공학석사학위논문

미세조류배양시스템의 글루코즈 농도 실시간 모니터링을 위한 라만 분광기 기반 농도 예측 기법

Development of soft sensor based on Raman spectroscopy for on-line monitoring of glucose concentrations in microalgal production system

2013년 2월

서울대학교 대학원
화학생물공학부
오세규

미세조류배양시스템의 글루코즈 농도 실시간 모니터링을 위한 라만 분광기 기반 농도 예측 기법

Development of soft sensor based on Raman spectroscopy for on-line monitoring of glucose concentrations in microalgal production system

지도교수 이종민

이 논문을 공학석사 학위논문으로 제출함
2012년 12월

서울대학교 대학원
화학생물공학부
오세규

오세규의 석사 학위논문을 인준함
2012년 12월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

초 록

미세조류는 광합성을 하는 수중 단세포 생물로 청정에너지 및 유용물질의 원료로 주목 받고 있다. 미세조류는 비타민, 천연색소, 카로테노이즈, 단백질, 탄수화물과 같은 물질 생산을 위해 많이 사용되며 최근에는 바이오디젤의 원료가 되는 Triacylglycerols (TAGs)로 인해 큰 주목을 받고 있다.

광생물반응기를 이용한 미세조류 배양에서 중요한 두 변수는 글루코즈의 농도와 빛의 세기이다. 따라서 최적제어를 위해서는 두 변수의 실시간 측정이 가능해야 한다. 빛의 세기의 경우 광도계를 이용하여 실시간 측정이 가능하고, 글루코즈 농도의 경우 continous glucose monitors (CGMs)나 고성능액체크로마토그래피 (High-performance liquid chromatography, HPLC)로 측정이 가능하다. 하지만 CGMs는 실시간 측정은 가능하나 범용성이 떨어지고 장기간 사용이 불가능하며, HPLC는 범용성이 좋고 장기간 사용이 가능하나 실시간 측정이 어려운 단점이 있다.

본 연구에서는 라만 분광기와 다변량 회귀 분석 기법을 이용하여 글루코즈 농도를 실시간으로 측정할 수 있는 통합 프레임워크를 제안한다. 제안된 프레임워크의 과정은 다음과 같다. 우선 Rolling-Circle Filter (RCF)를 이용하여 라만 스펙트럼의 배경을 제거한다. 그 다음 부분회귀분석 (Partial Least Squares, PLS)을 이용해 라만 스펙트럼과 글루코즈 농도의 관계를 찾고, PLS를 통해 예측된 글루코즈 농도를 Successive Savitzky-Golay filter를 이용해 보정한다. 본 연구에서 제안하는 프레임워크를 이용해 농도 예측이 가능한지 알아보기 위해 두 가지

의 실험을 진행하여 검증하였다. 첫 번째 실험의 경우 본 프레임워크를 통해 글루코즈 농도 예측성능은 R^2 기준 0.899에서 0.943으로 향상되었으며, 실제 미세조류배양액을 이용한 두 번째 실험의 경우 글루코즈 농도 예측성능은 R^2 기준 0.413에서 0.973으로 크게 향상되었다.

주요어 : 소프트센서, 라만 분광기, 미세조류, 다변량통계분석, 화학계량법, 온라인 모니터링

학번 : 2011-21048

목 차

그 림

표

제 1 장

서론

최근 바이오디젤의 원료로써 미세조류가 큰 주목을 받고 있다. 미세조류는 바이오디젤의 원료가 되는 지질의 일종인 triacylglycerols (TAGs)뿐만 아니라 비타민, 천연색소, carotenoid, 단백질, 탄수화물과 같은 유용한 물질을 생산하기 위해서도 사용된다[1, 2]. 또한 온실가스 문제를 해결하기 위해 미세조류를 이용한 이산화탄소 포집에 대한 연구도 진행 중에 있다[3].

미세조류는 일반적으로 광합성을 이용해 생장하는 광영양배양, 외부로 부터 유기물을 공급받는 종속영양배양, 광합성을 통해 스스로 무기물을 유기물로 변환하며 외부로부터도 유기물을 공급받는 혼합영양배양에서 모두 배양 가능한데, 유용물질의 원료가 되는 지질과 바이오매스는 혼합영양배양에서 가장 많이 생산된다[4]. 오랫동안 미세조류는 광합성만을 이용하는 open-pond에서 자연 상태로 배양되어 왔다. 이러한 환경에서 배양된 미세조류는 주로 건강보조식품의 원료로 사용된다[5]. 하지만 최근 미세조류 배양을 위해 광생물반응기가 주목받고 있다. 광생물반응기는 외부 환경에 크게 영향을 받는 open-pond와는 달리 외부와 단절되어 있고 빛의 세기, 외부에서 공급받는 유기물의 양, 유속 등 다양한 조건을 제어할 수 있다. 최적제어가 이루어진 광생물반응기는 다른 배양조건에 비해 높은 생산성을 보장하고 따라서 제약관련 물질이나 바이오디젤과 같은 고부가가치 제품 생산에 적

합하다[6].

미세조류 배양공정에서 가장 중요한 파라미터는 글루코즈 농도와 빛의 세기이다. 글루코즈 농도가 너무 낮거나 빛의 세기가 너무 약할 경우 세포의 생장은 제한되며, 반대일 경우에는 세포의 생장이 방해받게 된다[7]. 따라서 최적 제어를 위해서는 두 파라미터의 값을 실시간으로 알 수 있어야 한다. 또한 두 파라미터의 값을 실시간으로 알 수 있다면 공정이 예상대로 운전되는지도 알 수 있다. 빛의 세기의 경우 광도계를 이용하면 실시간으로 쉽게 측정이 가능하다. 글루코즈의 경우도 Continuous Glucose Monitors (CGMs)를 이용하면 글루코즈의 농도를 실시간으로 측정할 수 있다. 이 장비는 효소촉매 반응을 통해 소모된 산소의 양을 통해 글루코즈 농도를 예측하는 바이오센서의 일종이다[8]. CGMs의 경우 수명이 3∼7일 정도로 매우 짧고, 기본 장비를 제외한 센서 하나당 가격이 $35∼100정도로 고가이기 때문에 주로 의료용도로 사용되고 있다. 또한 범용성이 있는 장비가 아니기 때문에 글루코즈 농도 측정 이외에는 활용할 수 있는 곳이 없다. 범용성이 있는 장비 중에는 고성능액체크로마토그래피 (High-performance liquid chromatography, HPLC)가 주로 사용된다. HPLC를 이용해 글루코즈 농도를 측정하기 위해서는 우선 샘플링을 해야 한다. 그 후 원심분리기를 이용해 미세조류 세포와 상등액을 분리한 뒤 상등액만 채취한다. 그 다음 syringe filter를 이용해 상등액에 미량 남아있는 세포를 제거한 뒤, 여과된 상등액을 이용하여 HPLC로 글루코즈 농도를 측정한다. 이러한 과정은 1시간 30분에서 2시간 30분정도가 소요되며 많은 수작업이 필요하며 HPLC 보정이 필요한 경우 추가적인 시간 소모가 발생한다. 이를 통해 얻은 글루코즈의 농도가 예상과 다르거나 조정이 필요할 경우 공정변수 조작을 통해 이를 보정해 줘야 하는데 샘플링한 시점과

분석이 완료된 시점과의 시간차이로 인해 feedback 제어의 어려움이 발생한다.

본 연구에서는 라만 분광기를 이용하여 미세조류 배양공정의 글루코즈 농도를 실시간으로 예측할 수 있는 통합 프레임워크를 제안하였다. 라만 분광기는 분자의 진동방식을 측정해서 물질의 특성을 분석하는 장비로 조작이 간단하고 실시간으로 결과를 얻을 수 있다. 건조, 동결, 분쇄와 같은 샘플 전처리가 필요하지 않고 분석과정에서도 시료를 파괴하지 않기 때문에 실시간 모니터링에 사용하기 적절하며[9], 물에 의한 스펙트럼의 변화가 상대적으로 적기 때문에 수용액 상태로 존재하는 미세조류 배양액에 사용하기 적합하다[10]. 그리고 글루코즈 농도만을 측정할 수 있는 CGMs와는 달리 라만 스펙트럼에는 많은 화학적 정보가 있기 때문에 본 논문에서 제시하는 통합 프레임워크를 이용해 다양한 농도를 예측할 수 있다. 하지만 라만스펙트럼에는 스펙트럼의 모양을 전체적으로 왜곡시키는 형광 배경효과가 존재한다. 특히 미세조류와 같은 색소가 존재하는 생물학적 샘플의 경우 더욱 강한 형광 배경효과가 존재하게 되는데[10, 11], 정확한 분석을 위해서는 배경효과는 반드시 제거되야 한다. 배경효과를 제거하지 않을 경우 스펙트럼의 피크들이 강한 비선형성을 나타내 분석이 어려워 지거나 통계적으로 전혀 예측이 불가능하게 될 수 있다. 본 연구와 유사한 라만 분광기를 이용한 모니터링에 관련된 기존 연구들의 경우 fitting이 잘 되도록 회귀분석 기법이나 데이터 처리 기법에 초점을 맞추고 있다[12, 13, 14]. 반면에 라만분광기를 이용한 분광학적 또는 생물학적 연구에서는 피크를 직접적으로 분석해야 하기 때문에 배경효과 제거가 중요하게 다뤄지고 있다[15, 16, 17]. 본 연구의 목적은 라만 데이터를 이용한 글루코즈 농도 모니터링이 목적이지만 라만스펙트럼의

화학적 정보 왜곡의 근본적 원인인 배경효과를 제거하는데 중점을 두었다.

본 연구에서 제안된 통합 프레임워크는 다음과 같은 순서로 진행된다.

1. 전처리 (Pre-processing): Rolling-Circle Filter (RCF)를 이용해 라만 스펙트럼의 배경효과를 제거한다.

2. 모델링 (Modeling): 부분회귀분석 (Partial least squares regression, PLSR)을 이용해 글루코즈 농도와 샘플의 라만스펙트럼과의 관계를 찾는다.

3. 후처리 (Post-processing): PLS 모델을 이용해 예측된 글루코즈 농도를 Successive Savitzky-Golay smoothing filter를 이용해 보정한다.

본 연구에서 제안하는 프레임워크로 글루코즈 농도 예측이 가능한지 알아보기 위해 두 가지 실험을 진행하였다. 첫 번째 실험은 선행연구로서 글루코즈, 글라이신, 물, 콩기름을 혼합해 만든 34개의 혼합물 샘플을 이용하였고, 두 번째 실험은 본 실험으로서 미세조류를 광생물반응기에서 배양한 뒤 시간에 따라 36개의 샘플을 얻어 진행하였다.

제 2 장

실험

2.1 실험 장비

미세조류를 배양하기 위해 본 연구에서는 Satorious사의 광생물반응기인 BIOSTAT® PBR 2S를 사용했다. 최대 3L까지 배양이 가능하며 관형으로 제작되어 타 광생물반응기에 비해 표면적이 넓어 빛의 공급이 잘 되는 장점이 있다. 미세조류 배양액의 글루코즈 농도 측정을 위해서는 Agilent사의 1260 Infinity HPLC를 사용했다.

그림 1: 본 실험에서 사용한 광생물반응기인 BIOSTAT® PBR 2S

2.2 라만 분광법

라만 분광법은 분자의 진동과 회전 모드의 특성을 분석하여 분자의 특성을 분석하는 분광법으로 분석에 걸리는 시간이 짧고 사용하는 시료의 전처리가 필요없어 공정 모니터링 분야에 많이 활용되고 있다. 빛의 산란은 크게 탄성 산란인 레일레이 산란과 비탄성 산란인 라만 산란으로 구분된다. 라만 분광기는 산란의 대부분을 차지하는 레일레이 산란 사이에 존재하는 라만 산란을 측정하여 시료의 특성을 분석한다. 라만 스펙트럼은 물에 대해 민감하게 변화하지 않아 수용액 시료의 분석에 적절하고 Fourier transform infrared (FT-IR)이나 near IR에 비해 광섬유에 의한 파장 이동이 효율적이기 때문에 온도에 민감한 라만 분

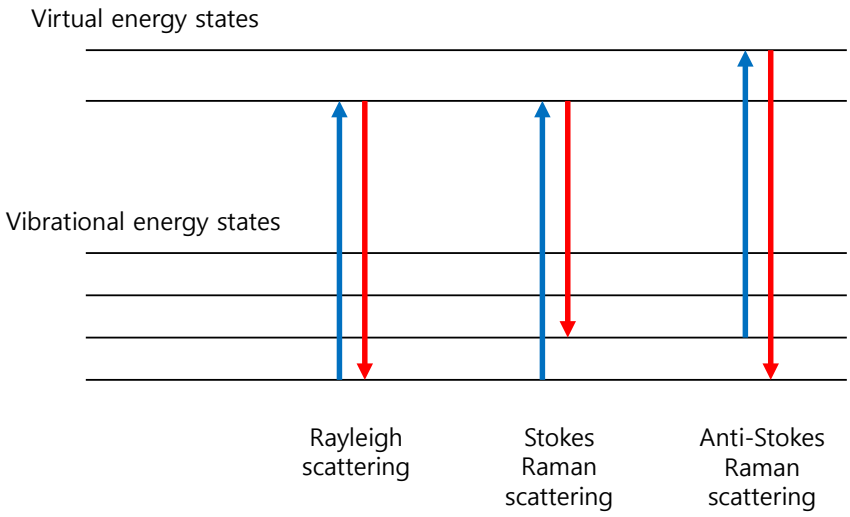

그림 2: 빛의 산란의 종류에 따른 에너지 준위 변화

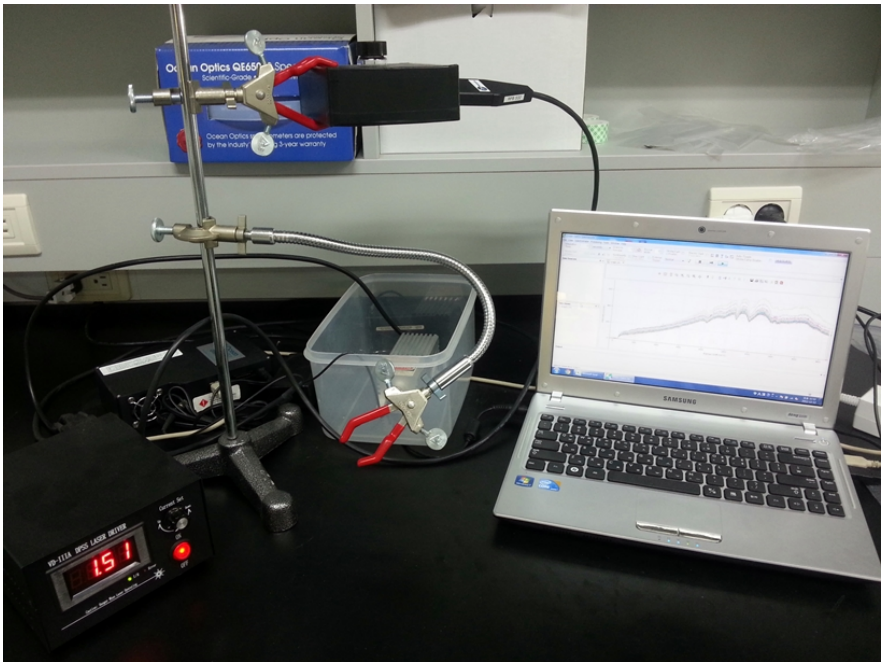

그림 3: 본 연구 진행에 사용된 라만 분광기

광 검출기를 측정 현장으로부터 먼 곳에 배치할 수 있어 가혹한 환경의 공정에도 사용할 수 있다[10, 18].

