(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0038263
(43) 공개일자 2020년04월10일

(51) 국제특허분류(Int. Cl.)
G16B 30/10 (2019.01) *C12N 15/67* (2006.01)
C12N 9/12 (2006.01) *G06N 3/12* (2006.01)
G16B 20/20 (2019.01) *G16B 40/20* (2019.01)
(52) CPC특허분류
G16B 30/10 (2019.02)
C12N 15/67 (2013.01)
(21) 출원번호 10-2020-7005489
(22) 출원일자(국제) 2018년07월25일
심사청구일자 없음
(85) 번역문제출일자 2020년02월25일
(86) 국제출원번호 PCT/CN2018/097040
(87) 국제공개번호 WO 2019/020054
국제공개일자 2019년01월31일
(30) 우선권주장
201710611752.5 2017년07월25일 중국(CN)

(71) 출원인
난징진시루이 사이언스 앤드 테크놀로지 바이올로지 코포레이션
중국 난징 211100, 지앙닝, 모링 애비뉴, 융시 로드 28
(72) 발명자
판, 롱
중국 지앙수 211100 난징 지앙닝 사이언스 파크 융시 로드 28
순, 얀
중국 지앙수 211100 난징 지앙닝 사이언스 파크 융시 로드 28
(뒷면에 계속)
(74) 대리인
최우성

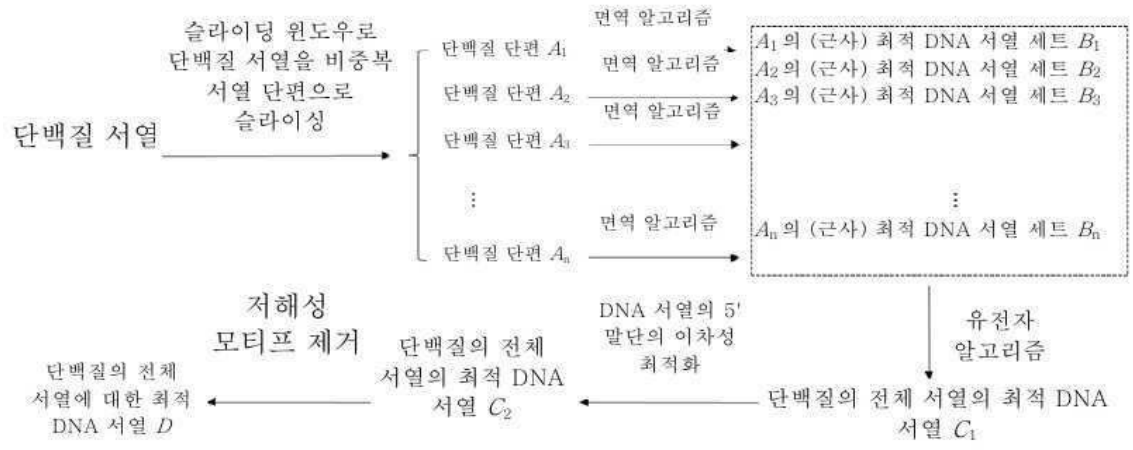
전체 청구항 수 : 총 8 항

(54) 발명의 명칭 **면역 알고리즘에 근거된 코돈 최적화 방법**

(57) 요 약

면역 알고리즘에 근거된 코돈 최적화 방법은 면역 알고리즘 및 유전자 알고리즘이 단백질 코딩 서열에서 국부 다목적 최적화 및 전역 다목적 최적화를 각각 수행하는데 연속적으로 이용되고, 그리고 이후 전면적 방법이 최적 발현 서열을 최대 정도까지 검색하기 위해, 상기 서열에서 미세 조정과 최적화를 수행하는데 이용되는 것으로 특징화된다. 본 발명은 유전자 알고리즘의 무작위 전역 병렬 검색의 특징을 유지할 뿐만 아니라, 성급한 수렴을 비교적 큰 정도로 방지하여 전역 최적 해법으로의 신속한 수렴을 담보한다. 본 발명은 최초로, 단계별 과정 (각각 순서대로 국부 최적화, 전역 최적화, 그리고 미세 조정과 최적화)을 통해 코돈 최적화를 실행하는 정확도 및 효율에서 면역 알고리즘 및 유전자 알고리즘의 이점을 조합하고, 그리고 실례 검사를 통해 코돈 최적화에서 상기 알고리즘의 높은 효율을 입증한다.

대 표 도 - 도1



(52) CPC특허분류
C12N 9/12 (2013.01)
C12Y 207/11024 (2013.01)
G06N 3/126 (2013.01)
G16B 20/20 (2019.02)
G16B 40/20 (2019.02)
C07K 2319/43 (2013.01)
(72) 발명자
우, 동밍
중국 지앙수 211100 난징 지앙닝 사이언스 파크 용시 로드 28
후앙, 샤오루오
중국 지앙수 211100 난징 지앙닝 사이언스 파크 용시 로드 28
창, 리후아
중국 지앙수 211100 난징 지앙닝 사이언스 파크 용시 로드 28
리우, 전유
중국 지앙수 211100 난징 지앙닝 사이언스 파크 용시 로드 28

명 세 서

청구범위

청구항 1

면역 알고리즘에 근거된 코돈 최적화 방법에 있어서, 면역 알고리즘 및 유전자 알고리즘이 단백질 코딩 서열에서 국부 다목적 최적화 및 전역 다목적 최적화를 각각 수행하는데 연속적으로 이용되고, 그리고 이후 전면적 방법이 최적 발현 서열을 최대 정도까지 검색하기 위해, 상기 서열에서 미세 조정과 최적화를 수행하는데 이용되는 것을 특징으로 하는 최적화 방법.

청구항 2

청구항 1에 있어서, 하기의 3 단계를 포함하는 것을 특징으로 하는 최적화 방법: 국부 최적화의 첫 번째 단계, 다시 말하면, 단백질 서열을 비중복 서열 단편 $A_1, A_2...A_n$으로 개열하고, 그리고 이후, 대략적으로 최적의 DNA 서열 세트 $B_1, B_2...B_n$를 산출하기 위해, 면역 알고리즘을 이용하여 각 서열 단편에 대한 코돈 최적화를 완결하는 단계; 전역 최적화의 두 번째 단계, 다시 말하면, 유전자 알고리즘을 활용하여 $B_1, B_2...B_n$에 근거된 단백질의 전장의 DNA 코딩 서열을 초기화하고, 그리고 단백질 서열의 최적 DNA 서열 C_1을 걸러내는 단계; 그리고 미세 조정과 최적화의 세 번째 단계, 이것은 인코딩된 단백질의 N 말단 영역에 상응하는 DNA 서열의 5' 말단에서 전면적 최적화를 수행하여 DNA 서열 C_2를 산출하고, 그리고 발현 저해성 모티프를 제거하여, 최적 발현 서열 D를 최종적으로 산출하는 것을 포함함.

청구항 3

청구항 1 또는 2에 있어서, 단백질은 20개보다 많은 아미노산으로 구성되는 화합물을 지칭하고; 단백질은 위치의 면에서 분비 단백질, 막 단백질, 세포질 단백질, 핵 단백질 등을 포함하고; 기능의 면에서 항체 단백질, 조절 단백질, 구조 단백질 등을 포함하고; 공급원의 면에서 상동성 발현 단백질 및 이종성 발현 단백질을 포함하고; 서열의 면에서 자연 단백질 및 인위적으로-변형된 단백질, 완전한 단백질/항체, 절두된 부분적인 단백질/항체, 그리고 2개 또는 그 이상의 단백질로부터 및 단백질과 펩티드 사슬로부터 형성된 융합 단백질을 포함하고; 본 발명에서 규정된 항체는 무손상 항체, 그리고 Fab, ScFV, SdAb, 키메라 항체, 이중특이적 항체, Fc 융합 단백질, 기타 등등을 포함하지만 이들에 한정되지 않는 것을 특징으로 하는 최적화 방법.

청구항 4

청구항 1 또는 2에 있어서, 면역 유전자 알고리즘은 단백질 단편에서 국부 최적화를 수행하기 위한 다목적 최적화 방법을 채택하고, 모집단 초기화는 고도로-발현된 단백질을 인코딩하는 서열의 이중 코돈 표에 근거되고, 그리고 각 유전자는 동의 코돈에 의해 직접적으로 인코딩되고; 그리고 최적화 과정에서, 항체 다양성이 담보되고, 그리고 알고리즘의 전역 검색 능력을 증가시키기 위해, 면역 유전자 알고리즘의 항체 정보 엔트로피, 항체 모집단 유사성, 항체 농도 및 중합화 적합도를 계산하고 기억 세포를 갱신함으로써 모집단 변성의 현상이 예방되는 것을 특징으로 하는 최적화 방법.

청구항 5

청구항 1 또는 2에 있어서, 유전자 알고리즘은 단백질의 전체 서열에서 전역 최적화를 수행하기 위한 다목적 최적화 방법을 채택하고, 초기화된 모집단은 국부 최적화에 종속되는 최적화된 단편에 근거하여 무작위로 산출되고, 그리고 각 유전자는 각 단백질 단편의 최적화된 서열 세트에 의해 직접적으로 인코딩되는 것을 특징으로 하는 최적화 방법.

