단백질 인코딩 문제에 대한 NSGA-II 와 NSGA-III 의 성능 비교

김동현, 김인서, 이주훈, 김진성*
중앙대학교

jeus5771@cau.ac.kr, inseo764@cau.ac.kr, huni4800@cau.ac.kr, *kimjsung@cau.ac.kr

Performance Comparison of NSGA-II and NSGA-III on Protein Encoding Problem

Donghyeon Kim, Inseo Kim, Juhun Lee, Jinsung Kim*
Chung-Ang Univ.

요 약

단백질 생산 극대화는 신약 개발, 백신 개발들이 포함되는 다양한 생명 공학 분야의 문제들을 해결하기 위해 필수적이다. 따라서, 단백질 발현 수준에 영향을 미치는 다양한 요소들을 고려하여 코딩 서열들을 설계하는 것이 중요하다. 다수의 목적 함수들을 포함하는 다목적 최적화 문제들을 해결하기 위해 NSGA-II 와 NSGA-III 방법이 널리 사용되고 있기 때문에, 두 방법 중 어떠한 방법이 단백질 인코딩 문제에서 더 우수한 결과를 제공할 수 있는지에 대한 비교 분석이 필요하다. 본 논문에서는 코딩 서열들의 설계 과정에 다수의 목적 함수들이 사용될 때, NSGA-II 와 NSGA-III 방법에 대한 성능을 비교 분석한다.

Ⅰ. 서 론

백신 개발, 진단 키트 개발들이 포함되는 생명공학 분야에서 단백질 생산은 필수적인 역할을 한다 [1]. 예를 들어, 신속 진단 검사 키트 생산에는 조립에 필요한 단백질을 대량으로 생산하고 정제하는 것이 필수적이다 [2]. 일반적으로, 단백질 생산을 극대화하기 위해 다수의 표적 유전자 사본들을 숙주 게놈에 통합하는 방법이 사용된다. 이러한 방법은 이론적으로 삽입한 유전자 수에 비례한 단백질 생산을 가능하게 하지만, 실제로 단백질로 번역되는 서열인 코딩 서열들(CoDing Sequences)이 어떻게 설계되었는지 따라 단백질 발현 수준이 달라진다. 결과적으로, 단백질 생산을 극대화하기 위해서는 코딩 서열들의 설계 과정에서 단백질 발현 수준에 영향을 미치는 여러 요소들이 고려되어야 한다.

다수의 목적 함수들을 포함하는 최적화 문제는 다목적 최적화(Multi-objective optimization) 문제로 귀결된다. 다목적 최적화 문제에서 하나의 목적 함수(Objective function)를 개선하는 것은 종종 다른 목적 함수의 악화를 야기한다 [3]. 이는 모든 목적 함수들이 동시에 향상되는 단일 최적 해를 찾는 것을 불가능 하게 한다. 따라서, 다목적 최적화 문제는 의사결정을 위한 파레토 최적해(Pareto-optimal solution)를 찾는 것을 목표로 한다. NSGA-II(Non-dominated Sorting Genetic Algorithm II)는 다목적 최적화 문제를 해결하기 위한 대표적인 유전 알고리즘이다. 하지만, 현실의 많은 문제들이 다수의 목적 함수들을 포함함에 따라 네 가지 이상의 목적 함수들을 가지는 문제들을 효과적으로 해결하기 위한 NSGA-III(Non-dominated Sorting Genetic Algorithm III)가 제안되었다. 비록 NSGA-III 가 NSGA-II 의 향상된 버전이지만, 많은 목적 함수들을 가지는 다양한 문제들에 대해 NSGA-III 방법이 NSGA-II 보다 항상 우수함을 보장하지는 않는다. 예를 들어, 기숙사 건물 설계 문제에서 방의 모양, 복도 모양 등과 같은 설계 변수들은 목적 함수 값으로 계산된다. 10 개 설계 변수를 포함한 기숙사 건물 타입 1 과 9 개 설계 변수를 포함한 기숙사 건물 타입 2 최적화 문제에서 사이클 수가 적을 때는 NSGA-II 가 NSGA-III 보다 우수했다 [4]. 따라서, 단백질 인코딩 문제에서 NSGA-II 와 NSGA-III 방법 중 어느 방법이 더 우수한 결과를 제공할 수 있는지에 대한 비교 분석이 필요하다.