본 실험에서는 Ocean Optics사의 OE65000 Scientfic-grade 분광기를 사용했다. 라만 분광기에 사용하는 탐측기 (Probe)의 경우 출력은 50mW이며 532nm 파장의 DPSS (Diode-Pumped Solid-State) 레이저를 사용했다.

2.3 혼합물 샘플

본 프레임워크를 이용해 글루코즈 농도 예측이 가능한지 알아보기 위해 진행한 사전실험에서는 인위적으로 제조된 34개의 혼합물 샘플을 이용하였다. 혼합물 샘플을 제조하기 위해 D-(+)-글루코즈 (ACS

표 1: 혼합물 실험계획법 (DOE)에 사용된 농도의 최댓값과 최솟값

(g/mL)	초순수	콩기름	글루코즈	글라이신
최솟값	0.7029	0.1344	0	0
최댓값	0.7920	0.2100	0.0385	0.0174

표 2: DOE를 통해 결정된 17개의 혼합물 샘플 농도

(g/mL)	초순수	콩기름	글루코즈	글라이신
1	0.7029	0.2100	0.0385	0.9688
2	0.7920	0.1680	0	0.9600
3	0.7920	0.1554	0	0.9648
4	0.7920	0.1470	0.0385	0.9775
5	0.7425	0.2100	0	0.9525
6	0.7277	0.2100	0	0.9551
7	0.7178	0.2100	0.0385	0.9663
8	0.7920	0.1344	0.0385	0.9823
9	0.7574	0.1806	0.0193	0.9659
10	0.7301	0.1953	0.0289	0.9674
11	0.7747	0.1743	0.0096	0.9630
12	0.7747	0.1680	0.0096	0.9654
13	0.7747	0.1638	0.0289	0.9717
14	0.7499	0.1953	0.0096	0.9592
15	0.7425	0.1953	0.0096	0.9605
16	0.7376	0.1953	0.0289	0.9661
17	0.7747	0.1575	0.0289	0.9741

reagent grade, Sigma-Aldrich Co.), 초순수, 콩기름 (CJ), 글라이신 (99.0%, Sigma-Aldrich Co.)이 사용되었다. 글루코즈는 미세조류 배양 시 탄소원으로 사용되며 글루코즈 농도 예측 성능을 평가하기 위해 사용되었다. 콩기름의 경우 물과 섞이지 않는 물질이 존재할 경우에도 예측이 가능한지 평가하기 위해 사용되었다. 미세조류 배양 시 질소원으로 사용되는 글라이신은 라만 스펙트럼 촬영 시 배경효과가 발생하는데[19], 스펙트럼의 배경효과를 더 강하게 할 의도로 사용하였으며

수용액 상태의 미세조류 배양액과 비슷한 환경을 만들기 위해 물을 기반으로 혼합물을 제조 하였다. 샘플의 농도는 혼합물 실험계획법 (Design of experiment, DOE)을 통해 결정했으며 extreme vertex design 을 사용하여 17가지의 다른 농도를 결정한 뒤 이를 1회 반복 제조하여 34개의 혼합물 샘플을 제조하였다. 샘플의 부피는 10mL이며 실험계획법은 MINITAB® 16 (Minitab Inc.)을 통해 적용되었다.

2.4 미세조류 샘플

미세조류는 광생물반응기에서 복합영양조건 하에 배양되었으며 미세조류는 지질이 풍부한 것으로 알려진[20] *Chlorella protothecoides* (UTEX B25)를 사용하였다. 배지는 Siegler et al.[21] 에 의해 수정된 Shihira-Ishikawa and Hase 배지[22] 를 사용하였으며 조성은: KH_2PO_4 (2.8 g), K_2HPO_4 (1.2 g), $MgSO_4{\cdot}7H_2O$ (1.2 g), $FeSO_4{\cdot}7H_2O$ (48 mg), H_3BO_3 (11.6 mg), $CaCl_2{\cdot}2H_2O$ (10 mg), $MnCl_2{\cdot}4H_2O$ (7.2 mg), $ZnSO_4{\cdot}7H_2O$ (0.88 mg), $CuSO_4{\cdot}5H_2O$ (0.32 mg), MoO_3 (72 mg), thiamine HCl (40μg), Glucose (40 g), Glycine (0.5 g) 이다. 배지는 고압증기 멸균기에서 121 °C 에서 15분간 멸균하여 사용하였다. 배양은 동일한 조건에서 배양시간을 늘려가며 (1일, 2.5일, 4일, 5.5일, 7일) 총 5회 진행하였다. 본 실험에서 첫 번째, 네 번째 그리고 다섯 번째 실험의 경우 12시간 간격으로 샘플링을 수행했으며, 두 번째와 세 번째 실험은 각각 5회와 8회의 샘플링을 수행하였다. 따라서 본 실험에서 36개의 미세조류 샘플을 얻었다.

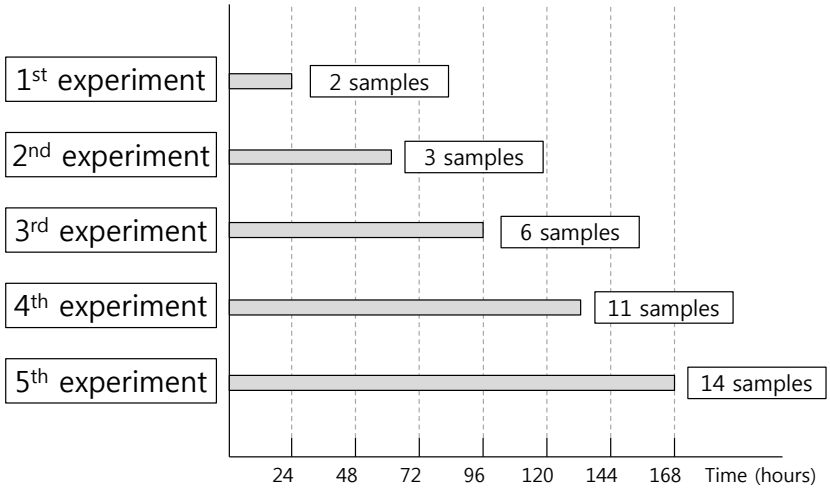

그림 4: 미세조류 배양을 위한 다섯 번의 실험과 샘플의 갯수

제 3 장

이론

3.1 데이터 처리 기법

라만 스펙트럼을 이용해 정확한 통계 모델을 만들기 위해서는 데이터의 불필요한 부분을 제거하거나 중요한 부분을 강조하는 등 데이터 변형 및 재배열이 필요하다. 라만 분광기를 이용한 화학 계량법에서 데이터 처리를 하는 주요 목적은 스펙트럼의 배경효과 제거, 노이즈 제거, 정규화다. 또한 본 논문의 경우 노이즈 제거 기법을 응용하여 통계 모델을 통해 예측된 글루코즈 농도를 재 보정하는데 사용하였다.

3.1.1 Rolling-Circle Filter (RCF)

라만 스펙트럼을 분석할 때 가장 큰 문제가 되는 부분은 형광 배경효과이다. 특히 미세조류의 경우 세포에 존재하는 클로로필이나 카로테노이드와 같은 색소 때문에 더 강한 배경효과가 발생하게 된다[10]. 더욱이 배경부분에는 어떠한 화학적 정보도 포함되어 있지 않기 때문에 분석에 사용하기 전에 반드시 제거를 해야 한다. 배경효과는 스펙트럼을 전체적으로 왜곡시키는 것 뿐만 아니라 그림 5에서 볼 수 있듯이 동일 농도의 샘플임에도 불구하고 스펙트럼 강도차이를 발생시키기도 한다.

이러한 배경효과를 제거하기 위해 본 논문에서 RCF를 사용하였다[11, 23]. RCF는 스펙트럼의 중요 피크와 배경간의 기하학적 차이를

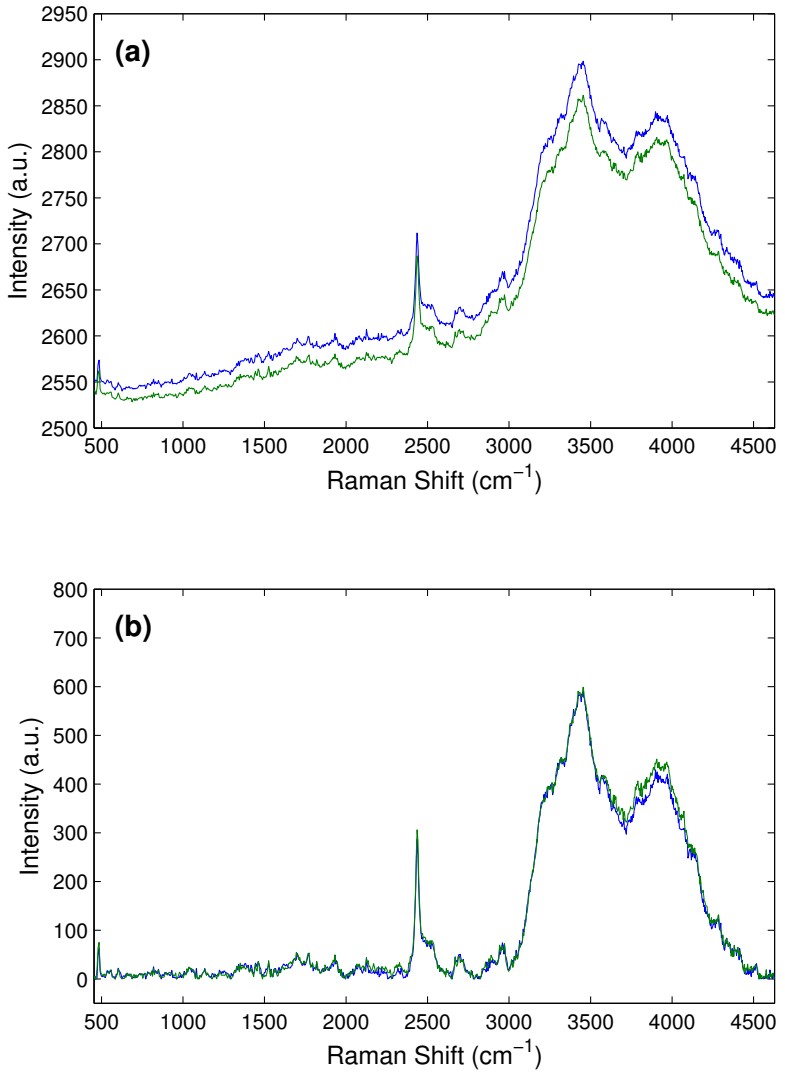

그림 5: (a) 동일 농도를 가진 두 샘플의 스펙트럼. (b) RCF를 이용하여 배경을 제거한 두 스펙트럼.

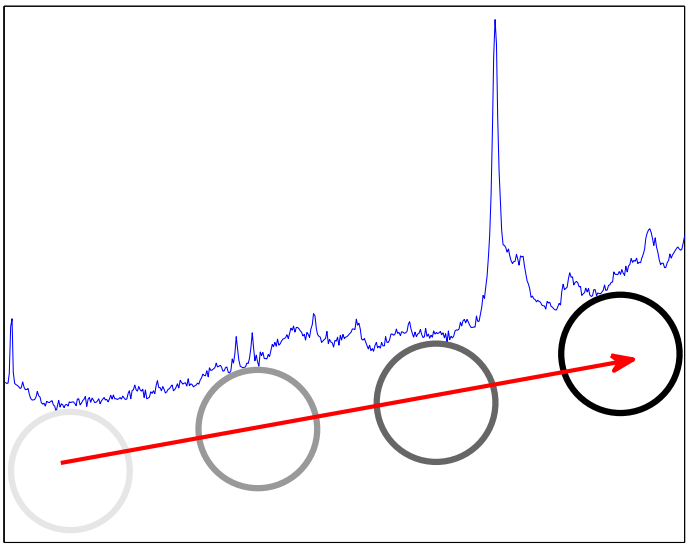

그림 6: RCF 알고리즘의 기하학적 표현

이용해 배경을 제거하는 기법이다. 이 기법의 적용방법은 다음과 같다. 우선 특정 반지름 R의 원을 생성한다. 그 다음 라만 스펙트럼 아래쪽에 원을 굴린 뒤 원이 들어가지 못하는 부분만 남기는 방법이다 (그림 6).

