다중 발화점을 이용한 Grassfire 스팟매칭 기법

류윤규*

Grassfire Spot Matching Method for multi-seed matched spot pair

Yun-Kyoo Ryoo

요 약

grassfire 스팟매칭 알고리즘은 중심 스팟을 기준으로 이웃 스팟들의 패턴 유사도에 따라 스팟을 매칭하는 알고리즘으로 잔디에 붙는 불이 사방으로 번져가는 방식을 이용한 grassfire 알고리즘을 이다. 씨드 스팟쌍은 매칭이 정확하게 확인된 스팟쌍으로써 매칭이 시작되는 발화점이며 이것으로부터 스팟매칭이 시작된다. grassfire 스팟매칭 알고리즘에서는 스팟매칭을 시작하는 발화점이 필요한데 기존 grassfire 스팟매칭 알고리즘에서는 한 개의 발화점을 이용하였다. 본 논문에서는 grassfire 알고리즘의 스팟매칭의 성능을 높이기 위하여 한 개의 발화점이 아닌 다중의 발화점을 선정하는 방법을 제안한다. 다중 발화점을 이용한 grassifre 알고리즘은 여러 개의 발화점을 선정한 후 개별 발화점으로부터 스팟매칭을 수행하고 결과들을 계산한다. 제안된 알고리즘은 한 개의 씨드스팟을 이용한 방법보다 스팟 검출율과 스팻매칭 정확도의 측면에서 좋은 성능을 보인다.

Abstract

Grassfire spot matching method is based on similarity comparison of topological patterns for neighbor spots. This is a method where spot matching is performed as if fire spreads all around on grass. Spot matching starts from a seed spot pair confirmed as a matched pair of spots and spot matching spreads to the direction where the best matching result is produced. In this paper, it is a bit complicated way of grassfire method where multi-seed matched spot pair are manually selected and spot matching is performed from each multi-seed matched spot pair. The proposed method shows better performance in detection rate and accuracy than that of the previous method.

▸ Keywords : neighbor spot, topological pattern, grassfire, multi-seed matched spot pair, normalized Hausdorff distance

• * 제1저자 : 대구보건대학교 보건의료전산과 교수
• 투고일 : 2014. 10. 30, 심사일 : 2014. 11. 30, 게재확정일 : 2014. 12. 30.

Ⅰ. 서 론

2차원 전기영동은 단백질체학에서 널리 쓰이는 단백질 분리방법이다. 전기영동의 기본원리는 단백질 분자의 등전점과 질량의 특성을 이용하여 샘플 속에 포함된 단백질을 2차원의 평면의 겔 위에서 분리하는 것이다[1].

단백질이 분리되면 2차원 겔 상에 반점을 관찰할 수 있는데 이러한 반점들이 개별적으로 분리되고 이러한 단백질이 어떤 단백질인지는 그것들이 위치한 2차원적 위치가 중요한 단서를 제공하게 된다.

단백질 연구를 위해서는 특정한 조직에서 어떠한 단백질들이 어떠한 양상으로 발현되는지를 조사하는 일이 매우 중요하다. 동일한 생체조직이라도 각각 다른 환경에서 서로 다른 단백질을 생성시키는데 이러한 단백질 발현의 차이를 추적하기 위하여 정상조직의 기준이 되는 참조겔(reference gel)과 테스트하고자 하는 목적겔(target gel)을 서로 비교하여 단백질 발현과 구성의 차이를 검출하게 된다[2].

한 겔 상에는 일반적으로 수 천개의 단백질이 포함되어 있으므로 단백질의 발현의 차이를 파악하는 것을 수작업으로는 거의 불가능하다. 이러한 이유 때문에 2차원 전기영동의 분석 작업은 자동화의 과정을 거치는 것이 일반적이다.

