DNA Microarray Data Analysis & Extracting Biological Meanings

이 성 근
ISTECH Inc.

Ⅰ. 들어가는 말

DNA 칩을 생물학 또는 의학의 도구로 사용함에 있어서 사용자를 어렵게 하는 점은, 고가의 비용 문제를 일단 접어 놓더라도, DNA 칩 실험이 다른 분자생물학 실험에 비해 여러 면에서 까다로울 수 있다는 점과 여전히 의생물학자에게는 칩 분석이 어렵게 느껴진다는 것이다. 하지만 실험과 분석에 얽힌 여러 문제들은 customized cDNA 칩의 보급, 정형화된 QC를 갖춘 상용 칩의 등장, 다양한 공용/상용 분석 소프트웨어의 등장, 기본 분석 방법의 보급 등으로 DNA 칩 초창기보다 많이 개선되고 있다.

해외의 경우 이런 하드웨어의 개선과 더불어 DNA 칩 데이터의 분석 수준이 점차 향상되고 있는데, 국내에는 아직 이에 미치지 못하는 경향이 있다. 이 글은 DNA 칩 분석의 전반적인 과정을 개략적으로 살펴보아 분석 방법의 뼈대를 이해할 수 있는데 초점을 두었으며, 이를 통해 보다 세밀한 연구를 위한 계기가 되었으면 한다.

Ⅱ. 그룹간 비교분석(Class Comparison)

보통 유의유전자(Differentially Expressed Gene: DEG) 추출이라 불리는 과정으로, 분석 대상 그룹(예를 들면, 대조군과 실험군 비교) 사이에서 통계적으로 의미 있는 발현 차이를 보이는 유전자를 찾기 위한 분석 과정이다. 분석 대상 그룹의 개수에 따라 2-class/multiclass test를 적용하며, 분석 데이터의 통계적 분포에 따른 가정에 따라 母數(parametric)/非母數(nonparametric) test를 적용하게 된다.

1. Fold change

DNA 칩 데이터 분석 초창기에 주로 사용된 방법이지만, 적용의 단순성과 결과 해석의 용이성이라는 장점으로 최근까지도 많이 사용되는 방법이다. 각 유전자에 대해 기준 샘플(control/reference sample)의 발현 수치와 처리 샘플(treatment sample) 발현 수치의 비(ratio) 값을 계산하여 기준 샘플보다 처리 샘플에서 상대적으로 어느 정도 발현 했느냐를 보는 것이다. 보통 비 값 자체를 fold change라고 하지만, 때로는 밑이 2인 로그($\log_2$) 값이 취해진 형태를 fold change라고도 한다.

Fold change는 일정 비(ratio) 이상의 유의유전자를 고르기 위해서 임계치(threshold)를 설정해야 하는데 일반적으로 2 fold가 기준이 되며, 데이터에 따라 1.5 fold로 낮추거나 4 fold 이상으로 올리기도 한다. 하지만, 이러한 일괄적인 임계치 적용은 문제가 될 수 있다. 예를 들어 2-fold를 적용했을 경우, 상대적으로 발현 수치가 낮은 영역에서는 해당 조건을 만족하는 유전자들이 많은 반면에, 발현 수치가 높은 영역에서는 2 fold 조건이 만족되기 어려운 것이다. 또한 fold change는 그룹간 비교를 할 때 유전자 발현수치의 그룹 내 변이 등의 통계적인 유의성이 고려되지 않는다. 예를 들어, 대조군(control class) 샘플 3개와 실험군(experiment class) 샘플 3개가 주어졌을 때, fold change 방법으로 유의유전자를 구하기 위해 각 유전자에 대해 총 9가지 경우의 fold change값을 구하여 평균을 취하게 된다. 하지만 통계적 신뢰 수준에 대한 언급 없이 이러한 평균값만으로는 9가지 값들을

대표한다고 말하기 힘들다. 9가지 값들 중에서 한 두 개의 이상치(outlier)만으로 평균값이 왜곡될 수 있기 때문이다. 이러한 이유로 fold change는 통계적 방법이라기 보다는 경험적 방법이라 할 수 있다.

2. Two-sample(unpaired) t-test

t-test는 fold change 방법과 더불어 유의유전자 분석에 널리 쓰이는 방법으로, 통계적인 유의성을 부여한다는 측면에서 fold change의 개선된 방법이다. 사실 t-test 수식(아래 수식 참조)을 잘 들여다 보면 fold change와의 연관성을 알 수 있는데, 본질적으로 t-test는(fold change가 대변하는) 그룹 사이의 평균 발현 차이(수식의 분자에 해당)를 각 그룹내의 변이(수식의 분모에 해당)로 나눈 것이라 할 수 있다. 따라서 각 그룹 내 편차가 작을수록 또한 두 그룹 사이의 평균 발현 차이가 클수록 t-score의 절대값이 커지게 되며, t-score 절대값이 커질수록 통계적 유의성이 더욱 보장된다. 이러한 통계적 유의성은 p-value를 통해 알 수 있는데, p-value를 구하는 방식은 데이터에 대한 가정에 따라 두 가지로 나눌 수 있다. 데이터가 정규분포를 따른다고 가정할 수 있는 경우는 보통 Welch approximation 방법이 사용되며, 별다른 확률 분포를 가정하지 않는 경우는 permutation test가 주로 쓰인다.

