(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0143551
(43) 공개일자 2020년12월24일

(51) 국제특허분류(Int. Cl.)
G16C 20/30 (2019.01) *G06N 3/00* (2019.01)
G16C 20/70 (2019.01)
(52) CPC특허분류
G16C 20/30 (2019.02)
G06N 3/004 (2013.01)
(21) 출원번호 10-2019-0069924
(22) 출원일자 2019년06월13일
심사청구일자 2019년06월13일

(71) 출원인
부경대학교 산학협력단
부산광역시 남구 용소로 45, 부경대학교내 (대연동)
(72) 발명자
유준
부산광역시 해운대구 좌동순환로217번길 19, 309동 1601호(좌동, 엘아이지건영아파트)
주벨라, 페타
싱가포르 주롱 웨스트 스트릿 52 블락 530 5-381
(뒷면에 계속)
(74) 대리인
김정수

전체 청구항 수 : 총 6 항

(54) 발명의 명칭 **화합물의 크로마토그래피 용출 순서를 예측하는 방법**

(57) 요 약

본 발명은 (a) 정량적 구조-체류 관계(quantitative structure-retention relationship, QSRR) 관계식을 모델링하는 단계; 및 (b) 수리 계획법(mathematical programming)을 이용해 상기 QSRR 관계식으로부터 혼합물 내 화합물의 크로마토그래피 용출 순서를 예측하는 단계;를 포함하며, 상기 수리 계획법은, (i) 화합물의 예측된 용출

(뒷면에 계속)

대 표 도 - 도2



순서를 제약조건(constraint)으로 사용하는 비선형 계획법(non-linear programming) 또는 (ii) 체류시간 예측 오차와 용출 순서 예측 오차를 목적함수로 사용하는 다목적 최적화(multi-objective optimization, MOO) 기법인 것을 특징으로 하는, 혼합물 내 화합물의 크로마토그래피 용출 순서를 예측하는 방법에 대한 것으로서, 본 발명에 의하면 오믹스 분석(프로테오믹스, 메타볼로믹스 등)에서 주어진 펩타이드 또는 대사물을 식별하는데 있어 보다 높은 확실성을 제공하는 개별 화합물(펩타이드 혹은 대사물)의 정확한 위치 확인을 가능케 하여 크로마토그래피 분야 특히, 역상 또는 이온 교환 모드의 크로마토그래피에서 복잡한 혼합물의 분리를 최적화할 수 있다.

(52) CPC특허분류
G16C 20/70 (2019.02)

(72) 발명자
본첵, 토마스
폴란드 그단스크 80-416 할러 107 그단스크 의과대학

알리풀리, 알함
부산광역시 남구 용소로 45(대연동)

명 세 서

청구범위

청구항 1

(a) 정량적 구조-체류 관계(quantitative structure-retention relationship, QSRR) 관계식을 모델링하는 단계; 및

(b) 수리 계획법(mathematical programming)을 이용해 상기 QSRR 관계식으로부터 혼합물 내 화합물의 크로마토그래피 용출 순서를 예측하는 단계;를 포함하며,

상기 수리 계획법은, (i) 화합물의 예측된 용출 순서를 제약조건(constraint)으로 사용하는 비선형 계획법(non-linear programming) 또는 (ii) 체류시간 예측 오차와 용출 순서 예측 오차를 목적함수로 사용하는 다목적 최적화(multi-objective optimization, MOO) 기법인 것을 특징으로 하는,

혼합물 내 화합물의 크로마토그래피 용출 순서를 예측하는 방법.

청구항 2

제1항에 있어서,

상기 단계 (a)에서 모델링해 얻어지는 QSRR 관계식은 하기 식으로 표시되는 선형 모델인 것을 특징으로 하는 혼합물 내 화합물의 크로마토그래피 용출 순서를 예측하는 방법:

$$t_{R,j} = a_1 x_{j,1} + a_2 x_{j,2} + \ldots + a_n x_{j,n}$$

(상기 식에서 $t_{R,j}$은 오름차순으로 정렬된 화합물 j의 체류시간, $x_{j,i}$ (i = 1, ..., n)는 화합물 j 의 분자 설명자(molecular descriptor), a_i (i = 1, ..., n)는 회귀계수(regression coefficient)임).

청구항 3

제1항에 있어서,

상기 단계 (a)에서 모델링해 얻어지는 QSRR 관계식은 인공신경망(artificial neural networks, ANN)을 이용해 얻어진 비선형 모델인 것을 특징으로 하는 혼합물 내 화합물의 크로마토그래피 용출 순서를 예측하는 방법.

청구항 4

제1항에 있어서,

상기 단계 (b)에서 아래 (II)의 부등식 제약조건(constraint) 하에 아래 (I)의 비선형 계획법을 적용해 혼합물 내 화합물의 크로마토그래피 용출 순서를 예측하는 것을 특징으로 하는 혼합물 내 화합물의 크로마토그래피 용출 순서를 예측하는 방법:

$$\min_{\mathbf{a}} \left\{ \sum_{j=1}^{m} \left(t_{R,j} - a_1 x_{j,1} - a_2 x_{j,2} - a_3 x_{j,3} \right)^2 + \sum_{j=1}^{m-1} \alpha_j \right\} \quad \text{(I)}$$

$$a_1 \left(x_{j,1} - x_{j+1,1} \right) + a_2 \left(x_{j,2} - x_{j+1,2} \right) + a_3 \left(x_{j,3} - x_{j+1,3} \right) - \alpha_j \leq 0 \quad \text{(II)}$$

(상기 식에서, α_j는 양수의 이완 매개변수, $\bar{\mathbf{a}}$ 는 $a1$, $a2$, $a3$와 α_j (j = 1, 2, …, m--으로 이루어진 결정변

수 벡터임).

청구항 5

제1항에 있어서,

상기 단계 (b)에서 하기 식 (III)로 표시되는 체류 시간 예측 오차를 나타내는 목적함수 및 하기 식 (IV)로 표시되는 용출 순서 예측 오차를 나타내는 목적함수에 기초해 다목적 최적화를 수행해 혼합물 내 화합물의 크로마토그래피 용출 순서를 예측하는 것을 특징으로 하는 혼합물 내 화합물의 크로마토그래피 용출 순서를 예측하는 방법:

$$\%RMSE(t_R) = \sqrt{\sum \left(\frac{t_R - \hat{t}_R}{t_R} \right)^2 \Big/ m} \times 100 \qquad \text{(III)}$$

(상기 식에서, t_R과 R $\hat{t}_R$는 각각 분석 실험에서 측정한 체류시간과 모델로 예측한 체류시간을 나타내며, m은 각 컬럼에 대해 체류시간을 측정한 혼합물의 개수를 나타냄),

$$\%RMSE(\text{order}) = \sqrt{\sum \left(\frac{\text{order}_{pred.} - \text{order}_{obs.}}{\text{order}_{obs.}} \right)^2 \Big/ m} \times 100 \qquad \text{(IV)}$$

(상기 식에서, $\text{order}_{obs.}$과 $\text{order}_{pred.}$는 각각 분석 실험에서 측정한 용출 순서와 예측된 체류시간으로부터 결정한 용출 순서를 나타냄).