RCF의 장점은 반지름이라는 단 하나의 파라미터만 존재함에도 불구하고 가장 많이 사용되는 polynomial fitting기법보다 우수하며, 배경의 모양에 상관없이 제거가 가능하다는 점이다[10]. RCF를 적용하기 위해서는 우선 라만 이동 (Raman shift)과 라만 강도가 같은 범위를 가지도록 식 (3.1)을 이용해 정규화 해야한다.

$$\mathbf{X}_{norm}(i) = \frac{\mathbf{X}(i) - min[\mathbf{X}]}{max[\mathbf{X}] - min[\mathbf{X}]} \cdot N \tag{3.1}$$

여기서 X는 스펙트럼 강도, N은 스펙트럼 포인트의 갯수 그리고 $i = 1,2,3\cdots N$이다.

RCF의 성능을 평가하기 위해 5개의 가우시안 함수(식 3.2)를 이용해 라만 스펙트럼의 특성 피크를 만들고(그림 7(a)) 2개의 가우시안 함수를 이용해 배경효과를 만든 뒤 노이즈를 추가해 가상의 라만 스펙트럼(그림 7(b))을 생성했다. 그 뒤 반지름이 각각 1500, 500, 100인 RCF를 이용하여 배경을 제거한 결과(그림 7(c)) 반지름이 100인 RCF를 사용할 경우 원래의 특성 피크를 잘 찾아내는 것을 확인하였다(그림 7(d)). RCF를 사용할 때 가장 중요한 것은 반지름의 크기를 결정하는 것인데 특성 피크보다는 크고 배경보다는 작은 크기의 원을 사용해야 한다.

$$F(x) = A \cdot exp[-\{(x - x_0)/\Delta x\}^2] \tag{3.2}$$

표 3: 가상 스펙트럼 생성에 사용된 가우시안 함수의 파라미터

	A	Δx	x_0
peak 1	60	10	100
peak 2	130	5	300
peak 3	20	5	450
peak 4	100	10	600
peak 5	80	40	900
background 1	300	700	600
background 2	300	500	900

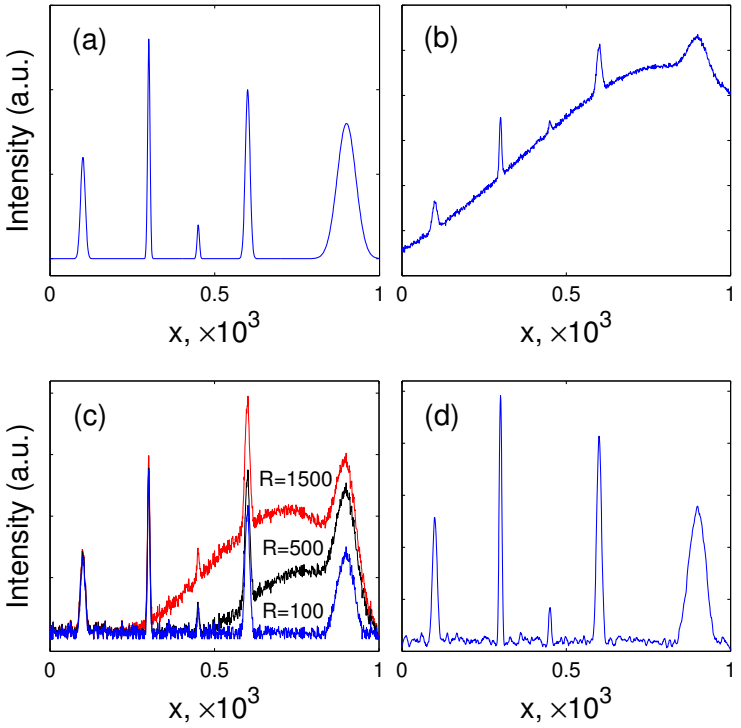

그림 7: 가우시안 함수를 이용해 만든 가상의 라만 스펙트럼과 RCF 적용 결과

3.1.2 Successive Savitzky-Golay smoothing filter

Savitzky-Golay smoothing filter (SG filter)[24] 는 일반적으로 스펙트럼의 노이즈를 제거하는 전처리 단계에 사용되는 기법이다. SG filter를 이용해 노이즈가 제거되는 원리는 다음과 같다. 우선 이동창 (moving window)의 크기를 결정한 뒤 이동창 크기의 데이터를 이용해 회귀선을 구한다. 그 다음 중간의 한 점을 회귀선 위로 옮기면 된다. 그리고 한 점을 이동하여 위와 같이 반복하면 된다. 하지만 노이즈를 제거하기 위해 사용되는 데이터는 많게는 수만 포인트까지 된다. 즉, 수만번의

회귀선을 구하는 작업을 반복해야 하므로 계산 시간이 상당히 오래 걸리게 된다. Savitkzy와 Golay는 어떠한 데이터를 사용하더라도 계수가 동일하다는 것을 발견했고 회귀선의 차수에 따른 계수를 정리하여 계산시간을 획기적으로 개선하였다[24]. 이러한 저장된 계수를 사용하여 SG filter를 사용하는 식은 다음과 같다.

$$y_k^* = \frac{1}{\sum_{i=-n}^{i=n} c_i} \sum_{i=-n}^{i=n} c_i \cdot y_{k+i} \tag{3.3}$$

여기서 y_k는 노이즈가 제거된 점, $\sum_{i=-n}^{i=n} c_i$는 정규화 인자, c_i는 계수 그리고 y_{k+i}는 원래의 데이터이다. 이는 MATLAB에서 `sgolayfilt` 함수를 이용하여 계산할 수 있다.

본 논문에서는 SG filter를 스펙트럼 노이즈를 제거하는 전처리 단계에서 사용하지 않고 다변량 회귀분석을 통해 예측된 시간에 따른 글루코즈 농도를 보정하는데 사용하였다. 배양시간에 따른 글루코즈 농도는 전 시간과 다음 시간의 농도와 상관관계를 가지고 있다. 따라서 중간의 한 값만 경향성을 크게 벗어나는 일은 일어나지 않는다. 하지만 실제 농도가 아닌 예측된 농도는 오차 범위 내에서 경향성을 벗어나는 일이 발생한다. 이러한 예측 오차를 노이즈로 간주하여 본 논문에서는 SG filter를 예측된 농도를 보정하는데 사용하였다. 또한 실시간 모니터링에 사용하여야 하므로 새로운 예측값이 들어올 때마다 연속적으로 SG filter를 적용하여 농도를 보정하였다.

3.1.3 Standard Normal Variate (SNV)

Standard Normal Variate (SNV)는 스펙트럼의 광산란 보정을 위해 가장 많이 사용되는 전처리 기법이다[25]. 보정 방법은 스펙트럼을 평

균으로 뺀 뒤 표준편차으로 나눠주면 된다. 식은 아래와 같다.

$$x_{ij,SNV} = \frac{(x_{ij} - \bar{x}_i)}{\sqrt{\frac{\sum (x_{ij} - \bar{x}_i)^2}{p-1}}} \tag{3.4}$$

여기서 $x_{ij,SNV}$는 전체 스펙트럼의 각 성분을 말하고, $\bar{x}_i$는 평균 스펙트럼이며 p는 스펙트럼의 갯수를 의미한다. SNV는 결과적으로 스펙트럼들의 기울기 편차를 보정하는 역할을 한다[26].

3.2 다변량 회귀 분석

3.2.1 다중회귀분석 (Multiple Linear Regression, MLR)

MLR은 가장 기본적인 다변량 회귀 분석기법으로 종속변수 **Y**와 독립변수 **X**를 유사 역행렬 (Pseudo inverse matrix)을 이용해 둘의 관계인 **B**를 직접적으로 찾는 기법이다[27].

$$\mathbf{Y} = \mathbf{XB} \tag{3.5}$$

$$\mathbf{B} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{3.6}$$

하지만 MLR의 경우 변수들간 강한 상관관계가 존재하는 것을 말하는 다중공선성 (Multicollinearity)과 특정 변수들의 조합으로 다른 변수를 완벽히 설명 가능한 특이성 (Singularity)이 존재하는 데이터에 대해서는 사용하기 힘든 단점이 있다. 스펙트럼 데이터의 경우 다중공선성과 특이성이 다른 데이터에 비해 심하게 나타나기 때문에 일반적으로 MLR은 화학계량법에 사용되지 않는다. 또한 데이터 특성을 고려하

고 않고 회귀분석을 시행하기 때문에 과적합 (Over-fitting)이 발생한다.

3.2.2 주성분분석 (Principal Component Analysis, PCA)

주성분분석 (Principal Component Anaylsis, PCA)은 1901년 Karl Pearson이 개발한 기법으로 현재까지도 가장 많이 사용되는 데이터 압축 기법이다[28]. 변수가 많은 고차원의 데이터는 변수간 상관관계가 심하며 시각적 표현이 불가능하여 분석이 쉽지 않다. PCA는 데이터의 공분산이 최대가 되는 방향으로 새로운 축을 설정한 뒤 이 축에 데이터를 사영시키는 방법으로 정보 손실을 최소화 하며 데이터를 압축한다 (그림 8)[29]. PCA에 의해 데이터는 원 데이터와 같은 차원의 데이터로 변형되며 이 변형된 데이터의 변수는 주성분(Principal Component, PC)

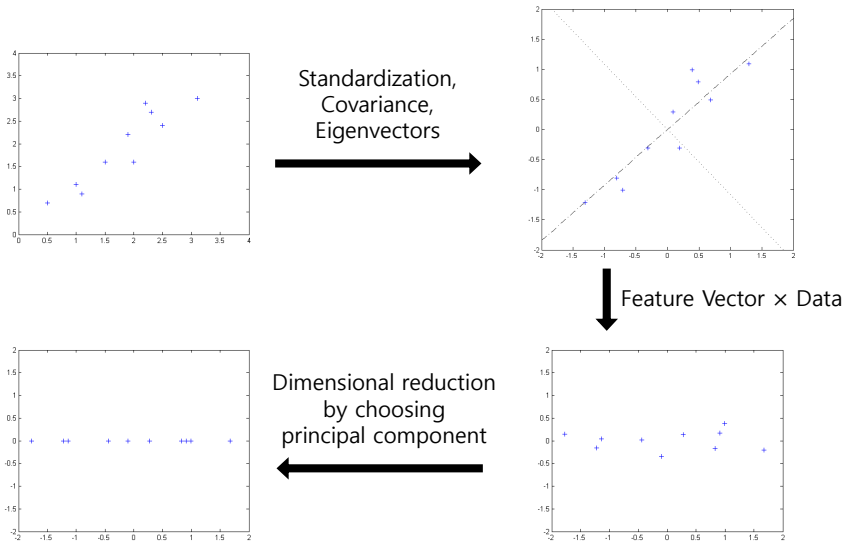

그림 8: PCA를 이용하여 2차원 데이터를 1차원으로 압축하는 과정

이라 부른다. 가장 많은 정보를 포함하고 있는 성분이 첫 번째로 오게 되며 점차 가진 정보량이 적을 수록 뒤쪽으로 배치된다. 일반적으로 시각적 분석을 위해서는 2개의 주성분을 취해 2차원 평면에 데이터를 표현하며 변수간 상관관계가 없어지도록 데이터를 변형 및 압축하는 것이 목적일 경우에는 주성분이 가진 정보량의 누적값을 통해 몇 개의 주성분을 분석에 사용할 지 판단하게 된다. 정보량의 경우 각 주성분의 고유값(eigenvalue)을 통해 알 수 있으며 전체 주성분의 고유값의 합 중에서 차지하는 비율이 전체 데이터에서 그 주성분이 포함하는 정보량이 된다.

PCA는 다음과 같은 식으로 표현된다.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{3.7}$$

여기서 **X**는 원 데이터, **T**는 주성분들이며 score라 부른다. **P**는 고유벡터(eigenvector)값들이며 PCA에서는 loading이라 부르며 **E**는 오차를 나타낸다. PCA를 계산하는데 가장 많이 사용되는 알고리즘은 Nonlinear Iterative Partial Least Squares (NIPALS) 알고리즘이다. NIPALS 알고리즘은 고유벡터를 찾기 위한 다양한 방법 중 하나인데 다른 방법으로는 많이 알려진 특이값분해 (Singular Value Decomposition, SVD)가 있다.

3.2.3 주성분회귀분석 (Principal Component Regression, PCR)

주성분회귀분석 (Principal Component Regression, PCR)은 PCA를 통해 변수간 상관관계가 없도록 변형된 X의 주성분을 이용하여 회귀분석을 수행하는 기법으로 PCA를 이용하여 데이터의 주성분을 찾은

Algorithm 1 NIPALS algorithm

1: **for** $i = 1 \rightarrow$ number of PCs **do**
2: $p = (E_{(i-1)}^T t)/(t^T t)$ % Project **X** onto t
3: $p = p(p^T p)^{-0.5}$ % Normalization
4: $t = (E_{(i-1)} p)/(p^T p)$ % Project **X** onto p
5: **if** $(t^T t)_{new} - (t^T t)_{old} > threshold \times (t^T t)_{new}$ **then**
6: Go to step 2 % Check for convergence
7: **end if**
8: $E_{(i)} = E_{(i-1)} - (t p^T)$ % Remove the estimated PC
9: **end for**

뒤 주성분을 이용하여 MLR을 진행하게 된다. PCA를 통해 구해진 주성분들은 서로 상관관계가 없기 때문에 다중공선성 문제가 발생하지 않는다. 하지만 PCA는 예측을 원하는 종속변수 y값을 고려하지 않고 독립변수 X만을 이용해 주성분을 구하고 또한 이 주성분들은 X를 가장 잘 설명하는 성분들이 선택되기 때문에 X의 가장 중요한 성분이 y값을 잘 설명하리란 보장은 없다.