본 논문에서는 코딩 서열들의 단백질 발현 수준을 평가하기 위한 여섯 가지 목적 함수들을 소개하고, NSGA-II 와 NSGA-III 방법의 절차를 소개한다. 그다음, 단백질 인코딩 문제에서 사용하는 목적 함수 개수에 따른 NSGA-II 와 NSGA-III 방법의 성능을 비교분석 한다.

Ⅱ. 본론

단백질 발현 수준이 높은 코딩 서열들을 설계하기 위해서는 코돈 사용 편향, 코돈 문맥 편향, 숨은 종결 코돈, 상동 재조합, 구아닌-사이토신 함량, 헤어핀 루프 구조들을 고려해야 한다. 각 요소들은 CAI, CPB, HSC, HD, GC3, SL 목적 함수들로 평가된다. CAI, CPB, HSC, HD 들의 값은 클수록 단백질 발현 수준이 높다는 것을 나타내고, GC3 와 SL 값은 0 으로 수렴할수록 단백질 발현 수준이 높다는 것을 나타낸다.

NSGA-II 와 NSGA-III 방법은 초기화 단계, 변이 단계, 선택 단계로 구성된다. 초기화 단계는 초기 N 개의 해들을 생성하는 단계이고, 변이 단계는 N 개의 원본 해들로 새로운 N 개의 해들을 생성하는 단계이다. 선택 단계는 $2N$ 개의 해들 중 좋은 N 개의 해들을 선택하는 단계이다. NSGA-II 와 NSGA-III 방법의 차이점은 선택 단계에서 NSGA-II 는 비지배 정렬(Non-dominated sorting)과 군집 거리 정렬(Crowding distance sorting)을 수행하지만, NSGA-III 방법은 비지배 정렬과 참조점 기반 정렬

(Reference point-based sorting)을 수행한다는 점이다. 참조점 기반 정렬을 위한 참조점들은 Das-Dennis 방법을 사용하여 N보다 작으면서 가장 가깝게 설정한다.

단백질 인코딩 문제에서 다양한 목적 함수의 개수에 대한 NSGA-II 와 NSGA-III 방법의 성능을 비교하기 위해, 네 개부터 여섯 가지 목적 함수를 사용하는 경우에 대한 실험을 수행한다. 네 개의 목적 함수들을 사용하는 경우 CAI, CPB, HSC, HD 를 사용하고, 다섯 가지 목적 함수 사용에는 GC3 를 추가한다, 마지막으로, 여섯 가지 목적 함수를 사용하는 경우에는 SL 를 추가한다. 모든 실험은 NVIDIA 의 CUDA 플랫폼을 사용하여 실행한다.

Ⅲ. 실험

본 연구에서 사용된 실험 환경은 NVIDIA GeForce RTX 4090 을 사용했다. 사용된 CUDA 버전은 12.4 이고 드라이버 버전은 550.54.15 이다.

단백질	CDSs	길이	CDSs x 길이
Q5VZP5	2	1158	2316
A4Y1B6	3	716	2148
B3LS90	4	679	2716
B4TWR7	5	505	2525
Q91X51	6	446	2676
Q89BP2	7	388	2716

[표 1] 실험에 사용된 단백질 인스턴스들

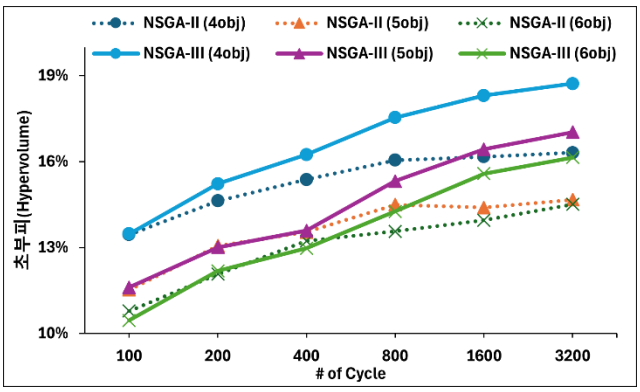
[표 1]은 실험에 사용된 여섯 가지 단백질 인스턴스들을 보여준다. 이들은 아미노산 서열의 길이와 코딩 서열의 개수를 기반으로 단백질 인코딩 문제에서 다양한 복잡성을 실험하기 위해 선택되었다. 또한, 실험에는 100 개의 해들과 100, 200, 400, 800, 1600, 3200 사이클 수, 그리고 20% 변이 확률을 사용했다. 각 단백질들에 대한 실험은 세 번씩 수행했다.

