trị trung bình. Nên chọn lựa PAM hay CLARA vào bài toán dự đoán HEG có những thuận lợi nhất định.

A. Thuật toán PAM

PAM (partition around medoids) - phân chia xung quanh các *medoid – trung tâm*: Đây là một giải thuật phân cụm kiểu k-medoids. Tìm *k* cụm trong *n* gen bằng cách: trước tiên tìm một gen đại diện ***gp*** hay PHEG (Predicted Highly Expressed Gene) hay medoid cho mỗi cụm. Tập các medoid ban đầu được lựa chọn tuỳ ý. Sau đó lặp lại các thay thế một trong số các medoid bằng một trong số những đối tượng không phải medoid miễn là tổng khoảng cách của kết quả phân cụm được cải thiện [6].

Gọi là ***gp*** hay là gọi là gen mediod.
Gọi ***gi***: gen khác với *gp* ($i = 1..n$)

Thuật toán PAM

Đầu vào: Tập hợp các chuỗi gen $G = \{g_1, g_2, \ldots, g_n\}$, số cụm k
Đầu ra: Tập hợp gen đã được phân vào *k* cụm.
(1) **Chọn** tùy ý *k* gen giữ vai trò là các *gp* ban đầu
(2) **Lặp lại**

(3) Với mỗi *gp*
Lần lượt xét các gen *gi* không là *gp*
Tính *S* là độ lợi khi hoán đổi *gp* với *gi*
S = Egp - E1gi
(4) **If** S < 0 **then** hoán vị *gi* với *gp*
(5) **Cho đến khi không thay đổi gi với *gp***

B. Thuật toán CLARA

Đối với bộ gen lớn việc dùng PAM để phân cụm tốn thời gian và chất lượng phân cụm thấp và quan trọng là khó xác định được tập các HEG; thuật toán CLARA khắc phục nhược điểm của thuật toán PAM trong trường hợp này. CLARA tiến hành trích mẫu cho tập dữ liệu có *n* phần tử, nó áp dụng thuật toán PAM cho mẫu này và tìm ra các đối tượng trung tâm medoid cho mẫu được trích ra từ dữ liệu này. Nếu mẫu dữ liệu được trích theo một cách ngẫu nhiên, thì các medoid của nó xấp xỉ với các medoid của toàn bộ tập dữ liệu ban đầu. Để tiến tới một xấp xỉ tốt hơn, CLARA đưa ra nhiều cách lấy mẫu và thực hiện phân cụm cho mỗi trường hợp, sau đó tiến hành chọn kết quả phân cụm tốt nhất khi thực hiện phân cụm trên mẫu này. Để đo chính xác, chất lượng của các cụm được đánh giá thông qua độ phi tương tự trung bình của toàn bộ các đối tượng dữ liệu trong tập đối tượng dữ liệu ban đầu [6]. Sau đây là thuật toán CLARA:

Gọi S là kích thước mẫu được trích từ tập gen $G = \{g1, g2, \ldots, gn\}$, trong đó:
k: số cụm,
n: số gen.
S: tập hợp các gen được đưa vào cụm.