청구항 6

청구항 1 또는 2에 있어서, 미세 조정과 최적화는 DNA 서열의 5' 말단에서 최소 자유 에너지 MFE, 코돈 콘텍스트 및 CAI를 계산하고 분류하고, 그리고 분류 결과에 따라서 단백질 서열의 N 말단에 대한 최적 코딩 서열을 선택하기 위한 전면적 방법을 이용하는 것을 특징으로 하는 최적화 방법.

청구항 7

청구항 1 또는 2에 있어서, 코돈 최적화 방법은 하기의 숙주 발현 시스템에 최소한 적용가능한 것을 특징으로 하는 최적화 방법: 1) 포유류 발현 시스템; 2) 곤충 발현 시스템; 3) 효모 발현 시스템; 4) 대장균 (*Escherichia coli*) 발현 시스템; 5) 바실루스 서브틸리스 (*Bacillus subtilis*) 발현 시스템; 6) 식물 발현 시스템, 그리고 7) 무세포 발현 시스템.

청구항 8

청구항 1 또는 2에 있어서, 코돈 최적화 방법은 하기의 발현 벡터에 최소한 적용가능한 것을 특징으로 하는 최적화 방법: 일시적인 발현 벡터 및 안정된 발현 벡터, 바이러스 발현 벡터 및 비바이러스 발현 벡터, 유도된 및 비-유도된 발현 벡터.

발명의 설명

기 술 분 야

[0001] **기술 분야**

[0002] 본 발명은 단백질 가공 기술, 그리고 특히, 단백질 가공에서 코돈 최적화 방법, 그리고 구체적으로 면역 알고리즘에 근거된 코돈 최적화 방법에 관계한다.

배 경 기 술

[0003] **배경**

[0004] 코돈 축중성은 아미노산이 단백질 번역 동안 복수의 상이한 코돈에 의해 인코딩될 수 있는 현상을 지칭한다. 동일한 아미노산을 인코딩하는 상이한 코돈은 동의 코돈으로 불린다. 길이에서 200개 아미노산으로 구성되는 단백질은 일반적으로, 10^{20}개 보다 많은 상이한 DNA 서열에 의해 인코딩될 수 있다. 상이한 종에서, 동의 코돈의 발생 빈도는 상이하고, 그리고 이런 현상은 코돈 선호로 불린다. 코돈 최적화는 주로, 인자, 예를 들면, 숙주 발현 시스템의 코돈 선호에 근거된다. 단백질의 아미노산 서열을 변화시키지 않는다는 것을 전제로 하여, 단백질을 숙주 발현 시스템에서 다수의 DNA 코딩 서열로부터 가장 효율적으로 발현할 수 있는 DNA 서열을 걸러내는데 컴퓨터 알고리즘이 이용된다.

[0005] 현재, 코돈 최적화의 과정에서 단백질 발현에 영향을 주는 것으로 종종 고려되는 주요 인자는 숙주 세포의 코돈 선호 (이의 통상적으로 이용되는 특징화 파라미터는 코돈 적응 지수 [CAI], 숙주 세포의 이중 코돈 선호 [코돈 콘텍스트], CBI [코돈 바이어스 지수], ENC [코돈의 효과적인 숫자], FOP [최적 코돈의 빈도], CPP [코돈 선호 파라미터], 그리고 tAI [tRNA 적응 지수]를 포함한다), 숨겨진 종결 코돈의 숫자, GC 함량, 희귀한 코돈 함량, mRNA 저해성 조절 모티프의 숫자, mRNA 이차 구조 (주로, 헤어핀 구조 및 최소 자유 에너지를 포함), 기계 학습에서 핵심 코돈 및 수학 모델의 채점, 마이크로RNA 결합 부위, G4 함량, 그리고 단백질 이차 구조의 코돈 선호를 포함한다 (Joshua B. Plotkin & Grzegorz Kudla, *Nature Reviews Genetics*, 2011). 코돈 최적화를 위해 현재 이용가능한 소프트웨어 및 알고리즘은 DNAWorks, Jcat, Synthetic gene designer, GeneDesign 2.0, OPTIMIZER, Eugene, mRNA Optimizer, COOL, D-Tailor, UpGene, GASCO, Codon Harmonization, QPSO, GeMS 및 ATGME (Evelina Angov, *Biotechnology Journal*, 2011; Nathan Gould et al., Frontiers in Bioengineering and Biotechnology, 2014)를 포함한다.

[0006] 코돈 최적화 알고리즘에서 이용된 휴리스틱 알고리즘 (가령, 입자 스웜 및 유전자 알고리즘)과 비교하여, 면역 알고리즘은 독특한 이점을 갖는다. 면역 알고리즘은 생물학적 면역 기전에 근거된 향상된 유전자 알고리즘이다. 이것은 해결되어야 하는 실제 문제의 목적 함수가 항원에 상응하는 것을 가능하게 하고, 그리고 상기 문제의 해법이 항체에 상응하는 것을 가능하게 한다. 생물학적 면역성의 원리에 따라서, 생물학적 면역계는 살아있는 생물체를 침입하는 항원에 저항하는 상응하는 항체를 세포 분열 및 분화를 통해 자동적으로 산출하는 것으로 목격될 수 있다. 이런 과정은 면역 반응으로서 지칭된다. 면역 반응의 과정에서, 일부 항체는 기억 세포로서 보존되고, 그리고 동일한 유형의 항원이 다시 한 번 침입할 때, 기억 세포는 활성화되고 다수의 항체를 신속하게 생산하는데, 이것은 재반응을 초기 반응보다 더욱 빠르고 강하게 만들고, 이것은 면역계의 기억 기능을 반영한다. 항원과의 결합 후, 항체는 일련의 반응을 통해 항원을 파괴한다. 이와 동시에, 상이한 항체는 또한, 서로를 증진하고 저해하여 항체의 다양성 및 면역 균형을 유지시킨다. 이런 균형은 농도 기전에 따라서 달성된다, 다시

말하면, 항체의 농도가 더욱 높을수록, 이들 항체는 더욱 저해되고; 그리고 항체의 농도가 더욱 낮을수록, 이들 항체는 더욱 증진되는데, 이것은 면역계의 자기 조절 기능을 반영한다.

발명의 내용

해결하려는 과제

[0007] **요약**

[0008] 본 발명의 목적은 기존의 코돈 최적화 방법의 긴 주기 및 불량한 발현 정확도의 문제점을 해결하고, 그리고 한정된 시간 내에 코돈 최적화 공간의 완전한 대규모 검색을 효과적으로 할 수 있는, 다시 말하면, 단백질 코딩 서열 세트로부터 가장 효과적인 발현을 갖는 DNA 서열을 걸러낼 수 있는 면역 알고리즘에 근거된 코돈 최적화 방법을 발명하는 것이다.

[0009] 본 발명의 기술적인 해법은 하기와 같다.

[0010] 면역 알고리즘에 근거된 코돈 최적화 방법은 면역 알고리즘 및 유전자 알고리즘이 단백질 코딩 서열에서 국부 다목적 최적화 및 전역 다목적 최적화를 각각 수행하는데 연속적으로 이용되고, 그리고 이후 전면적 방법이 최적 발현 서열을 최대 정도까지 검색하기 위해, 상기 서열에서 미세 조정과 최적화를 수행하는데 이용된다는 것을 포함한다.

[0011] 특히, 본 발명의 방법은 하기의 3 단계를 포함한다: 국부 최적화의 첫 번째 단계, 다시 말하면, 단백질 서열을 비중복 서열 단편 $A_1, A_2...A_n$으로 개열하고, 그리고 이후, 대략적으로 최적의 DNA 서열 세트 $B_1, B_2...B_n$를 산출하기 위해, 면역 알고리즘을 이용하여 각 서열 단편에 대한 코돈 최적화를 완결하는 단계; 전역 최적화의 두 번째 단계, 다시 말하면, 유전자 알고리즘을 활용하여 $B_1, B_2...B_n$에 근거된 단백질의 전장의 DNA 코딩 서열을 초기화하고, 그리고 단백질 서열의 최적 DNA 서열 C_1을 걸러내는 단계; 그리고 미세 조정과 최적화의 세 번째 단계, 이것은 인코딩된 단백질의 N 말단 영역에 상응하는 DNA 서열의 5' 말단에서 전면적 최적화를 수행하여 DNA 서열 C_2를 산출하고, 그리고 발현 저해성 모티프를 제거하여, 최적 발현 서열 D를 최종적으로 산출하는 것을 포함한다.