단, 일반적으로 t-test는 각 그룹 내 replication이 적어도 5~6개 이상일 때 적용해야 좋은 결과를 얻을 수 있다는 점에 유의해야 한다.

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

→ approximately t-distributed with d.o.f, v

$$\text{where } v = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{(S_1^2/n_1)^2/(n_1-1) + (S_2^2/n_2)^2/(n_2-1)}$$

3. Volcano plot

화산에서 용암을 분출하는 것처럼 보인다 하여 붙여진 이름이며, Fold change와 t-test 방법으로 추출된 유전자들의 분포를 한눈에 알 수 있는 유용한 시각화 방법이다(그림 1 참조). 예를 들어 2-fold 이상의 유의

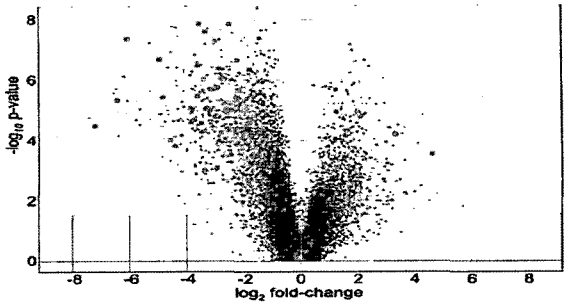

그림 1. Volcano Plot 예시 화면

유전자들 중에서 통계적으로도 유의한(P-value가 작은) 유전자들이 관심사가 되는데, 이 유전자들을 추출하려면 아래 그림에서 좌우상단 모서리에 있는 유전자들을 고르면 될 것이다(그림에서 회색 부분).

4. Significance Analysis of Microarrays(SAM)

SAM은 t-test의 약점을 보완하고자 Stanford 통계 그룹(V.G. Tusher et al, 2001)이 개발한 방법이다. 사실 유의유전자를 찾기 위한 대부분의 母數 통계적 방법(parametric methods)들이 t-test의 변형된 형태인데, 그 중에서도 SAM은 가장 널리 알려져 있다. SAM 방법에 의한 t-test 개선 효과는 크게 두 가지로 생각할 수 있는데, 첫째는 t-test의 분모에 fudge factor라고 불리는 상수를 도입함으로써 낮은 발현수치 영역에서 생기는 false positive를 감소시키고자 한 것이며, 둘째는 permutation test를 이용해 False Discovery Rate(FDR)을 산출하고자 한 것이다. FDR은 선별된 유전자 목록에 존재하는 false positive 비율을 예측한 것이다.

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{S(i) + S_0} \qquad s_0 : \text{fudge factor}$$

그림 2(좌)는 Northern blot으로 검증해 보았을 때, fold change로 뽑힌 유의유전자보다 SAM 방법으로 뽑힌 유의유전자의 발현 양상이 일관된 결과를 보였다는 것이며, 그림 2(우)는 fold change로 뽑힌 유의유전자의 false positive 비율이 SAM의 결과보다 상당히 높다는 것을 보여주고 있다.

SAM 방법 이후에도 많은 유의유전자 분석 방법들

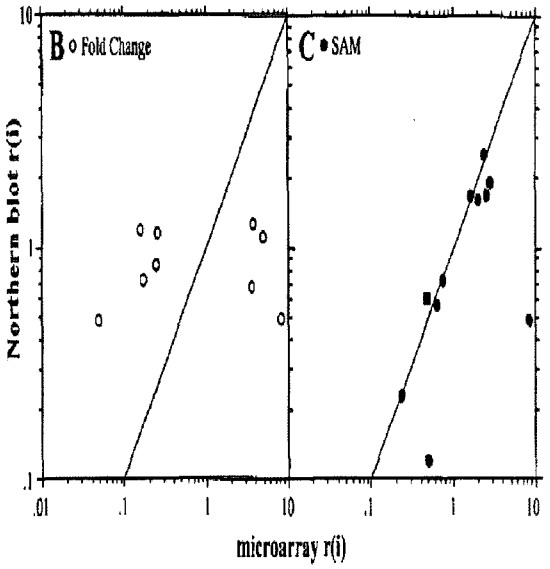

Table 1. Comparison of methods for identifying changes in gene expression

Parameter	Number falsely significant	Number called significant	FDR
SAM			
Δ = 0.4	134.9	288	47%
Δ = 0.5	78.1	192	41%
Δ = 0.6	56.1	162	35%
Δ = 0.9	19.1	80	24%
Δ = 1.2	8.4	46	18%
Δ = 1.2; R = 1.5	4.5	34	12%
Fold change			
R = 2.0	283.1	348	81%
R = 2.5	137.8	169	82%
R = 3.0	76.8	99	78%
R = 3.5	46.7	64	73%
R = 4.0	29.3	35	84%
Pairwise fold change			
R = 1.2	245.6	355	69%
R = 1.3	155.4	220	71%
R = 1.5	76.2	118	65%
R = 1.7	44.8	70	64%
R = 2.0	22.8	38	60%

To increase the stringency for calling significant changes in gene expression, parameters for each method (Δ and R) were increased, as described in the text. The false discovery rate (FDR) was defined as the percentage of falsely significant genes compared to the genes called significant.