2차원 전기영동은 실험방법이 간단하지만 실험결과는 매우 많은 실험적 오차가 포함된다. 동일한 샘플을 가지고 같은 실험실에서 동일한 실험기구를 이용하여 2회 실험했을 때 단백질의 분리결과는 상당한 차이를 발견할 수 있는데 이것은 주로 단백질들의 분리 위치에 많은 변이가 나타나기 때문이다. 만약, 각각 다른 두 실험실에서 전기영동을 수행 한다면 결과는 더욱 더 큰 변이가 나타난다[3]. 이러한 2차원 전기영동의 특성이 자동화를 힘들게 하는 주된 요인이다.

이러한 특성을 고려하여 2차원 전기영동을 자동화시키기 위해서 현재 많은 알고리즘들이 제안되고 있다. 2차원 전기영동의 자동화 단계는 크게 스팟검출과 스팟매칭의 단계로 구분할 수 있다. 스팟검출 단계는 2차원 전기영동의 결과영상에서 배경이 되는 부분과 단백질로 추정되는 스팟을 분리해 내는 과정이며 스팟매칭 단계는 참조 겔과 목적 겔를 1:1로 매칭하여 동일한 단백질쌍을 검출하는 과정이다.

2차원 전기영동 영상 안에서 단백질의 종류 및 특성을 검출하고 식별하는 것은 매우 중요한 작업이다. 영상 안의 단백질 스팟이 어떤 단백질인지 분석하기 위해서 그 스팟을 추출하여 질량 분석 작업이 필요하다. 그러나, 이것은 시간과 노력이 많이 소요되기 때문에 분석 비용이 스팟 수에 따라 기하급수적으로 증가하는 단점이 있다. 따라서 스팟매칭 방법을 이용하여 단백질 분석의 기준이 되는 참조 겔 영상와 분석하고자하는 획득된 목적겔 영상을 비교 분석하여 매칭함으로써 많은 문제점들을 해결할 수 있다.

스팟매칭을 이용한 방법은 그래프를 이용한 방법, 반복점진적 방법, 주변스팟의 유사성을 이용한 방법, 이웃스팟의 위상 유사도를 이용한 방법 등의 많은 연구가 현재까지 활발하게 연구되고 있다.

본 논문에서는 grassfire 스팟매칭 알고리즘[4]에 기반하여 한 개의 발화점이 아닌 다중의 발화점을 이용한 grassifre 알고리즘을 제안한다. 여러 개의 발화점을 선정한 후 실험한 결과 기존의 알고리즘 보다 검출율 및 정확도 측면에서 좋은 성능을 보인다.

Ⅱ. 관련연구

2.1 스팟매칭의 정의

스팟매칭은 참조 겔과 목적 겔의 영상에서 동일한 단백질 스팟쌍을 매칭함으로써 스팟의 정보를 얻어내고 단백질의 종류를 식별할 수 있는 단서를 제공한다. 따라서 두 겔의 차이를 분석하여 특이한 단백질의 발현이나 특정 단백질의 속성, 역할, 변이 등을 추적할 수 있다. 참조 겔과 목적 겔에서 검출한 스팟의 집합을 각각 $P=\{p_1, p_2, \cdots, p_m\}$, $Q=\{q1, q2,\cdots,qn\}$이라 한다. 각각의 스팟은 $p_i=(x_i, y_i)$ 그리고 $q_j=(x_j, y_j)$와 같이 2차원 공간에서 스팟의 중심 좌표로 표현된다. 스팟매칭은 참조 겔의 스팟과 목적 겔의 스팟로 부터 동일한 단백질 스팟을 찾아내어 1:1로 매칭하여 정합쌍의 집합 $M=\{(p_{i1}, q_{j1}), (p_{i2}, q_{j2}), \cdots, (p_{il}, q_{jl})\}$을 찾는 것이다($p_{il}\in P$, $q_{jl}\in Q$, $m\neq n$, $l\leq m$ 또는 $l\leq n$).