[0012] 단백질은 20개보다 많은 아미노산으로 구성되는 화합물을 지칭하고; 단백질은 위치의 면에서 분비 단백질, 막 단백질, 세포질 단백질, 핵 단백질 등을 포함하고; 기능의 면에서 항체 단백질, 조절 단백질, 구조 단백질 등을 포함하고; 공급원의 면에서 상동성 발현 단백질 및 이종성 발현 단백질을 포함하고; 서열의 면에서 자연 단백질 및 인위적으로-변형된 단백질, 완전한 단백질/항체, 절두된 부분적인 단백질/항체, 그리고 2개 또는 그 이상의 단백질로부터 및 단백질과 펩티드 사슬로부터 형성된 융합 단백질을 포함한다. 본 발명에서 규정된 항체는 무손상 항체, 그리고 Fab, ScFV, SdAb, 키메라 항체, 이중특이적 항체, Fc 융합 단백질, 기타 등등을 포함하지만 이들에 한정되지 않는다.

[0013] 면역 유전자 알고리즘은 단백질 단편에서 국부 최적화를 수행하기 위한 다목적 최적화 방법을 채택하고, 모집단 초기화는 고도로-발현된 단백질을 인코딩하는 서열의 이중 코돈 표에 근거되고, 그리고 각 유전자는 동의 코돈에 의해 직접적으로 인코딩되고; 그리고 최적화 과정에서, 항체 다양성이 담보되고, 그리고 알고리즘의 전역 검색 능력을 증가시키기 위해, 면역 유전자 알고리즘의 항체 정보 엔트로피, 항체 모집단 유사성, 항체 농도 및 중합화 적합도를 계산하고 기억 세포를 갱신함으로써 모집단 변성의 현상이 예방된다.

[0014] 유전자 알고리즘은 단백질의 전체 서열에서 전역 최적화를 수행하기 위한 다목적 최적화 방법을 채택하고, 초기화된 모집단은 국부 최적화에 종속되는 최적화된 단편에 근거하여 무작위로 산출되고, 그리고 각 유전자는 각 단백질 단편의 최적화된 서열 세트에 의해 직접적으로 인코딩된다.

[0015] 미세 조정과 최적화는 DNA 서열의 5' 말단에서 최소 자유 에너지 MFE, 코돈 콘텍스트 및 CAI를 계산하고 분류하고, 그리고 분류 결과에 따라서 단백질 서열의 N 말단에 대한 최적 코딩 서열을 선택하기 위한 전면적 방법을 이용한다.

[0016] 코돈 최적화 방법은 하기의 숙주 발현 시스템에 최소한 적용가능하다: 1) 포유류 발현 시스템; 2) 곤충 발현 시스템; 3) 효모 발현 시스템; 4) 대장균 (*Escherichia coli*) 발현 시스템; 5) 바실루스 서브틸리스 (*Bacillus subtilis*) 발현 시스템; 6) 식물 발현 시스템, 그리고 7) 무세포 발현 시스템.

[0017] 코돈 최적화 방법은 하기의 발현 벡터에 최소한 적용가능하다: 일시적인 발현 벡터 및 안정된 발현 벡터, 바이

러스 발현 벡터 및 비바이러스 발현 벡터, 유도된 및 비-유도된 발현 벡터.

[0018] 본 발명의 유익한 효과는 하기와 같다.

[0019] 면역 알고리즘은 유전자 알고리즘으로부터 향상된 알고리즘이다. 최적화에서 성급한 국부 수렴을 예방하는데 있어서 면역 알고리즘의 이점에 비추어, 본 발명은 최초로, 국부 최적화를 위한 코돈 최적화를 실행하기 위해 면역 알고리즘을 도입하고, 그리고 차후 유전자 알고리즘을 통해 전역 최적화를 실행하고 미세 조정과 최적화를 최종적으로 실행하고, 그리고 따라서, 상이한 알고리즘의 이점을 조합하는 완전히 새로운 3-단계 하이브리드 최적화 알고리즘을 개발하고; 그리고 코돈 최적화에서 상기 알고리즘의 높은 효율은 하기 실시예에 의해 더욱 입증된다.

[0020] 유전자 알고리즘과 비교하여, 본 발명의 면역 알고리즘은 하기의 특징을 갖는다: 첫째로, 면역 알고리즘은 검색 속도를 가속화하고 유전자 알고리즘의 전반적인 검색 능력을 향상시킬 수 있는 면역 기억 기능을 갖고; 두 번째로, 면역 알고리즘은 항체의 다양성을 유지하는 기능을 갖는데, 이것은 유전자 알고리즘의 국부 검색 능력을 향상시키는데 활용될 수 있고; 그리고 최종적으로, 면역 알고리즘은 자기 조절 기능을 갖는데, 이것은 유전자 알고리즘의 전역 검색 능력을 향상시키고 국부 해법에 빠지는 것을 방지하는데 이용될 수 있다. 이런 이유로, 면역 유전자 알고리즘은 유전자 알고리즘의 무작위 전역 병렬 검색의 특징을 유지할 뿐만 아니라, 성급한 수렴을 비교적 큰 정도로 방지하여 전역 최적 해법으로의 신속한 수렴을 담보한다. 본 발명은 최초로, 단계별 과정 (각각 순서대로 국부 최적화, 전역 최적화, 그리고 미세 조정과 최적화)을 통해 코돈 최적화를 실행하는 정확도 및 효율에서 면역 알고리즘 및 유전자 알고리즘의 이점을 조합하고, 그리고 실례 검사를 통해 코돈 최적화에서 상기 알고리즘의 높은 효율을 입증한다.

[0021] 본 발명은 고속 및 높은 효율의 이점을 갖는다.

도면의 간단한 설명

[0022] **도면의 간단한 설명**

도 1은 본 발명의 최적화 알고리즘의 계통 흐름도이다.

도 2는 본 발명의 면역 알고리즘의 계통 흐름도 (다시 말하면, 국부 최적화 흐름)이다.

도 3은 본 발명의 유전자 알고리즘의 흐름 (다시 말하면, 전역 최적화 흐름)을 보여준다.

도 4는 본 발명의 DNA 서열의 5' 말단을 최적화하는 흐름을 보여준다.

도 5는 본 발명의 검사 단백질의 유전자 서열 설계의 계통도이다.

도 6은 본 발명의 pTT 발현 벡터 지도이다.

도 7은 본 발명의 웨스턴 블롯팅 결과의 계통도이다.

발명을 실시하기 위한 구체적인 내용

[0023] **상세한 설명**

[0024] 하기는 첨부된 도면 및 특정한 실례를 참고로 하여 본 발명을 더욱 설명한다.

[0025] 이것은 도 1-7에서 도시된 바와 같다.

[0026] 면역 알고리즘에 근거된 코돈 최적화 방법은 도 1에서 도시된 바와 같이, 면역 알고리즘 및 유전자 알고리즘이 단백질 코딩 서열 (서열 번호 3 및 서열 번호 4)에서 국부 다목적 최적화 및 전역 다목적 최적화를 각각 수행하기 위해 연속적으로 이용되고, 그리고 이후, 전면적 방법이 최적 발현 서열 (서열 번호 5 및 서열 번호 6)을 최대 정도로 검색하기 위해 서열에서 미세 조정과 최적화를 수행하는데 이용된다는 것을 포함하고, 여기서:

[0027] I. 면역 알고리즘 (다시 말하면, 국부 최적화, 흐름에 대해 도 2를 참조한다).

[0028] 이러한 단계에서 최적화 변수 L의 숫자는 2이고, 다시 말하면, 2가지 특질, 코돈 콘텍스트 및 CAI가 각 단편에 대해 최적화되고 (상세한 설명을 위해 하기를 참조한다), 이것은 다목적 최적화에 속한다. 면역계가 N 항체 (다시 말하면, 모집단 크기가 N이다)로 구성된다고 가정할 때, 각 항체 유전자는 M (이와 동등하게, 단백질 서열에서 아미노산의 숫자가 M이다)의 길이를 갖고, 그리고 각 유전자는 동의 코돈으로 직접적으로 인코딩된다.

[0029] (1) 상이한 숙주 발현 시스템의 기본 데이터 세트 (다시 말하면, 고도로-발현된 단백질의 코딩 서열)에 따라서, 코돈 출현빈도 표 및 이중 코돈 출현빈도 표가 서열을 산출하고 코돈 콘텍스트 및 CAI를 계산하기 위해 계산된다.

[0030] (2) 초기 반응에서, 초기 항체가 이중 코돈 출현빈도에 따라서 산출된다. 단백질 서열 $a_1a_2...a_m$을 실례로 하면, a_1에 대한 동의 코돈은 c_{11} 및 c_{12}이고, 그리고 a_2에 대한 동의 코돈은 c_{21}, c_{22} 및 c_{23}인 것으로 가정된다. 첫 번째 아미노산 a_1에 대한 코돈은 코돈 출현빈도 표에서 c_{11} 및 c_{12}의 빈도에 따라서 선택된다. 이중 아미노산 a_1a_2에 상응하는 이중 코돈은 $c_{11}c_{21}$, $c_{11}c_{22}$, $c_{11}c_{23}$, $c_{12}c_{21}$, $c_{12}c_{22}$ 및 $c_{12}c_{23}$이고, 여기서 [$c_{11}c_{21}$, $c_{11}c_{22}$, $c_{11}c_{23}$] 및 [$c_{12}c_{21}$, $c_{12}c_{22}$, $c_{12}c_{23}$]을 포함하는, 이중 동의 코돈의 2가지 세트가 있다. a_1에 대해 선택된 코돈이 C_{11}이라고 가정할 때, 아미노산 a_2에 대한 코돈은 $c_{11}c_{21}$, $c_{11}c_{22}$ 및 $c_{11}c_{23}$의 빈도에 따라서 c_{21}, c_{22} 및 c_{23} 중에서 한 가지에서 선택된다. 만약 a_1에 대해 선택된 코돈이 C_{12}이면, 아미노산 a_2에 대한 코돈은 $c_{12}c_{21}$, $c_{12}c_{22}$ 및 $c_{12}c_{23}$의 빈도에 따라서 c_{21}, c_{22} 및 c_{23} 중에서 한 가지에서 선택된다. 간단히 말하면, 다른 아미노산에 대한 코돈의 선택은 이의 이전 아미노산에 대한 코돈의 선택에 관련되고, 그리고 첫 번째 아미노산에 대한 코돈이 코돈 출현빈도 표에 따라서 직접적으로 선택된다는 점을 제외하고, 이들의 이중 동의 코돈의 빈도에 의해 결정된다.