청구항 6

제5항에 있어서,

상기 단계 (b)에서 하기 단계를 실시해 파레토 (Pareto) 최적해를 선택하는 다목적 최적화를 수행하는 것을 특징으로 하는 혼합물 내 화합물의 크로마토그래피 용출 순서를 예측하는 방법:

(b-1) 파레토 해들로 구성된 파레토 프론트(front)에서 체류시간과 용출 순서의 예측 오차의 최적 절충점인 굴곡점(knee point)을 선택하는 단계;

(b-2) 용출 순서 예측 오차를 감소시키는 다음 파레토 해로 이동하는 단계;

(b-3) 적용 구역(applicability domain)을 이용하여 해를 검증하는 단계; 및

(b-4) 체류시간 예측 오차의 증가가 미리 정한 임계값에 도달하거나 적용구역에서 이상점이 미리 정한 임계값을 넘을 때까지 상기 단계 (b-1) 및 (b-2)를 반복하는 단계.

발명의 설명

기 술 분 야

[0001] 본 발명은 혼합물 내 화합물의 용출 순서를 예측하는 방법에 대한 것에 대한 것으로서, 보다 상세하게는 수리계획법을 이용하여 정량 구조-체류 관계로부터 혼합물 내 화합물의 용출 순서를 정확하게 예측하는 방법에 대한 것에 대한 것이다.

배 경 기 술

[0002] 정량적 구조-체류 관계(quantitative structure-retention relationships, QSRR)를 이용한 분석물 내 화합물의 체류시간 예측은 크로마토그래피를 이용한 분리방법 개발에서 가장 널리 보급된 강력한 도구 중 하나이다. 종래의 많은 연구와 발명은 분석 대상인 화합물의 컬럼 내 체류시간 예측에 대한 여러 장점을 설명하지만, 용출 순

서의 예측에 대한 연구는 매우 드물다. 일반적으로 예측력이 높은 QSRR 모델은 이론적 예상은 물론 실험적으로도 관찰된 바와 같이 용출된 화합물의 순서를 잘 예측한다고 간주되지만, 간단한 혼합물(예를 들어 조성이 알려진 혼합물)에 대해서만 이러한 가정이 유효하며, 가스 및 키랄 크로마토그래피에 한해 용출 순서를 예측하는 방법들이 존재한다.

[0003] 한편, 복잡한 혼합물(약제나 단백질 등)의 분리를 90% 이상 차지하는 역상 액체 크로마토그래피(reversed-phase liquid chromatography, RP-LC)의 경우 그 메커니즘은 매우 복잡하여 아직까지도 완전히 이해되지 않고 있다. 따라서, QSRR 모델이 적당한 오차 이내로 체류시간을 예측할지라도, 수백 또는 수천 개의 매우 가깝거나 겹치는 피크들의 용출 순서 예측은 종종 매우 나쁜 결과를 보여준다. 더구나 RP-LC에서 이 복잡한 크로마토그래피 문제를 해결하기 위한 접근법은 연구 문헌에서도 사실상 전무하다.

선행기술문헌

특허문헌

[0004] (특허문헌 0001) 미국 공개특허 2010-0161530 A1 (공개일: 2010. 06. 24.)

(특허문헌 0002) 미국 등록특허 5,958,784 A (등록일: 1999. 09. 28.)

비특허문헌

[0005] (비특허문헌 0001) Kaliszan, R. QSRR: Quantitative Structure-(Chromatographic) Retention Relationships. Chem. Rev. 2007, 107, 3212-3246.

(비특허문헌 0002) H뱽berger, K. Quantitative structure-(chromatographic) retention relationships. J. Chromatogr. A 2007, 1158, 273-305.

(비특허문헌 0003) Amos, R. I. J.; Haddad, P. R.; Szucs, R.; Dolan, J. W.; Pohl, C. A. Molecular modeling and prediction accuracy in Quantitative Structure-Retention Relationship calculations for chromatography. TrAC Trends Anal. Chem. 2018 , 105, 352-359.

발명의 내용

해결하려는 과제

[0006] 본 발명은 상기한 바와 같은 종래기술의 문제점을 해결하기 위해 고안된 것으로, 역상 또는 이온 교환 모드의 크로마토그래피를 통해 복잡한 혼합물(약제나 단백질 등)의 분리를 최적화할 수 있도록 혼합물 내 화합물의 용출 순서를 정확하게 예측하는 방법을 제공하고자 한다.

과제의 해결 수단

[0007] 상기한 바와 같은 기술적 과제를 달성하기 위해서, 본 발명은 (a) 정량적 구조-체류 관계(quantitative structure-retention relationship, QSRR) 관계식을 모델링하는 단계; 및 (b) 수리 계획법(mathematical programming)을 이용해 상기 QSRR 관계식으로부터 혼합물 내 화합물의 크로마토그래피 용출 순서를 예측하는 단계;를 포함하며, 상기 수리 계획법은, (i) 화합물의 예측된 용출 순서를 제약조건(constraint)으로 사용하는 비선형 계획법(non-linear programming) 또는 (ii) 체류시간 예측 오차와 용출 순서 예측 오차를 목적함수로 사용하는 다목적 최적화(multi-objective optimization, MOO) 기법인 것을 특징으로 하는, 혼합물 내 화합물의 크로마토그래피 용출 순서를 예측하는 방법을 제공한다.

[0008] 또한, 상기 단계 (a)에서 모델링해 얻어지는 QSRR 관계식은 하기 식으로 표시되는 선형 모델인 것을 특징으로 하는 혼합물 내 화합물의 크로마토그래피 용출 순서를 예측하는 방법을 제안한다:

[0009] $$t_{R,j} = a_1 x_{j,1} + a_2 x_{j,2} + \ldots + a_n x_{j,n}$$

[0010] (상기 식에서 $t_{R,j}$은 오름차순으로 정렬된 화합물 j의 체류시간, $x_{j,i}$ (i = 1, ..., n)는 화합물 j 의 분자 설명자(molecular descriptor), a_i (i = 1, ..., n)는 회귀계수(regression coefficient)임).

[0011] 또한, 상기 단계 (a)에서 모델링해 얻어지는 QSRR 관계식은 인공신경망(artificial neural networks, ANN)을 비롯한 비선형 회귀식을 이용해 얻어진 비선형 모델인 것을 특징으로 하는 혼합물 내 화합물의 크로마토그래피 용출 순서를 예측하는 방법을 제안한다.