3.2.4 부분회귀분석 (Partial Least Squares, PLS)

PLS는 X의 score를 구할 때 Y의 score와 공분산이 최대가 되도록 계산을 하는 방식으로 X와 Y를 연관시킨다. 즉, 종속변수 Y를 동시에 고려하여 잠재변수(주성분)를 계산하기 때문에 일반적으로 PCR보다 예측 성능이 우수하다. PLS를 계산하기 위해서 일반적으로 많이 사용되는 알고리즘은 NIPALS 알고리즘이다[30]. 하지만 계산속도가 향상된 알고리즘인 SIMPLS 알고리즘이 많이 사용되는 추세이다[31]. SIMPLS 알고리즘을 사용하는 대표적인 프로그램으로는 MATLAB과 SAS가 있다. SIMPLS 알고리즘의 기본 구조는 Algorithm 2와 같다. 이렇게 구해진 $\mathbf{B}_{PLS}$는 다음 식을 통해 새로운 독립변수 $\mathbf{x}^*$로 부터 예측 값 $\hat{\mathbf{y}}^*$

Algorithm 2 SIMPLS algorithm

1: $\mathbf{S} = \mathbf{X}_0^T \mathbf{Y}_0$ % Compute cross-product
2: **for** $a = 1 \rightarrow A$ **do**
3: **if** a=1 **then**
4: $\mathbf{S}$ % compute SVD of $\mathbf{S}$
5: **end if**
6: $\mathbf{S} - \mathbf{P}(\mathbf{P}^T\mathbf{P})^{-1}\mathbf{P}^T\mathbf{S}$ % compute SVD of $\mathbf{S}$
7: $\mathbf{r}$ =first left singular vector % Get weights
8: $\mathbf{t} = \mathbf{X}_0\mathbf{r}$ % Compute scores
9: $\mathbf{p} = \mathbf{X}_0^T\mathbf{t}/(\mathbf{t}^T\mathbf{t})$ % Compute loadings
10: Store $\mathbf{r}, \mathbf{t}$, and $\mathbf{p}$ into $\mathbf{R}, \mathbf{T}$, and $\mathbf{P}$, respectively.
11: **end for**
12: $\mathbf{B}_{PLS} = \mathbf{RT} - \mathbf{Y}_0$ % Compute regression coefficients

를 구하는데 사용된다.

$$\hat{\mathbf{y}}^* = \bar{\mathbf{y}} + (\mathbf{x}^* - \bar{\mathbf{x}})\mathbf{B}_{PLS} \tag{3.8}$$

여기서 $\bar{\mathbf{x}}$와 $\bar{\mathbf{y}}$는 각각의 평균을 의미한다. 라만스펙트럼의 경우 다중공선성과 특이성 문제가 심각하게 발생하고 라만스펙트럼의 주성분이 글루코즈 농도를 잘 설명한다는 보장이 없기 때문에 MLR이나 PCR보다 PLS를 적용하는 것이 합리적이다. PLS를 사용하는데 중요한 것은 잠재변수의 갯수를 결정하는 것이다. 많은 잠재변수를 사용할 경우 트레이닝 데이터는 잘 설명하나 테스트 데이터를 잘 설명하지 못하는 과적합이 발생하고 너무 적은 잠재변수를 사용할 경우 트레이닝 데이터조차 예측하지 못하는 모델이 된다. 이를 위해 본 논문에서는 10-묶음 교차검증 (10-fold cross validation)을 이용하여 잠재변수의 갯수를 결정하였다.

3.2.5 Radial Basis Function PLS (RBF-PLS)

Radial Basis Function (RBF)망은 입력층, 출력층 그리고 숨은층(hidden layer)로 구성되어 있다.

입력층은 들어온 입력을 분배하는 역할을 하며 숨은층의 각각의 유닛은 radial function을 나타낸다. 그리고 출력 유닛의 입력값은 숨은 유닛의 가중치 합이다. RBF망에서 X의 비선형 변환을 수행하고 activation 행렬 A를 만들기 위해 일반적으로 가우시안 함수가 사용된다.

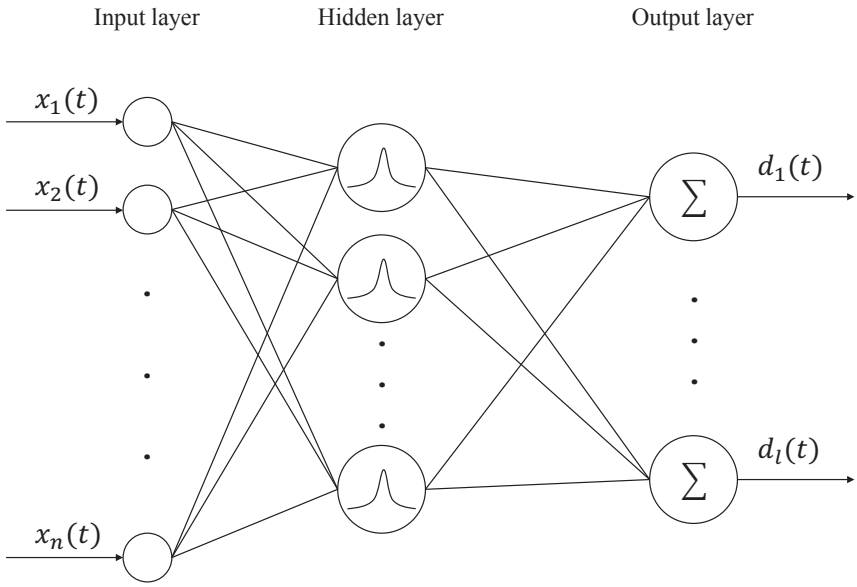

그림 9: RBF망의 구조

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{pmatrix} \in R^{m \times m} \tag{3.9}$$

$$a_{ij} = exp(-\|c_j - x_i\|^2 / \sigma_j^2), \quad i, j = 1, 2, ..., m \tag{3.10}$$

여기서 x_i는 i번째 샘플로부터 얻어진 변수의 값으로 이루어진 벡터이며, a_{ij}는 행렬 A의 성분이다. 그리고 c_j와 σ_j는 각각 j번째 RBF의 중심과 넓이를 의미하며, 파라미터 c_j는 다음과 같이 정의된다.

$$c_j = x_j, \quad j = 1, 2, ..., m \tag{3.11}$$

따라서 activation 행렬 A는 대칭 행렬이 된다.

그 다음 행렬 A와 Y에 대하여 PLS를 수행한다. PLS 모델은 다음과 같이 표현된다.

$$Y = TR + F = AWR + F \tag{3.12}$$

여기서 $T(m \times n_T)$는 회귀 계수 행렬을 나타내는 $A, R(n_T \times l)$의 score 행렬이며 A와 Y의 공분산이 최대가 되도록하는 가우시안 함수의 선형 조합이다. $W(m \times n_T)$는 A의 변환행렬이며 $F(m \times l)$는 오차를 의미한다.

이렇게 PLS와 RBF망을 결합함으로써 A와 Y의 비선형 문제를 선형 대수 문제로 변환할 수 있다. 트레이닝 데이터로 RBF-PLS 모델을 만든 뒤 아래 식을 통해 새로운 샘플에 대한 출력의 예측값을 구할 수 있다.

$$Y_p = A_p WR \tag{3.13}$$

여기서 A_p는 새로운 독립변수 X_p의 activation 행렬이며 Y_p는 이로부터 계산된 예측값이다.

제 4 장

결과

4.1 혼합물 샘플의 글루코즈 농도 예측

제조한 혼합물 샘플의 라만 스펙트럼을 얻기 위해 34개의 샘플을 각각 10mL 바이알에 담아 스펙트럼을 얻었다. 라만 스펙트럼은 외부의 빛에 민감하기 때문에 그림 3의 상단에 보이는 빛이 투과되지 않는 상자를 만든 뒤 그 안에서 촬영이 진행되었다. 5초간 누적된 라만 산란을 통해 하나의 스펙트럼을 얻었으며 이렇게 4번 얻은 스펙트럼의 평균을 최종 스펙트럼으로 사용하였다. 즉 하나의 스펙트럼을 얻는데 20초의 시간이 걸렸다. 본 실험에 사용된 라만 분광기를 이용하여 스펙트럼을 얻으면 그림 10과 같이 양 끝의 불필요한 부분까지 출력된다. 그림 10의 Raman shift의 전체 구간은 -83.50cm^{-1} $\sim$ 4690.2cm^{-1}인데 불필요한 부분을 제거한 뒤 453.86cm^{-1} $\sim$ 4628.2cm^{-1} 구간만을 분석에 사용하였다 (그림 12(a)).

전처리 여부에 따른 PLS 모델의 예측 성능을 비교하기 위해 우선 전처리를 하지 않은 34개의 라만 스펙트럼(그림 12(a))을 이용해 PLS 모델을 만들었다. 최적의 잠재변수의 갯수는 10-묶음 교차검증을 통해 얻은 교차검증의 평균 제곱근 오차 (Root mean square error of cross-validation, RMSECV)가 가장 작을 때의 잠재변수의 갯수를 사용하였다. 여기서 10-묶음 교차검증은 전체 데이터를 10개의 셋으로 나눈 뒤 하나의 셋을 테스트 데이터로 나머지 아홉개의 셋을 트레이닝 데이

터로 사용하는 방법이며 이러한 과정을 10번 반복하여 계산된 10개의 테스트 데이터의 예측력을 기준으로 잠재변수의 갯수를 결정하였다. 10-묶음 교차검증의 과정은 그림 11과 같다. 또한 예측력을 평가하기 위해 사용한 RMSECV는 다음과 같이 정의된다.

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \tag{4.1}$$

여기서 n은 샘플의 갯수, y_i는 실제 글루코즈 농도이며 $\hat{y}_i$는 PLS 모델을 통해 예측된 글루코즈의 농도이다. 그림 12(b)는 RMSECV가 가장 작은 값을 가질 때의 잠재변수의 갯수인 7개의 잠재변수를 이용해 만든 PLS 모델의 결과이다.

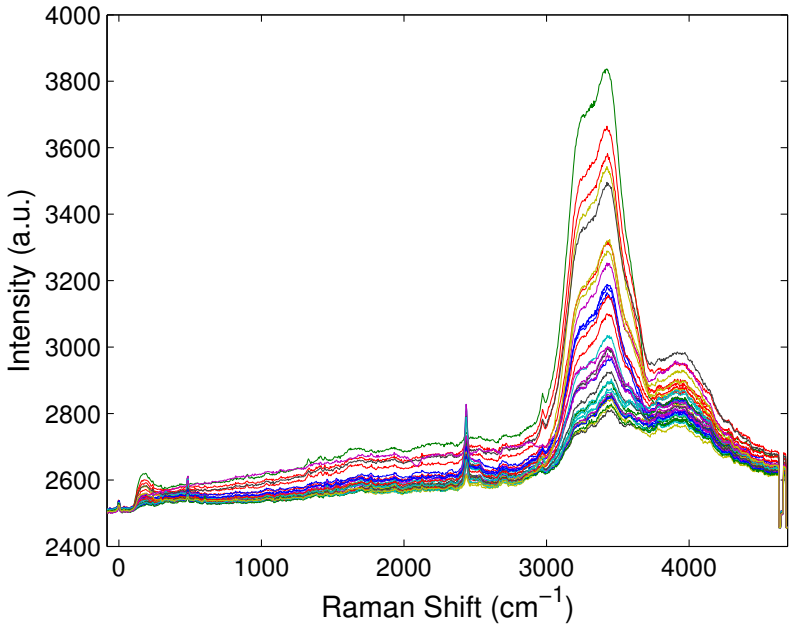

그림 10: 혼합물 샘플로부터 얻은 전체구간의 라만 스펙트럼

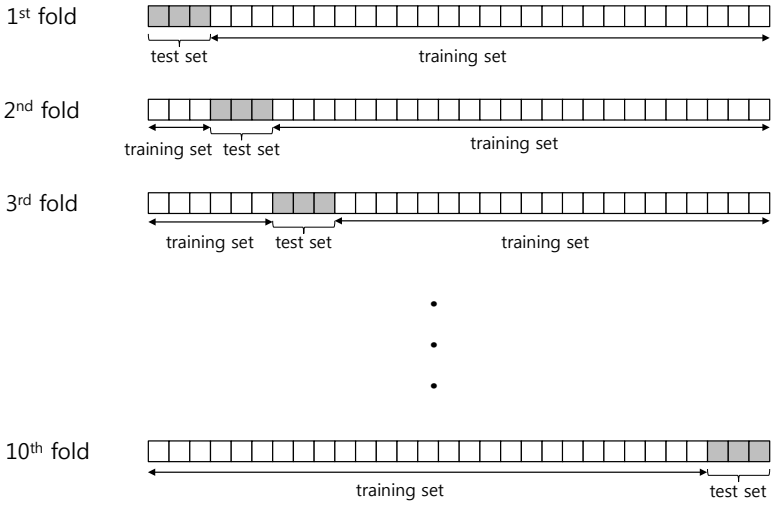

그림 11: 10-묶음 교차검증의 과정

전처리를 하지 않은 라만 스펙트럼을 이용했음에도 예측 결과가 꽤 우수한 것을 확인할 수 있는데 이는 혼합물 샘플에 포함된 물질이 4가지로 단순하고 배경효과가 강하지 않기 때문이다. RCF를 이용해 라만 스펙트럼의 배경을 제거했을 때의 PLS 모델의 성능을 평가하기 위해서는 우선 RCF의 반지름을 결정해야 한다. RCF의 최적 반지름을 결정하기 위한 과정은 다음과 같으며 이에 대한 결과는 그림 14에서 볼 수 있다.

1. 반지름 *r*의 RCF를 이용해 라만 스펙트럼의 배경을 제거한다.

2. 10-묶음 교차검증을 통해 PLS 모델을 만든다.