[0031] (3) 비-초기 반응에서, 모집단은 부모 개체 및 기억 세포에서 보관된 K 항체로 구성된다. 기억 세포의 항체는 최적화 이력에서 나타났던 K 최적 항체를 기록하는데, 여기서 낮은 적합도를 갖는 항체는 최적화 과정에서 더욱 높은 적합도를 갖는 개체에 의해 점진적으로 대체된다.

[0032] (4) 항체의 적합도 F ($F_{[코돈\ 콘텍스트]}$ 및 $F_{[CAI]}$ 포함)가 계산되고, N 자손 개체가 다목적 최적화에 따라서 선택되고, 그리고 교차 및 변이 작업이 새로운 모집단에 대해 완결된다. 여기에서 변이는 코돈의 무작위 돌연변이이다.

[0033] (5) 항체 모집단 유사성 S의 계산

[0034] 본 발명은 모집단 유사성 S를 계측하기 위해 새논의 평균 정보 엔트로피 $H(N)$를 이용한다.

[0035] 먼저, P_{ij}는 동의 코돈 i가 아미노산 j에서 나타날 확률, 다시 말하면:

[0036] $$P_{ij}=\frac{N_{ij}}{N},$$ 이고

[0037] 여기서 N_{ij}는 모집단 내에 모든 개체의 j-번째 아미노산 위치에서 나타나는 동의 코돈 i의 총수이다. 이후, $Hj(N)$는 j-번째 유전자의 정보 엔트로피 (다시 말하면, 단백질 서열의 j-번째 아미노산)이고, 그리고 하기와 같이 규정된다:

[0038] $$H_j(N)=-\sum_{i=1}^{N}P_{ij}\log_2 P_{ij}$$

[0039] 전체 모집단의 평균 정보 엔트로피는

[0040] $$H(N)=-\frac{1}{M}\sum_{j=1}^{M}H_j(N)$$ 이다.

[0041] 모집단 유사성 S는 하기와 같이 규정된다:

[0042] $$S=\frac{1}{1+H(N)}$$

[0043] (6) 최적화가 진행됨에 따라서, 모집단 내에 항체의 유사성이 연속적으로 향상된다. 항체의 균질성을 방지하고 항체의 다양성을 향상시키고, 그리고 따라서, 전역 검색 능력을 향상시키고 성급한 수렴을 예방하기 위해, 모집단 유사성 S가 역치 S_0보다 클 때, 면역계 세포의 물질대사 기능이 P 새로운 항체를 산출하도록 모의되고, 그리

고 산출 과정은 상기 (2)와 동일하고, 따라서 항체의 총수는 $P+N$ 에 도달한다. 만약 모집단 유사성 S가 역치 S_0 보다 적으면, 모집단은 진화의 차세대에 계속해서 직접적으로 들어가고, 그리고 기억 세포가 갱신된다.

[0044] (7) $S>S_0$일 때, 항체 농도 및 중합화 적합도가 항체 모집단 $P+N$에 대해 계산된다. 항체 농도는 모집단 내에서 각 항체와 유사한 항체의 백분율, 다시 말하면,

[0045] $$C_i = \frac{A_i}{N-1}$$ 을 지칭하고,

[0046] 여기서 Ai는 항체 i에 대한 유사성이 유사성 상수 λ보다 큰 항체의 숫자를 지칭한다. λ는 2개의 개체가 비교될 때 M 코돈 사이에서 동일한 코돈의 숫자를 지칭한다.

[0047] 중합화 적합도 F'는 항체 적합도 F가 항체 농도에 따라서 교정된 후 획득된 값, 다시 말하면:

[0048] $$F_i^{'} = \alpha \frac{F_i}{\sum_i^N F_i} + (1-\alpha)\frac{A_i}{\sum_i^N A_i} (0<\alpha<1)$$ 이다.

[0049] 중합화 적합도에 따라서, 자손 모집단이 선택되고, 그리고 기억 세포가 갱신되고, 그리고 그 다음 라운드의 최적화가 실행된다. 우리가 2가지 서열 특질, 코돈 콘텍스트 및 CAI를 동시에 고려하기 때문에, $F'_{[\text{코돈 콘텍스트}]}$는 $F_{[\text{코돈 콘텍스트}]}$에 근거하여 계산되고, 그리고 $F'_{[CAI]}$는 $F_{[CAI]}$에 근거하여 계산된다. 만약 종결 대수가 도달되면, 진화가 중지되고, 그리고 단일 단백질 단편의 최적화된 서열 세트가 출력된다.

[0050] II. 유전자 알고리즘 (다시 말하면, 전역 최적화, 흐름에 대해 도 3을 참조한다).

[0051] 면역 알고리즘을 통한 최적화에 의해 산출된 모든 단백질 단편의 최적화된 서열 세트에 근거하여, 초기화된 모집단 N이 무작위로 산출된다. 유전자 알고리즘의 흐름에 따라서, 적합도 계산, 자손 모집단의 선택, 교차, 변이 및 기억 갱신이 완결된다. 만약 종결 대수가 도달되면, 진화가 중지되고, 그리고 단백질의 전체 서열에 대한 최적 DNA 코딩 서열이 출력된다. 전체 흐름은 다목적 최적화에 속한다. 최적화 과정에서, 우리는 각 유전자를 인코딩하기 위해 각 단백질 단편의 최적화된 서열 세트를 직접적으로 이용한다.

[0052] III. 미세 조정과 최적화.

[0053] 미세 조정과 최적화는 2 단계로 구성된다: 첫 번째, DNA의 5' 말단을 최적화하고, 그리고 이후, 발현 저해성 모티프를 제거. DNA의 5' 말단의 최적화 과정은 도 4에서 도시된 바와 같다. 전면적 방법이 단백질의 N 말단 아미노산 서열 (8-15개 아미노산)의 모든 가능한 DNA 코딩 서열을 열거하고, 그리고 이들의 코돈 콘텍스트 및 CAI를 계산하는데 이용된다. 이후, 단백질 서열의 개시 코돈의 상류에 위치된 50 bp (50 bp의 디폴트 값, 그리고 0-50 bp의 선택가능 길이 범위)의 벡터 서열이 DNA 코딩 서열에 연속적으로 연결되고, 그리고 연결된 서열의 최소 자유 에너지 (MFE)가 소프트웨어 mfold에 의해 계산된다. 최고 5'-말단 서열을 선택하기 위해, 최소 자유 에너지 (값이 더욱 클수록, 더욱 우수하다), 코돈 콘텍스트 (값이 더욱 클수록, 더욱 우수하다) 및 CAI (값이 더욱 클수록, 더욱 우수하다)에 따라서, 신호 펩티드의 코딩 서열이 분류된다.

[0054] IV. 상기 흐름의 상세

[0055] (1) 기본 데이터 세트 및 이중 코돈 표의 산출

[0056] 기본 데이터 세트는 상이한 숙주 발현 시스템에서 고도로-발현된 단백질 및 이들의 상응하는 DNA 코딩 서열을 지칭한다. 이중 코돈 표는 기본 데이터 세트에서 모든 이중 코돈의 상대적 적합도를 지칭한다 (계산 방법에 대해 하기를 참조한다).

[0057] (2) 코돈 콘텍스트 및 CAI의 계산 흐름

[0058] a) 코돈 상대적 적합도 w_{ij}:

$$w_{ij} = \frac{x_{ij}}{x_{i최대}}$$

[0059]

[0060] 여기서 x_{ij}는 기본 데이터 세트에서 나타난 아미노산의 i-번째 유형의 j-번째 동의 코돈의 숫자를 나타내고, 그리고 $x_{i최대}$는 기본 데이터 세트에서 나타난 아미노산의 i-번째 유형에 대해 가장 높은 이용 빈도를 갖는 동의 코돈의 숫자를 나타낸다.