그림 2. fold change와 SAM의 비교(좌); FDR(우)(Tusher et al., 2001)

이 개발되고 있는데, 특히 replication이 적을 때(주로 2-3장) 사용할 수 있는 유용한 통계 방법들이 주목을 받고 있다(Jain et al., 2003). 또한 Pan의 논문은 유의 유전자 분석 방법들을 비교한 좋은 리뷰이니, 칩 데이터 분석을 제대로 느끼고자 하는 분들에게 필독을 권한다.

5. Analysis of Variance(ANOVA)

유의유전자 추출을 위한 그룹 비교(class comparison) 실험 디자인을 할 때, 비교하고자 하는 그룹이 항상 2개만 있는 것은 아니다. 비교 그룹이 2개인 경우에는, 예를 들어, t-test 방법을 적용하면 되지만, 비교 그룹이 3개 이상인 경우에는 어떻게 할 것인가? 이에 대해 2가지의 해결 방법이 제시될 수 있는데, 첫째는 t-test를 모든 가능한 2개의 그룹 쌍(pair)에 적용하는 것이며, 둘째는 ANOVA 방법을 모든 그룹에 한번에 적용하는 것이다.

비교해야 할 그룹이 7개라고 해보자. 그럼 첫째 방법으로 분석해야 할 쌍은 21개 쌍이며, 각 쌍에 대해 많은 수의 유의유전자들이 포함된 목록이 나올 것이다. 7개 그룹 사이에서 통계적으로 의미 있는 발현 차이를 보이는 유전자를 추출하기 위한 것이라면 위 접근 방법은 그다지 좋은 방법이 아닐 것이다. 더구나 21개 각각의 t-test에 대해 p=0.05로 신뢰 수준을 설정했다 하더라도 대략 21*0.05 ≅1개의 테스트 결과는 false positive일 거라고 예상할 수 있는 것이다.

따라서, 비교 분석 그룹이 3개 이상인 경우에는 사용자가 설정한 신뢰 수준(statistical significance level)이 보장되면서 한번에 분석할 수 있는 ANOVA등의 multiclass 방법이 유용한 것이다.

Ⅲ. 군집 분석(Class Discovery)

Clustering이라 불리는 과정을 의미한다. 발현 패턴이 유사한 정도에 따라 유전자끼리 또는 샘플끼리 묶어주는 과정이며 전자의 경우는 gene clustering, 후자의 경우는 sample clustering이라 한다. Gene clustering의 경우는 유전자의 기능 탐색 등에 이용되며, sample clustering의 경우는 의료 분야의 진단이나 예후 예측을 위해 주로 연구되고 있다.

1. Hierarchical Clustering(HC)

외부의 자극에 대한 효모의 분자유전학적 반응을 DNA Chip을 통해 조사한 Eisen et al.의 논문 이후에 널리 쓰이게 된 방법이며, 특히 논문과 함께 무료로 배포된 'Cluster'와 'TreeView' 소프트웨어는 지금까

지도 가장 많이 쓰이는 DNA 칩 소프트웨어중의 하나이다.

원래 HC 방법은 분할 방식(divisive approach)과 융합 방식(agglomerative approach)이 있는데, DNA 칩 분석에는 융합 방식이 주로 쓰인다. 전자의 경우는 전체 집합에서 점점 세부 그룹으로 나누어 가는 것이고, 후자는 개별 대상들을 가까운 것끼리 묶어서 점차 큰 그룹으로 만들어 나가는 것이다. 이런 이유로 전자를 top-down 방식, 후자를 bottom-up 방식이라고도 일컫는다.

HC 방법으로 gene clustering 하는 것은 다음과 같다. 예를 들어 1,000개의 유전자가 있다면, 우선 각각의 유전자를 하나의 그룹(cluster)으로 인식하게 된다.

제1step : 1,000개의 그룹에서 발현 패턴이 가장 유사한 두 개의 그룹을 찾아서 이들을 하나의 그룹(cluster)으로 묶는다. 이 과정으로 999개의 그룹이 남는다.

제2step : 999개의 그룹 중에서 발현 패턴이 가장 유사한 두 개의 그룹을 찾아서 이들을 하나의 그룹(cluster)으로 묶는다. 이 과정으로 998개의 그룹이 남는다.