3. 잠재변수의 갯수를 늘려가며 RMSECV를 계산한 뒤 가장 작은

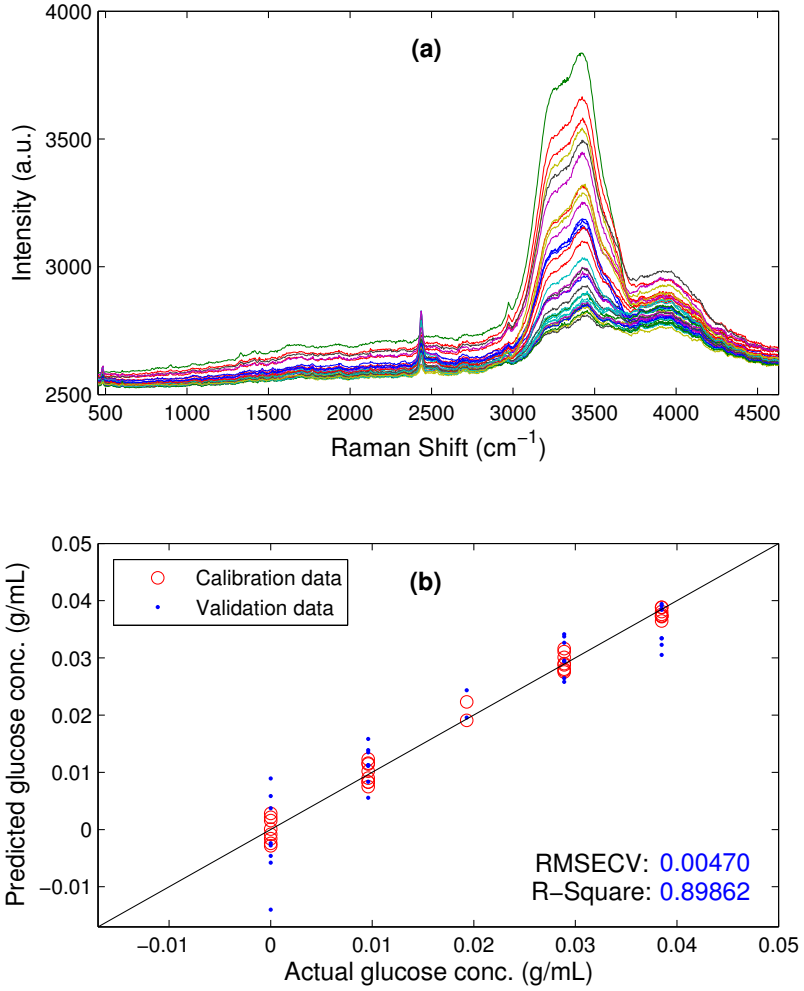

그림 12: 34개의 혼합물 샘플로부터 부터 얻은 라만 스펙트럼과 PLS 모델의 성능

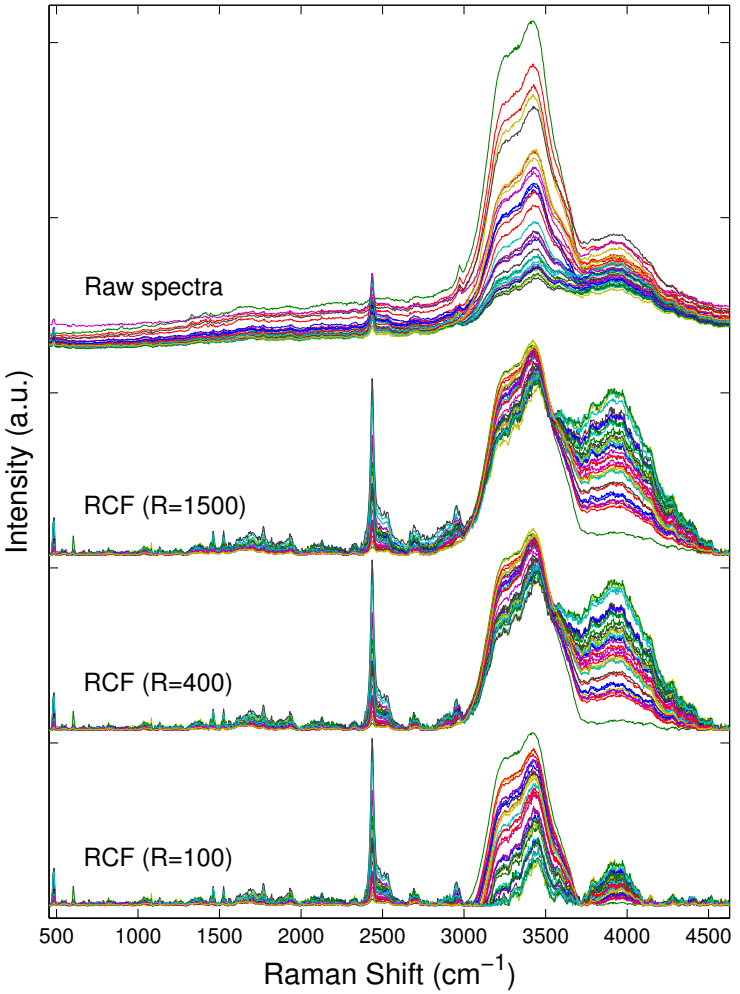

그림 13: RCF의 반지름 크기에 따른 스펙트럼의 변화

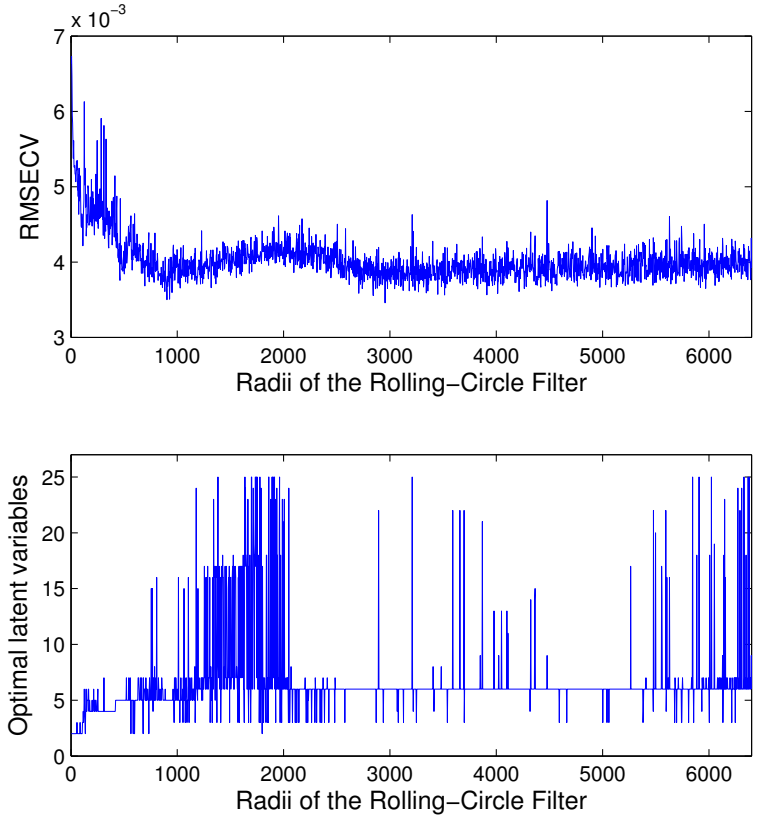

그림 14: RCF의 반지름의 따른 최적의 잠재변수의 갯수와 그 때의 RMSECV 값

RMSECV값을 가지는 잠재변수의 갯수를 결정한다.

4. 위와 같은 과정을 $1 \leq r \leq 6400$에 대하여 반복 수행한다. 그 다음 가장 작은 RMSECV를 가지는 반지름과 잠재변수의 갯수를 결정한다. 본 연구에서 사용한 라만 스펙트럼의 경우 1 포인트는 3.2cm^{-1}이기 때문에 이 과정은 총 2000번 반복된다.

그림 14을 보면 반지름에 따른 RMSECV값이 심한 노이즈를 가지게 되

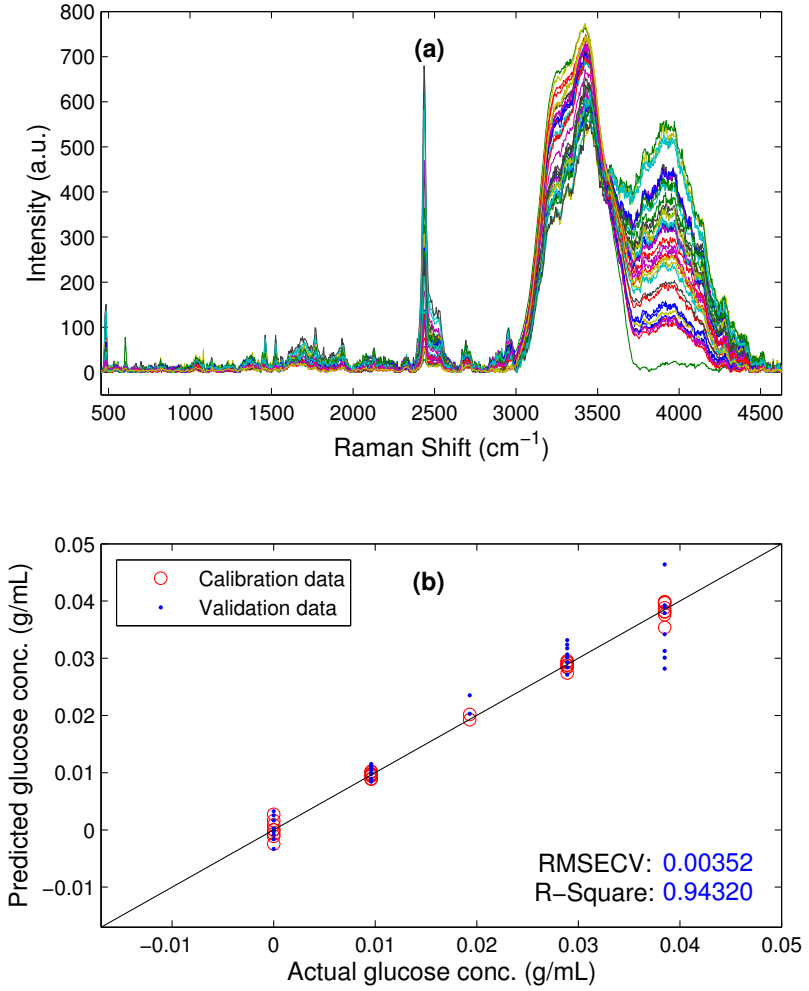

그림 15: RCF를 이용해 배경을 제거한 라만 스펙트럼과 PLS 모델의 성능

는데 이는 10-묶음 교차검증의 특징 때문에 발생한 결과이다. 본 연구에서는 교차검증을 수행할 때 10개의 묶음이 임의적으로 형성되도록 하였다.

따라서 교차검증을 수행할 때 마다 트레이닝과 테스트에 사용되는 데이터가 바뀌게 된다. 교차검증의 신뢰도를 높이기 위해 일반적으로 이러한 무작위로 추출하는 방법을 사용한다. 이 경우 반지름이 800에서 1000사이일 때 가장 작은 RMSECV값을 가지게 된다. 최종적으로 최적 반지름은 899.2(281 포인트)로 최적의 잠재변수 갯수는 5개로 결정되었다. 그림 15는 이러한 과정을 통해 결정된 파라미터를 이용해 PLS를 수행한 결과이다. 비록 강한 배경효과가 존재하지 않는 샘플이지만 RCF의 적용은 큰 성능향상을 가져오는 것을 확인할 수 있다.

4.2 미세조류 샘플의 글루코즈 농도 예측

4.2.1 RCF적용 전 후의 PLS 모델 성능 비교

미세조류 샘플의 라만 스펙트럼을 얻기 위해 혼합물 샘플과 동일하게 20초간 얻은 라만 스펙트럼의 평균값을 사용하였다. 그림 16은 이렇게 얻은 전체구간의 라만 스펙트럼인데 Raman shift가 3800 이후인 부분의 상단이 잘린 모습을 볼 수 있다. 이는 본 실험에서 사용한 라만 분광기의 Intensity (a.u.)의 측정 한계가 65535이기 때문에 이 값을 넘는 Raman intensity는 측정되지 않는다. 따라서 미세조류로 부터 얻은 라만 스펙트럼의 경우 453.86cm^{-1} ~ 3759.7cm^{-1} 구간을 분석에 사용하였다. 그림 17(a)는 전처리를 하지 않는 36개 미세조류 샘플의 라만 스펙트럼이다. 혼합물 샘플과는 달리 매우 심한 배경효과가 발생하는 것을 볼 수 있다. 우선 배경효과 제거 없이 7개의 잠재변수를 이용해

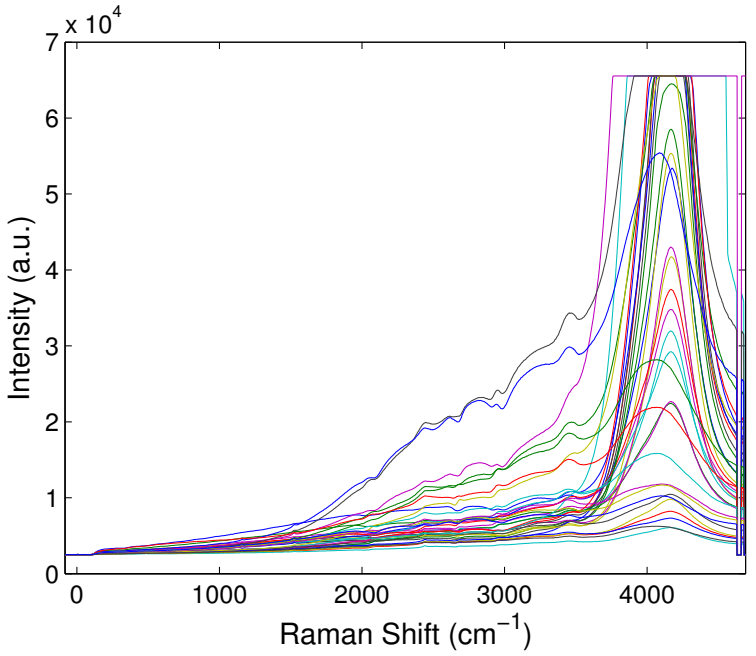

그림 16: 미세조류 샘플로부터 얻은 전체구간의 라만 스펙트럼

PLS 모델을 만든 뒤 예측 성능을 평가하였다.