[0061] b) 표적 서열의 코돈 적응 지수 (CAI):

$$CAI = \left(\prod_{k=1}^{L} w_k \right)^{\frac{1}{L}}$$

[0062]

[0063] 여기서 L은 표적 서열 (다시 말하면, 단백질 서열 또는 단편)의 아미노산의 숫자를 지칭하고, w_k는 각 아미노산 코돈에 의해 이용된 코돈에 상응하는 기본 데이터 세트의 코돈 상대적 적합도이다. CAI는 0 및 1 사이에 값을 갖는다. 최적화 과정에서, 우리는 인코딩 DNA의 CAI 값을 증가시키기 위해 최선의 노력을 다한다.

[0064] c) 이중 코돈의 상대적 적합도 p_k:

$$p_{\vec{k}} = \frac{\alpha_{CC}^{\vec{k}}}{\alpha_{AA}^{j(\vec{k})}}, \forall k \subseteq \{1,2,\ldots,3721\}$$

[0065]

[0066] 여기서 3,721가지 종류의 이중 코돈 (61x61 = 3721, 종결 코돈을 고려하지 않음)이 있고, α_{CC}^{k}는 단백질 서열 기본 데이터 세트 또는 표적 서열 (다시 말하면, 단백질 서열 또는 이의 단편)에서 나타난 이중 코돈의 k-번째 유형의 숫자를 나타내고, 그리고 $\alpha_{AA}^{j(k)}$ 는 나타난 바와 같은 이중 코돈에 상응하는 이중 아미노산의 숫자를 나타낸다.

[0067] d) 표적 서열의 코돈 콘텍스트 (CC):

$$CC = 1 - \frac{\sum_{k=1}^{3721} \left| p_0^{\vec{k}} - p_1^{\vec{k}} \right|}{3721}$$

[0068]

[0069] 여기서 $p_0^{\vec{k}}$ 는 표적 서열의 이중 코돈의 k-번째 유형의 상대적 적합도를 나타내고, 그리고 $p_1^{\vec{k}}$ 는 기본 데이터 세트의 이중 코돈의 k-번째 유형의 상대적 적합도를 나타낸다. CC는 0 및 1 사이에 값을 갖는다. 최적화 과정에서, 우리는 인코딩 DNA의 CC 값을 증가시키기 위해 최선을 다한다.

[0070] (3) NSGA2 및 SPEA2 알고리즘 (NSGA2는 디폴트에 의해 이용된다)이 면역 알고리즘 및 유전자 알고리즘의 다목적 최적화 과정에서 자손 모집단의 선택에 이용될 수 있고, 그리고 2-포인트 교차가 교차에 이용된다.

[0071] 하기는 실례에 의해 본 발명의 이점을 더욱 예증한다.

[0072] 검사에서 이용된 숙주 발현 시스템은 CHO 세포주이고, 그리고 2개의 단백질은 전체적으로 최적화되고 염기서열 결정된다 (유관한 정보를 위해 표 1을 참조한다). JNK3 단백질 서열은 서열 번호 1에서 나타나 있는 바와 같고, 그리고 GFP 단백질 서열은 서열 번호 2에서 나타나 있는 바와 같고; 최적화 전 JNK3 단백질 및 GFP 단백질의 코딩 서열은 각각, 서열 번호 3 및 서열 번호 4에서 나타나 있는 바와 같고, 그리고 최적화 후 JNK3 단백질 및 GFP 단백질의 코딩 서열은 각각, 서열 번호 5 및 서열 번호 6에서 나타나 있는 바와 같다.

[0073] 표 1: 최적화된 검사 단백질 서열의 정보

표 1

[0074]

단백질	GenBank 수탁 번호 (야생형)	태그	태그의 위치
JNK3	U34820.1	Flag 태그	C 말단
GFP	AY174111.1	Flag 태그	C 말단

[0075] 도 5에서 도시된 바와 같이, 검사 단백질을 인코딩하는 유전자 단편이 합성되고, 그리고 각각, EcoR I 및 Hind III 개열 부위를 통해 pTT5 발현 벡터 (NRC로부터 구입됨, 그리고 플라스미드 지도는 도 6에서 도시된 바와 같다) 내로 클로닝된다.

[0076] **CHO 3E7 세포의 일시적인 발현 단계:**

[0077] 1. 대수증식기에서 CHO 3E7 현탁 세포가 신선한 FreeStyle CHO 배지로 5 x 10^5 세포 /mL로 희석되고, 그리고 30 mL의 세포 현탁액이 각 125 mL 삼각형 플라스크에서 접종된다.

[0078] 2. 이들 세포는 37℃ 및 5% CO_2의 조건 하에 현탁 배양에 종속된다.

[0079] 3. 세포 밀도가 1-1.2 x 10^6 세포 /mL에 도달할 때, 클로닝된 표적 유전자를 보유하는 플라스미드 벡터가 PEI 형질감염 시약에 의해 1 ug/ml의 용량에 따라서 CHO 3E7 세포 내로 각각 형질감염된다.

[0080] 4. 48 시간의 형질감염 후, 세포를 수확하기 위해 배양 배지가 1500 회전/분에서 원심분리된다. 표본은 냉장고에서 -80 ℃에서 보관될 수 있다.

[0081] **웨스턴 블롯 실험 단계:**

[0082] 항-Flag 태그 항체를 이용하여, 세포 용해물 내에 표적 단백질의 발현량이 웨스턴 블롯팅에 의해 검출되었다. 베타-액틴 단백질은 내부 참조로서 이용된다. 각 플라스미드의 발현 실험은 3회 반복된다. 웨스턴 블롯팅의 결과는 도 7에서 도시된다.

[0083] 상술된 단계는 하기와 같다.

[0084] 1. CHO 세포가 세포 용해 완충액을 이용하여 용해되고, 그리고 단백질 농도가 결정된다.

[0085] 2. 단백질 용액은 5X SDS-PAGE 단백질 부하 완충액이 첨가되고, 그리고 항온 수조에서 10 분 동안 가열된다.

[0086] 3. 단백질 표본이 미량피펫으로 SDS-PAGE 겔의 표본 부하 웰 내로 첨가되고, 그리고 각 웰이 20 ul의 표본으로 부하된다.

[0087] 4. 140 V에서 일정한 전압 전기이동이 60 분 동안 이용되고, 그리고 브로모페놀 블루가 겔의 바닥에 가깝게 도달할 때 전기이동이 중지된다.

[0088] 5. 막 전달 전압은 100 V이고, 그리고 낮은 온도에서 막 전달 시간은 60 분이다.

[0089] 6. 막 전달이 완결된 후, 단백질 막이 사전에 제조된 세척액에 배치되고, 그리고 막 상에서 막 전달 액체를 제거하기 위해 1-2 분 동안 헹굼된다.

[0090] 7. 이것은 실온에서 45 분 동안 진탕기에서 천천히 진탕함으로써 차단된다.

[0091] 8. 이것은 희석된 일차 항체가 첨가되고, 그리고 천천히 진탕하면서 실온에서 1 시간 동안 배양된다.

[0092] 9. 이것은 세척액이 첨가되고, 그리고 총 3 회 5 분 동안 진탕기에서 세척을 위해 천천히 진탕된다.

[0093] 10. 이것은 희석된 이차 항체가 첨가되고, 그리고 천천히 진탕하면서 실온에서 1 시간 동안 배양된다.

[0094] 11. 이것은 세척액이 첨가되고, 그리고 총 3 회 5 분 동안 진탕기에서 세척을 위해 천천히 진탕된다.

[0095] 12. 화학발광 검출.

[0096] 13. 웨스턴 블롯팅 결과 사진은 소프트웨어 Image J로 정량적으로 분석된다.

[0097] 표 2: 최적화 전후에 단백질의 상대적 발현량 (웨스턴 블롯팅에 의해 검출될 때)

표 2

[0098]

	GFP (상대적 발현량 ± 표준 편차)	JNK3 (상대적 발현량 ± 표준 편차)
최적화 후	22.06 ± 1.78	8.01 ± 0.21
야생형	1.19 ± 0.16	1.09 ± 0.10
비율	18.37 ± 2.90	7.42 ± 0.58

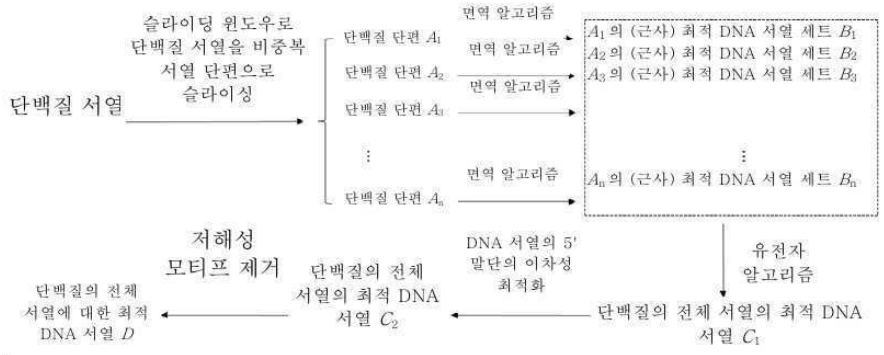
[0099] *상대적 발현량: 야생형 서열의 3회 반복된 실험에서 발현량의 최소 값에 의해 나눗셈된 단백질 발현량

[0100] 표 2로부터 목격될 수 있는 바와 같이, 본 특허의 3-단계 하이브리드 코돈 최적화에 종속된 후 JNK3 및 GFP 단백질의 발현량은 각각, 야생형 서열의 것과 비교하여 7.42 ± 0.58 배 및 18.37 ± 2.90 배 증가되는데, 이것은 새로운 알고리즘의 높은 효율을 완전히 입증한다. 컴퍼니 (company)의 실제 생산에서, 우리는 또한, 복수의 단백질에 대한 이러한 알고리즘 및 다른 알고리즘의 최적화 효과를 비교하고 검사하는데, 이것 역시 이러한 알고리즘이 더욱 안정되고 효율적이라는 것을 입증한다.