제3step : 998개의 그룹 중에서 발현 패턴이 가장 유사한 두 개의 그룹을 찾아서 이들을 하나의 그룹(cluster)으로 묶는다. 이 과정으로 997개의 그룹이 남는다.

이런 식으로 제999step까지 시행하면 최종 하나의 클러스터가 남게 된다(그림 3 참조).

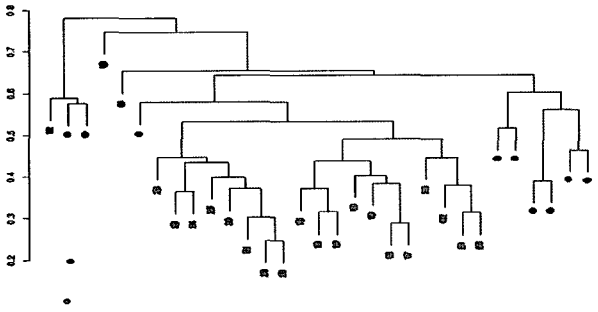

그림 3. 수형도(dendrogram)

이렇게 묶여지는 것을 tree 형태로 보여준 그림이 바로 수형도(dendrogram)이다. 그런데 위의 알고리즘에서 하나 주목해야 할 부분이 있다. 바로 두 그룹 사이의 가까운 정도, 즉 거리를 어떻게 정의하느냐이다. 이 정의에 따라 두 그룹의 연결 방식(linkage type)이 정해지는데, single linkage, complete linkage, average linkage, Ward's method 등이 주로 쓰인다.

HC 방법의 장점은 의생물학자들에게 익숙한 시각화 이외에도, 소프트웨어에서 흔히 동반되는 번잡한 패러미터(parameter)를 직접 입력할 필요가 없다는 것이다. 즉, K-means나 SOM처럼 클러스터 개수를 미리 예상하여 입력할 필요가 없다. 또한 수형도(dendrogram)에서 사용자가 원하는 대로 클러스터의 크기와 개수를 정할 수 있다는 장점이 있다. HC는 여러 장점 못지않게 단점 역시 부각되어 왔는데, 첫째로 각 step에서 한번 그룹으로 묶여지면 이후 step에서도 refinement 과정을 거치지 않고 계속 유지된다는 것이다. 이런 이유로 각 그룹(cluster)의 묶인 정도(tightness)가 K-means 등의 방법보다는 다소 덜한 편이다. 둘째는 분석 데이터가 hierarchical structure라는 '딱딱한' 구조를 가지지 않는 경우에는 다른 방법보다 clustering 결과가 좋지 않을 수 있다는 것이다.

2. K-means

반복 계산 과정을 통해 최적의 K개 그룹(cluster)을 찾는 과정이다. 각 그룹의 중심(centroid; 수학적으로 평균벡터를 의미함)에 대해 각 그룹의 구성원들(gene clustering이라면 각 유전자들)이 얼마나 밀집해 있는지를 평가 척도로 하여 일정 수준에 이를 때까지 반복 과정을 계속 수행한다.

K-means의 장점은 반복 과정을 통해 수학적인 최적화를 거치므로 결과적으로 나오는 그룹들이 비교적 잘 밀집해 있다는 것이다. 하지만, 클러스터 개수 K를 사용자가 미리 예상해서 입력해야 한다는 것과 처음 K개의 centroid를 주는 방법에 따라 결과가 많이 달라질 수도 있다는 점 등이 단점으로 지적될 수 있다.

3. Self Organizing Map(SOM)

SOM은 전산학 분야에서 비교적 최근에 개발되어

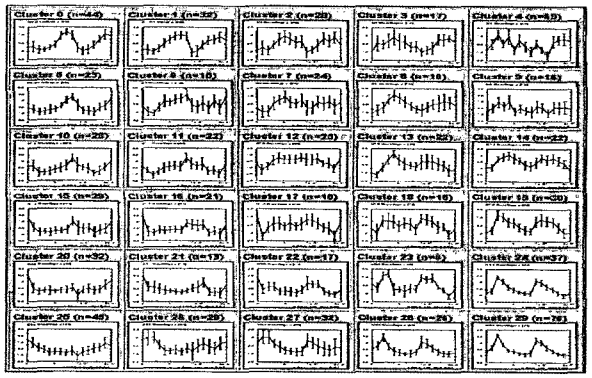

그림 4. SOM의 2D profile

다른 분야에 널리 쓰이던 방법인데, DNA 칩 분석에서는 Tamayo et al., Golub et al. 등의 발표 이후에 보편적으로 쓰이게 되었다. SOM의 가장 큰 특징은 고차원 데이터를 저차원(주로 2차원)으로 변환시켜 시각적으로 이해할 수 있도록 해주는 것인데, 이러한 특징이 고차원 DNA칩 데이터를 분석하는데 도움을 주었다. 또한 SOM은 K-means를 일반화한 형태로 여길 수도 있는데, 사용자에게 패러미터(parameter)를 조정할 수 있는 권한을 부여하여 원하는 형태의 결과를 산출하게 해준다. 하지만, 이러한 점이 의생물학자들에게는 오히려 성가신 일이 될 수도 있을 것이다.