[0012] 또한, 상기 단계 (b)에서 아래 (II)의 부등식 제약조건(constraint) 하에 아래 (I)의 비선형 계획법을 적용해 혼합물 내 화합물의 크로마토그래피 용출 순서를 예측하는 것을 특징으로 하는 혼합물 내 화합물의 크로마토그래피 용출 순서를 예측하는 방법:

$$\min_{\mathbf{a}} \left\{ \sum_{j=1}^{m} \left(t_{R,j} - a_1 x_{j,1} - a_2 x_{j,2} - a_3 x_{j,3} \right)^2 + \sum_{j=1}^{m-1} \alpha_j \right\} \qquad \text{(I)}$$

$$a_1 \left(x_{j,1} - x_{j+1,1} \right) + a_2 \left(x_{j,2} - x_{j+1,2} \right) + a_3 \left(x_{j,3} - x_{j+1,3} \right) - \alpha_j \le 0 \quad \text{(II)}$$

[0013]

[0014] (상기 식에서, α_j는 양수의 이완 매개변수, $\bar{\mathbf{a}}$ 는 $a1$, $a2$, $a3$와 α_j (j = 1, 2, …, m-1)으로 이루어진 결정변수 벡터임).

[0015] 상기 단계 (b)에서 하기 식 (III)로 표시되는 체류 시간 예측 오차를 나타내는 목적함수 및 하기 식 (IV)로 표시되는 용출 순서 예측 오차를 나타내는 목적함수에 기초해 다목적 최적화를 수행해 혼합물 내 화합물의 크로마토그래피 용출 순서를 예측하는 것을 특징으로 하는 혼합물 내 화합물의 크로마토그래피 용출 순서를 예측하는 방법을 제안한다:

$$\%RMSE(t_R) = \sqrt{\sum \left(\frac{t_R - \hat{t}_R}{t_R} \right)^2 \Big/ m} \times 100 \qquad \text{(III)}$$

[0016]

[0017] (상기 식에서, t_R과 $\hat{t}_R$ 는 각각 분석 실험에서 측정한 체류시간과 모델로 예측한 체류시간을 나타내며, m은 각 컬럼에 대해 체류시간을 측정한 혼합물의 개수를 나타냄),

$$\%RMSE(\text{order}) = \sqrt{\sum \left(\frac{\text{order}_{pred.} - \text{order}_{obs.}}{\text{order}_{obs.}} \right)^2 \Big/ m} \times 100 \qquad \text{(IV)}$$

[0018]

[0019] (상기 식에서, $\text{order}_{obs.}$과 $\text{order}_{pred.}$는 각각 분석 실험에서 측정한 용출 순서와 예측된 체류시간으로부터 결정한 용출 순서를 나타냄).

[0020] 상기 단계 (b)에서 하기 단계를 실시해 파레토 (Pareto) 최적해를 선택하는 다목적 최적화를 수행하는 것을 특징으로 하는 혼합물 내 화합물의 크로마토그래피 용출 순서를 예측하는 방법을 제안한다:

[0021] (b-1) 파레토 해들로 구성된 파레토 프론트(front)에서 체류시간과 용출 순서의 예측 오차의 최적 절충점인 굴곡점(knee point)을 선택하는 단계;

[0022] (b-2) 용출 순서 예측 오차를 감소시키는 다음 파레토 해로 이동하는 단계;

[0023] (b-3) 적용 구역(applicability domain)을 이용하여 해를 검증하는 단계; 및

[0024] (b-4) 체류시간 예측 오차의 증가가 미리 정한 임계값에 도달하거나 적용구역에서 이상점이 미리 정한 임계값을

넘을 때까지 상기 단계 (b-1) 및 (b-2)를 반복하는 단계

발명의 효과

[0025] 본 발명은 정량적 구조-체류 관계(quantitative structure-retention relationships, QSRR)를 기반으로, 화합물의 예측된 용출 순서를 제약조건으로 하거나 화합물의 체류시간 및 용출 순서의 오차를 목적함수로 사용해 수리 계획법(비선형 계획법(non-linear programming) 또는 다목적 최적화(multi-objective optimization, MOO))를 통해 혼합물 내 화합물의 용출 순서를 정확하게 예측하는 방법에 대한 것으로서, 본 발명에 의하면 오믹스 분석(프로테오믹스, 메타볼로믹스 등)에서 주어진 펩타이드 또는 대사물을 식별하는데 있어 보다 높은 확실성을 제공하는 개별 화합물(펩타이드 혹은 대사물)의 정확한 위치 확인을 가능케 하여 크로마토그래피 분야 특히, 역상 또는 이온 교환 모드의 크로마토그래피에서 복잡한 혼합물의 분리를 최적화할 수 있다.

도면의 간단한 설명

[0026] 도 1은 다목적 최적화(multi-objective optimization, MOO) 기법을 이용한 화합물의 크로마토그래피 용출 순서 예측 방법에 대한 개념 설명도이다.

도 2는 예측된 용출 순서를 제약조건으로 사용하는 비선형 계획법 결과로서,두가지 실시예에 대한 MLR과 NLP의 컬럼별 %RMSE 값 비교 결과이다[A) 체류 시간, B) 용출 순서, C) MLR과 MLR-MOO 모델의 %RMSE 값 차이]

도 3은 예측된 용출 순서를 제약조건으로 사용하는 비선형 계획법 결과로서,선형 QSRR(MLR)에 대한 본 발명(실시예 1 (CS1))의 성능을 보여주는 결과이다[A) 체류시간 예측, B) 용출 순서 예측, C) 적용 구역].

도 4는 체류시간 예측 오차와 용출 순서 예측 오차를 목적함수로 사용하는 다목적 최적화 결과로서, 두가지 실시예 대한 MLR과 MLR-MOO의 컬럼별 %RMSE 값 비교 결과이다[A) 체류 시간, B) 용출 순서, C) MLR과 MLR-MOO 모델의 %RMSE 값 차이].

도 5는 체류시간 예측 오차와 용출 순서 예측 오차를 목적함수로 사용하는 다목적 최적화 결과로서, 선형 QSRR(MLR)에 대한 본 발명의 성능을 보여주는 결과이다[(실시예 1 (CS1): A) 체류시간 예측, B) 용출 순서 예측, C) 적용 구역; 실시예2 (CS2): D) 체류시간 예측, E) 용출 순서 예측, F) 적용 구역].

도 6은 MLR-MOO 파레토 프론트를 나타낸 그래프이다[A) 실시예1(CS1), B) 실시예2(CS2)].

발명을 실시하기 위한 구체적인 내용

[0027] 본 발명을 설명함에 있어서 관련된 공지 기능 또는 구성에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략할 것이다.