2.2 그래프와 이웃스팟의 정의

스팟매칭은 그래프 이론을 이용하여 중심스팟에 대한 이웃스팟을 정의한 후 이웃스팟들이 구성하는 위상들의 유사도를 비교하여 스팟의 매칭여부를 결정한다. 정점의 집합으로 구성된 스팟의 집합에 어떠한 그래프를 적용하

는가는 이웃스팟의 정의에 가장 직접적인 영향을 준다. 그림 1은 점의 집합 V에 5-NNG를 적용한 결과이다. 그래프는 정점의 집합과 간선의 집합으로 정의되는 것으로서 식(1), (2), (3)과 같이 기술될 수 있다.

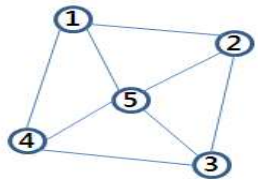

그림 1 5개의 노드로 구성된 그래프의 예

$$V=\{1,2,3,4,5\} \tag{1}$$

$$E=\begin{Bmatrix}(1,2),(1,3),(1,4),(1,5),\\(2,1),(2,3),(2,4),\\(3,1),(3,2),(3,4)(3,5),\\(4,1),(4,2),(4,3),(4,5),\\(5,1),(5,3),(5,4)\end{Bmatrix} \tag{2}$$

$$G=(V,E) \tag{3}$$

이렇게 정점의 집합 V에 특정 그래프가 적용되면 스팟들 사이에 간선이 생성되고 이 간선들이 이웃스팟의 여부를 결정짓게 된다. 이웃스팟의 정의는 식(4)와 같이 기준스팟 v에서 임의의 스팟 u로의 간선이 생성되어 있다면 이웃스팟으로 인정하는 것으로 정의할 수 있다. 또 이웃스팟의 개수를 차수(degree)라고 명명하며 식(5)와 같이 수학적으로 기술한다.

$$N_G(v)=\{u\,|\,vu\in E\} \tag{4}$$

$$\deg_G(v)=|N_G(v)| \tag{5}$$

그림 1에서 정점 5의 이웃스팟을 식(4) 그리고 식(5)와 같이 수학적으로 표현해 보면 식(6) 그리고 식(7)과 같이 기술할 수 있다.

$$N_{3-NNG}(5)=\{1,3,4\} \tag{6}$$

$$\deg_{3-NNG}(5)=|N_{3-NNG}(5)|=3 \tag{7}$$

이웃스팟을 정의하기 위한 그래프 이론은 다양하지만 스팟매칭 문제를 해결을 위해 주로 이용되는 그래프는 Delaunay triangulation, Gabriel graph, Relative Neighbor graph 그리고 k-NNG 등이 있다. 본 논문에서는 참조 논문[5]에 기반하여 이웃스팟을 정의하기 위하여 5-NNG 그래프를 이용한다.

2.3 grassfire에 의한 스팟매칭

grassfire는 잔디에 불이 붙은 후 사방으로 불이 번져가는 모습에서 착안한 알고리즘 방식이다. grassfire에 의한 스팟매칭은 잔디밭에 불을 붙이기 위하여 매칭이 정확히 확인된 한 쌍의 매칭쌍이 필요하다. 이 발화점부터 매칭을 시작하여 잔디밭에 불이 번져가듯이 주변으로 매칭을 확대해 가는 방식으로 매칭을 시도한다. 이렇게 하면 이전의 매칭결과가 다음 매칭에 힌트로 사용되기 때문에 아주 빠르고 정확한 매칭결과를 얻을 수 있다[6].

우선 씨드 스팟쌍은 매칭이 확인된 스팟쌍이므로 “매칭쌍 테이블“에 매칭쌍으로 등록하고 씨드 스팟쌍으로부터 매칭을 수행한다. 그러면 이웃스팟쌍들이 도출되는데 이들을 불이 번져갈 수 있는 잠재 통로로 생각하여 이들을 중심스팟쌍으로 하여 매칭하고 매칭된 결과를 “매칭 정보테이블”에 등록한다. 매칭된 결과는 참조 겔 중심스팟번호, 참조 겔 피벗 스팟번호, 목적 겔 중심스팟 번호, 목적 겔 피벗스팟 번호, 매칭쌍의 개수, 매칭되지 못한 스팟의 수, 정규화된 하우스도르프 거리로 구성된다.