SOM 시각화의 장점은 비슷한 패턴을 보이는 그룹(cluster)들을 이웃에 배치시키는 것이다. 그림 4에서 각각의 그룹(cluster) 주위를 보면, 해당 그룹과 비슷한 패턴의 그룹들이 주변에 배치되어 있는 것을 볼 수 있을 것이다.

Ⅳ. 샘플 판별 분석(Sample Class Prediction)

DNA 칩은 실험만 잘되면 결과는 별다른 수고 없이 기본 과정만으로 쉽게 해석될 수 있다고 생각하시는 분들이 있다. 굳이 비유를 하자면, 음식 재료만 좋으면 음식이 맛있을 것이란 얘기이다. 하지만 좋은 재료에 좋은 요리사의 훌륭한 요리 솜씨가 곁들어지게 된다면, 재료의 풍미가 살면서 한층 맛깔스러운 요리가 될 것이다. DNA 칩의 경우도 마찬가지다. 세심한 분석자의 손길이 닿아야 하는 것이다. 그 예를 잘 보여주는 것이 일반적으로 classification으로 불리는 샘플 판별 분석 과정이다. 그림 5는 같은 데이터에 대해 분석을 어떻게 하느냐에 따라 결과가 얼마나 달라질 수 있는가를 명확히 보여준다.

샘플 판별 분석의 최종 목표는 보다 적은 수의 유전자로 보다 정확한 판별 결과를 얻는 것이다. 이를 위해 판별하고자 하는 그룹의 특징을 잘 나타내는 유전자를 골라야 하며(gene selection 과정), 샘플을 잘 판별하여야 하고(classifier 선택 과정), 마지막으로 제일 중요하다고 할 수 있는 신뢰할 만한 판별 정확도를 얻는 것이다(generalization error estimation 과정).

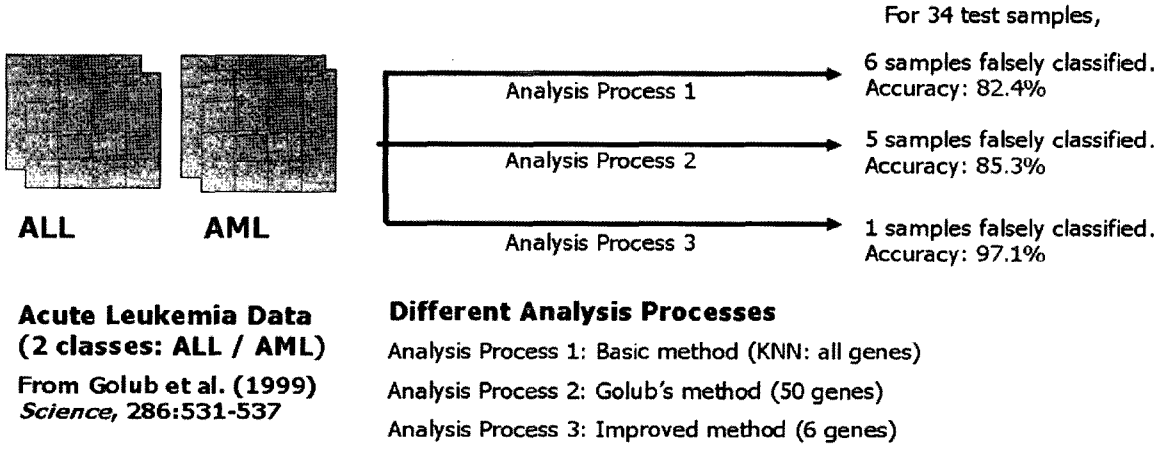

그림 5. 분석과정에 따른 결과의 차이

1. 판별 유전자 추출(Gene selection)

간단히 말하면, 각 그룹을 잘 구별할 수 있는 또는 각 그룹을 특징적으로 나타낼 수 있는 최상의 유전자를 찾아내는 것이다. 보통 DNA칩에는 수 천 내지 수 만개의 유전자 발현 수치가 주어지는데 이 중에서 수십 내지 수백 개, 또는 단 몇 개의 마커 유전자를 찾아내는 것이 이 과정의 목표가 된다.

판별 유전자(discriminatory gene) 추출 방법은 크게 두 가지로 나눌 수 있는데, 개별 유전자 계산 방식(univariate approach)과 복합 유전자 계산 방식(multivariate approach)이 있다. 전자는 각 개별 유전자의 판별 능력을 독립적으로 계산하여 이를 비교해 가장 높은 판별 능력을 지니는 유전자를 선택하는 방식이며, 후자는 유전자 사이의 상관 관계를 고려하여 판별 능력이 높은 유전자 그룹을 선택하는 방식이다. 개별 유전자 계산 방식은 계산 실행 시간이 짧으면서도 효과적인 판별 결과를 기대할 수 있기 때문에 현재 판별 분석에서 주로 쓰이는 방법이다. 하지만 유전자 사이의 상호관계를 고려하지 않으므로 이를 보완하기 위해서 복합 유전자 방식을 채택하기도 한다. 복합 유전자 방식으로는 PCA, SVD 등의 차원 감소 방법이 주로 쓰인다.