[0028] 본 발명의 개념에 따른 실시예는 다양한 변경을 가할 수 있고 여러 가지 형태를 가질 수 있으므로 특정 실시예들을 도면에 예시하고 본 명세서 또는 출원에 상세하게 설명하고자 한다. 그러나 이는 본 발명의 개념에 따른 실시 예를 특정한 개시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.

[0029] 본 명세서에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 설시된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0030] 본 발명은 QSRR에 기반한 체류시간 및 용출 순서의 오차가 목적함수의 일부로 정의된 수리 계획법을 통하여 이 문제를 해결하는 방법을 포함한다. 핵심은 다음 방정식으로 정의된 QSRR 모델의 일반적인 식이다(선형 모델의 경우):

[0031] $$t_{R,j} = a_1 x_{j,1} + a_2 x_{j,2} + \ldots + a_n x_{j,n} \qquad (1)$$

[0032] 위 식에서 $t_{R,j}$은 오름차순으로 정렬된 화합물 j의 체류시간, $x_{j,i}$ (i = 1, ..., n)는 화합물 j 의 분자 설명자(molecular descriptor), 그리고 a_i (i = 1, ..., n)는 회귀식의 계수(regression coefficient)이다. (여기서

$t_{R,j}$ 과 $x_{j,i}$ 는 평균이 0으로 조정되었다.) a_i 는 다중선형회귀식(multiple linear regression, MLR)과 같은 방법으로 구할 수 있으며, 만약 QSRR이 비선형이라면 인공신경망(artificial neural networks, ANN) 같은 비선형 모델링 기법이 사용될 수 있다.

[0033] **(1) 예측된 용출 순서를 제약조건으로 사용하는 비선형 계획법 (C-NLP)**

[0034] QSRR을 수리 계획법의 공식으로 나타내면 다음과 같다:

$$\min_{\mathbf{a}} \sum_{j} \left(t_{R,j} - \hat{t}_{R,j} \right)^2 \qquad (2)$$

[0035]

[0036] 여기서 , $\hat{t}_{R,j} = f(\mathbf{x})$ 이다. 모델링된 QSRR 관계식 $f(\mathbf{x})$ 는 a_i 를 매개변수로 갖는 임의의 함수 형태를 갖는다. 후술할 실시예에서처럼 n = 3일 때, 그리고 관계식으로 MLR을 사용하면 식(1)은 다음과 같이 된다:

$$\min_{\mathbf{a}} \sum_{j} \left(t_{R,j} - \hat{t}_{R,j} \right)^2 = \min_{\mathbf{a}} \sum_{j} \left(t_{R,j} - a_1 x_{j,1} - a_2 x_{j,2} - a_3 x_{j,3} \right)^2 \qquad (3)$$

[0037]

[0038] 즉 QSRR 문제는 전형적인 비선형 계획법(non-linear programming)이다. 예측된 체류시간을 오름차순으로 정리하면 그로부터 용출 순서를 예측하는 것은 매우 간단하다. 하지만, 위에서 설명한 바와 같이 식(2)를 통해 구한 QSRR 모델로 복잡한 수많은 피크들의 체류시간을 적절한 수준의 오차 이내로 예측할 수 있을지라도 용출 순서의 예측 오차는 매우 나쁠 수 있다. 이는 회귀식에서의 '지렛대 효과'를 생각하면 자명하다.

[0039] 수리 계획법 측면에서 위에서 언급한 문제점은 다음과 같이 부등호 제약조건을 도입하면 해결될 것처럼 보인다:

$$\min_{\mathbf{a}} \sum_{j} \left(t_{R,j} - a_1 x_{j,1} - a_2 x_{j,2} - a_3 x_{j,3} \right)^2$$

$$\text{subject to:} \qquad (4)$$

$$\hat{t}_{R,j} \le \hat{t}_{R,j+1} \text{ or}$$

$$a_1 x_{j,1} + a_2 x_{j,2} + a_3 x_{j,3} \le a_1 x_{j+1,1} + a_2 x_{j+1,2} + a_3 x_{j+1,3}$$

[0040]

[0041] m 개의 화합물에 대해 부등식 제약조건을 벡터-행렬 표기로 나타내면 $\bar{\mathbf{X}} \cdot \mathbf{a} \le \mathbf{0}$인데, 여기서 $\bar{\mathbf{X}}$는 $\left[\left(x_{j,1} - x_{j+1,1} \right) \quad \left(x_{j,2} - x_{j+1,2} \right) \quad \left(x_{j,3} - x_{j+1,3} \right) \right]$ 을 j번째 행으로 갖는 ((m-1)×3) 행렬이고 (j = 1, 2, …, m-1), $\mathbf{a} = \left[a_1 \quad a_2 \quad a_3 \right]^T$ 이다.

[0042] 하지만, 몇 가지 수치적 실험을 통해 위와 같은 제약조건은 너무나 과도하여 간단한 혼합물의 경우에도 나쁜 결과를 보였다. 결과적으로 다음과 같은 이완된 부등식 제약조건을 사용하여 낮은 체류시간 예측 오차와 용출 순서 예측 오차를 동시에 얻을 수 있었다.

$$\min_{\mathbf{a}} \left\{ \sum_{j=1}^{m} \left(t_{R,j} - a_1 x_{j,1} - a_2 x_{j,2} - a_3 x_{j,3} \right)^2 + \sum_{j=1}^{m-1} \alpha_j \right\}$$

$$\text{subject to:} \qquad (5)$$

$$a_1 \left(x_{j,1} - x_{j+1,1} \right) + a_2 \left(x_{j,2} - x_{j+1,2} \right) + a_3 \left(x_{j,3} - x_{j+1,3} \right) - \alpha_j \le 0$$

[0043]

[0044] 여기서 α_j는 양수의 이완 매개변수, $\bar{\mathbf{a}}$는 $a1$, $a2$, $a3$와 α_j (j = 1, 2, …, m-1)으로 이루어진 결정변수 벡터이다. m 개의 화합물에 대해 부등식 제약조건을 벡터-행렬 표기로 나타내면 다음과 같다:

$$\begin{bmatrix} x_{1,1}-x_{2,1} & x_{1,2}-x_{2,2} & x_{1,3}-x_{2,3} & -1 & 0 & 0 & \cdots & \cdots & 0 \\ x_{2,1}-x_{3,1} & x_{2,2}-x_{3,2} & x_{2,3}-x_{3,3} & 0 & -1 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & 0 & 0 & -1 & \cdots & \cdots & 0 \\ x_{j,1}-x_{j+1,1} & x_{j,2}-x_{j+1,2} & x_{j,3}-x_{j+1,3} & \vdots & \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \ddots & \vdots \\ x_{m-1,1}-x_{m,1} & x_{m-1,2}-x_{m,2} & x_{m-1,3}-x_{m,3} & 0 & 0 & 0 & \cdots & 0 & -1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_j \\ \vdots \\ \alpha_{m-1} \end{bmatrix} \leq \mathbf{0} \qquad (6)$$