Thuật toán CLARA

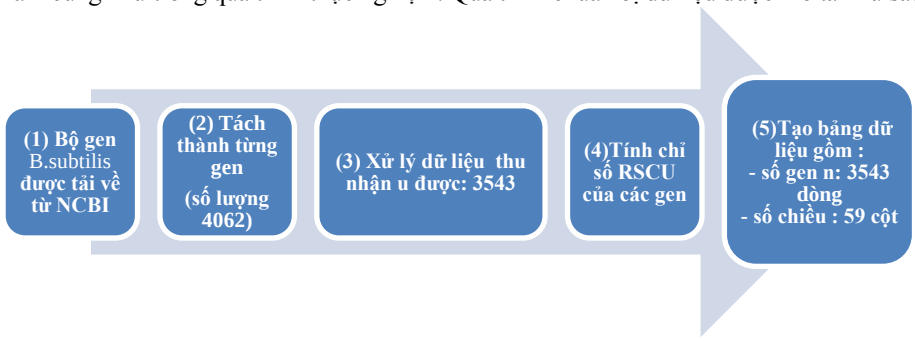
Đầu vào: Tập hợp các chuỗi gen $G = \{g1, g2, \ldots, gn\}$, số cụm *k*
Đầu ra: Tập hợp gen đã được phân vào *k* cụm.
(1) For *i* = 1 to *S* do
(2) Lấy một mẫu có *Sj* gen ngẫu nhiên từ tập dữ liệu *G*. Áp dụng thuật toán PAM cho mẫu dữ liệu này nhằm để tìm các gen medoid đại diện cho các cụm.
(3) Đối với mỗi đối tượng trong tập dữ liệu ban đầu, xác định gen medoid tương tự nhất trong số k đối tượng medoid.
(4) Tính độ phi tương tự 2 trung bình cho phân hoạch các đối tượng thu được ở bước trước. Nếu giá trị này bé hơn giá trị tối thiểu hiện thời thì sử dụng giá trị này thay cho giá trị tối thiểu ở trạng thái trước, như vậy, tập *k* đối tượng medoid xác định ở bước này là tốt nhất cho đến thời điểm này.
(5) EndFor

IV. THỰC NGHIỆM

Bài viết sử dụng phần mềm Rstudio cùng gói thư viện Cluster có chứa các thuật toán PAM và CLARA. Thuật toán được thử nghiệm trên máy tính cá nhân có RAM 4 GB, Intel Core i3. Tiến hành thực nghiệm trên bộ dữ liệu B.Subtilus như đã mô tả với các tùy chọn *k* = 2 đến ^ = 12 cho cả hai thuật toán PAM và CLARA. Kết quả phân cụm thu được từ hai thuật toán như sau:

A. Thu nhận, xử lý số liệu thực nghiệm

Bên cạnh *E.coli*, vi khuẩn chủng Bacillus subtilis gọi là *B.subtilis* cũng là một trong số các hệ thống biểu hiện đang được quan tâm nghiên cứu và sử dụng hiện nay trong lĩnh vực sản xuất protein tái tổ hợp. Hiện nay, hệ thống biểu hiện trên *B. subtilis* đang thu hút các nhà nghiên cứu thuộc lĩnh vực protein tái tổ hợp bởi tính an toàn khi sử dụng trực tiếp trong lĩnh vực y dược, mỹ phẩm cũng như trong quá trình thực nghiệm. Quá trình chuẩn bị dữ liệu được mô tả như sau:



Hình 3. Quy trình xử lý dữ liệu tổng quát

[1] E_{gp}, E_{gi} lần lượt là giá trị hàm mục tiêu trước và sau khi thay **gp** bởi **gi** $E=\sum_{i=1}^{k}\sum_{p\in c_i} d(gp,gi)^2$
[2] Công thức tính (5)

Dữ liệu sau khi tải về từ ngân hàng gen quốc tế NCBI, tách thành từng gen, loại bỏ những gen không bắt đầu bằng những codon khởi đầu phiên mã (ATG, GTG, TTG) và những gen có chiều dài không phải là bội số của ba để thu nhận tập các gen có khả năng là HEG. Số lượng mẫu thu được là 3543 gen. Từ trình tự thu được của các gen tính RSCU cho từng gen, tạo bảng dữ liệu cho mẫu gồm 3543 gen và 59 codon như đã mô tả ở trên.