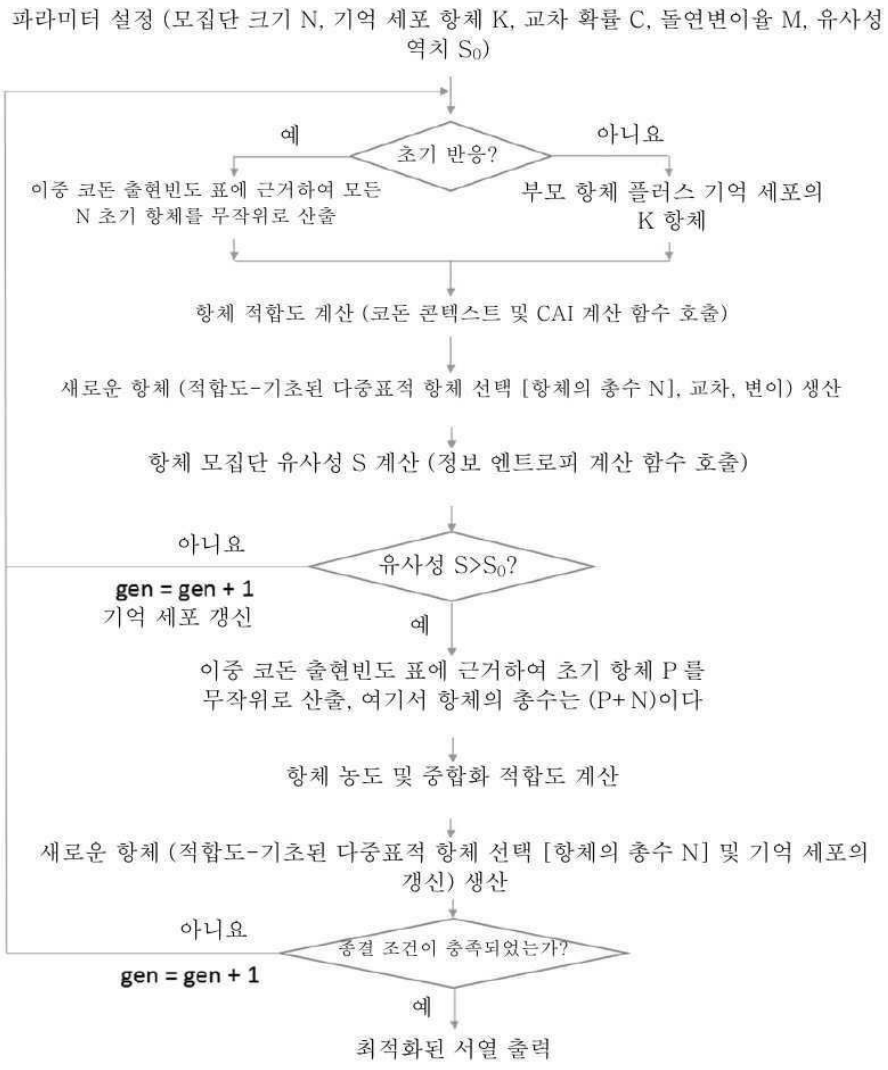
[0101] 본 발명에 관련되지 않은 부분은 선행 기술에서 것들과 모두 동일하거나, 또는 선행 기술을 이용함으로써 실현될 수 있다.

도면

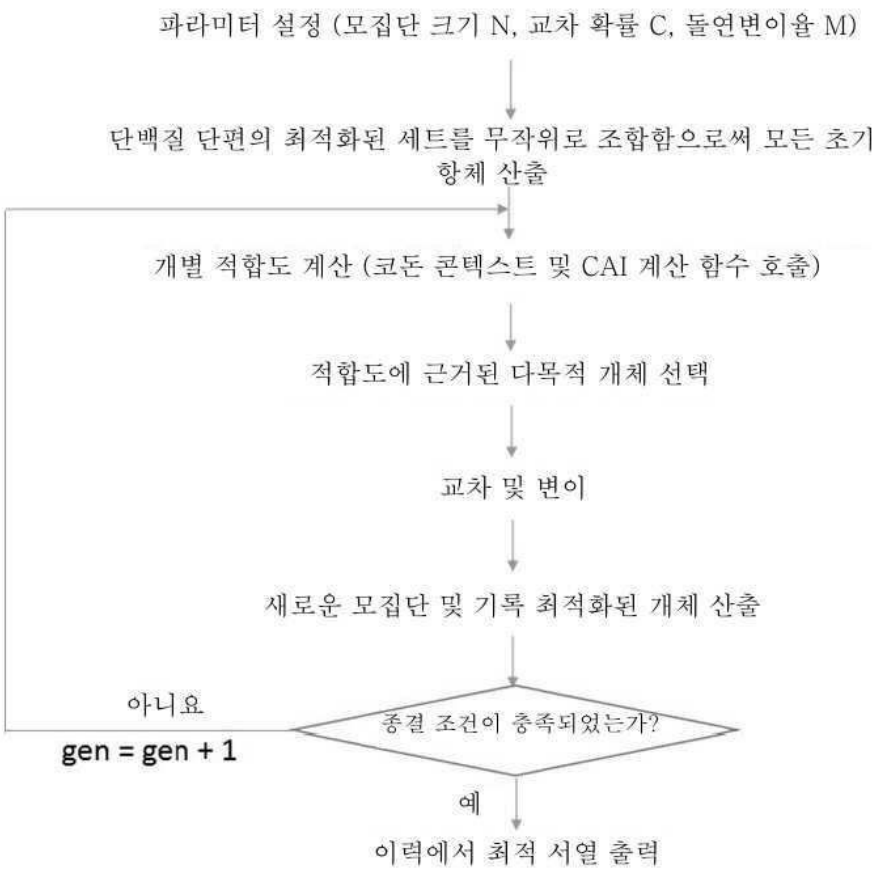
도면1



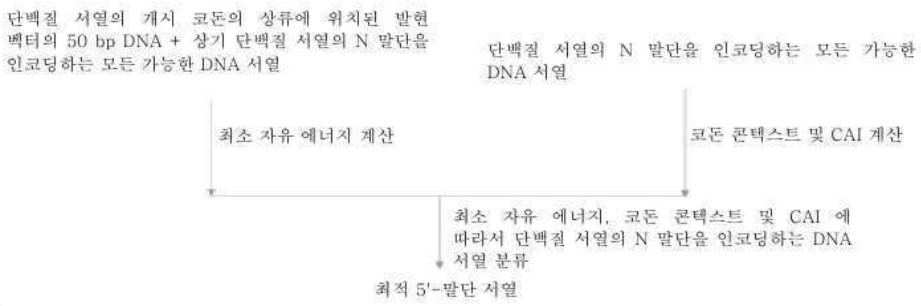
도면2


파라미터 설정 (모집단 크기 N, 기억 세포 항체 K, 교차 확률 C, 돌연변이율 M, 유사성 역치 S_0)
초기 반응?
예
아니요
이중 코돈 출현빈도 표에 근거하여 모든 N 초기 항체를 무작위로 산출
부모 항체 플러스 기억 세포의 K 항체
항체 적합도 계산 (코돈 콘텍스트 및 CAI 계산 함수 호출)
새로운 항체 (적합도-기초된 다중표적 항체 선택 [항체의 총수 N], 교차, 변이) 생산
항체 모집단 유사성 S 계산 (정보 엔트로피 계산 함수 호출)
유사성 $S>S_0$?
아니요
gen = gen + 1
기억 세포 갱신
예
이중 코돈 출현빈도 표에 근거하여 초기 항체 P 를 무작위로 산출, 여기서 항체의 총수는 (P+N)이다
항체 농도 및 중합화 적합도 계산
새로운 항체 (적합도-기초된 다중표적 항체 선택 [항체의 총수 N] 및 기억 세포의 갱신) 생산
종결 조건이 충족되었는가?
아니요
gen = gen + 1
예
최적화된 서열 출력