매칭정보테이블에 등록된 매칭쌍 중에서 매칭 유사도가 가장 높고 “매칭상 테이블”에 등록되지 않은 한 개의 매칭쌍을 선정하여 그곳으로 불이 번져가도록 한다. 불이 번져간다는 말은 선정된 한 개의 매칭쌍을 중심스팟으로 매칭을 수행하고 그 이웃스팟쌍을 “매칭 정보테이블”에 등록시킨다는 의미이다. 이미 “매칭 정보테이블”에 매칭결과가 등록되어 경우는 중복으로 매칭결과를 등록하지 않도록 한다. 이러한 과정을 반복하면 스팟매칭이 씨드 스팟쌍 주위로 마치 불이 번져가

듯이 수행되게 되면 모든 스팟쌍의 매칭이 완료되면 불이 번져갈 곳이 없어 스팟매칭이 끝나게 된다.

2.4 다중 발화점에 의한 grassfire에 의한 스팟 매칭

grassfire 스팟매칭 알고리즘에서는 스팟매칭을 시작하는 발화점이 필요한데 기존 grassfire 스팟매칭 알고리즘에서는 한 개의 발화점을 실험자가 수작업으로 선정하여 수행하고 있다. 좋은 발화점의 요건은 가능한 많은 이웃스팟을 가지고 있어야 하고 이웃스팟들의 위상의 유사도가 높은 것이 좋다.

단일 발화점의 경우 수작업으로 선정될 경우 이런한 요건을 만족하지 못할 수 있다. 즉, 나쁜 발화점은 스팟매칭의 결과에도 나쁜 영향을 미치게 된다. 잘못 매칭된 스팟이 다수 존재하거나 참조 겔과 목적 겔의 유사도가 매우 낮을 경우 발화점의 선정을 더욱 중요해진다.

이러한 단일 발화점을 이용한 grassfire 스팟매칭 방법의 단점을 극복하기 위하여 본 논문에서는 다중 발화점을 이용한 grassfire 스팟매칭 방법을 제안한다. 다중 발화점이란 발화점을 복수 개를 이용하는 것이다. 만일 5개의 발화점을 선정하였다면 각각의 발화점을 이용하여 5번의 grassfire 스팟매칭을 수행한다. 그리고 각각의 결과들을 취합하여 최종의 결과를 계산한다.

III. 실험결과 및 분석

3.1 실험 데이터 생성

본 논문에서는 멀티 발화점을 이용한 grassfire 스팟매칭 방법의 개념과 가능성을 입증하기 위하여 실제의 참조 겔과 목적 겔 데이터를 사용하지 않고 시뮬레이션 방법에 의하여 500개의 스팟쌍을 생성하였고 이를 이용하여 실험을 수행하였다.

우선 겔의 크기를 512x512로 가정하였고 500개의 스팟을 가진 참조 겔을 생성한다. 난수를 이용하여 스팟의 x 좌표와 y좌표를 생성하고 생선된 스팟을 먼저 생성된 스팟과 비교한다. 스팟을 겹쳐지거나 거리가 매우 짧은 스팟이 생성하지 않도록 최소거리를 5픽셀로 설정한다. 겹쳐지거나 스팟간의 거리가 5픽셀 이하인 스팟은 폐기하고 이러한 조건을 만족할 때까지 계속 생성한다.

일단 참조 겔이 생성되면 참조 겔을 변형하여 목적 겔을 생성한다. 난수를 이용하여 개별 스팟의 x와 y의 좌표값에 소량의 변위를 추가한다. 이 때 소량의 변위는 정규분포 난수를 이용하여 실제 데이터와 유사한 데이터를 얻도록 한다.

3.2 실험방법

우선 단일 발화점을 이용한 grasfire 스팟매칭을 수행하기 위하여 수작업으로 1개의 씨드 스팟을 먼저 생성하고 다중 발화점을 이용한 grassfire 스팟매칭을 실험하기 위하여 5개의 씨드 스팟을 수동으로 선정하여 각각의 씨드 스팟을 이용함으로써 5번 grassfire 스팟매칭을 차례대로 수행하고 최종적으로 이들 결과를 취합하여 계산한다.