B. Áp dụng thuật toán PAM, CLARA cho bộ dữ liệu B. Subtilus.

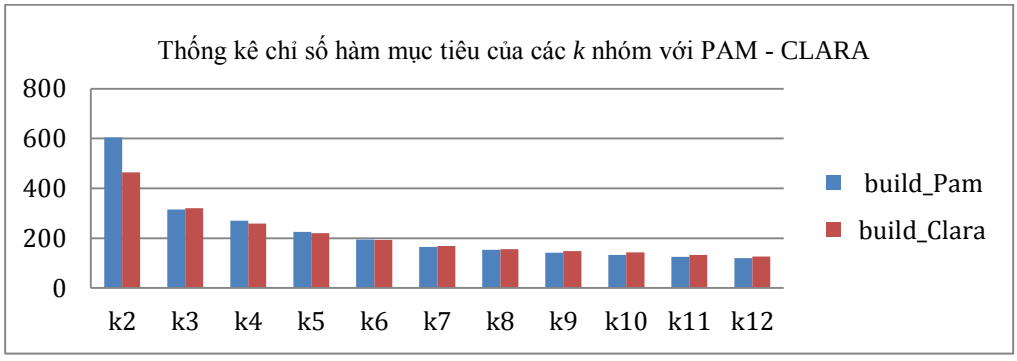
Các tiêu chí để nhận dạng chất lượng của các thuật toán phân cụm dựa vào việc đánh giá và chọn kết quả gom cụm tối ưu như: Độ nén (compactness) đối tượng trong cụm nên gần nhau có nghĩa là các gen trong cùng một cụm nên càng gần nhau điều này được thể hiện qua chỉ số Silhouette (Silhouette Index) cho ước lượng tính chia cắt và chặt của một sự phân chia cụm. Độ phân tách (separation): *Tiêu chí này cho các cụm nên xa nhau*. Ngoài ra còn có các tham số khác để so sánh mức độ hiệu quả của phân cụm như: hàm mục tiêu, thời gian thực thi các thuật toán.

1. *So sánh về hàm mục tiêu*

Áp dụng PAM, CLARA cho các với k = 2 đến k = 12, thu được hai cụm với số lượng khá tương đồng với độ. Chất lượng phân cụm được đánh giá thông qua hàm mục tiêu, chất lượng phân cụm tốt nhất khi hàm mục tiêu đạt giá trị tối thiểu. Như thống kê ở hình 4 và bảng số liệu ở bảng 1, ta thấy giá trị hàm mục tiêu của các cụm giảm dần, cả hai đạt giá trị bằng nhau ở phân nhóm k = 8, và k = 12.

Bảng 1. Bảng thống kê giá trị của hàm mục tiêu cuả hai thuật toán PAM và CLARA

Nhóm / Thuật toán	k2	k3	k4	k5	k6	k7	k8	k9	k10	k11	k12
PAM	604.4944	315.5504	270.2221	225.8861	194.2379	164.3636	152.7764	142.1114	132.9562	124.7681	120.5503
CLARA	464.702	320.3063	258.0315	219.5571	193.5844	168.9772	155.3532	148.2193	142.772	132.9559	125.8634



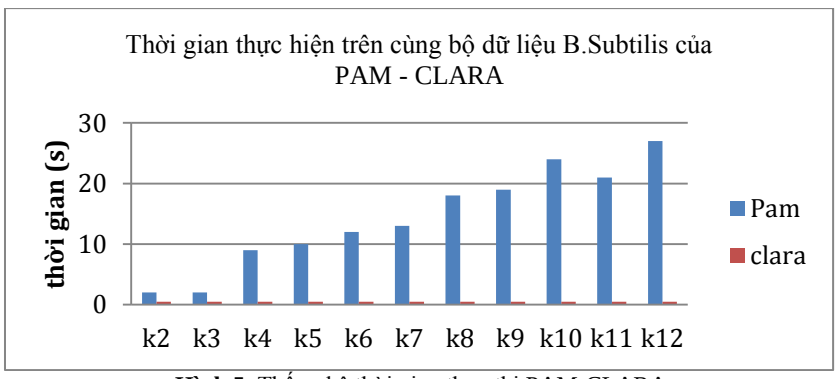
Hình 4. Thống kê chỉ số hàm mục tiêu ứng với số phân nhóm k của hai thuật toán

2. *Thời gian thực thi PAM-CLARA trên bộ dữ liệu B.subtilis*

Tiêu chí thời gian thực nghiệm rất được quan tâm khi chọn lựa các thuật toán áp dụng tính toán với các bộ dữ liệu lớn. Qua thực nghiệm nhận thấy khi áp dụng CLARA trên bộ gen B. Subtilis gần như thực hiện ngay tức thì và đồng đều giữa các tham số k (xem bảng 2). Khi áp dụng PAM lên cùng bộ dữ liệu gen B. Subtilis có sự thay đổi về thời gian thực hiện biến thiên theo xu hướng tăng theo k (xem hình 5).

