Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)"; Cần Thơ, ngày 4-5/8/2016
DOI: 10.15625/vap.2016.000103

XÁC ĐỊNH THAM SỐ QUAN TRỌNG CHO VIỆC THIẾT KẾ GEN DÙNG TRONG TÁI TỔ HỢP

Dương Thị Kim Chi[1], Trần Văn Lăng[2,3,*], Lê Mậu Long[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, Trường Đại học Nguyễn Tất Thành

chidtk@tdmu.edu.vn, langtv@vast.vn, lmlong@ntt.edu.vn

TÓM TẮT— *Việc sử dụng các gen tự nhiên trong quá trình sản xuất các sản phẩm tái tổ hợp dùng trong y học, dược học, hay cải tạo giống cây trồng trong nông nghiệp thường cho kết quả biểu hiện thấp. Việc thiết kế gen hay tối ưu hóa gen đã được tiến hành nghiên cứu của nhóm Hugo G Menzella năm 2011. Vào năm 2014, nhóm Agnieszka Zylicz-Stachula năm 2014 chứng minh khả năng gia tăng mức độ biểu hiện của gen mục tiêu sau khi tối ưu hóa so với gen tự nhiên ban đầu. Bài viết này trình bày kết quả khảo sát các tham số ảnh hưởng đến tối ưu hóa gen từ các chương trình đang được sử dụng hiện nay như: Eugene, GeneOptimizer, VisualGeneDeveloper, OptimumGene. Đánh giá các tiêu chí tối ưu hóa của chương trình giữa gen tự nhiên và gen tối ưu hóa trên cùng bộ gen MHA5. Các kết quả thống kê được dùng để xác định các tham số quan trọng cho việc thiết kế gen dùng trong tái tổ hợp.*

Từ khóa— *Tối ưu, sinh tin học*

I. GIỚI THIỆU

Phân tử DNA tái tổ hợp (recombinant DNA technology) đầu tiên trong ống nghiệm (*in vitro*) được ra đời từ những năm đầu của thập niên 1970; đó là cơ sở cho sự ra đời của công nghệ sinh học hiện đại: *kỹ thuật di truyền* (genetic engineering) [1]. Sự ra đời và phát triển nhanh chóng của lĩnh vực này không những đã đưa lại sự hiểu biết sâu sắc về cấu trúc và các cơ chế hoạt động của các gen và bộ gen; mà còn trở thành lực lượng sản xuất trực tiếp của xã hội, góp phần giải quyết những vấn đề thực tiễn đặt ra trong y dược học, nông nghiệp và môi trường.

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 *vector tạo dòng* (plasmid) và đưa vào một vật chủ; ở đó đoạn gen này được dịch mã thành một protein đặc biệt [2], protein đó được gọi là protein tái tổ hợp. Khi tuyển chọn gen tự nhiên vào quá trình sản xuất thường cho kết quả biểu hiện thấp vì các gen khi đưa vào hệ thống biểu hiện sẽ có thể xuất hiện sự không tương thích về xu hướng sử dụng codon hay thành phần GC của gen, trình tự lặp lại. Từ đó làm giảm khả năng biểu hiện ra protein mục tiêu. Chọn lựa một gen tốt cho việc sản xuất sẽ làm gia tăng biểu hiện ra protein mục tiêu, điều này đã được nghiên cứu của nhóm *Hugo G Menzella* và cộng sự năm 2011 [3] hay *AgnieszkaZylicz-Stachula* và cộng sự năm 2014 [4]..

Việc thiết kế lại gen tự nhiên hay tối ưu hóa gen dựa trên cơ sở đánh giá các tiêu chí sinh học sẽ làm nâng cao biểu hiện gen mục tiêu [5]. Đã có nhiều phần mềm hỗ trợ cho nhà sinh học việc tối ưu hóa gen này. Nhìn chung các phần mềm dựa trên một số nghiên cứu của các nhà sinh học để chọn lựa các tiêu chí cho việc xây dựng chương trình tối ưu hóa gen. Có ba phương pháp tối ưu hóa gen được các phần mềm này áp dụng như sau:

- Nhóm giải pháp ***Một amino acid – một codon (One amino acid – one codon)*:** Đây là phương pháp được phát triển sớm nhất. Phương pháp này sử dụng codon ưa thích nhất cho mỗi amino acid dựa vào bảng thống kê xu hướng sử dụng codon cho mỗi loài. Từ trình tự amino acid của protein mục tiêu, chương trình sẽ thay thế amino acid bằng codon ưa thích tương ứng. Phương pháp được chương trình GenOptimizer áp dụng dựa trên giải pháp của Puigbò P., Guzmán E.Romeu A. and Garcia-Vallvé S. 2007 [6].
- Nhóm giải pháp ***Một amino acid – nhiều codon (One amino acid – one randomization)*:** phương pháp này xét tất cả các codon có thể mã hóa cho amino acid tương ứng trong trình tự, kết hợp với các tiêu chí khác như %GC, trình tự nhận biết của enzyme cắt giới hạn, trình tự lặp lại, ... để từ đó có thể mã hóa ra trình tự protein mục tiêu. Phương pháp này sử dụng hàm mục tiêu để tìm ra các gen tối ưu. Đại diện cho phương pháp này là chương trình Eugene dựa trên đề xuất của Paulo Gaspar [3, 7].
- ***Phương pháp kết hợp* (*Hybrid construct*):** Đây là phương pháp kết hợp từ hai phương pháp trên, chỉ xét các amino acid được mã hóa bởi các codon có tần suất sử dụng cao; về nguyên tắc có thể rút ngắn thời gian xử lý nhưng có thể bỏ qua một số trình tự tốt. Chương trình áp dụng phương pháp này là DNA Words, sử dụng giải pháp của Hoover and Lubkowski [8].