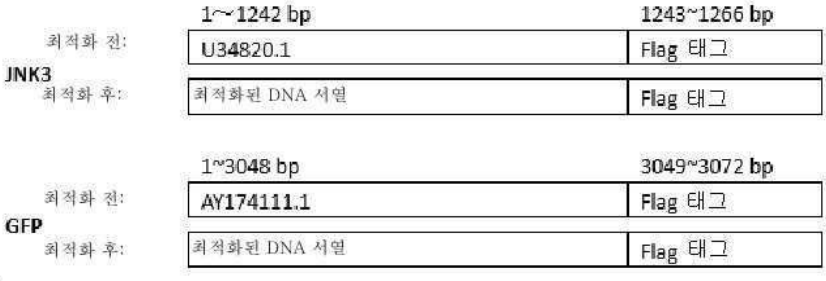
도면3


파라미터 설정 (모집단 크기 N, 교차 확률 C, 돌연변이율 M)
단백질 단편의 최적화된 세트를 무작위로 조합함으로써 모든 초기 항체 산출
개별 적합도 계산 (코돈 콘텍스트 및 CAI 계산 함수 호출)
적합도에 근거된 다목적 개체 선택
교차 및 변이
새로운 모집단 및 기록 최적화된 개체 산출
종결 조건이 충족되었는가?
아니요
gen = gen + 1
예
이력에서 최적 서열 출력

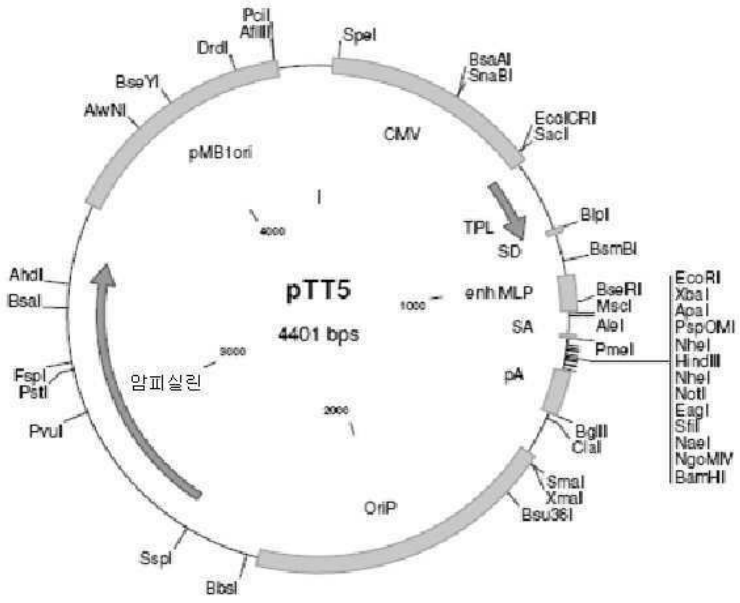
도면4


단백질 서열의 개시 코돈의 상류에 위치된 발현 벡터의 50 bp DNA + 상기 단백질 서열의 N 말단을 인코딩하는 모든 가능한 DNA 서열
단백질 서열의 N 말단을 인코딩하는 모든 가능한 DNA 서열
최소 자유 에너지 계산
코돈 콘텍스트 및 CAI 계산
최소 자유 에너지, 코돈 콘텍스트 및 CAI 에 따라서 단백질 서열의 N 말단을 인코딩하는 DNA 서열 분류
최적 5'-말단 서열

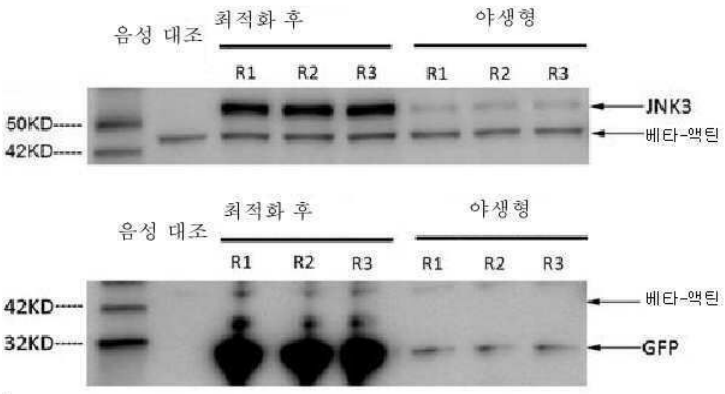
도면5


1~1242 bp
1243~1266 bp
최적화 전:
U34820.1
Flag 태그
JNK3
최적화 후:
최적화된 DNA 서열
Flag 태그
1~3048 bp
3049~3072 bp
최적화 전:
AY174111.1
Flag 태그
GFP
최적화 후:
최적화된 DNA 서열
Flag 태그

도면6



도면7



서 열 목 록

<110> Nanjingjinsirui Science & Technology Biology Corp.

<120> CODON OPTIMIZATION METHOD BASED ON IMMUNE ALGORITHM

<160> 6

<210> 1

<211> 430

<212> PRT

<213> Artificial sequence

<220><223> JNK3 protein sequence

<400> 1

Met Ser Leu His Phe Leu Tyr Tyr Cys Ser Glu Pro Thr Leu Asp Val

1 5 10 15

Lys Ile Ala Phe Cys Gln Gly Phe Asp Lys Gln Val Asp Val Ser Tyr
20 25 30
Ile Ala Lys His Tyr Asn Met Ser Lys Ser Lys Val Asp Asn Gln Phe

35 40 45
Tyr Ser Val Glu Val Gly Asp Ser Thr Phe Thr Val Leu Lys Arg Tyr
50 55 60
Gln Asn Leu Lys Pro Ile Gly Ser Gly Ala Gln Gly Ile Val Cys Ala
65 70 75 80
Ala Tyr Asp Ala Val Leu Asp Arg Asn Val Ala Ile Lys Lys Leu Ser
85 90 95
Arg Pro Phe Gln Asn Gln Thr His Ala Lys Arg Ala Tyr Arg Glu Leu

100 105 110
Val Leu Met Lys Cys Val Asn His Lys Asn Ile Ile Ser Leu Leu Asn
115 120 125
Val Phe Thr Pro Gln Lys Thr Leu Glu Glu Phe Gln Asp Val Tyr Leu
130 135 140
Val Met Glu Leu Met Asp Ala Asn Leu Cys Gln Val Ile Gln Met Glu
145 150 155 160
Leu Asp His Glu Arg Met Ser Tyr Leu Leu Tyr Gln Met Leu Cys Gly

165 170 175
Ile Lys His Leu His Ser Ala Gly Ile Ile His Arg Asp Leu Lys Pro
180 185 190
Ser Asn Ile Val Val Lys Ser Asp Cys Thr Leu Lys Ile Leu Asp Phe
195 200 205
Gly Leu Ala Arg Thr Ala Gly Thr Ser Phe Met Met Thr Pro Tyr Val
210 215 220
Val Thr Arg Tyr Tyr Arg Ala Pro Glu Val Ile Leu Gly Met Gly Tyr

225 230 235 240
Lys Glu Asn Val Asp Ile Trp Ser Val Gly Cys Ile Met Gly Glu Met
245 250 255
Val Arg His Lys Ile Leu Phe Pro Gly Arg Asp Tyr Ile Asp Gln Trp

260 265 270

Asn Lys Val Ile Glu Gln Leu Gly Thr Pro Cys Pro Glu Phe Met Lys

275 280 285

Lys Leu Gln Pro Thr Val Arg Asn Tyr Val Glu Asn Arg Pro Lys Tyr

290 295 300

Ala Gly Leu Thr Phe Pro Lys Leu Phe Pro Asp Ser Leu Phe Pro Ala

305 310 315 320

Asp Ser Glu His Asn Lys Leu Lys Ala Ser Gln Ala Arg Asp Leu Leu

325 330 335

Ser Lys Met Leu Val Ile Asp Pro Ala Lys Arg Ile Ser Val Asp Asp

340 345 350

Ala Leu Gln His Pro Tyr Ile Asn Val Trp Tyr Asp Pro Ala Glu Val

355 360 365

Glu Ala Pro Pro Pro Gln Ile Tyr Asp Lys Gln Leu Asp Glu Arg Glu

370 375 380

His Thr Ile Glu Glu Trp Lys Glu Leu Ile Tyr Lys Glu Val Met Asn

385 390 395 400

Ser Glu Glu Lys Thr Lys Asn Gly Val Val Lys Gly Gln Pro Ser Pro

405 410 415

Ser Ala Gln Val Gln Gln Asp Tyr Lys Asp Asp Asp Asp Lys

420 425 430

<210> 2

<211> 246

<212> PRT

<213> Artificial sequence

<220><223> GFP protein sequence

<400> 2

Met Ser Lys Gly Glu Glu Leu Phe Thr Gly Val Val Pro Ile Leu Val

1 5 10 15

Glu Leu Asp Gly Asp Val Asn Gly Gln Lys Phe Ser Val Ser Gly Glu

20 25 30

Gly Glu Gly Asp Ala Thr Tyr Gly Lys Leu Thr Leu Lys Phe Ile Cys

35 40 45

Thr Thr Gly Lys Leu Pro Val Pro Trp Pro Thr Leu Val Thr Thr Phe
50 55 60
Ser Tyr Gly Val Gln Cys Phe Ser Arg Tyr Pro Asp His Met Lys Gln
65 70 75 80
His Asp Phe Phe Lys Ser Ala Met Pro Glu Gly Tyr Val Gln Glu Arg
85 90 95
Thr Ile Phe Tyr Lys Asp Asp Gly Asn Tyr Lys Thr Arg Ala Glu Val
100 105 110