Bảng 2. Thống kê thời gian thực thi của PAM và CLARA trên cùng tập gen B. Subtilis

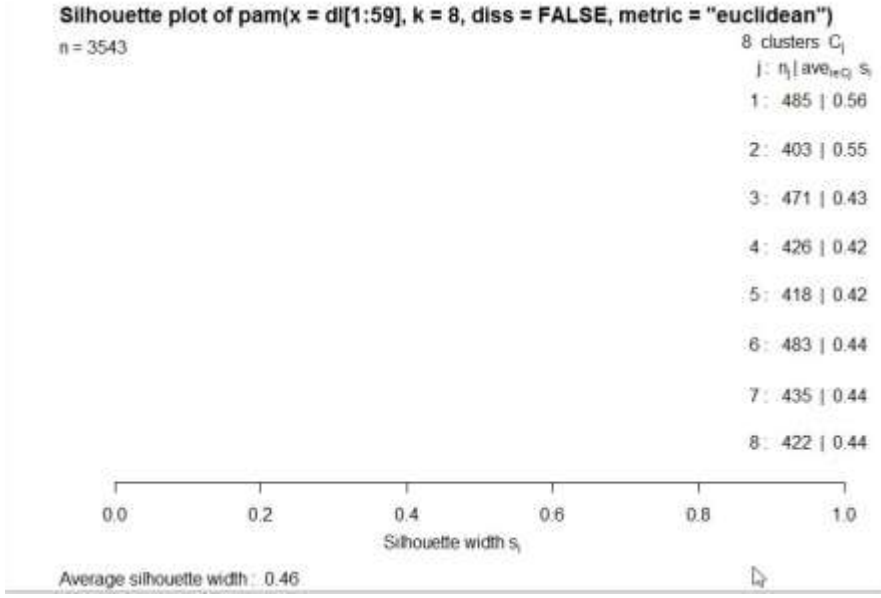
Nhóm	k2	k3	k4	k5	k6	k7	k8	k9	k10	k11	k12
PAM	2	2	9	10	12	13	18	19	24	21	27
CLARA	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5



Hình 5. Thống kê thời gian thực thi PAM-CLARA

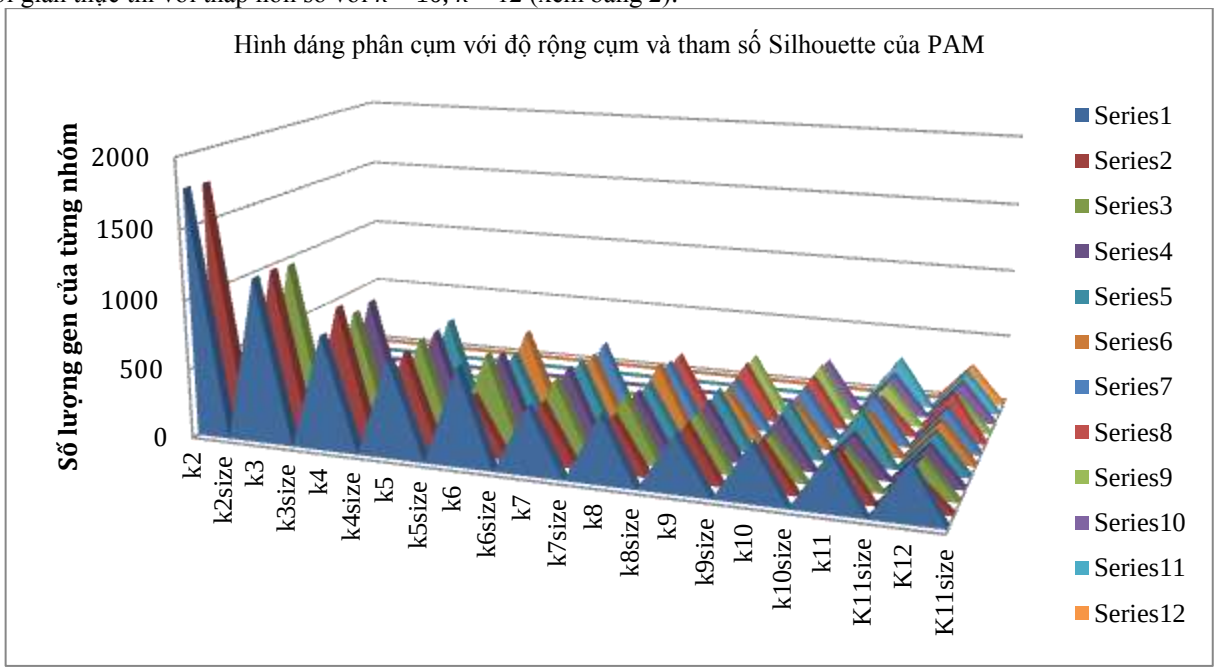
3. Chỉ số Silhouette (Silhouette Index)[4]