그림 17(b)에서 볼 수 있듯이 배경효과를 제거하지 않은 라만 스펙트럼을 사용할 경우 글루코즈 농도를 전혀 예측하지 못하는 것을 볼 수 있다.

라만 스펙트럼의 배경을 제거한 뒤 PLS 모델을 만들기 위해 혼합물 샘플의 PLS 모델을 만들 때와 동일한 과정으로 본 샘플에 대한 RCF의 최적 반지름과 PLS의 잠재변수의 최적 갯수를 결정하였다. 그림 18에서 볼 수 있듯이 이 경우도 RCF의 반지름이 800과 1000사이일 때 가장 우수한 예측 성능을 보였다. 흥미로운 사실은 전혀 다른 두 샘플임에도 RCF의 반지름에 따른 RMSECV값의 경향이 비슷하다는 점이다. 따라서 본 논문에서 제시하는 RCF의 최적 반지름은 다른 샘플에

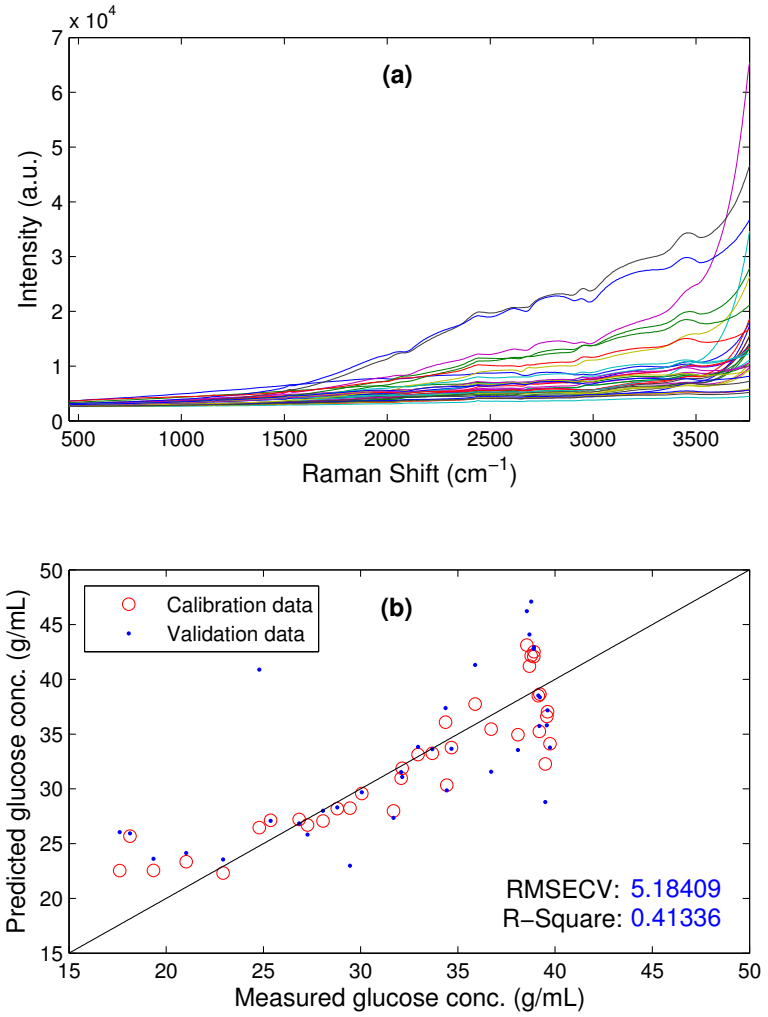

그림 17: RCF를 이용해 배경을 제거한 라만 스펙트럼과 PLS 모델의 성능

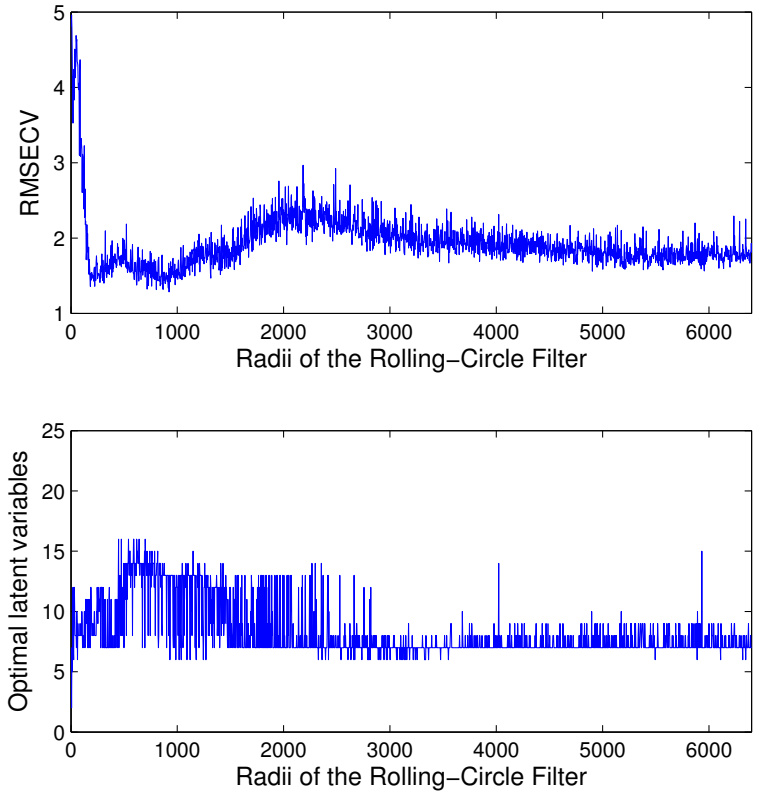

그림 18: RCF의 반지름의 따른 최적의 잠재변수의 갯수와 그 때의 RMSECV 값

대해서도 동일하게 적용될 것으로 보인다. 미세조류 샘플의 글루코즈 농도 예측을 위한 PLS 모델 생성에서도 혼합물 샘플과 동일하게 899.2를 RCF의 최적 반지름으로 결정하였으며 이 때 가장 작은 RMSECV 값을 가지도록 하는 잠재변수의 갯수는 7개 였다. 그림 19(a)는 RCF를 이용해 배경을 제거한 라만 스펙트럼이다. 그림 17(a)와 비교해볼 경우 잘 보이지 않던 특성 피크들이 크게 강조된 것을 볼 수 있다. 그림 19(b)는 RCF를 이용해 배경을 제거한 라만 스펙트럼을 이용한 PLS 모

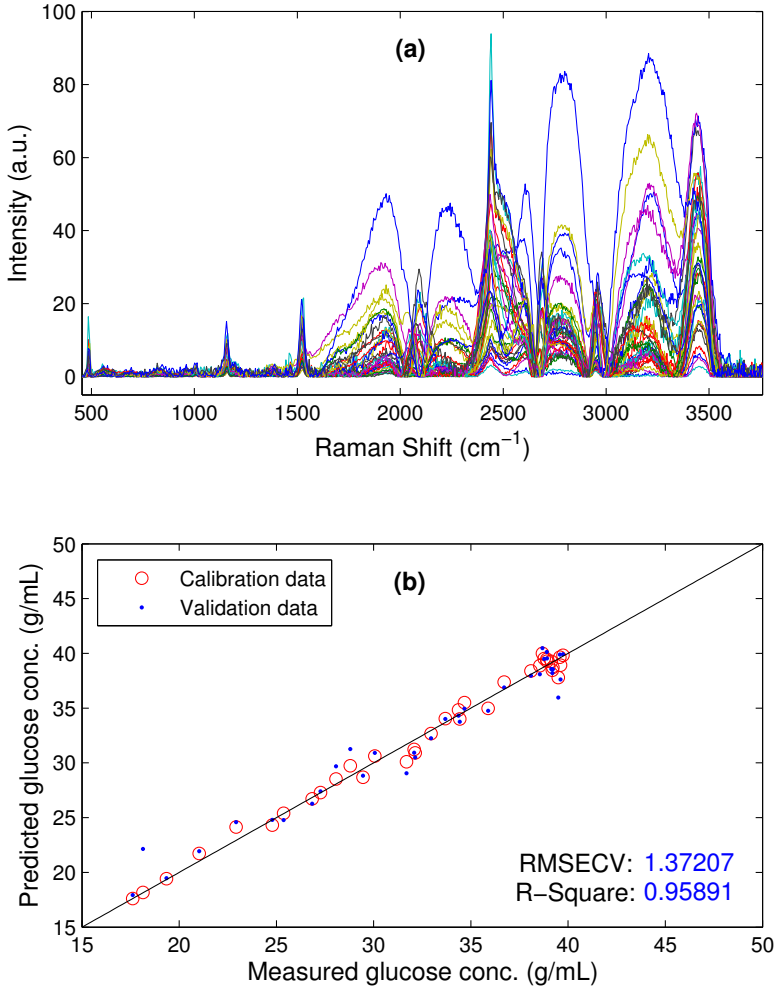

그림 19: RCF를 이용해 배경을 제거한 라만 스펙트럼과 PLS 모델의 성능

델의 예측 결과인데 글루코즈 농도를 상당히 정확히 예측하는 것을 볼 수 있다. 이처럼 미세조류와 같이 배경효과가 강한 샘플에 대해서는 배경효과 제거는 필수적이다.

4.2.2 독립적인 실험군의 글루코즈 농도 예측

지금까지 5개의 미세조류 배양 실험세트의 데이터와 10-묶음 교차검증을 사용하여 PLS 모델 생성에 필요한 최적 파라미터를 결정하였다. 10-묶음 교차검증을 사용할 때 5개의 실험세트의 모든 정보중 무작위로 90%의 정보가 모델 생성에 사용되며 나머지 10%가 검증에 사용된다. 즉, 하나의 실험세트에서 모델 생성과 검증이 모두 이루어지는 것이다. 본 논문에서 제시하는 통합 프레임워크의 성능을 엄격하게 검증하게 위해서 이번 절에서는 4개의 실험세트만을 이용해 PLS 모델을 생성한 뒤 독립적인 나머지 실험세트의 모든 글루코즈 농도를 예측해보려 한다. 이 과정은 그림 20에 요약되어 있다. PLS와 같은 통계 모델의 경우 외삽(extrapolation)을 할 경우 많은 문제가 생긴다. 따라서 전 범위의 데이터를 포함하기 위해 5번째 실험세트는 모델 생성에 사용하고 그 다음으로 많은 데이터를 가지고 있는 4번째 실험세트의 글루코즈 농도를 예측하였다. RCF의 반지름과 PLS의 잠재변수의 갯수는 앞에서 결정된 899.2와 7개를 사용하였다. 4번째 실험세트의 실제 농도와 예측된 농도는 그림 21에서 볼 수 있다. 모델 생성에 사용한 실험세트와 검증에 사용한 실험세트는 독립적이지만 농도 예측 성능이 우수한 것으로 보아 본 PLS 모델은 안정적이라 할 수 있다.

4.2.3 Successive SG filter 적용

본 연구에서는 SG filter를 예측된 글루코즈 농도를 보정하기 위한 후처리 기법으로 사용하였다. 이 기법을 사용하기 위해선 다항식의 차수와 이동창의 크기를 결정해야 한다. 본 논문에서는 2차 다항식과 크기가 5인 이동창을 사용하여 Successive SG filter를 그림 21의 데이터에 적용하였다. 적용과정은 다음과 같다.

1. 이동창이 5일 경우 최소한 5개의 데이터가 있어야 보정이 가능하므로 최초 네 값이 누적되는 동안은 아무런 보정이 이루어지지 않는다. 여섯 번째 값이 들어오면 이 5개의 값을 이용해 2차 회귀곡선을 만든다.

2. 첫 값과 끝 값을 제외한 중간의 세 값중 하나를 선택하여 회귀곡선 위로 이동시킨다. 본 논문에서는 네 번째 값을 보정하는데 사용하였는데 그 이유는 새로운 값이 들어오는 시간과의 시간지

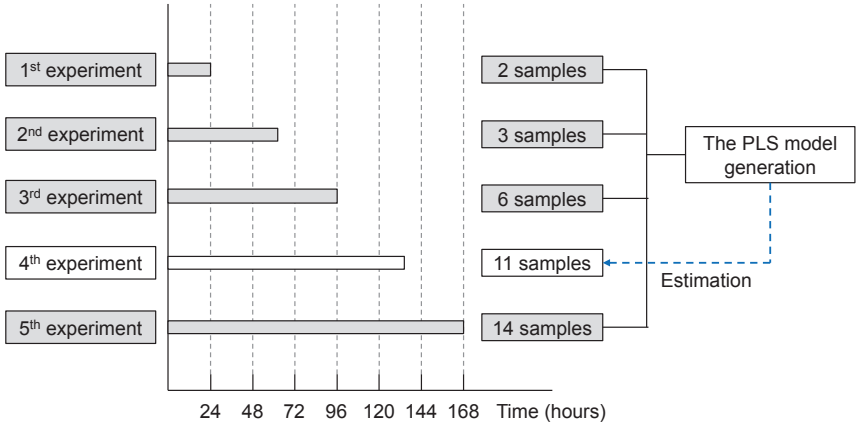

그림 20: 독립적인 실험세트에 대한 PLS 모델 검증 과정

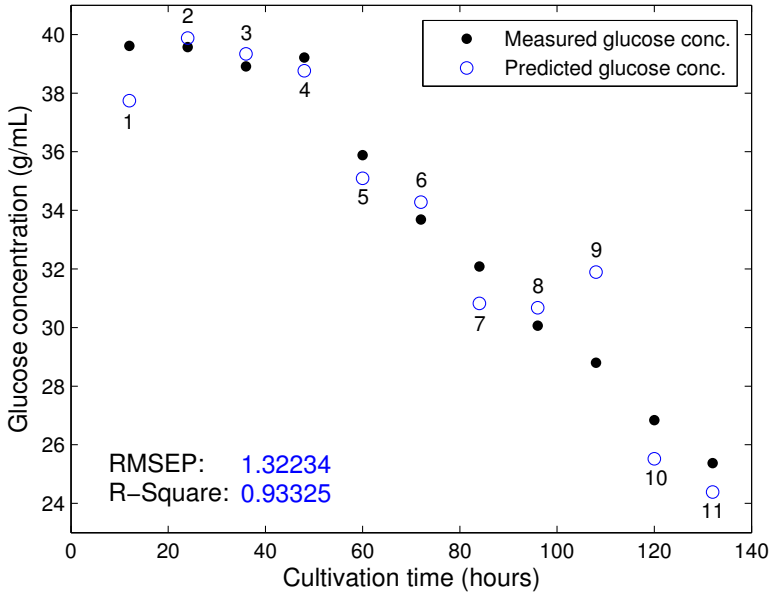

그림 21: 4번째 실험세트의 배양 시간에 따른 글루코즈 농도 예측 결과

연(time delay)을 최소화 하기 위함이다.