[0045]

[0046] **(2) 체류시간 예측 오차와 용출 순서 예측 오차를 목적함수로 사용하는 다목적 최적화 (MOO)**

[0047] 다목적 최적화 (multi-objective optimization, MOO) 문제는 다수의 목적함수를 갖는 최적화 문제로서 일반적인 식은 다음과 같다:

$$\min\left(g_1(\boldsymbol{\alpha}_1), g_2(\boldsymbol{\alpha}_2), \ldots, g_k(\boldsymbol{\alpha}_k)\right) \qquad (7)$$
$$\text{subject to: } \boldsymbol{\alpha}_i \in \mathrm{A}$$

[0048]

[0049] 위에서 정수 k (≥2)는 목적함수 g의 개수를 나타내며 집합 A는 결정 벡터 α의 가능한 집합이다. 다목적 최적화에서는 일반적으로 모든 목적함수를 최소화하는 해가 존재하지 않는다. 따라서 파레토 (Pareto) 최적해에 주의를 기울이는데, 파레토 최적해란 여러 목적함수들 중 적어도 하나를 악화시키지 않고서는 나머지 다른 목적함수를 향상시킬 수 없는 해이다.

[0050] 이 발명에서는 두 개의 목적함수가 사용되는데 하나는 체류시간 예측의 오차를, 다른 하나는 용출 순서 예측의 오차를 나타낸다. 그리고 파레토 최적해는 다음과 같은 방법을 사용하여 선택된다: (1) 파레토 해들로 구성된 파레토 프론트(front)에서 체류시간과 용출 순서의 예측 오차의 최적 절충점인 굴곡점(knee point)을 선택한다. (2) 용출 순서 예측 오차를 감소시키는 다음 파레토 해로 이동한다. (3) 적용 구역(applicability domain)을 이용하여 해를 검증한다(validate). (4) 체류시간 예측 오차의 증가가 미리 정한 임계값에 도달하거나 적용구역에서 이상점이 미리 정한 임계값을 넘을 때까지 (2)와 (3)을 반복한다. 도 1에 이를 개념적으로 나타내었다.

[0051] 이하, 본 발명을 실시예를 들어 더욱 상세히 설명하도록 한다.

[0052] 제시된 실시예는 본 발명의 구체적인 예시일 뿐이며, 본 발명의 범위를 제한하기 위한 것은 아니다.

[0053] **<실시예>**

[0054] 다음의 2가지 실시예를 통해 본 발명의 적용 가능성을 입증하였다: (i) 62개의 유기 화합물의 혼합물(CS1)과 (ii) 98개의 합성 펩타이드의 혼합물(CS2). 첫번째 실시예(CS1)를 위한 분석은 35℃에서 10분의 기울기 시간을 갖는 Supelcosil LC 컬럼을 사용하여 수행되었다. 두번째 실시예(CS2)를 위해 각기 다른 기울기 설정과 온도에서 7개의 크로마토그래피 칼럼 (Xterra, Licrospher, PRP, Discovery RP-Amide C-16, Licrospher CN, Discovery HS F5-3 및 Chromolith)을 평가하였다. 크로마터그래피 분석 데이터는 참고문헌에서 구하였다.

[0055] QSRR 관계식 모델링을 위해 각 실시예에서 사용한 분자 설명자는 아래 표 1에 기재하였다.

[0056] [표 1] 실시예(CS1 및 CS2)에서 사용한 분자 설명자

분자 설명자		설명
CS1	μ	쌍극자(dipole) 모멘트
	δ_{min}	가장 음으로 하전된 원자의 초과 전하
	$SASA$	용매-접근가능 표면적
CS2	Sum_{AA}	20개의 자연 발생 아미노산의 개별적인 체류 시간의 합
	$vDW_{vol.}$	반 데르 발스 부피
	$c\log P$	전산 옥탄올-물 계수

[0057]

[0058] 두 실시예 모두 QSRR 관계식의 구체적인 형태로 선형모델이 고려되었으며, 비교를 위해 control MLR 모델 계수는 최소자승법을 이용하여 구하였다. 식 (5)의 이완된 제약조건을 가지는 비선형 계획법 문제의 해는 interior-point 법을 이용하여 구하였다. 식 (7)의 다목적 최적화 문제의 해는 유전 알고리즘(genetic algorithm)을 이용하여 구하였다. 두 방법 모두 비교를 위해 구한 control MLR 모델의 계수를 최적화에서 초기값으로 사용하였다.

[0059] 다목적 최적화의 경우 체류 시간 예측 오차를 나타내는 목적함수로는 체류시간의 백분율 평균 제곱근 오차(percentage root mean square error, %RMSE)를 사용하였다:

$$\%RMSE(t_R) = \sqrt{\sum\left(\frac{t_R - \hat{t}_R}{t_R}\right)^2 \Big/ m} \times 100 \qquad (8)$$

[0060]

[0061] 여기서 t_R과 $\hat{t}_R$는 각각 분석 실험에서 측정한 체류시간과 모델로 예측한 체류시간을 나타내며, m은 각 컬럼에 대해 체류시간을 측정한 혼합물의 개수를 나타낸다. 용출 순서의 예측은 QSRR 모델로부터 예측된 체류시간을 오름차순으로 정렬한 후 가능하며 용출 순서의 정확도를 나타내는 목적함수로는 역시 %RMSE를 사용하였다:

$$\%RMSE(\text{order}) = \sqrt{\sum\left(\frac{\text{order}_{pred.} - \text{order}_{obs.}}{\text{order}_{obs.}}\right)^2 \Big/ m} \times 100 \qquad (9)$$

[0062]

[0063] 여기서 $\text{order}_{obs.}$과 $\text{order}_{pred.}$는 각각 분석 실험에서 측정한 용출 순서와 예측된 체류시간으로부터 결정한 용출 순서를 나타낸다.

[0064] 두 방법(C-NLP, MOO) 모두 선형 QSRR 모델을 사용한 경우(MLR) 두 실시예 모두에서 체류시간 예측의 정확도를 어느 정도 희생하여 (도 2A 및 도 4A) 용출 순서 예측의 정확도를 상당한 수준으로 높일 수 있었다 (도 2B 및 도 4B). 도 2C 및 도 4C에서 보듯이, %RMSE(tR)의 최대 증가가 각각 약 20%와 15% 정도인 반면, %RMSE(order)의 감소는 각각 최대 260%와 80 %나 되었다.