Nằm trong nhóm các độ đo đánh giá nội (internal validation measures) [10], với k = 8, ta thấy PAM cho kết quả tốt ở các cụm 1 và 2 với chỉ số silhouette $s(i)$[3] đạt cao nhất có nghĩa là giá trị của các gen trong cụm này tương đồng nhất, có khả năng tìm được HEG ở gần tâm của hai cụm này nhiều nhất (xem hình 6).



Hình 6. Minh họa thông số Silhouette của PAM với k = 8

Với những thông tin trong hình 7, cung cấp góc nhìn tổng quan về hình dạng các cụm qua thông số số lượng gen của các cụm và chỉ số Silhouette trung bình của các cụm với thuật toán PAM. Những thông tin này hỗ trợ việc quyết định chọn giá trị k phù hợp cho việc phân cụm. Theo thực nghiệm cho thấy kết quả phân cụm ổn định với k = 11, thời gian thực thi với thấp hơn so với k = 10, k = 12 (xem bảng 2).



Hình 7. Minh họa hình dáng tổng quan các cụm qua thông số Silhouette của PAM với k=2 đến k=12

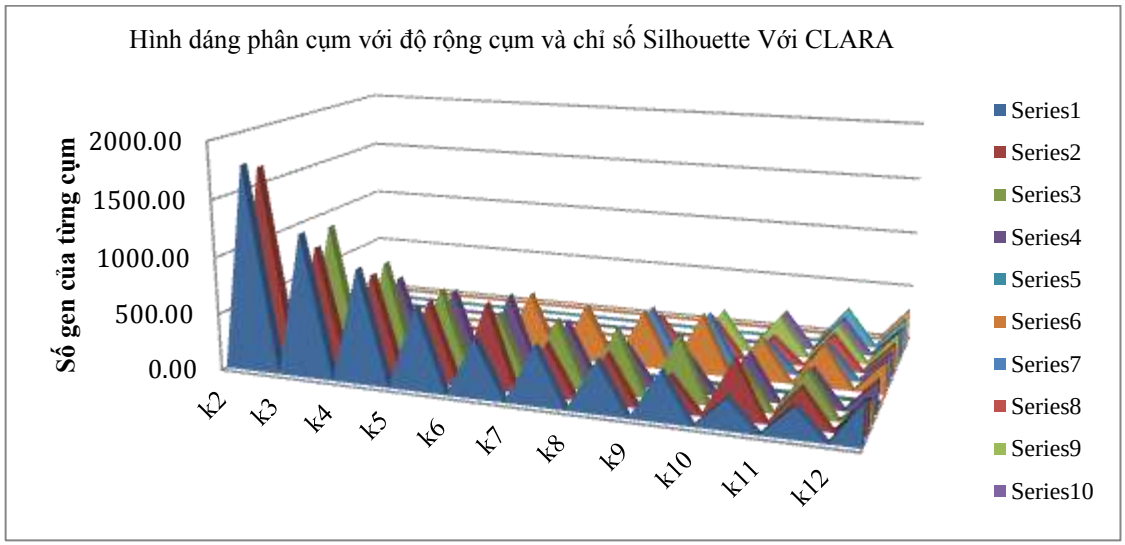
Bài viết cũng tiến hành thống kê tương tự cho thực nghiệm với CLARA và thu được kết quả như hình 8. Với CLARA chất lượng các cụm được cải tiến rõ nét về chất số lượng cho từng cụm và chỉ số Silhouette tốt nhất tập trung ở k = 7, k = 8. Giá trị cao nhất của tham số này đạt cao nhất trong tập các chỉ số Silhouette của cả PAM và CLARA. Bên cạnh đó các HEG tốt nhất được đề xuất bởi CLARA có xu hướng hội tụ với hai giá trị k này (xem hình 9) với số lượng HEG thu được giống nhau hơn 90%.