2. 판별 알고리즘(classifier)

판별 유전자가 정해지면 이를 바탕으로 샘플을 판별해야 할 것이다. 이렇게 직접 샘플을 판별하는 역할을 맡는 것이 판별 알고리즘이며, Fisher's Linear Discriminant Analysis(FLDA) 등의 전통적으로 많이 쓰이던 통계 판별 방법부터 최근의 Support Vector Machine(SVM), 인공신경망(artificial neural network)를 비롯한 기계학습법에 이르기까지 다양한 방법들이 시도되고 있다.

판별 알고리즘의 역할을 그림 6에서 간단히 이해해보자. 빨간색 동그라미들이 그룹 1의 구성원(샘플)들이고 검은색 사각형들이 그룹 2의 구성원(샘플)들이다. 이제 하나의 직선을 그어서 두 그룹의 경계를 설정하고자 한다. 이 설정된 경계에 따라 미래의 샘플들이 판별될 것이다. 그럼 어떻게 직선을 그어야

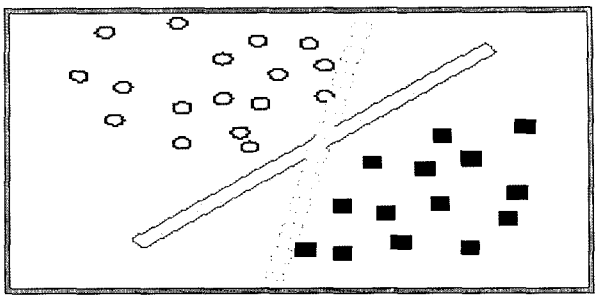

그림 6. 판별 알고리즘의 역할

두 그룹의 영역을 제대로 나누었다고 할 수 있을 까? 점선과 실선 중에서 어떤 것이 그룹 1과 2를 구별하는 좀 더 나은 경계선일까? 이런 질문들에 답하고자 하는 것이 판별 알고리즘, 즉 판별기(classifier)라고 생각하면 된다.

3. 일반 정확도 측정(Generalization Error Estimation)

판별을 실질적으로 수행하는 부분은 아니지만 믿을만한 정확도 예측을 위한 평가 척도가 된다는 측면에서, 판별 분석에서 가장 중요하다고 할 수 있는 부분이다. 판별 분석의 핵심은 판별 유전자와 선택한 판별 방법으로 과연 어느 정도의 정확도를 얻을 수 있느냐이기 때문이다. 특히 샘플 판별 분석의 주된 응용 분야가 진단이나 예후 예측 등의 의료분야인 만큼 신뢰성 있는 일반 정확도 계산은 매우 중요하다 할 것이다. 여기서 일반 정확도란 현재 주어진 데이터에만 국한되는 정확도가 아니라, 이 데이터가 속한 모집단에 대한 정확도를 일컫는다. 대체로 논문에 실린 소량의 DNA 칩 데이터에 대해 높은 정확도로 보고된 방법들을 이후 동일 성격의 다른 독립적인 데이터로 검증해보면 원래 보고된 값보다 낮은 정확도를 보이는 경우가 많다. DNA 칩 데이터의 경우, 분석 대상 샘플 수가 적기 때문에, 이러한 적은 수의 샘플로 신뢰할만한 일반 정확도를 얻는다는 것이 쉽지 않기 때문이다. 그래서, 이런 상황에 맞는 적절한 정확도 예측 방법이 요구되는 것이다.

그림 7의 그래프는 적절한 정확도 예측 방법이 적용되지 않으면 일반 정확도가 부풀려질 수도 있다는 것을 보여준다. Ambroise et al.는 일반적으로 많이 적

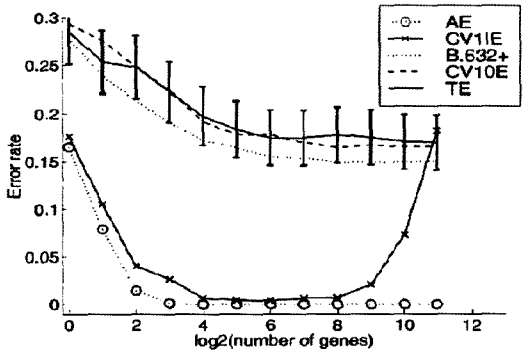

Fig. 1. Error rates of the SVM rule with RFE procedure averaged over 50 random splits of the 62 colon tissue samples into training and test subsets of 31 samples each. TE, test error.