Xu hướng ứng dụng khoa học tính toán hỗ trợ công việc thiết kế gen cho sản xuất protein tái tổ hợp đang được các nhóm nghiên cứu và các công ty về công nghệ sinh học rất quan tâm. Các sản phẩm phần mềm này thường được hỗ trợ miễn phí trên các website hay các phần mềm ứng dụng. Những người nghiên cứu sinh học khi cần tối ưu hóa gene để nâng cao khả năng biểu hiện protein tái tổ hợp đều sử dụng các chương trình tối ưu hóa gene đã phát triển trên thế giới như GeneOptimizer, OptimunGene hay Eugene, hoặc đặt mua các gen đã được tối ưu hóa từ công ty sinh học.

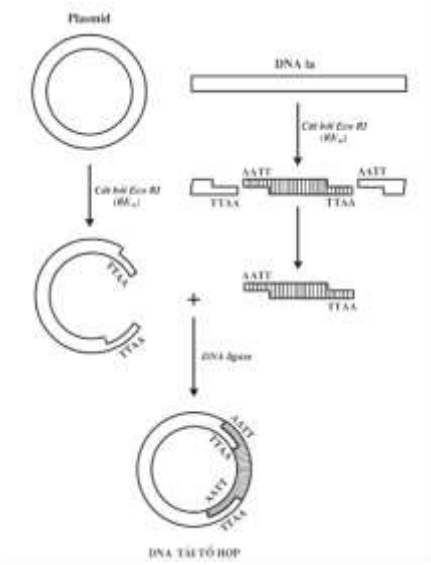
Trong bài báo này với mục đích khảo sát các phần mềm thiết kế gen Eugene, GeneOptimizer, Visual Gene Developer, OptimumGene. Tiêu chí khảo sát đó là quy trình thiết kế, tham số được sử dụng cho tối ưu hóa gen, thuật toán áp dụng, cũng như công nghệ thiết kế. Các kết quả thống kê sẽ được so sánh với sản phẩm Gen *mha*5.1 – gen tái tổ hợp được tổng hợp từ hãng (Genscript), qua đó xác định các tham số quan trọng cho việc thiết kế gen tái tổ hợp.

Các phần còn lại của bài báo được trình bày theo các nội dung như sau: phần 2 đưa ra phân tích và khảo sát tham số ảnh hưởng, phần 3 trình bày kết quả thực nghiệm Gen ha5 trên từng phần mềm thiết kế gen và so sánh với gen mha5.1, và phần cuối cùng là kết luận.

A. *Bài toán tìm các tham số quan trọng cho thiết kế gen tái tổ hợp*

a) 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 vector tách dòng. Những vectơ tách dòng mang DNA tái tổ hợp này có thể biểu hiện thành các protein tái tổ hợp trong các sinh vật như *Hình* 1 [9].



Hình 1. Minh họa hình thành DNA tái tổ hợp từ Plasmid (vectơ tách dòng) và DNA lạ

2. Codon đồng nghĩa (Synonymous Condon)

Cơ sở khoa học của việc tối ưu hóa gen dựa trên hiện tượng codon đồng nghĩa. Một codon gồm ba nucleotide sẽ có $4^3 = 64$ loại codon. Tuy nhiên 64 loại này nhưng do chỉ có 20 amino acid, vì vậy một amino acid có thể được mã hóa bởi ít nhất hai loại codon khác nhau, từ đó dẫn đến các codon đồng nghĩa [11].

Ví dụ: Như minh họa ở Hình 2 - được đóng khung, Amino acid *Ala (hay A)* có bốn codon đồng nghĩa là GCA, GCC, GCG, GCU.

	CGU									UUA					UCU					
	CGC									UUG					UCC					
GCU	CGA						GGU			CUU				CCU	UCA	ACU			GUU	
GCC	CGG						GGC		AUU	CUC				CCC	UCG	ACC			GUC	UAA
GCA	AGA	AAU	GAU	UGU	CAA	GAA	GGA	CAU	AUC	CUA	AAA		UUU	CCA	AGU	ACA		UAU	GUA	UAG
GCG	AGG	AAC	GAC	UGC	CAG	GAG	GGG	CAC	AUA	CUG	AAG	AUG	UUC	CCG	AGC	ACG	UGG	UAC	GUG	UGA
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	
A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	

Hình 2. Minh họa các codon đồng nghĩa.