[3] *s(i) được tính ở công thức 6*

Bảng 3. Thống kê độ rộng của các cụm và chỉ số Silhouette của các cụm với $k = 7$, $k = 8$ trong CLARA

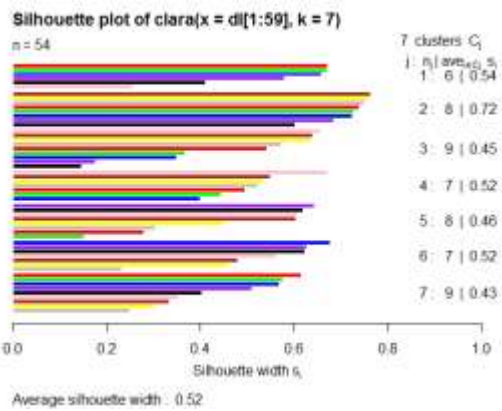
STT	size_k7	Sil_k7	size_k8	Sil_k8
1	509	0.5405087	449	0.6346691
2	451	0.7187359	425	0.6254570
3	584	0.4536771	562	0.5724334
4	490	0.5168400	355	0.5705324
5	469	0.4569732	476	0.5089383
6	490	0.5226611	493	0.3738123
7	50	0.4344639	459	0.4973842
8			324	0.3717753

Hình 8 là hình dạng các cụm qua thông số số lượng gen của các cụm và chỉ số Silhouette của các cụm với thuật toán CLARA. Với thông tin nhận được từ thống kê này ta thấy chất lượng phân cụm với CLARA ổn định với các $k = 6$, $k = 7$, $k = 8$ với số HEG thu được của ba tập này giống nhau hơn 90% và chỉ số Silhouette có giá trị cao nhất.

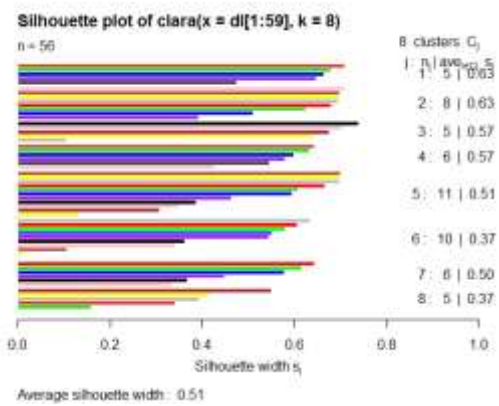


Hình 8. Minh họa hình dáng tổng quan các cụm qua thông số Silhouette của CLARA với k=2 đến k=12

(**a).** HEG với k=7

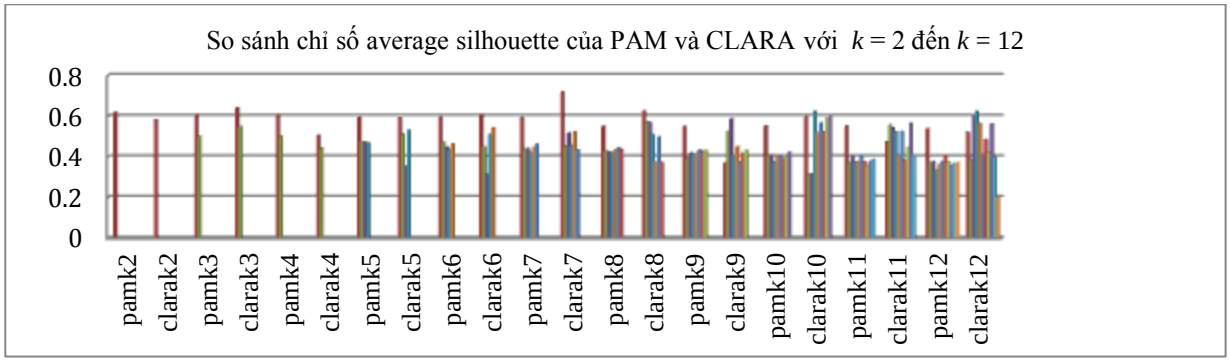


(**b).** HEG với k=8



Hình 9. Hình dáng các gen được dự đoán là HEG tốt nhất với $k = 7$, $k = 8$ bằng thuật toán CLARA