3. 여섯 번째 값이 들어오면 두 번째 값부터 여섯 번째 값까지의 5개의 데이터를 이용해 회귀곡선을 만든다. 그 다음 다섯 번째 값을 회귀곡선 위로 이동시킨다.

위의 과정이 반복되면 표본화 시간(sampling time)만큼의 시간지연이 발생하며 농도가 보정된다. 일반적으로 표본화 시간의 간격은 판단 시간(decision time)의 간격보다 짧기 때문에 실시간 모니터링 데이터로 사용이 가능하다.

그림 22은 글루코즈 농도의 실제값, 예측값 그리고 보정값을 모두 비교한 결과이다. 보정 작업이 이루어지지 않은 첫 번째부터 네 번째까지의 값과 마지막 값을 제외하면 전반적으로 예측 성능이 크게 향상된

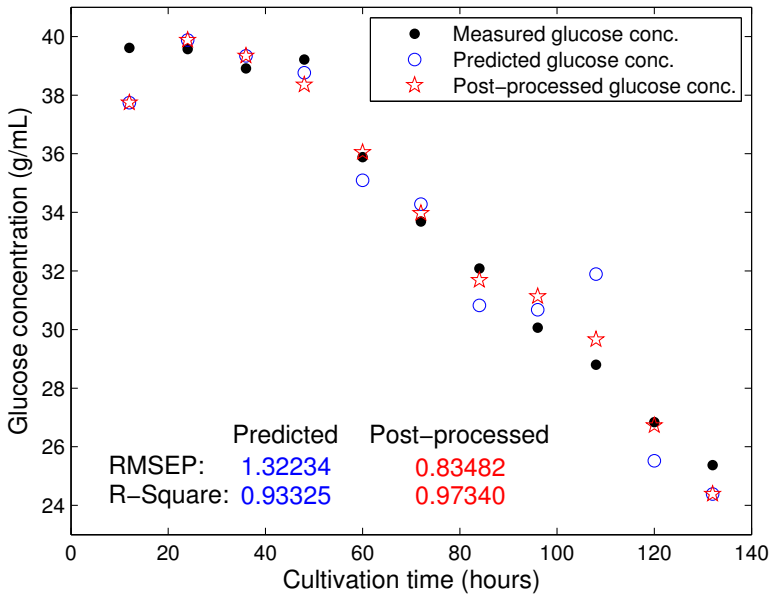

그림 22: Successive SG filter를 적용한 후의 글루코즈 농도

것을 볼 수 있다.

4.2.4 전처리 기법과 회귀분석 기법에 따른 예측 모델 성능 비교

라만 스펙트럼은 배경뿐만 아니라 노이즈와 산란효과도 존재한다. 따라서 이러한 문제를 해결하기 위해 다양한 전처리 기법이 사용된다. 노이즈는 일반적으로 SG filter를 통해 제거되며 산란효과는 SNV를 이용해 제거된다. 따라서 전처리 기법의 경우 RCF, SG filter 그리고 SNV 세 가지 전처리 기법을 비교하였다. 회귀분석 기법을 결정하기 위해서는 MLR, PCR, PLS, RBF-PLS가 사용되었다. RBF-PLS를 사용하기 위해서는 σ값을 결정해야 한다. 최적의 σ값을 결정하기 위해서

Toolbox for multivariate calibration techniques (TOMCAT)[32]을 사용하였다. 통계 모델의 성능 비교는 4.2.4절과 동일한 과정으로 진행하였다. 예측 성능을 비교하는 지표로는 트레이닝 데이터를 예측할 경우의 최소 제곱근 오차(root mean square error of calibration, RMSEC)와 테스트 데이터를 예측할 경우의 최소 제곱근 오차(root mean square error of prediction, RMSEP)를 사용하였다. 예측 성능을 비교한 결과는 표 4에 나타내었다.

RCF를 이용해 배경을 제거한 선형 PLS와 RBF-PLS 모델의 예측 성능이 가장 우수한 것을 볼 수 있다. PLS와 RBF-PLS 모델의 RMSEP의 차이는 0.28%로 무시할 수 있는 수준이다. 선형 모델과 비선형 모델의 예측 성능 차이가 미미할 경우 선형 모델을 사용하는 것이 더 안정적인 모니터링이 가능하기 때문에 본 논문에서는 PLS 모델을 사용하였다. MLR의 경우 트레이닝 데이터는 완벽하게 예측하지만 새로운 데이터에 대해서는 전혀 예측하지 못하는 과적합이 심하게 발생하는 것을 볼 수 있다. 또한 스펙트럼의 노이즈를 제거하는 SG filter의 경우 예측 성능이 큰 영향을 미치지 못하는 것도 확인할 수 있다.

표 4: 전처리 기법과 다변량 회귀분석 기법에 따른 예측성능 비교

		MLR	PCR	PLS	RBF-PLS
Raw spectra	RMSEC	0.0000	4.1758	3.0241	2.2026
	RMSEP	6.7156	2.2768	5.3726	3.9841
RCF	RMSEC	0.0000	1.0122	0.6726	0.9307
	RMSEP	24.1593	1.6270	**1.3223**	**1.3186**
SG filter	RMSEC	0.0000	4.2140	3.0553	2.7105
	RMSEP	8.3996	2.3508	5.3937	4.4173
SNV	RMSEC	0.0000	2.6500	1.5212	1.5308
	RMSEP	4.1143	2.5201	1.6675	1.6375
RCF + SG filter	RMSEC	0.0000	1.0286	0.7437	1.1346
	RMSEP	25.8540	1.6758	1.3279	1.3767
RCF + SNV	RMSEC	0.0000	3.3459	1.4438	1.9974
	RMSEP	4.9046	4.6268	2.6982	2.9678

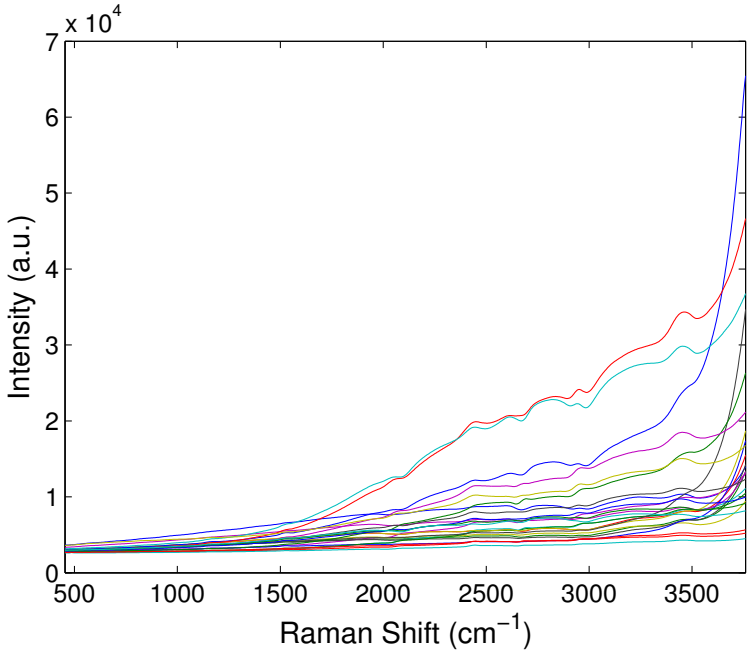

그림 23: SG filter를 적용한 라만 스펙트럼

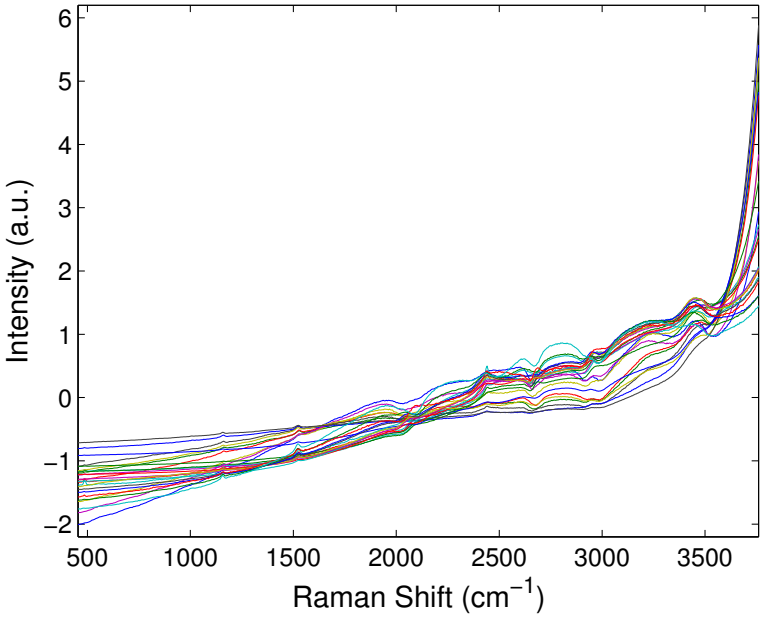

그림 24: SNV를 적용한 라만 스펙트럼

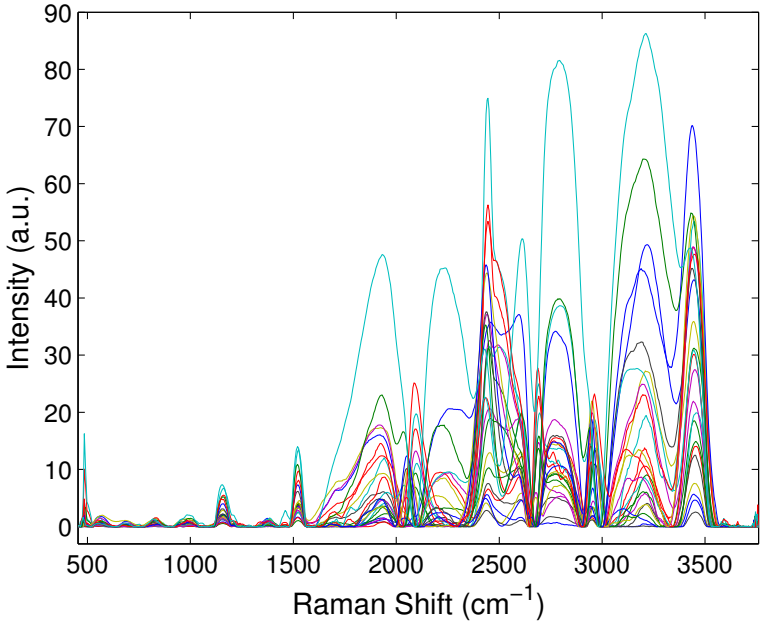

그림 25: SG filter와 RCF를 모두 적용한 라만 스펙트럼

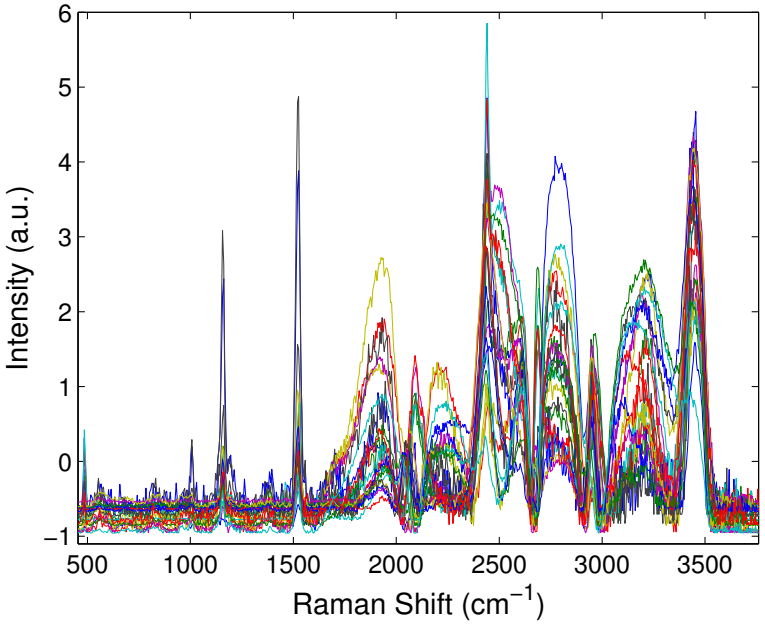

그림 26: RCF와 SNV를 모두 적용한 라만 스펙트럼

제 5 장

결론 및 제안

본 연구에서는 미세조류 배양 공정의 글루코즈 농도를 실시간으로 예측할 수 있도록 라만 분광기와 다변량 회귀분석을 이용한 통합 프레임워크를 제안하였다. 라만 분광기는 건조, 파쇄, 추출과 같은 전처리 없이 샘플로 부터 직접적으로 스펙트럼을 얻을 수 있어 실시간 모니터링에 적합하다. 이렇게 얻은 라만 스펙트럼은 배경효과를 큰 문제점을 가지고 있다. 이러한 배경을 제거하기 위해 본 연구에서는 RCF를 사용하여 라만 스펙트럼의 배경을 제거하였다. 다변량 회귀 분석을 위해서는 다중공선성과 특이성이 강한 라만 스펙트럼의 특성을 고려하여 PLS를 사용하였다. 이러한 과정을 통해 농도 예측이 가능한지 알아보기 위해 두 가지의 실험을 진행하였다. 첫 번째 실험은 글루코즈, 글라이신, 물, 콩기름을 섞은 혼합물 샘플을 이용하였으며, 두 번째 실험에서는 미세조류를 배양시간에 따라 5회 배양하여 얻은 샘플로 진행하였다. RCF의 최적 반지름과 PLS의 최적 잠재변수의 갯수를 결정하기 위해 10-묶음 교차검증을 사용하였다. 두 가지 실험 모두 라만 스펙트럼의 배경을 제거 할 경우 예측 성능이 향상 되었으며, 특히 미세조류 샘플의 경우 색소에 의한 형광배경효과가 크게 발생하기 때문에 RCF 적용 전 후의 예측성능이 큰 차이를 보이는 것을 볼 수 있다. 미세조류 샘플의 경우 글루코즈 농도가 시간에 의존하는데 이러한 특성을 고려하여 Successive SG filter를 이용하여 예측된 농도를 보정하였다.