Lys Phe Glu Gly Asp Thr Leu Val Asn Arg Ile Glu Leu Lys Gly Ile
115 120 125
Asp Phe Lys Glu Asp Gly Asn Ile Leu Gly His Lys Met Glu Tyr Asn
130 135 140
Tyr Asn Ser His Asn Val Tyr Ile Met Ala Asp Lys Pro Lys Asn Gly
145 150 155 160
Ile Lys Val Asn Phe Lys Ile Arg His Asn Ile Lys Asp Gly Ser Val
165 170 175

Gln Leu Ala Asp His Tyr Gln Gln Asn Thr Pro Ile Gly Asp Gly Pro
180 185 190
Val Leu Leu Pro Asp Asn His Tyr Leu Ser Thr Gln Ser Ala Leu Ser
195 200 205
Lys Asp Pro Asn Glu Lys Arg Asp His Met Ile Leu Leu Glu Phe Val
210 215 220
Thr Ala Ala Gly Ile Thr His Gly Met Asp Glu Leu Tyr Lys Asp Tyr
225 230 235 240

Lys Asp Asp Asp Asp Lys
245

<210> 3
<211> 1290
<212> DNA
<213> Artificial sequence

<220><223> JNK3 protein coding sequence before optimization

<400> 3

atgagcctcc atttcttata ctactgcagt gaaccaacat tggatgtgaa aattgccttt 60

tgtcagggat tcgataaaca agtggatgtg tcatatattg ccaaacatta caacatgagc 120

aaaagcaaag ttgacaacca gttctacagt gtggaagtgg gagactcaac cttcacagtt 180

ctcaagcgct accagaatct aaagcctatt ggctctgggg ctcagggcat agtttgtgcc 240

gcgtatgatg ctgtccttga cagaaatgtg gccattaaga agctcagcag acccctttcag 300

aaccaaacac atgccaagag agcgtaccgg gagctggtcc tcatgaagtg tgtgaaccat 360

aaaaacatta ttagtttatt aaatgtcttc acaccccaga aaacgctgga ggagttccaa 420

gatgtttact tagtaatgga actgatggat gccaacttat gtcaagtgat tcagatggaa 480

ttagaccatg agcgaatgtc ttacctgctg taccaaatgt tgtgtggcat taagcacctc 540

cattctgctg gaattattca cagggattta aaaccaagta acattgtagt caagtctgat 600

tgcacattga aaatcctgga ctttggactg gccaggacag caggcacaag cttcatgatg 660

actccatatg tggtgacacg ttattacaga gcccctgagg tcatcctggg gatgggctac 720

aaggagaacg tggatatatg gtctgtggga tgcattatgg gagaaatggt tcgccacaaa 780

atcctctttc caggaaggga ctatattgac cagtggaata aggtaattga acaactagga 840

acaccatgtc cagaattcat gaagaaattg caacccacag taagaaacta tgtggagaat 900

cggcccaagt atgcgggact caccttcccc aaactcttcc cagattccct cttcccagcg 960

gactccgagc acaataaact caaagccagc caagccaggg acttgttgtc aaagatgcta 1020

gtgattgacc cagcaaaaag aatatcagtg gacgacgcct tacagcatcc ctacatcaac 1080

gtctggtatg acccagccga agtggaggcg cctccacctc agatatatga caagcagttg 1140

gatgaaagag aacacacaat tgaagaatgg aaagaactta tctacaagga agtaatgaat 1200

tcagaagaaa agactaaaaa tggtgtagta aaaggacagc cttctccttc agcacaggtg 1260

cagcaggact acaaggatga tgatgacaaa 1290

<210> 4

<211> 738

<212> DNA

<213> Artificial sequence

<220><223> GFP protein coding sequence before optimization

<400> 4

atgagtaaag gagaagaact tttcactgga gttgtcccaa ttcttgttga attagatggc 60

gatgttaatg ggcaaaaatt ctctgtcagt ggagagggtg aaggtgatgc aacatacgga 120

aaacttaccc ttaaatttat ttgcactact gggaagctac ctgttccatg gccaacactt 180

gtcactactt tctcttatgg tgttcaatgc ttttcaagat acccagatca tatgaaacag 240

catgactttt tcaagagtgc catgcccgaa ggttatgtac aggaaagaac tatattttac 300

aaagatgacg ggaactacaa gacacgtgct gaagtcaagt ttgaaggtga tacccttgtt 360

aatagaatcg agttaaaagg tattgatttt aaagaagatg gaaacattct tggacacaaa 420

atggaataca actataactc acataatgta tacatcatgg cagacaaacc aaagaatgga 480

atcaaagtta acttcaaaat tagacacaac attaaagatg gaagcgttca attagcagac 540

cattatcaac aaaatactcc aattggcgat ggccctgtcc ttttaccaga caaccattac 600

ctgtccacac aatctgccct ttccaaagat cccaacgaaa agagagatca catgatcctt 660

cttgagtttg taacagctgc tgggattaca catggcatgg atgaactata caaagactac 720

aaagatgatg atgacaag 738

<210> 5

<211> 1290

<212> DNA

<213> Artificial sequence

<220><223> JNK3 protein coding sequence after optimization

<400> 5

atgtctctgc acttcctgta ctactgttct gagcccaccc tggacgtgaa gattgccttc 60

tgccagggct ttgacaagca ggtggatgtg agctacatcg ccaagcacta caacatgtcc 120

aagagcaagg tggacaacca gttctacagc gtggaggtgg gagacagcac cttcacagtg 180

ctgaagagat accagaacct gaagccaatt ggctctggag cccagggcat tgtgtgtgct 240

gcctatgatg ctgtgctgga cagaaatgtg gccatcaaga agctgagcag acccttccag 300

aaccagacac atgccaagag agcctacaga gagctggtgc tgatgaagtg tgtgaaccac 360

aagaacatca tcagcctgct gaatgtgttc acccctcaga agacactgga ggagttccag 420

gatgtgtacc tggtgatgga gctcatggat gccaacctgt gccaggtgat ccagatggag 480

ctggaccatg agaggatgag ctacctgctg taccagatgc tgtgtggcat caagcacctg 540

cacagtgctg gaatcatcca cagagacctg aagccaagca acattgtggt gaagtctgac 600

tgtacactga agatcctgga ctttggactg gccagaacag ccggcacatc ttttatgatg 660

acaccatacg tggtgacaag atactacaga gcccctgagg tgatcctggg catgggctac 720

aaggagaacg tggacatctg gtctgtgggc tgcatcatgg gagagatggt gagacacaag 780

atcctgtttc ctggaagaga ctacattgac cagtggaaca aggtgattga gcagctgggc 840

acccccttgtc ctgagttcat gaagaagctg cagccaactg tgaggaacta tgtggagaac 900

agaccaaagt atgctggcct gaccttcccc aagctcttcc ctgacagcct gtttcctgct 960

gattctgagc acaacaagct gaaggccagc caggccagag acctgctgag caagatgctg 1020

gtgattgatc ctgccaagag aatctctgtg gatgatgccc tgcagcaccc ctacatcaat 1080

gtgtggtacg acccagctga ggtggaggcc ccacctccac agatctatga caagcagctg 1140

gatgagagag agcacacaat tgaagagtgg aaggagctga tctacaaaga agtgatgaac 1200

tctgaggaga agaccaagaa tggagtggtg aagggccagc cctctccaag cgcccaggtg 1260

cagcaggact acaaggatga tgatgacaaa 1290

<210> 6

<211> 738

<212> DNA

<213> Artificial sequence

<220><223> GFP protein coding sequence after optimization

<400> 6

atgagcaagg gagaggaact gttcacagga gtggtgccca tcctggtgga gctggatgga 60

gatgtgaatg gccagaagtt ttctgtgtct ggggaaggag aaggcgatgc cacctatggc 120

aagctgacac tgaagttcat ctgcaccaca gggaagctgc ctgtgccctg gccaacactg 180

gtgaccacct tctcctatgg agtccagtgc ttcagcagat acccagacca catgaagcag 240

catgacttct tcaagagtgc catgcctgag ggctatgtgc aggagagaac catcttctat 300

aaggatgatg gaaactacaa gacaagagct gaggtgaagt ttgagggaga caccctggtg 360

aacagaattg agctgaaggg cattgacttc aaggaggatg gcaacatcct gggccacaag 420

atggagtaca attacaacag ccacaatgtg tacatcatgg ctgataagcc aaagaatgga 480

atcaaggtga acttcaagat tagacacaac atcaaagacg gatctgtgca gctggctgac 540

cattaccagc agaacacacc cattggagat ggcccagtgc tgctgcccga caaccactac 600

ctgagcacac agtctgccct gagtaaggac cctaatgaga agagggacca catgattctg 660

ctggagtttg tgacagctgc tggcatcacc catggcatgg atgagctgta caaggactac 720

aaagatgatg atgacaag 738