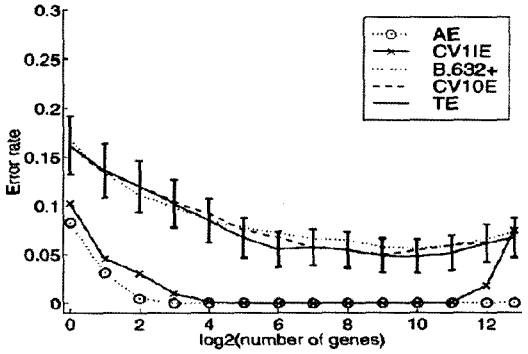

Fig. 2. Error rates of the SVM rule with RFE procedure averaged over 50 random splits of the 72 leukemia tissue samples into training and test subsets of 38 and 34 samples, respectively. TE, test error.

그림 7. 일반 정확도 측정의 중요성

용하는 leave-one-out cross validation(LOOCV)의 external validation과 internal validation의 차이를 언급하였으며, 이를 보완하기 위한 Bootstrap, 10-fold CV 등의 방법을 비교하였다(그래프 참조).

V. 유전자 그룹의 생물학적 해석

지금까지는 유전자의 발현 수치를 다루는 수학적, 통계적 분석 작업에 대해 얘기하였다. 수치만을 다루는 과정이라 자연스럽게 수학, 통계, 전산학 분야에 관련된 내용이었지만, DNA 칩도 결국은 생체 시스템을 이해하기 위한 하나의 도구일 뿐이므로 최종 결과는 생물학적으로 의미가 있어야 한다. 앞에서 다룬 세 가지의 분석 과정, 즉 유의유전자 추출(DEG finding), 군집 분석(clustering), 샘플 판별 분석(sample classification)에서 나오는 산물들은 유전자 목록 또는 유전자 그룹이다. 따라서 이들에 대한 생물학적 해석이야말로 중요한 분석 과정의 하나이다.

유전자를 생물학적으로 해석하기 위해서는 관련 정보가 있어야 한다. 이 정보를 얻기 위해서 PubMed 등을 통해 직접 문헌 정보를 찾아 볼 수도 있고, 생물학 데이터베이스에서 유전자 관련 주석 정보를 얻을 수도 있다. 이런 정보를 취합하여 주어진 유전자 그룹의 의미를 파악하고자 하는 것이 이 과정에서 수행되는 일이다.

유전자들에 대한 생물학적 해석은 개별 유전자 단위로 주석 정보를 참조하는 일차원적 해석과, 유전자 그룹을 하나로 묶어서 그룹 내 유전자들 사이의 관계를 고려하여 집단 전체에 생물학적 의미를 부여하는 고차원적 해석이 있다. 전자의 경우는 각 유전자에 대해, 염색체상의 위치, 관련 경로(pathway) 정보, 알려진 기능 등의 정보를 취합, 정리하는 것이며, 후자의 경우는 그렇게 모은 정보를 활용하여 주어진 그룹 내 유전자 사이의 공통성을 찾아내는 것이다. 이런 공통성을 찾기 위해서는 각 유전자에 대한 주석 정보가 일관적인 논리 체계를 갖춘 용어에 의해 기술되어야 할 것이다. 최근 Gene Ontology가 이런 역할을 해내고 있다.

1. Gene Ontology(GO)

기존의 각 생물학 데이터베이스는 주석을 어떤 식의 용어로 할 것인가에 대해 나름대로의 규칙을 정해 놓았었다. 하지만, 이러한 용어 체계가 각 데이터베이스에 따라 또는 종(species)에 따라 제각각이라는 게 문제였다. 이러한 생물학 용어의 다변성(heterogeneity)으로 인해 나타나는 정보 공유의 비효율성을 지양하고자 출범한 것이 Gene Ontology Consortium이다.

GO의 장점은 잘 갖추어진 체계성에 있다. 첫째, GO 용어 사이의 관계가 잘 정립되어 있고(Directed Acyclic Graph; is a, part-of relationships), 둘째, Biological Process(BP), Molecular Function(MF), Cellular

Component(CC)의 세 가지 특징 별로 GO 용어가 구별되어 있다. 이러한 체계성으로 인해 현재 대부분의 생물학 데이터베이스에서 GO 주석을 제공하고 있으며, GO 주석 정보를 활용하여 유전자 그룹의 생물학적 해석을 시도하는 노력이 점차 증가되고 있다.

2. 주석 정보를 활용한 데이터 마이닝

주석 정보를 활용하여 주어진 유전자 그룹에 내재된 생물학적 연관성을 찾아내는 분석 과정이다. 통계적인 유의성을 통해 찾는 방법(Al-Shahrour et al. 2004)과 GO 구조를 이용한 패턴 인식 방법(Lee et al. 2004) 등이 널리 쓰이고 있다. 예를 들어, 그림 8에서 보듯이 군집 분석(clustering analysis)의 결과를 그룹(cluster)별로 보기 쉬운 tree형태로 생물학적으로 해석, 요약한 것을 볼 수 있다(Lee et al. 2004).