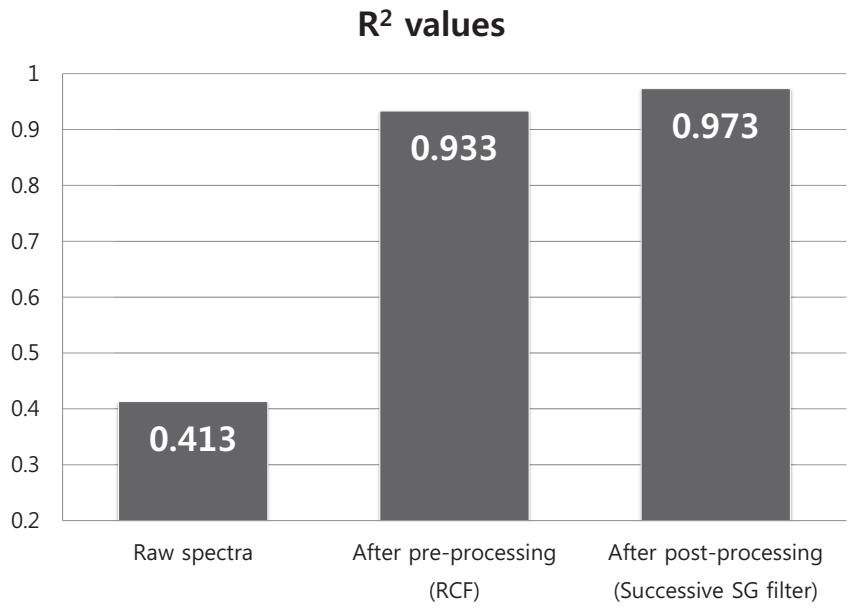

그림 27: 미세조류 샘플로부터 얻은 라만 스펙트럼의 전처리 및 후처리 적용에 따른 PLS 모델의 R^2값 비교

Successive SG filter를 적용할 경우 예측 성능은 RMSEP 기준 36.8% 향상 되었다. 그림 27에서 첫 번째 R^2값은 아무런 처리를 하지 않은 라만 스펙트럼을 이용하여 농도를 예측한 성능을 나타낸다. 그 다음은 RCF를 이용하여 배경효과를 제거한 뒤 농도를 예측할 경우의 성능이며 마지막은 예측한 농도를 Successive SG filter로 후처리까지 한 경우의 예측성능이다. 전처리와 후처리를 통해 큰 예측성능 향상이 있는 것을 확인할 수 있다. 또한 다른 전처리 기법과 다변량 회귀분석 기법들을 비교한 결과 라만 스펙트럼의 특성을 잃어버리지 않도록 전처리를 잘 해줄 경우 선형 PLS 모델로도 예측이 잘 되는 것을 확인할 수 있었다.

본 연구에서 제안한 통합 프레임워크를 이용해 광생물반응기와

라만 분광기를 연결하여 실제 온라인 모니터링에 적용하여 검증하는 과정이 필요할 것이다. 이 과정은 라만 분광기의 탐측기 (probe)를 immersion probe로 변경한 뒤 광생물반응기 상단의 남은 포트를 이용하면 설치가 가능하다. 또한 본 연구의 경우 하나의 라만 스펙트럼을 통해 하나의 물질의 농도를 예측했지만 좀 더 많은 데이터가 축적될 경우 하나의 라만 스펙트럼을 이용하여 여러 물질의 농도를 동시에 예측하는 것도 가능할 것으로 보인다.

참고 문헌

[1] R. Raja, S. Hemaiswarya, N. A. Kumar, S. Sridhar, R. Rengasamy, A perspective on the biotechnological potential of microalgae, Critical Reviews in Microbiology 34 (2) (2008) 77–88.

[2] Q. Hu, M. Sommerfeld, E. Jarvis, M. Ghirardi, M. Posewitz, M. Seibert, A. Darzins, Microalgal triacylglycerols as feedstocks for biofuel production: perspectives and advances, Plant Journal 54 (4) (2008) 621–639.

[3] J. Pires, M. Alvim-Ferraz, F. Martins, M. Simões, Carbon dioxide capture from flue gases using microalgae: Engineering aspects and biorefinery concept, Renewable and Sustainable Energy Reviews 16 (5) (2012) 3043–3053.

[4] T. Heredia-Arroyo, W. Wei, B. Hu, Oil accumulation via heterotrophic/mixotrophic chlorella protothecoides, Applied Biochemistry and Biotechnology 162 (7) (2010) 1978–1995.

[5] Y. K. Lee, Microalgal mass culture systems and methods: Their limitation and potential, Journal of Applied Phycology 13 (4) (2001) 307–315.

[6] R. Harun, M. Singh, G. M. Forde, M. K. Danquah, Bioprocess engineering of microalgae to produce a variety of consumer products, Renewable & Sustainable Energy Reviews 14 (3) (2010) 1037–1047.

[7] X. W. Zhang, Y. M. Zhang, F. Chen, Application of mathematical models to the determination optimal glucose concentration and light intensity for mixotrophic culture of spirulina platensis, Process Biochemistry 34 (5) (1999) 477–481.

[8] J. Wang, Electrochemical glucose biosensors, Chemical Reviews 108 (2) (2008) 814–825.

[9] R. J. Swain, M. M. Stevens, Raman microspectroscopy for non-invasive biochemical analysis of single cells, Biochemical Society Transactions 35 (2007) 544–549.

[10] Y. Y. Huang, C. M. Beal, W. W. Cai, R. S. Ruoff, E. M. Terentjev, Micro-raman spectroscopy of algae: composition analysis and fluorescence background behavior, Biotechnol Bioeng 105 (5) (2010) 889–98.

[11] O. D. Paraschuk, N. N. Brandt, O. O. Brovko, A. Y. Chikishev, Optimization of the rolling-circle filter for raman background subtraction, Applied Spectroscopy 60 (3) (2006) 288–293.

[12] F. Estienne, D. L. Massart, N. Zanier-Szydlowski, P. Marteau, Multivariate calibration with raman spectroscopic data: a case study, Analytica Chimica Acta 424 (2) (2000) 185–201.

[13] N. C. Dingari, I. Barman, J. W. Kang, C. R. Kong, R. R. Dasari, M. S. Feld, Wavelength selection-based nonlinear calibration for transcutaneous blood glucose sensing using raman spectroscopy, Journal of Biomedical Optics 16 (8).

[14] N. R. Abu-Absi, B. M. Kenty, M. E. Cuellar, M. C. Borys, S. Sakhamuri, D. J. Strachan, M. C. Hausladen, Z. J. Li, Real time monitoring of multiple parameters in mammalian cell culture bioreactors using an in-line raman spectroscopy probe, Biotechnol Bioeng 108 (5) (2011) 1215–1221.

[15] I. Ruisanchez, P. M. Ramos, Noise and background removal in raman spectra of ancient pigments using wavelet transform, Journal of Raman Spectroscopy 36 (9) (2005) 848–856.

[16] O. Samek, P. Zemanek, A. Jonas, H. H. Telle, Characterization of oil-producing microalgae using raman spectroscopy, Laser Physics Letters 8 (10) (2011) 701–709.

[17] F. Bonnier, S. M. Ali, P. Knief, H. Lambkin, K. Flynn, V. McDonagh, C. Healy, T. C. Lee, F. M. Lyng, H. J. Byrne, Analysis of human skin tissue by raman microspectroscopy: Dealing with the background, Vibrational Spectroscopy 61 (2012) 124–132.

[18] Q. Ye, Q. Xu, Y. Yu, R. Qu, Z. Fang, Rapid and quantitative detection of ethanol proportion in ethanol–gasoline mixtures by raman spectroscopy, Optics Communications 282 (18) (2009) 3785–3788.

[19] K. Krishnan, R. A. Plane, Raman study of glycine complexes of zinc(2) cadmium(2) and beryllium(2) and formation of mixed complexes in aqueous solution, Inorganic Chemistry 6 (1) (1967) 55–60.

[20] H. Xu, X. L. Miao, Q. Y. Wu, High quality biodiesel production from a microalga chlorella protothecoides by heterotrophic growth in fermenters, Journal of Biotechnology 126 (4) (2006) 499–507.

[21] H. D. Siegler, A. Ben-Zvi, R. E. Burrell, W. C. McCaffrey, The dynamics of heterotrophic algal cultures, Bioresource Technology 102 (10) (2011) 5764–5774.

[22] I. Shihira-Ishikawa, E. Hase, Nutritional control of cell pigmentation in chlorella-protothecoides with special reference to the degeneration of chloroplast induced by glucose, Plant and Cell Physiology 5 (2) (1964) 227–240.

[23] I. K. Mikhailyuk, A. P. Razzhivin, Background subtraction in experimental data arrays illustrated by the example of raman spectra and fluorescent gel electrophoresis patterns, Instruments and Experimental Techniques 46 (6) (2003) 765–769.

[24] A. Savitzky, M. J. E. Golay, Smoothing + differentiation of data by simplified least squares procedures, Analytical Chemistry 36 (8) (1964) 1627–&.

[25] R. J. Barnes, M. S. Dhanoa, S. J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Applied Spectroscopy 43 (5) (1989) 772–777.

[26] A. Candolfi, R. De Maesschalck, D. Jouan-Rimbaud, P. A. Hailey, D. L. Massart, The influence of data pre-processing in the pattern recognition of excipients near-infrared spectra, Journal of Pharmaceutical and Biomedical Analysis 21 (1) (1999) 115–132.

[27] M. Otto, Chemometrics: statistics and computer application in analytical chemistry, Vch Verlagsgesellschaft Mbh, 2007.

[28] K. Pearson, On lines and planes of closest fit to systems of points in space, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2 (11) (1901) 559–572.

[29] L. Smith, A tutorial on principal components analysis, Cornell University, USA 51 (2002) 52.

[30] H. Wold, Soft modelling by latent variables: the non-linear iterative partial least squares (nipals) approach, Perspectives in Probability and Statistics, In Honor of MS Bartlett (1975) 117–144.

[31] S. Dejong, Simpls - an alternative approach to partial least-squares regression, Chemometrics and Intelligent Laboratory Systems 18 (3) (1993) 251–263.

[32] M. Daszykowski, S. Serneels, K. Kaczmarek, P. Van Espen, C. Croux, B. Walczak, Tomcat: A matlab toolbox for multivariate calibration techniques, Chemometrics and Intelligent Laboratory Systems 85 (2) (2007) 269–277.

Abstract

Development of soft sensor based on Raman spectroscopy for on-line monitoring of glucose concentrations in microalgal production system

Se-kyu Oh
School of Chemical and Biological Engineering
The Graduate School
Seoul National University

Microalgal cultivation process has recently attracted much attention due to biotechnological and chemical potential. Microalgae can be used to produce a diverse range of valuable compounds such as vitamins, natural pigments, carotenoid, protein and carbohydrates. They also produce triacylglycerols (TAGs) as feedstocks for biodiesel production. In the algal production process using photo-bioreactor, two parameters, glucose concentrations and light intensities, are very improtant for optimal control. Therefore, two parameters should be measured in real-time. In case of light intensity, photometer can be used to measure light intensity in real-time. In case of glucose concentration, continuous glucose monitors (CGMs) and

High-performance liquid chromatography (HPLC) can be used to measure the concentration; however, these equipments have a lot of drawbacks.

In this work, we will present an integrated framework to estimate glucose concentration in real-time using Raman spectroscopy. The proposed framework proceed to the following steps. First, backgound effect on the Raman spectra will be removed by Rolling-Circle Filter (RCF). Secondly, we will find the relationship between Raman spectra taken from the samples and glucose concentrations using Partial Least Squares (PLS). In the last step, we will adjust the predicted values using Successive Savitzky-Golay smoothing filter. Two experiment were carried out in order to show that the proposed framework is able to estimate glucose concentration. In case of the first experiment, prediction performance (R^2) of glucose concentrations improved from 0.899 to 0.943 using the proposed framework. Also, in case of the second experiment, prediction performance (R^2) of glucose concentrations greatly improved from 0.413 to 0.973 using this framework.

Keywords : Soft sensor, Raman spectroscopy, Microalgae, Multivariate analysis, Chemometrics, On-line monitoring

Student Number : 2011-21048