B. *Hàm mục tiêu*

Một protein được cấu tạo từ gen đồng nghĩa được tạo từ nhiều amino acid, có thể được mã hóa bởi nhiều gen khác nhau, các gen này được gọi là *gen đồng nghĩa*. Việc tối ưu hóa gen sẽ lựa chọn gen đồng nghĩa tốt nhất hay gen mang lại hiệu suất biểu hiện cao nhất thỏa mãn các yêu cầu về tối ưu. Tập hợp các yêu cầu về tối ưu được gọi là *tham số quan trọng cho quá trình thiết kế gen* và đây cũng chính là các tham số được dùng cho hàm mục tiêu của bài toán tối ưu hóa. Các chương trình thiết gen hay tối ưu hóa gen hiện nay cũng tập trung xây dựng và cải tiến hàm mục tiêu này.

Các cách xây dựng hàm mục tiêu của các chương trình tối ưu hóa gen hiện nay dựa vào ba cách tiếp cận về thiết kế gen như đã nêu ở trên để xây dựng thuật toán cho hàm tính điểm cho chương trình của mình.

1. Các hàm tính điểm riêng lẻ

Trong phương pháp này, đa số các tiêu chí được xem xét bằng cách so sánh với một giá trị ngưỡng cho trước. Nếu trình tự xem xét thỏa mãn điều kiện sẽ được thu nhận, ngược lại sẽ bị loại bỏ. Bên cạnh đó, những tiêu chí không xác định được ngưỡng loại bỏ có thể dùng hàm tính điểm riêng lẻ để xếp thứ tự của các trình tự xem xét. Trình tự thực hiện của thuật toán sẽ xem xét lần lượt từng tiêu chí theo thứ tự được quy định trước tuỳ theo mức độ quan trọng của các tiêu chí này. Chương trình Visual Gene Developer áp dụng phương pháp này để thiết kế gen tối ưu hóa.

2. Gọi các tiêu chí lần lượt là:

3. x_1 : giá trị độ thích nghi tương đối (Wc) của từng codon.

4. x_2 : là khoảng giá trị %GC min-max.

5. x_3 : chiều dài trình tự lặp lại cùng chiều và ngược chiều tối thiểu

6. …

x_n : giá trị tiêu chí *n*

Hàm tính điểm riêng lẻ có dạng

$$f(x) = f(x) = x_1 + x_2 + \cdots + x_n \quad \rightarrow Max(Min), \tag{1.1}$$

với các điều kiện ràng buộc

$$\begin{cases} x_1 & \leq & b_1 \\ x_2 & \leq & b_2 \\ \cdots & \cdots & \cdots \\ x_n & \leq & b_n \\ x_1, x_2, .. x_n \geq 0 \end{cases}$$

7. Hàm tính điểm tuyến tính

Hàm mục tiêu được xây dựng với mục tiêu tối ưu hoá gen theo nhiều tiêu chí, các chương trình có thể lựa chọn tối ưu đồng thời hay tối ưu lần lượt các tiêu chí này. Khi chọn tối ưu hoá đồng thời, thuật toán cần xây dựng một hàm tính điểm tổng để có đánh giá tổng hợp về tất cả các tiêu chí cần xem xét.

Các chương trình tối ưu hoá gen hiện nay như GeneOptimizer sử dụng hàm tính điểm có dạng tổ hợp tuyến tính của các hàm tính điểm cho từng tiêu chí thành phần. Để thể hiện mức độ ảnh hưởng, hay còn gọi là độ ưu tiên cho từng tiêu chí, hệ số đứng trước từng tiêu chí trong hàm tuyến tính sẽ được chương trình hoặc người dùng lựa chọn để tìm kiếm kết quả phù hợp yêu cầu. Chương trình GeneOptimizer sử dụng hàm tính điểm dạng này.

Ký hiệu các tham số quan trọng:

x_1: Giá trị độ thích nghi tương đối (Wc) của từng codon:

x_2: Khoảng giá trị %GC min-max:

x_3: Chiều dài trình tự lặp lại cùng chiều và ngược chiều tối thiểu:

…

x_n : giá trị tiêu chí *n*
$c_1, c_2, c_3 \ldots c_n$: là các hệ số cho từng tiêu chí
Hàm tính điểm tuyến tính [12] có dạng như sau:

$$f(x) = cf(x) = c_1x_1 + c_2x_2 + \cdots + c_nx_n \quad \rightarrow Max(Min), \tag{1.2}$$

với các điều kiện ràng buộc

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n & \leq & b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n & \leq & b_2 \\ \cdots & \cdots & \cdots \\ a_{mn}x_1 + a_{mn}x_2 + \cdots + a_{mn}x_n & \leq & b_m \\ x_1, x_2, .. x_n \geq 0 \end{cases}$$

8. Hàm tính điểm phi tuyến

Cơ chế ảnh hưởng của các yếu tố lên sự biểu hiện protein mục tiêu rất phức tạp. Các yếu tố này có thể ảnh hưởng (hỗ trợ hoặc xung đột) lẫn nhau nên cần một dạng hàm tính điểm khác hơn để biểu thị. Tùy vào thuật toán áp

dụng của từng phần mềm mà chọn hàm mục tiêu này dạng, như phần mềm EuGene sử dụng thuật giải di truyền (Genetic Algorithm) để dự đoán gen tối ưu [13].