스팟매칭을 위한 데이터는 시뮬레이션 방법으로 생성하였으므로 매칭 여부에 대한 ground truth은 이미 알고 있는 상태이며 임의로 동일한 번호를 부여하여 매칭 여부를 보다 쉽게 확인할 수 있도록 하였다. 즉 참조 겔과 목적 겔이 동일한 번호를 가지고 스팟매칭이 되었다면 이 스팟쌍은 올바르게 매칭된 결과임을 알 수 있도록 하였다.

3.3 실험결과

표 1은 각각의 씨드 스팟에 의한 개별 결과를 통합하여 표시한 결과이다. 스팟매칭쌍은 5개의 씨드 스팟을 이용하여 매칭한 결과 1회 이상 매칭된 모든 쌍을 등록하였다. 씨드 스팟쌍 1은 첫 번째 발화점을 의미한다. “T” 표시는 씨드 스팟쌍을 이용한 grassfire 스팟매칭에서 해당 스팟매칭쌍이 매칭되었음을 나타내며 “F”로 표기된 부분은 해당 스팟매칭쌍이 해당 씨드 스팟쌍 조건하에서 매칭되지 못했음을 의미한다. 스팟매칭 정확도는 해당 스팟매칭쌍이 5개의 씨드 스팟쌍에 의하여 몇 번 매칭되었는지의 백분율을 나타낸다. 표 1은 실제로 500개의 스팟쌍에 대하여

모든 결과를 수록하여야 하지만 지면이 짧은 관계로 결과 중 일부만 표시하였다. 표 2는 단일 발화점을 이용한 경우와 5개의 발화점을 이용한 경우를 비교한 것이다. 검출율은 총 스팟쌍 중에서 실제로 총 검출된 스팟쌍의 비율을 표시한 것이며 스팟매칭 정확도는 매칭된 총 스팟쌍 중에서 올바르게 매칭된 스팟쌍의 비율을 나타낸 것이다.

발화점 5개를 이용한 방법은 검출율과 스팟매칭 정확도가 각각 100%와 99.3%로 단일 발화점을 이용한 경우보다 성능이 개선된 것을 확인할 수 있다. 표 2의 스팻매칭 정확도는 실제 데이터를 통한 실험 상황하에서 계산하여야 하지만 실제 데이터에서는 ground truth를 확보하는 것이 불가능하기 때문에 시뮬레이션 방법에 의하여 500개의 스팟쌍을 생성하였고 이를 이용하여 실험을 수행하였다. 이 같은 방법은 매우 의미있는 것으로 매칭된 결과를 손쉽게 검증할 수 있는 유용한 방법이다.

표 1. 다중 발화점을 이용한 스팟매칭 결과

발화점 / 스팟매칭쌍	씨드 스팟쌍 1	씨드 스팟쌍 2	씨드 스팟쌍 3	씨드 스팟쌍 4	씨드 스팟쌍 5	총 매칭 회수	스팟매칭 정확도
(301,301)	T	T	T	T	T	5	100%
(335,335)	T	T	T	T	T	5	100%
(225,225)	T	T	T	T	T	5	100%
(393,393)	T	T	T	T	T	5	100%
(15,15)	T	T	T	T	T	5	100%
	:						
(94,94)	T	T	T	F	T	4	80%
(165,165)	T	T	F	T	T	4	80%
(96,96)	T	F	T	T	T	4	80%
(463,463)	T	F	T	T	T	4	80%
(21,21)	T	T	T	T	F	4	80%
(26,26)	T	T	T	F	T	4	80%
	:						
(198,198)	T	T	T	F	F	3	60%
(111,111)	F	T	F	T	T	3	60%

표 2. 단일 발화점을 이용한 방법과 발화점 5개를 이용한 스팟매칭 방법의 비교

항 목 / 방 법	검 출 율	스팟매칭 정확도
발화점 1개를 이용한 방법	98%	96.5%
발화점 5개를 이용한 방법	100%	99.3%

IV. 결론

본 논문에서는 기존의 단일 발화점을 이용한 grassfire 스팟매칭을 보완할 수 있는 다중 발화점을 이용한 grassfire 스팟매칭 방법에 대하여 제안하고자 한다. 이 방법은 단일 발화점을 이용한 방법보다 검출율과 스팟매칭 정확도의 측면에서 향상된 성능을 나타낸다.