Hình 10 là một minh chứng khác để khẳng định hiệu quả của việc áp dụng thuật toán CLARA so với PAM.



Hình 10. Minh họa hình dáng tổng quan các cụm qua thông số Silhouette của CLARA với k = 2 đến k = 12.

Và theo thống kê từ dữ liệu công bố gen HEG-DB thì tỉ lệ HEG trên bộ gen thường vào khoảng 5% [11]. Như vậy khi áp dụng thuật toán PAM với các tùy chọn k = 2 đến 12, cho chất lượng phân cụm tốt với k = 12. Nhưng thuật toán này không chỉ ra được các giá trị cụ thể có thể là HEG. Và thời gian thực thi trên cùng một dữ liệu B.Subtilus cho kết quả chậm hơn khoảng 20 lần với k = 12 so với thuật toán CLARA. Bên cạnh ưu điểm về tốc độ thực hiện phân cụm CLARA còn đề xuất kết quả mẫu tốt nhất có thể là HEG với k = 7 là 54. Nếu nâng giá trị k lên k = 75, số gen tốt nhất đề nghị là 190 trên tổng số 3543 gen chiếm khoảng 5,3% được cho là phù hợp với số lượng HEG cần tìm.

```
Objective function:      83.84559
Clustering vector:       int [1:3543] 1 2 3 4 5 6 7 3 8 1 9 10 11 9 11 9 9 12 .
..
Cluster sizes:           86 53 60 72 56 66 75 33 65 20 17 53 51 94 75 27 42 76
21 15 25 31 44 48 20 72 70 76 56 46 38 110 14 46 50 74 60 69 26 39 71 27 59 60
99 70 29 54 48 44 11 30 26 41 56 36 13 40 40 10 50 21 26 20 80 46 57 50 5 35 55
 52 76 30 5
Best sample:
  [1]   17   45   51   79   82   84   89  138  152  191  219  242  259  277
 [15]  281  285  287  300  306  329  358  368  397  450  461  473  529  540
 [29]  563  606  633  638  647  649  676  685  712  744  751  769  781  786
 [43]  797  803  837  866  883  897  914  955 1004 1020 1028 1033 1065 1071
 [57] 1081 1089 1139 1153 1159 1172 1174 1195 1196 1205 1206 1210 1235 1244
 [71] 1246 1279 1282 1301 1306 1320 1321 1322 1328 1355 1379 1417 1451 1469
 [85] 1482 1485 1490 1502 1528 1550 1561 1621 1640 1651 1669 1675 1685 1687
 [99] 1739 1774 1786 1790 1820 1829 1832 1838 1841 1937 1944 1959 1975 1983
[113] 2063 2087 2098 2137 2146 2147 2148 2150 2170 2253 2260 2300 2326 2347
[127] 2352 2360 2366 2418 2438 2454 2455 2462 2510 2532 2553 2567 2582 2599
[141] 2621 2626 2645 2669 2692 2710 2717 2762 2765 2784 2785 2836 2838 2898
[155] 2906 2916 2944 2953 2996 2997 3006 3020 3025 3088 3102 3109 3161 3164
[169] 3213 3214 3254 3285 3347 3357 3376 3378 3381 3400 3425 3458 3461 3490
[183] 3492 3497 3521 3524 3528 3536 3540 3543
```

Hình 11. Minh họa kết quả thực nghiệm tìm HEG của CLARA với k=75

V. KẾT LUẬN

Bài báo trình bày các cách thức thiết kế gen tái tổ hợp, việc tìm HEG là một công đoạn quan trọng để thiết kế gen tái tổ hợp đạt hiệu quả cao. Cả ribosomal gen và chaperone protein gen đều là HEG và có xu hướng sử dụng codon cao. Nên việc phân nhóm các loại gen này dựa trên đặc tính sử dụng condon tương đồng RSCU được chọn để việc xác định các gen có khả năng là HEG. Dựa vào tập dữ liệu có thể là HEG này, bài báo áp dụng hai thuật toán phân cụm PAM và CLARA để phân hoạch tìm ra những gen có xu hướng gần gũi nhất để gom vào cụm và từ các cụm này để dự đoán HEG.