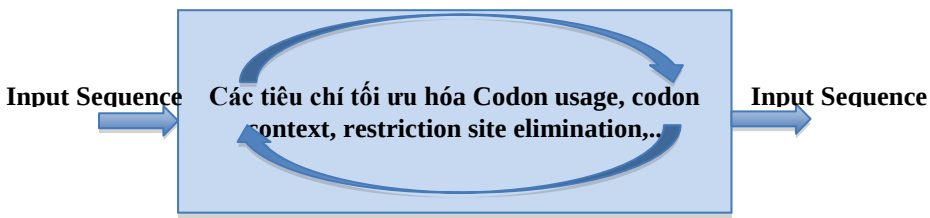
II. PHÂN TÍCH CÁC TIÊU CHÍ TỐI ƯU HÓA GEN CỦA MỘT SỐ PHẦN MỀM

Có nhiều phần mềm hỗ trợ công cụ tối ưu hóa gen đang được sử dụng như: Visual Gen Developer, OPTIMIZER, OptimumGene, EUGene, DNAWork, Jcat, Synthetic gene designer, GeneDesign, Gene Designer 2.0, mRNA Optimizer, COOL, D-Tailor, … Các phần mềm này thường cung cấp miễn phí và có hai hình thức người dùng có thể sử dụng dạng website hoặc dạng ứng dụng. Các tham số đầu vào cho chương trình được sử dụng như thế nào có thể tùy theo yêu cầu của người sử dụng hoặc không cần cung cấp. Mỗi phần mềm đều có những đặc trưng riêng và có tiêu chí riêng về tối ưu hóa gen. Nhìn chung các phần mềm đều có những đặc điểm chung như: ngôn ngữ thiết kế cho phép nâng cấp hay không, các thuật toán áp dụng, quy trình thực hiện, trong phạm vi bài viết này chúng tôi trình bày các thống kê về đặc điểm chung này của các nhóm phần mềm đại diện.

A. Phần mềm Eugene

Eugene [13] là một chương trình ứng dụng kết hợp nhiều thuật toán phục vụ cho việc tối ưu hóa gene được phát triển bởi Paulo Gaspar cùng các đồng sự và công bố năm 2012. Chức năng chính của chương trình là phân tích và thiết kế lại gen sử dụng nhiều phương pháp tối ưu hóa nhằm mục tiêu tăng tối đa hiệu quả mã hóa của gen. Sử dụng kết hợp 2 thuật toán: thuật toán mô phỏng luyện kim (Simulated Annealing Alogorithm) và thuật toán di truyền (Genetic Algorithm). Eugene thực hiện tối ưu hóa đa tiêu chí dựa trên một số tiêu chí như xu hướng sử dụng codon (codon usage), thành phần codon (codon context), số phần trăm GC (%GC), mã kết thúc ẩn (Hidden Stop Codons), trình tự lặp lại, trình tự Shine – Dalgarno.

Đây là phần mềm thuộc dạng đóng gói, chọn lựa một hay nhiều tiêu chí tối ưu hóa gen và chờ kết quả hiển thị kết quả của phần mềm với thiết kế tối ưu hóa đa mục tiêu nên khi chọn càng nhiều tiêu chí thì thời gian thực thi càng lớn. Sau đây hoạt động tổng quát của Eugene. Sơ đồ như Hình 3.

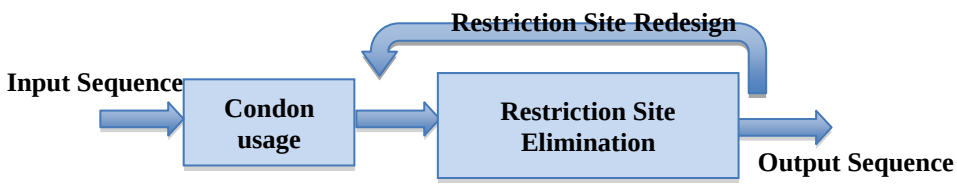


Hình 3. Sơ đồ hoạt động nhóm phần mềm đóng gói dạng đóng gói Eugene

B. GeneOptimizer

GeneOptimizer [8] là một website lớn lưu trữ các cơ sở dữ liệu quan trọng như cơ sở dữ liệu các gen biểu hiện cao (HEG Database), cơ sở dữ liệu xu hướng sử dụng codon của hơn 150 hệ thống biểu cùng với nhiều công cụ tính toán trong tối ưu hóa gene như CAIcal, E-CAI, Optimizer [8]. Chương trình tối ưu hóa GeneOptimizer cung cấp cho người dùng nhiều phương pháp tối ưu hóa gen: một amino acid - một codon, một amino acid – nhiều codon; trong đó sử dụng Phương pháp Monte Carlo. Các tiêu chí như xu hướng sử dụng codon, trình tự nhận biết của enzyme cắt giới hạn, %GC cũng được chương trình đưa ra cho người dùng tùy chọn trong quá trình tối ưu hóa.