[0065] 두 방법에 대해 두 실시예에서 사용된 7개의 RP-LC 컬럼 중, 각 실시예의 대표적인 컬럼 2개(CS1: Supelcosil LC, tG = 10 min, T = 35 ℃; CS2: Xterra, tG = 20 min, T = 40 ℃)에 대한 상세한 결과-체류시간과 용출 순서의 예측력과 해당 적용 구역-를 도 3 및 도 5에 나타내었다. 두 실시예 CS1(도 3A 및 3B, 도 5A 및 5B)과 CS2(도 5D 및 5E) 모두, 합리적인 수준의 체류시간 및 용출 순서 예측 성능을 보여준다. 두 실시예 모두에 포함된 거의 모든 분석 화합물은 잘 예측되고 훈련 세트에 포함된 구조적으로 중요한 분석물이 각각의 적용 가능성 영역 (도 3C, 5C 및 5F) 내에 있다. 따라서 개발된 모델은 구조적으로 먼 분석 화합물에 대해서도 안정적이고 강건한 것으로 간주된다.

[0066] 발명의 요약에서 설명한 굴곡점(knee point)에서 시작하여 파레토 최적해를 찾는 절차에 따라 찾은 최적해는 도

6A(CS1)과 6B(CS2)에 나타내었다. 미리 정한 임계값은 각각 굴곡점에서의 %RMSE(tR)의 10%, 적용구역에서 이상점 2개 이었다. 또한, 아래 표 2에서 보는 바와 같이 다목적 최적화 방법의 경우에도 두 실시예 모두 약간의 체류시간 예측 오차의 증가로 큰 용출 순서 예측의 감소를 얻을 수 있었다.

[0067] **[표 2] 굴곡점과 최적점의 %RMSE**

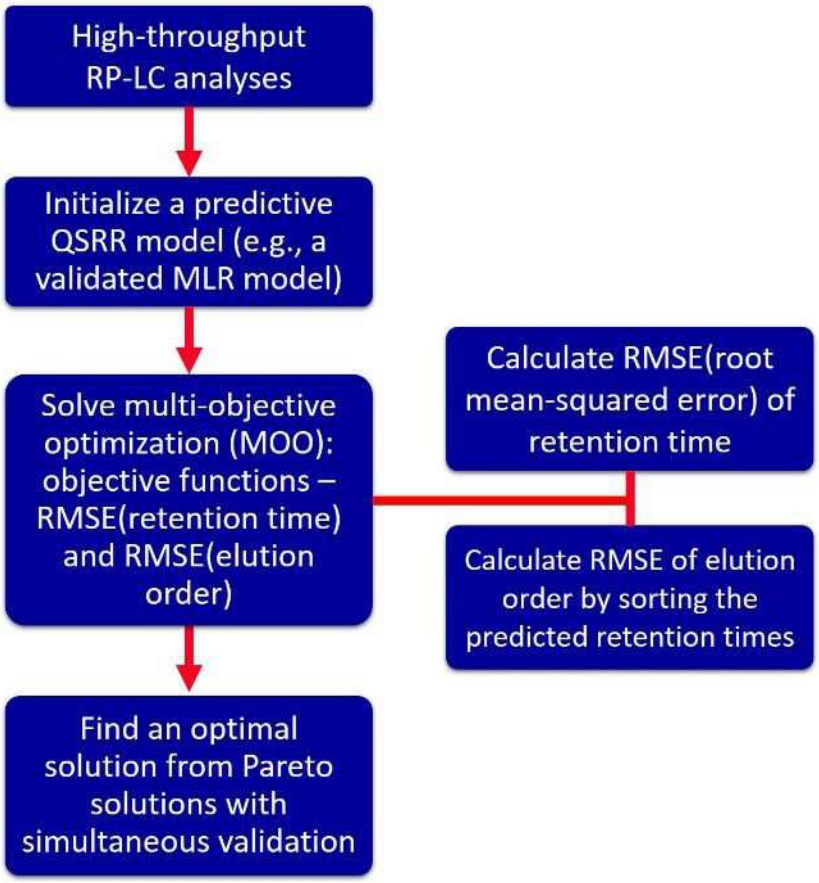
		%RMSE(t_R)	%RMSE(order)
CS1	Knee point	8.67	43.7
	Optimal point	9.33	42.0
CS2	Knee point	11.6	19.8
	Optimal point	12.1	18.1

[0068]

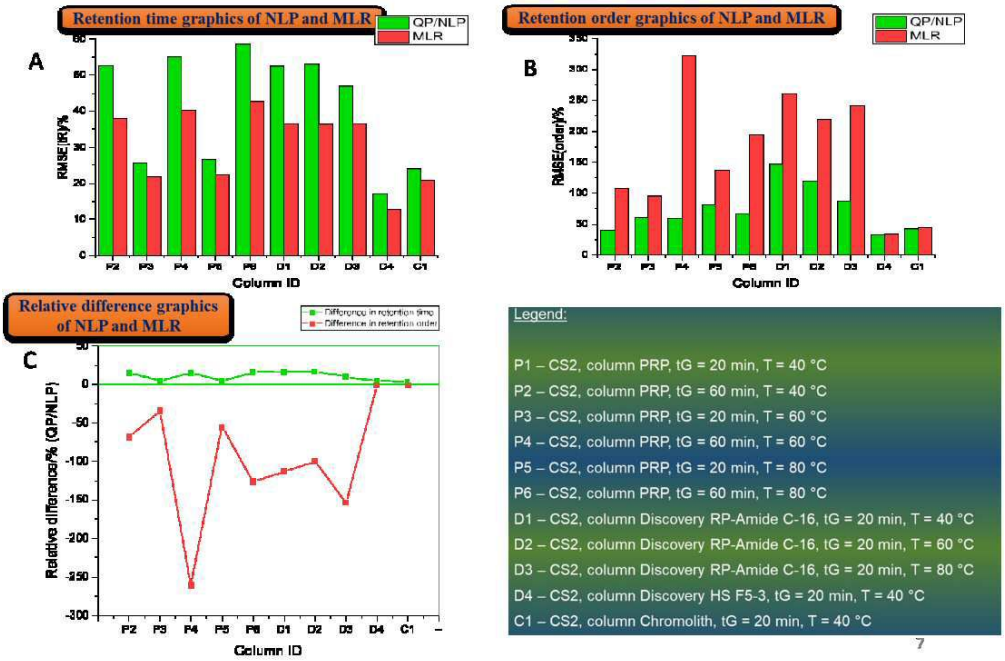
[0069] 이상, 첨부된 도면을 참조하여 본 발명의 실시예를 설명하였지만, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자는 본 발명이 그 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 실시될 수 있다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예에는 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다.