Bằng thực nghiệm với k = 10 đến k = 11 bằng PAM thì giá trị về độ lớn giữa các nhóm không thay đổi lớn chỉ dịch chuyển vị trí các mediod khi và hàm mục tiêu giảm chậm. Trong bài viết này chỉ chọn k = 11 cho việc biểu diễn kết quả phân cụm gen.

Đối với CLARA thì gian thực nghiệm trên cùng bộ dữ liệu giảm rõ rệt khoảng 20 lần khi so sánh với PAM cùng chọn giá trị k = 11. Chất lượng phân cụm và thời gian thực thi của CLARA tốt hơn trên cùng tập dữ liệu B.subtilis, CLARA sẽ là thuật toán được khuyến khích nên áp dụng cho bài toán tìm HEG.

VI. TÀI LIỆU THAM KHẢO

[1]. Menzella, H.G., "*Comparison of two codon optimization strategies to enhance recombinant protein production in Escherichia coli"*, Microbial cell factories, 2011.

[2]. The R Development Core Team, "*R: A Language and Environment for Statistical Computing*", 2014.

[3]. Gupta, S., *"Project report Codon optimization"*, 2003.

[4]. Pere Puigbo, E.G., Antoni Romeu1 and Santiago Garcia-Vallve, "*A web server for optimizing the codon usage of DNA sequences"* . Nucleic Acids Research, p. W126–W131, 2007.

[5]. Sharp, P. M, Tuohy, .TM, Mosurski, K. R, "*Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed gene"*, Nucleic Acids Res , 1987.

[6]. Jiawei HanUniver sity of Illinois, Micheline Kamber Jian Pei, "*Data Mining Concepts and Techniques"*., Elsevier, p. 443-494, 2012.

[7]. N. A.CampBell, J.B.R.y., L.A Urry, M. L. C, Rain, S. A.Wasserman, P.V.Minorsky, R.B. Jackson, "*Sinh Học",* GDVN, p. 10-15, 2014.

[8]. Võ Viết Cường , L. T. H., Đỗ Thị Huyền, Lê Quỳnh Giang, Nguyễn Thị Quý, Trương Nam Hải, "B*iểu hiện gen ha5.1 được cải biến mã có hoạt tính sinh học trong nấm men pichia pastoris x3".* Tạp chí sinh học, p. 35, 2013

[9]. A. Carbone, A. Zinovyev, and F. Képès, "*Codon adaptation index as a measure of dominating codon bias"*. Oxford University Press, 2003.

[10]. Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu. "*Understanding of internal clustering validation measures"* In: Proc. of the 2010 IEEE International Conference on Data Mining, pp. 911-916, 2010.

[11]. Puigbò, P., Guzmán, E., Romeu, A. and Garcia-Vallvé, S. "*OPTIMIZER: a web server for optimizing the codon usage of DNA sequences*", Nucleic Acids Research, 35(suppl 2), W126–W131,(2007).

PREDICTING HIGH EXPRESSION GENE FOR RECOMBINANT DNA DESIGN

Duong Thi Kim Chi, Tran Van Lang, Huynh Xuan Hiep

ABSTRACT*— Predicting high expression gene HEG (Highly Expressed Gene) is an important step in finding the optimal gene recombination process. The high expression gene in normal cells tend to have similar characteristics, mainly featured on codon usage trends. This article proposes a new approach to clustering application data to identify groups of genes with similar characteristics on codon usage trends to predict HEG. The experiment was deployed on two algorithms PAM (Partitioning Around Medoids), CLARA (Clustering for Large Applications) for clustering predicted HEG. The results showed that beter CLARA than PAM on time, the quality of clustering.*