다중 발화점을 이용한 grassfire 스팟매칭 방법은 잘못된 단일 발화점에 의한 스팟매칭의 실패를 보완해 주는 아주 유용한 수단이며 다중 발화점에서 스팟매칭 신뢰도의 계산방법은 스팟매칭의 정확도를 객관적으로 진단하여 더욱 스팟매칭의 확률을 높이는 아주 유용한 방법이다.

제안한 스팟매칭 방법은 참조 겔과 목적 겔의 유사도가 매우 낮거나 아웃라이어가 다수 존재하는 경우에 더욱 더 큰 성능향상을 기대할 것으로 생각되는데 이를 입증하기 위하여 향후에는 유사도가 매우 낮은 경우와 아웃라이어가 매우 많은 경우에서 다중 발화점을 이용한 방법을 이용하여 실험하는 것이 필요하다. 또한, 발화점을 자동으로 선정하여 가장 좋은 매칭성능을 확보할 수 있는 체계적인 연구가 수행되어야 한다.

본 논문에서는 합성데이터를 이용하여 성능을 평가하였는데 왜곡의 생성에 있어서 실제 단백질 전기영동에서 왜곡과는 다소 차이가 있을 수 있기 때문에 대량의 실제 단백질 전기영동 데이터를 통하여 성능을 계속적으로 증명하는 노력도 필요할 것이다. 또한 기존의 상용 스팟매칭 소프트웨어에 다중 발화점을 이용한 grassfire 스팟매칭 알고리즘을 적용하여 단백체학의 병목단계인 스팟매칭 과정을 효율적으로 개선하여 단백체학이 더욱 발전하도록 노력하는 것도 우선적으로 수행하여야 할 과제이다.

참 고 문 헌

[1] P. H. O'Farrell, "High Resolutin Two-dimensional Electrophoressis of Proteins," Journal of Biological Chemistry, Vol. 250, No. 10, pp. 4007-4021, May 1975.

[2] K. Kaczmarek, B. Walczak, S. de Jong and B. G. M. Vandeginste, "Matching 2D Gel Electrophoresis Images," Journal of Chemical Information and Computer Science, Vol. 43, pp. 978-986, 2003.

[3] K. C. Dukka Bahadur, Tatsyua Akutsu, Etsuji Tomita, Tomokazu Seki, and Asao Fujiyama, "Point Matching under Non-uniform Distortions and Protein Side Chain Packing Based on An Efficient Maximum Clique Algorithm," Genome Informatics Vol. 13, pp. 143-152, 2002.

[4] 류윤규, 2차원단백질 전기영동영상에서 Grassfire 기법을 이용하는 강건한 스팟 매칭 알고리즘, 영남대학교 대학원 박사학위논문, 2013.

[5] Chan-Myeong Han, Dae-Seong Jeoune, Hwoi-Won Kim, and Young-woo Yoon, "A Spot Matching Method using Topological Patterns of Neighbor Spots in 2-DE," Proceedings of the 7th International Conference on Information Security and Assurance (ISA2013), ASTL Vol. 21, pp. 156-159, SERSC, Cebu, Philippines, April 2013.

[6] Yun-Kyoo Ryoo, Chan-Myeong Han, Ja-Hyo Ku, Dae-Seong Jeoune, and Young-Woo Yoon, "Grassfire Spot Matching Algorithm in 2-DE" International Journal of Bio-Science and Bio-Technology, Vol. 5, No. 4, pp. 162-174, August 2013.