Ⅵ. 맺음말

바이오칩 기술은 생명공학기술의 새로운 주류로 최근 급속하게 떠오르고 있다. 바이오칩 중에서도 선두주자격이라 할 수 있는 DNA 칩은 이제 태동기를 벗어나 점차 성숙기의 길을 향해 가고 있다. 이러한 칩 제작 기술의 발달과 더불어 칩 분석 방법도 지속적으로 개선되고 있다. 학문적인 관점에서 해결해야 할 문제들이 아직 남아 있기는 하지만, 주요 실험디자인에 대한 기본 분석의 뼈대는 잘 갖추어져 있다. 사실 DNA 칩 데이터를 처음으로 대면하는 의생물학자들은 적지않게 당황하게 된다. 하지만 주위을 둘러보면 칩데이터 분석을 도와줄 수 있는 환경이 많이 개선되었다. 이를 통해 데이터와 자주 접하다 보면, DNA칩이라는 최신 기술을 통한 모래속의 진주알 찾기가 그리 어렵지 만은 아닐거라 여겨진다.

Ⅶ. 참고 문헌

1. V.G. Tusher et al.(2001) Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *Proc. Natl. Acad. Sci.* USA, **98**, 5116-5121.
2. N. Jain, J. Thatte, T. Braciale, K. Ley, M. O'Connell,

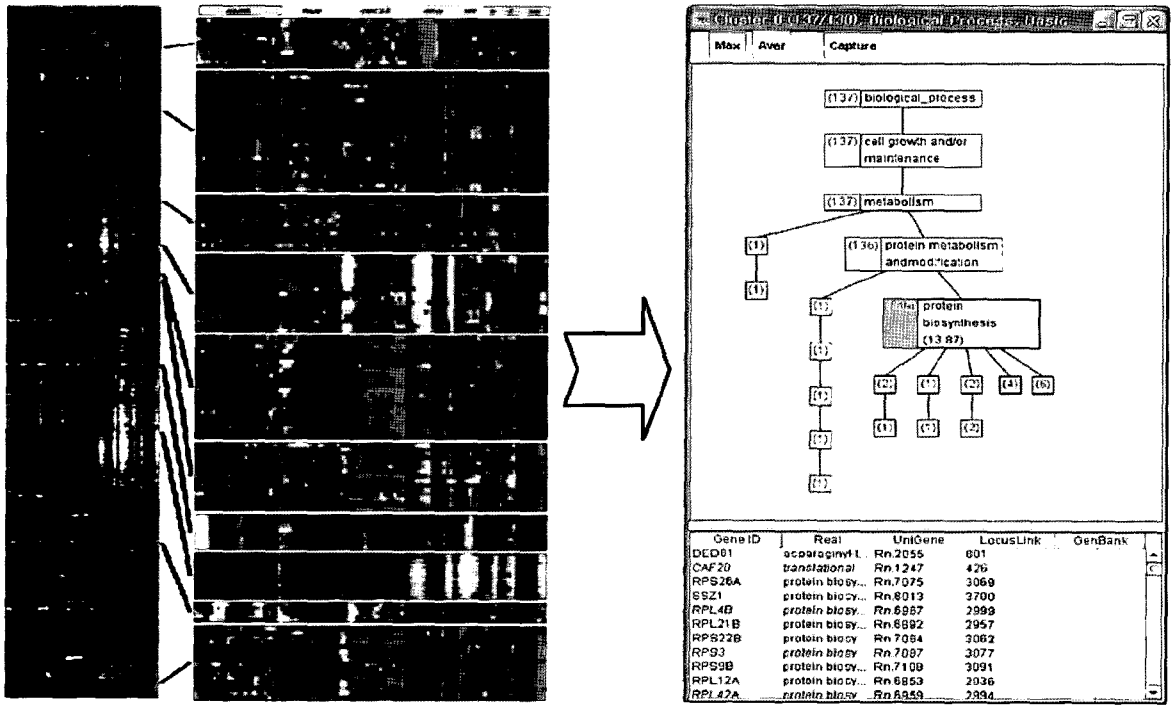

그림 8. Gene Ontology를 이용한 유전자 그룹의생물학적 해석

J.K. Lee(2003) Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics*, **19**(15):1945-1951.

3. W. Pan(2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, **18**(4):546-554.
4. M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein(1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci.* USA, **95**, 14863-14868.
5. P. Tamayo *et al.*(1999) Interpreting patterns of gene expression with SOMs - methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci.* USA, **96**, 2907-2912.
6. T.R. Golub *et al.*(1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-537.
7. C. Ambroise and G.J. McLachlan(2002) Selection bias in gene extraction on the basis of microarray gene expression data. *Proc. Natl Acad. Sci.* USA. **99**, 6562-6566.
8. The Gene Ontology Consortium(2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**:25-29.
9. F. Al-Shahrour, R. Diaz-Uriarte, J. Dopazo(2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**:578-580.
10. S.G. Lee, J.U. Hur, Y.S. Kim(2004) A graph-theoretic modeling on GO space for biological interpretation of gene clusters. *Bioinformatics*, **20**(3):381-388.