Sơ đồ hoạt động của GeneOptimizer được mô tả như Hình 4: từ trình tự ban đầu chương trình tính các chỉ số về sử dụng codon, có thể chọn lựa các phương án tối ưu hóa một amino acid-một codon, một amino acid – nhiều codon mà tính lại và cập nhật kết quả thiết kế gen.

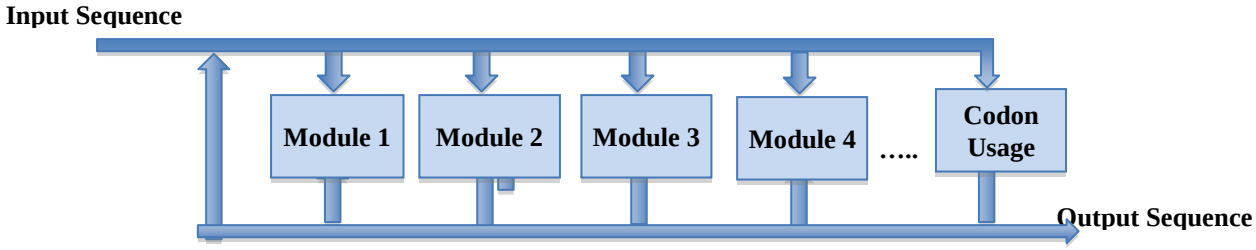


Hình 4. Sơ đồ hoạt động nhóm phần mềm GeneOptimizer

C. Visual Gene Developer

Visual Gene Developer[14] là phần mềm thiết kế chuyên ngành gen có nhiều chức năng để phân tích, thiết kế và tối ưu hóa gen. Ban đầu, phần mềm đã được phát triển để tối ưu hóa chuỗi DNA (chủ yếu là để tối ưu hóa codon) của các gen mục tiêu đã được nâng cấp để có các gói phần mềm nói chung kể từ năm 2008. Để tận dụng các công nghệ lập trình mới nhất, phần mềm áp dụng ngôn ngữ lập trình Visual Studio .Net FrameWork và thiết kế lại tất cả các mã nguồn của chương trình để có thêm tính năng mới như phát triển module người dùng, thuật toán di truyền được sử dụng. Visual Gene Developer áp dụng hàm tính điểm là riêng lẻ và tính hàm tính điểm phi tuyến để dự đoán gen tối ưu hóa.

Với quy tắc nhập vào trình tự một lần, người dùng chọn các module tính toán phù hợp, phần mềm sẽ tính toán và trả về kết quả gen tối ưu phù hợp (Hình 5).



Hình 5. Sơ đồ hoạt động phần mềm Visual Gene Deverloper

D. *OptimumGene*

OptimumGene [15] là phần mềm của công ty hàng đầu thế giới về dịch vụ tổng hợp gen. Các thuật toán OptimumGene đưa vào xem xét một loạt các yếu tố quan trọng liên quan đến giai đoạn khác nhau của biểu hiện protein, chẳng hạn như khả năng thích ứng codon, cấu trúc mRNA, và yếu tố phiên mã và dịch mã.

Quy trình hoạt động của phần mềm khá đơn giản như hình 6, chỉ cần nhập vào đoạn gen cần tối ưu, phần mềm sẽ tính toán và trả về kết quả trình tự mong muốn của gen.

Hình 6. Sơ đồ hoạt động phần mềm OptimumGene

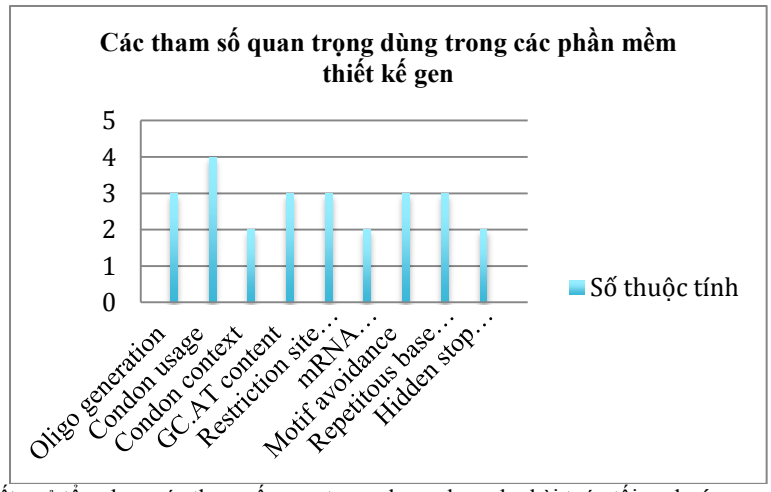
E. *Xác định các kết quả hệ thống tham số dựa trên kết quả thống kê*

Các tham số được sử dụng trong các phần mềm được mô tả về giá trị sinh học như sau [1]:

- Oligo generation: các đoạn trình tự nhỏ, chỉ số này cho biết trong bộ gene có chứa các codon hiếm hoặc ít được "ưa thích" đối sẽ làm làm giảm hiệu suất của hệ thống biểu hiện.
- Condon usage: xu hướng sử dụng codon. Chỉ số thích nghi codon (Codon Adaptation index – CAI, giá trị dao động từ 0 đến 1. Giá trị CAI = 1 chỉ các gene có xu hướng sử dụng các codon phổ biến nhất đối với hệ thống biểu hiện và giá trị CAI = 0 chỉ các gene sử dụng các codon không được dùng trong hệ thống biểu hiện.
- Condon context: thành phần codon.
- GC.AT content: Hàm lượng GC phân bố dọc theo chiều dài của gen. Tỉ lệ GC cao thì lực liên kết giữa hai mạch sẽ lớn, gây khó khăn cho sự cắt đứt liên kết trong quá trình tách mạch.
- Restriction site manipulation: bảng ghi nhận các vị trí thay đổi vùng trình tự.
- mRNA secondary structure: quyết định tính bền của phân tử, còn ảnh hưởng trực tiếp đến khả năng tham gia vào quá trình dịch mã tạo protein.
- Motif avoidance: loại bỏ motif trong trình tự.
- Repetitous base removal: trình tự lặp lại.
- Hidden stop codons: mã kết thúc ẩn, Việc xuất hiện các mã kết thúc này giúp quá trình dịch mã tránh được các sản phẩm dịch mã lệch khung.

Bảng 1. Tham số tham số được dùng trong các phần mềm thống kê

Tham số / Tên phần mềm	Oligo generation	Condon usage	Condon context	GC.AT content	Restriction site manipulation	mRNA secondary structure	Motif avoidance	Repetitous base removal	Hidden stop codons
Optimizer	x	x		x	x		x		
Visual Gene Developer	x	x			x	x	x	x	x
EUgene	x	x	x	x	x			x	x
OptimumGene		x	x	x		x	x		



Hình 7. Kết quả tổng hợp các tham số quan trọng được chọn cho bài toán tối ưu hoá gen có tái tổ hợp

Bảng 2. Các thông số kỹ thuật và địa chỉ tải các phần mềm

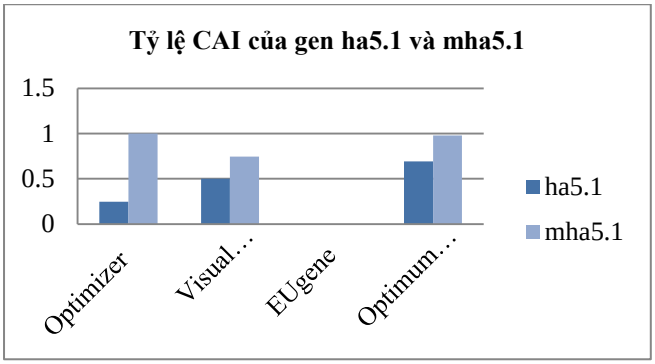
Tên chương trình	Địa chỉ tải và sử dụng phần mềm	Các thuật toán tham khảo chính	Ngôn ngữ lập trình	Khả năng nâng cấp người dùng
Optimizer	http://genomes.urv.cat/OPTIMIZER/	Puig bò et al (2007), Monte Carlo	PHP	Không
Visual Gene Developer	http://visualgenedeveloper.net	Jung and McDonald (2011)	.Net Framework	Có khả năng nâng cấp một số modules, nhưng không chỉnh sửa những module đã thiết kế
EUgene	http://bioinformatics.ua.pt/eugene/	Gaspar et al (2003)	Java	Không
OptimumGene	http://www.genscript.com/codon-opt.html	OptimumGene™ (2011)	PHP	Không

III. KẾT QUẢ THỰC NGHIỆM

Trong bài viết sử dụng gen *ha*5.1 (mã số AJ867074) từ ngân hàng gen NCBI, dùng trình tự gen này kiểm tra cho các phần mềm đã nêu trên và thu nhận các bảng so sánh về các tham số như: CAI; hàm lượng GC; vị trí tương đối các nucleotide; tỷ lệ phần trăm mã bộ ba. Gen tối ưu được thiết kế lại được gọi là mha5.1 do hãng Genscript tổng hợp. Các chỉ số này đã kiểm định bằng thực nghiệm của nhóm nghiên cứu Võ Việt Cường, Lê Thị Huệ, Đỗ Thị Huyền, Lê Quỳnh Giang, Nguyễn Thị Quý, Trương Nam Hải đã chứng minh rằng gen cải biến mha5.1 có khả năng biểu hiện tốt hơn gen ha5.1[10]. Kết quả kiểm định của các phần mềm trên từng tiêu chí so sánh với hai gen *ha5.1, mha5.1*:

A. Tham số codon usage - Chỉ số thích nghi codon- CAI

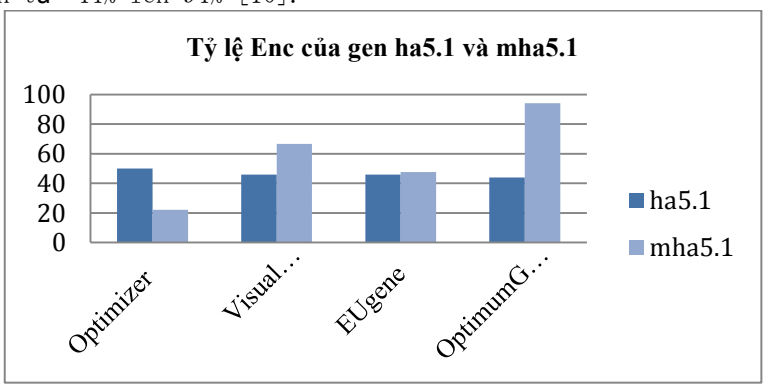
Gen *ha*5.1 trước cải biến có chỉ số CAI thấp, CAI = 0,69, Sau khi cải biến mã, gen *mha*5.1 có chỉ số phù hợp mã CAI đạt 0,98 [10]. Các phần mềm được kiểm tra trên cùng gen ha5.1 cho kết quả cũng khác biệt và trường hợp phần mềm Eugene chỉ dùng chỉ số codon đồng nghĩa - RSCU (Relative Synonysmous Codon Usage) để áp dụng cho tham số *codon usage* nên giá trị của phần mềm này là 0 cho biểu đồ thống kê. Do áp dụng thuật toán đơn giản nên phần mềm Optimizer cho kết quả cao nhất 100% và thấp nhất là Visual Gene Deverloper 66,7%,



Hình 8. Kết quả kiểm tra của các phần mềm khi so sánh tham số *codon usage*- CAI

B. *Tham số sử dụng condon đồng nghĩa Enc- Efective Number of Codons.*

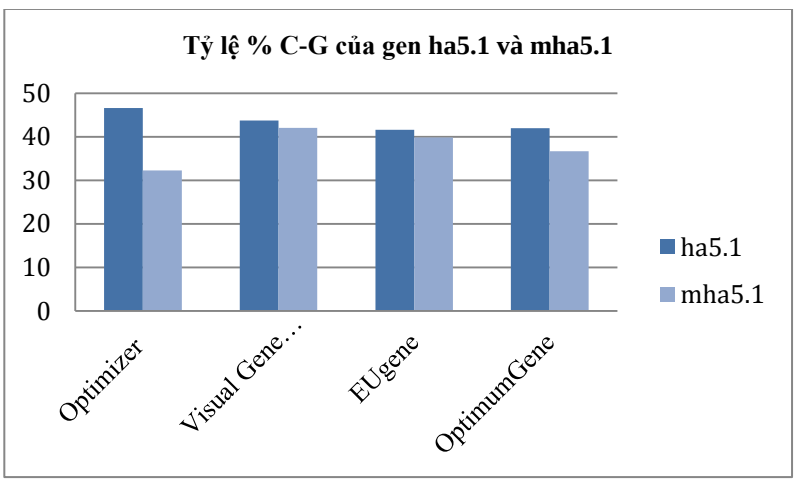
Phần mềm được thiết kế theo phương pháp *Một amino acid – một codon* như Optimizer cho kết quả tỷ lệ này ở gen tối ưu thấp hơn so với gen tự nhiên. Các phần mềm sử dụng *Một amino acid – nhiều codon* sẽ cho tỷ lệ này cao hơn ở gen được thiết kế. Đặc biệt kết quả Enc của phần mềm Optimungene tăng mạnh từ 44% lên 94% [10].



Hình 9. Kết quả kiểm tra của các phần mềm khi so sánh tham số Enc

C. *Tham số GC.AT content*

Hàm lượng GC cao và phân bố không đều dễ tạo cấu trúc thứ cấp, như cấu trúc kẹp tóc. Cấu trúc thứ cấp được hình thành hay mất đi gần vùng mRNA không dịch mã và gần mã bộ ba khởi đầu có ảnh hưởng đến tốc độ suy thoái mRNA và ảnh hưởng tới khởi đầu dịch mã. Hàm lượng GC trung bình giảm xuống của gen sau khi tối ưu của các phần mềm đều giảm xuống.



Hình 10. Kết quả kiểm tra của các phần mềm khi so sánh tham số GC.AT content

Theo kết quả phân tích của nhóm tác giả khảo sát [10] thì "sau khi cải biến mã, gen *mha*5.1 có chỉ số phù hợp mã CAI để biểu hiện trong *P. pastoris* đạt 0,98. Các mã bộ ba đều có tần số sử dụng ở mức cao trên 60% và không còn các mã bộ ba hiếm. Hàm lượng GC trung bình giảm xuống còn 36,64%, lượng GC phân bố đều hơn so với trước cải biến. Đặc biệt tần suất sử dụng các mã bộ ba ở mức 91-100% tăng mạnh từ 44% lên 94%, còn lại 2% các mã bộ bap hân bố ở

mức 51-60%, 2% phân bố ở mức 71-80% . Trình tự nucleotide gen *ha*5.1 và *mha*5.1 trước và sau cải biến có độ tương đồng 77%. Trình tự axit amin do hai gen *ha*5.1 và *mha*5.1 mã hóa có độ tương đồng100%. Kết quả nổi bật của nghiên cứu là gen sau cải biến đã được biểu hiện ra protein có hoạt tính HA cao mà chúng tôi không thấy ở dòng gen chưa được cải biến"[10].