단백질	NSGA-II			NSGA-III		
	4obj	5obj	6obj	4obj	5obj	6obj
Q5VZP5	2.3	2.5	58.1	2.1	2.3	56.0
A4Y1B6	2.2	2.5	32.9	2.0	2.2	31.2
B3LS90	2.3	2.6	36.9	2.1	2.3	35.2
B4TWR7	2.2	2.5	26.8	2.1	2.2	25.1
Q91X51	2.3	2.6	24.8	2.1	2.3	23.1
Q89BP2	2.4	2.6	23.1	2.3	2.4	21.4
평균	2.3	2.5	33.8	2.1	2.3	32.0

[표 2] NSGA-II 와 NSGA-III 방법의 실행 시간(초)

[표 2]는 실험한 단백질들에 대해 NSGA-II 와 NSGA-III 방법이 3200 사이클까지 수행했을 때의 실행 시간을 보여준다. 네 개와 다섯 가지 목적 함수들을 사용했을 때는 NSGA-III 방법이 NSGA-II 보다 평균적으로 0.2 초 정도 빨랐고, 여섯 가지 목적 함수들을 사용했을 때는 NSGA-III 방법이 NSGA-II 보다 평균적으로 1.8 초 정도 빠르게 실행되었다. 결과적으로, GPU 환경에서 NSGA-III 방법은 NSGA-II 보다 조금 더 빠르게 실행되었다.

[그림 1]은 [표 1]의 단백질들에 대한 NSGA-II 와 NSGA-III 방법의 평균 초부피(Hypervolume) 값들을 보여준다. 초부피는 다목적 최적화 문제에서 가장 널리 사용되는 평가 지표로 목적 공간에서 파레토 해가 차지하는 부피를 계산함으로써 결과로 얻은 해들의 품질을 평가한다. 더 높은 초부피의 값은 더 뛰어난 품질의 해들을 생성했다는 것을 나타낸다. 여섯 가지 목적 함수를 사용



[그림 1] 실험한 단백질들에 대한 평균 초부피

하는 100 사이클인 경우를 제외하면, 모든 실험 결과에 대해서 NSGA-III 방법이 NSGA-II 보다 높은 초부피 값을 보여 더 우수함을 나타냈다.

Ⅳ. 결론

본 논문에서는 단백질 인코딩 문제에서 사용하는 목적 함수 개수에 따른 NSGA-II 와 NSGA-III 방법의 성능을 비교 분석하였다. 실험 결과, NSGA-III 방법은 NSGA-II 보다 항상 빠른 실행 시간을 보였다. 게다가, 여섯 가지 목적 함수들을 사용한 100 번째 사이클을 제외하면, 모든 경우에서 NSGA-III 방법이 NSGA-II 보다 더 높은 초부피 값을 나타냈다. 따라서, 단백질 인코딩 문제에서 NSGA-III 방법을 사용하는 것이 NSGA-II 보다 거의 모든 경우에서 좋은 결과를 제공할 수 있음을 보여준다.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2022R1G1A1013586).

참 고 문 헌

[1] Ahmad, M., Hirz, M., Pichler, H., & Schwab, H. "Protein expression in Pichia pastoris: recent achievements and perspectives for heterologous protein production," Applied Microbiology and Biotechnology, vol. 98, pp. 5301-5317, 2014.

[2] Huleani, S., Roberts, M. R., Beales, L., and Papaioannou, E. H., "Escherichia coli as an antibody expression host for the production of diagnostic proteins: significance and expression," Critical Reviews in Biotechnology, vol. 42, no. 5, pp. 756-773, 2022.

[3] Zhou, A., Qu, B.-Y., Li, H., Zhao, S.-Z., Suganthan, P. N., and Zhang, Q., "Multiobjective evolutionary algorithms: A survey of the state of the art," Swarm and Evolutionary Computation, vol. 1, no. 1, pp. 32-49, 2011.

[4] Razmi, A., Rahbar, M., and Bemanian, M., "PCA-ANN integrated NSGA III framework for dormitory building design optimization: Energy efficiency, daylight, and thermal comfort," Applied Energy, vol. 305, p. 117828, 2022.