도면

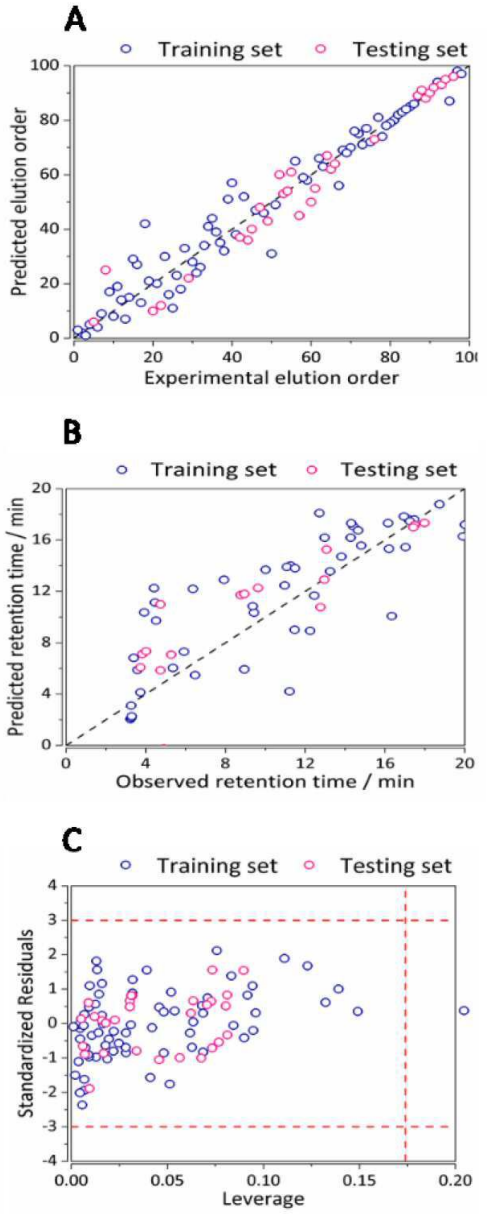
도면1



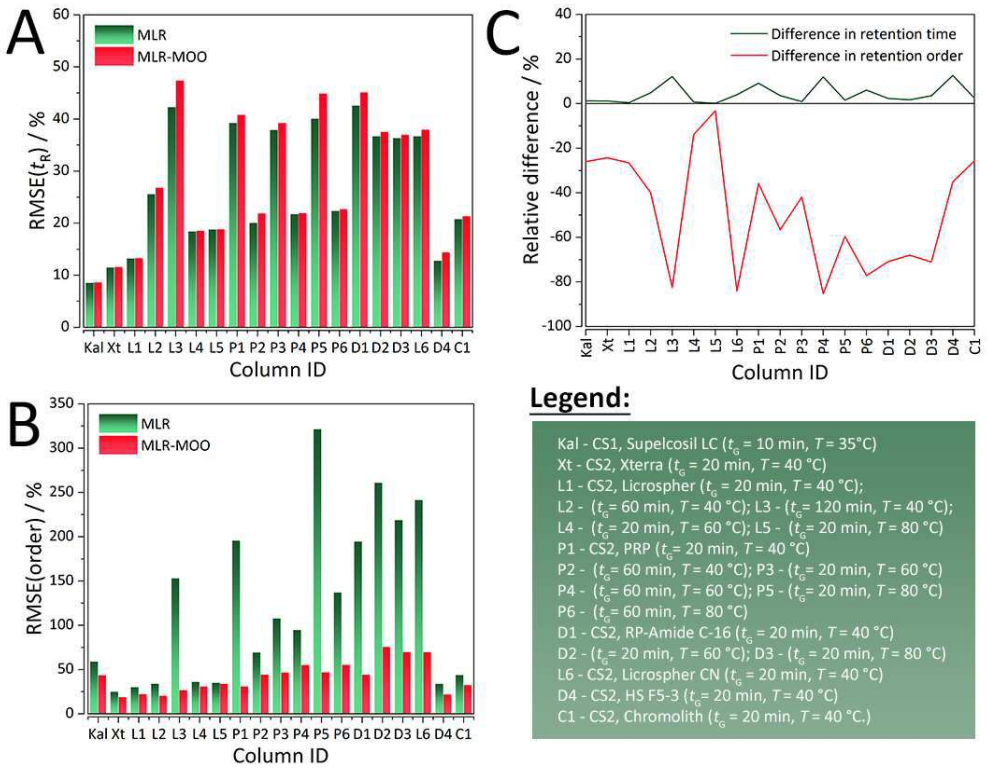
도면2


Retention time graphics of NLP and MLR
QP/NLP
MLR
A
RMSE(RI)/%
Column ID
Retention order graphics of NLP and MLR
QP/NLP
MLR
B
RMSE(order)/%
Column ID
Relative difference graphics of NLP and MLR
C
Relative difference/% (QP/NLP)
Column ID
Legend:
P1 – CS2, column PRP, tG = 20 min, T = 40 °C
P2 – CS2, column PRP, tG = 60 min, T = 40 °C
P3 – CS2, column PRP, tG = 20 min, T = 60 °C
P4 – CS2, column PRP, tG = 60 min, T = 60 °C
P5 – CS2, column PRP, tG = 20 min, T = 80 °C
P6 – CS2, column PRP, tG = 60 min, T = 80 °C
D1 – CS2, column Discovery RP-Amide C-16, tG = 20 min, T = 40 °C
D2 – CS2, column Discovery RP-Amide C-16, tG = 20 min, T = 60 °C
D3 – CS2, column Discovery RP-Amide C-16, tG = 20 min, T = 80 °C
D4 – CS2, column Discovery HS F5-3, tG = 20 min, T = 40 °C
C1 – CS2, column Chromolith, tG = 20 min, T = 40 °C
7

도면3


A
Training set
Testing set
Predicted elution order
Experimental elution order
B
Training set
Testing set
Predicted retention time / min
Observed retention time / min
C
Training set
Testing set
Standardized Residuals
Leverage