Như vậy phần mềm OptimunGene với các tham số *Condon usage, Condon context GC.AT, content mRNA, secondary structure, Motif avoidance*, kết hợp hợp hàm tính điểm đa điểm cho quá trình tối ưu hóa gen nên cho kết quả gen sửa đổi có chất lượng tương đối tốt và rất hữu ích cho các nhà sinh học khi thực nghiệm. Tuy nhiên đây là một phần mềm có trả phí khi áp dụng tối ưu hóa gen có trình tự lớn. Việc khảo sát các phần mềm hỗ trợ cho thiết kế gen để chọn lựa các tiêu chí quan trọng Oligo generation Condon usage, *Condon context, GC.AT content, Restriction site manipulation, mRNA secondary structureMotif avoidance, Repetitous base removal , Hidden stop codons* và các giá trị ràng buộc cho các tham số này được dùng xây dựng hàm mục tiêu và áp dụng các giải thuật phù hợp để xây dựng phần mềm thiết kế gen tái tổ hợp.

IV. KẾT LUẬN

Bài báo trình bày các cách thức thiết kế gen tái tổ hợp cũng như cách tiếp cận các phần mềm có hỗ trợ thiết kế gen và các kết quả khảo sát về công nghệ thiết kế các phần mềm này, quy trình thực hiện tối ưu hóa gen, giải thuật hay phương pháp tiếp cận hàm mục tiêu, thống kê so sánh hiệu quả của gen tái tổ hợp được áp dụng thực nghiệm trên gen ha5.1. Các kết quả bước đầu đã khẳng định các tham số xác định trong bài viết này là những tham số quan trọng cho quá trình thiết kế gen cho tái tổ hợp.

TÀI LIỆU THAM KHẢO

[1]. Nguyễn Hoàng Lộc, L. V. D., Trần Quốc Dung, "*Giáo Trình Công nghệ DNA tái Tổ hợp.* ĐH Quốc gia, TP Hồ Chí Minh", 2007.

[2]. Hoàng Trọng Phán, Trương Thị Bích Phượng., "*Giáo trình Di truyền học, vi sinh vật và ứng dụng"*. ĐH Huế, 2008.

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

[4]. Agnieszka Zylicz-Stachula, O.Z., Katarzyna Sliwinska, Joanna Jezewska-Frackowiak and a.P.M. Skowron, "*Modified 'one amino acid-one codon' engineering of high GC content TaqII-coding gene from thermophilic Thermus aquaticus results in radical expression increase"*, Microbial Cell Factories, 2014

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

[6]. 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.

[7]. Gaspar, P., J. Carbonell, and J. L. Oliveira, "*On the parameter optimization of Support Vector Machines for binary classification",* J Integr Bioinform, **9**(3): p. 201, 2012.

[8]. Hoover, D. M. and J. Lubkowski, "*DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis"*. Nucleic Acids Research, **30**(10): p. e43-e43, 2002.

[9]. 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.

[10]. 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 tron nấm men pichia pastoris x3".* tạp chí sinh học, p. 35, 2013

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

[12]. Thanh, N. H., "*Giáo trình tối ưu hóa"*. Nhà xuất bản Bách khoa - Hà Nội, p. 16, 2006.

[13]. Paulo Gaspar1, Jose´ Lui´s Oliveira , Jo¨rg Frommlet, Manuel A.S. Santos and Gabriela Moura, "*EuGene: maximizing synthetic gene design for heterologous expression",* Bioinformatics applications note, 2012.

[14]. Jung, S.-K. and K. McDonald (). *"Visual Gene Developer: a fully programmable bioinformatics software for synthetic gene optimization",* BMC Bioinformatics 12(1): 340, 2011.

[15]. D H Aggen, A S Chervin, T M Schmitt, B Engels, J D Stone, S A Richman, K H Piepenbrink, B M Baker, P D Greenberg, H Schreiber andD M Kranz. "*Single-chain VαVβ T-cell receptors function without mispairing with endogenous TCR chains",* Gene Therapy. July, 2011.

[16]. Gene-Wei Li, Eugene Oh, and Jonathan S. Weissman, "*The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteri"*, Nature , 484(7395): 538–541, 2012.

IDENTIFYING IMPORTANT PARAMETERS FOR DESIGN FOR RECOMBINANT DNA DESIGN

Duong Thi Kim Chi, Tran Van Lang, Le Mau Long

ABSTRACT*— The use of natural genes in the manufacturing process of recombinant products used in medicine, pharmacology, or breeding of agricultural crops often result in low expression. The design of the gene or genes that have been optimized for the study*

of group G Menzella Hugo et al 2011 or Agnieszka Zylicz-Stachula et al in 2014 demonstrated the ability to increase the expression level of target genes after optimizing natural than the original gene. This article presents the results of the survey parameters optimization affects genes from the program being used now as Eugene, GeneOptimizer, VisualGeneDeveloper, OptimumGene. Assessment criteria of program optimization between natural and genetically optimized gene on the genome MHA5. The statistical results are used to determine the important parameters for the design used in the recombinant DNA Design.