XÂY DỰNG MÔ HÌNH DỰ ĐOÁN ĐỘ MẠNH CỦA PROMOTER PHỤ THUỘC NHÂN TỐ SIGMA A Ở BACILLUS SUBTILIS

Phạm Trung Nghĩa[1,2], Nguyễn Phan Anh Tú[1,2], Võ Trí Nam[1,2], Nguyễn Đức Hoàng[1,2*]
[1]Trường Đại học Khoa học Tự Nhiên, ĐHQG-HCM
[2]Đại học Quốc gia thành phố Hồ Chí Minh
phamtrungnghiamap@gmail.com, npatu07@gmail.com, vtnam@hcmus.edu.vn, ndhoang@hcmus.edu.vn

Tóm tắt

Để hỗ trợ cho các nghiên cứu về sản xuất protein tái tổ hợp, các trình tự promoter đang được nghiên cứu nhằm kiểm soát được mức độ cũng như thời điểm biểu hiện. Trong báo cáo này, chúng tôi tập trung vào xây dựng mô hình dự đoán mức độ biểu hiện của các trình tự promoter phụ thuộc nhân tố sigma A ở *B. subtilis*. Dữ liệu trình tự promoter được thu nhận và phân thành 4 nhóm theo mức độ biểu hiện, sau đó thiết kế mô hình CNN (Convolutional neural network) kết hợp với đặc trưng PseDNC (pseudo dinucleotide composition) để dự đoán độ mạnh. Kết quả cho thấy mô hình phân nhóm promoter thành hai nhóm mạnh và yếu có độ chính xác 70%. Tiếp theo, kết quả hai nhóm promoter mạnh và yếu được lần lượt đưa vào 2 mô hình mới có độ chính xác lần lượt là 75% và 58% giúp phân thành bốn nhóm promoter yếu, trung bình, mạnh và mạnh hơn. Đây là mô hình phân nhóm đầu tiên được xây dựng để phân chia trình tự promoter theo độ mạnh trên chủng vi sinh vật mô hình *B. subtilis*.

Từ khóa: *Bacillus subtilis*, CNN, máy học, promoter, PseDNC

CONSTRUCTING MODELS FOR PREDICTING STRENGTH OF SIGMA A FACTOR DEPENDENT PROMOTER IN BACILLUS SUBTILIS

Pham Trung Nghia[1,2], Nguyen Phan Anh Tu[1,2], Nam Vo[1,2], Nguyen Duc Hoang[1,2*]
[1]University of Science, VNU-HCM
[2]Vietnam National University Ho Chi Minh City
phamtrungnghiamap@gmail.com, npatu07@gmail.com, vtnam@hcmus.edu.vn, ndhoang@hcmus.edu.vn

Abstract

To support researches on recombinant protein production, promoter sequences are being studied to effectively control the expressions. In this report, we focus on constructing models to predict the expression levels of the sigma A-dependent promoters in *B. subtilis*. The promoter sequence data was collected and classified into 4 groups according to their expression levels, then the model based on CNN (Convolutional neural network) and PseDNC (pseudo dinucleotide composition) features was built to predict their strengths. The result showed that the model for classifying strong and weak groups has accuracy 70%. Next, the strong and weak groups were put into two new models with accuracy 75% and 58% respectively, to classify into four groups: weaker, medium, strong, and stronger promoters. This is the first model for classifying the strengths of promoters of the model microorganism, *B. subtilis*.

Key words: *Bacillus subtilis*, CNN, machine learning, promoter, PseDNC