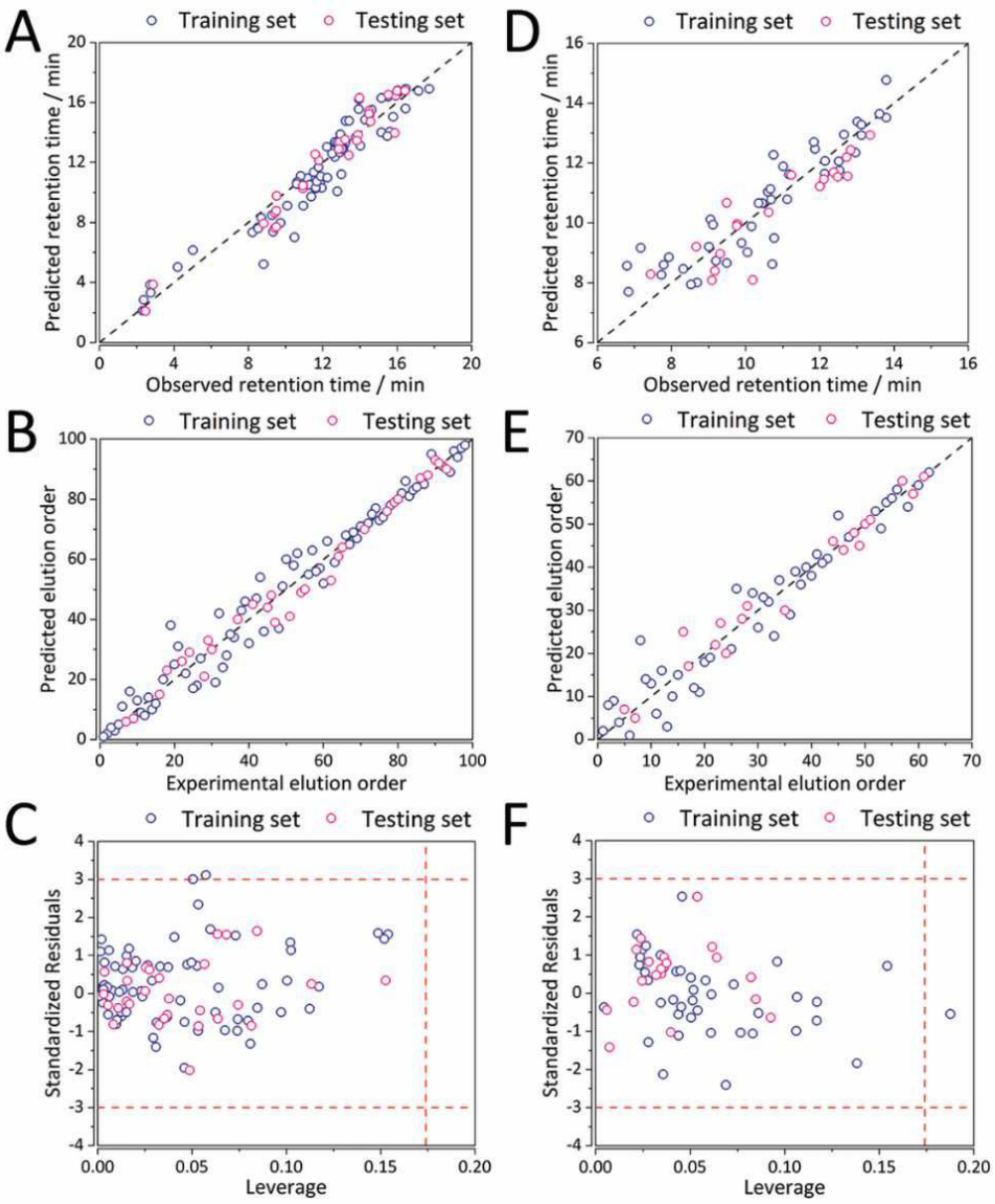
도면4



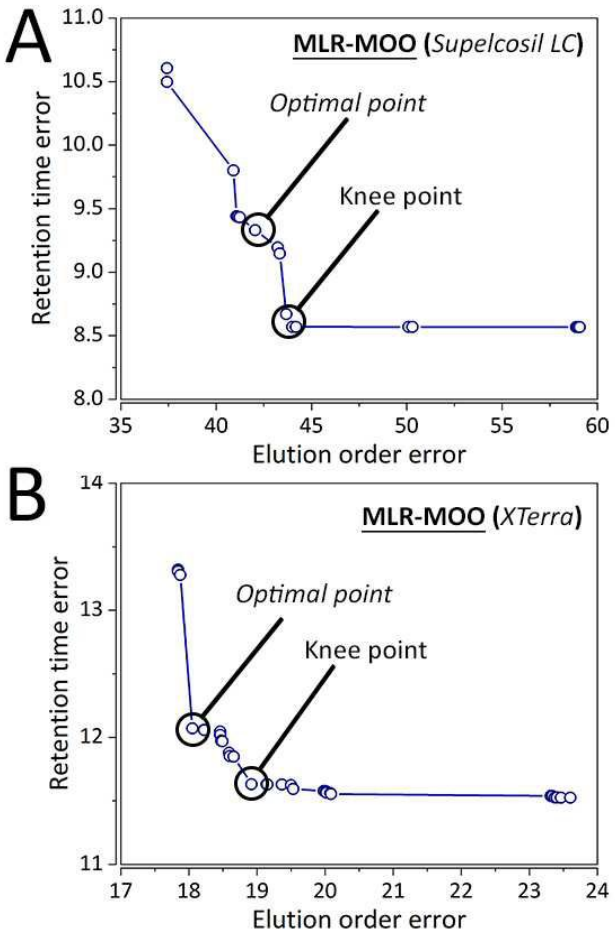
Legend:

Kal - CS1, Supelcosil LC (t_G = 10 min, T = 35°C)
Xt - CS2, Xterra (t_G = 20 min, T = 40 °C)
L1 - CS2, Licrospher (t_G = 20 min, T = 40 °C);
L2 - (t_G = 60 min, T = 40 °C); L3 - (t_G = 120 min, T = 40 °C);
L4 - (t_G = 20 min, T = 60 °C); L5 - (t_G = 20 min, T = 80 °C)
P1 - CS2, PRP (t_G = 20 min, T = 40 °C)
P2 - (t_G = 60 min, T = 40 °C); P3 - (t_G = 20 min, T = 60 °C)
P4 - (t_G = 60 min, T = 60 °C); P5 - (t_G = 20 min, T = 80 °C)
P6 - (t_G = 60 min, T = 80 °C)
D1 - CS2, RP-Amide C-16 (t_G = 20 min, T = 40 °C)
D2 - (t_G = 20 min, T = 60 °C); D3 - (t_G = 20 min, T = 80 °C)
L6 - CS2, Licrospher CN (t_G = 20 min, T = 40 °C)
D4 - CS2, HS F5-3 (t_G = 20 min, T = 40 °C)
C1 - CS2, Chromolith (t_G = 20 min, T = 40 °C.)

도면5


A
Training set
Testing set
Predicted retention time / min
Observed retention time / min
D
Training set
Testing set
Predicted retention time / min
Observed retention time / min
B
Training set
Testing set
Predicted elution order
Experimental elution order
E
Training set
Testing set
Predicted elution order
Experimental elution order
C
Training set
Testing set
Standardized Residuals
Leverage
F
Training set
Testing set
Standardized Residuals
Leverage

도면6


A
MLR-MOO (Supelcosil LC)
Optimal point
Knee point
Retention time error
Elution order error
11.0
10.5
10.0
9.5
9.0
8.5
8.0
35
40
45
50
55
60
B
MLR-MOO (XTerra)
Optimal point
Knee point
Retention time error
Elution order error
14
13
12
11
17
18
19
20
21
22
23
24