Single Nucleotide Polymorphism(SNP) 데이타와 Support Vector Machine(SVM)을 이용한 만성 간염 감수성 예측

(Prediction of Chronic Hepatitis Susceptibility using Single Nucleotide Polymorphism Data and Support Vector Machine)

김 동 회 † 엄 상 용 †† 함 기 백 ††† 김 진 ††††
(Donghoi Kim) (Saangyong Uhmn) (Kibaik Hahm) (Jin Kim)

요 약 본 논문에서는 한국인의 대표질환 중 하나인 만성 간염에 대한 질환 감수성을 예측하기 위해서 Single Nucleotide Polymorphism 데이타와 대표적인 기계학습 기술인 Support Vector Machine을 이용하였다. 실험을 위한 데이타로 만성간염 환자 173명과 정상인 155명의 SNP 데이타를 사용하였으며, 평가를 위한 방법으로는 Leave-One-Out Cross Valication을 사용하였다. 실험결과 SNP 데이타만으로는 67.1%의 예측 결과를 얻었으며 기본적인 건강요소인 나이와 성별을 특징요소로 사용함으로서 74.9%의 예측 결과를 보였다. 향후 보다 많은 SNP 데이타와 건강관련정보 그리고 생활패턴에 대한 요소들을 특징요소로 감수성 예측에 함께 사용한다면, SVM은 만성 간염 예측을 위한 보다 효과적인 도구가 될 것이다.

키워드 : SVM, SNP, 간염

Abstract In this paper, we use Support Vector Machine to predict the susceptibility of chronic hepatitis from single nucleotide polymorphism data. Our data set consists of SNP data for 328 patients based on 28 SNPs and patients classes(chronic hepatitis, healthy). We use leave-one-out cross validation method for estimation of the accuracy. The experimental results show that SVM with SNP is capable of classifying the SNP data successfully for chronic hepatitis susceptibility with accuracy value of 67.1%. The accuracy of all SNPs with health related feature(sex, age) is improved more than 7%(accuracy 74.9%). This result shows that the accuracy of predicting susceptibility can be improved with health related features. With more SNPs and other health related features, SVM prediction of SNP data is a potential tool for chronic hepatitis susceptibility.

Key words : SVM, SNP, chronic hepatitis

1. 서 론

인간 개개인의 유전적 다양성을 이루는데 가장 중요한 영향을 미치는 요인은 인간 유전체내의 이산적으로 (discrete) 퍼져있는 수많은 단일 염기변이라고 알려져 있다. 이러한 DNA내의 단일 위치내의 염기 변이는 Single Nucleotide Polymorphism(SNP)라 불린다. 이 SNP의 차이는 사람 개개인의 피부색, 혈액형, 체질뿐만 아니라 각종 암, 당뇨, 치매등과 같은 질환에 대한 감수성에도 관여하는 것으로 알려져 있다. 개개인간의 유전적 차이를 알기위해 특정 인간의 모든 염기서열을 해독하는 것은 매우 어려운 작업이지만, 상대적으로 SNP를 해독하는 것은 쉬운 작업이다[1].

간염은 바이러스(HAV, HBV, HCV, HDV, HEV)감염에 의하여 간의 염증과 간조직의 파괴가 일어나는 질환으로, 6개월 이상 계속되는 경우를 만성 간염이라 일컫고, 원인과 심한 정도가 다양한 일련의 간질환들을 총칭한다. 경한 형은 비진행성이거나 천천히 진행되는데

· 이 논문은 보건복지부 보건의료기반 진흥사업의 지원으로 연구되었음 (A010383)
· 이 논문은 2004 한림대학교 교비학술연구비 지원으로 연구되었음 (HRF-2004-40)
† 정 회 원 : 한림대학교 컴퓨터공학
kdh@hallym.ac.kr
†† 정 회 원 : 한림대학교 정보통신공학부
uhmn@hallym.ac.kr
††† 정 회 원 : 아주대학교 의과대학 간 및 소화기 질환 연구센터
hahmkb@hotmail.com
†††† 정 회 원 : 한림대학교 정보통신공학부 교수
jinkim@hallym.ac.kr
논문접수 : 2006년 7월 11일
심사완료 : 2007년 5월 15일

반하여 더 심한 형은 계속 진행되어 결국 간경변(간경화)에 이르기도 한다. 이러한 간 질환은 다른 민족에 비해 한국인에게 많이 발생하고 있음에도 불구하고 한국인 의 유전체에 대한 연구가 없다.

본 논문에서는 Support Vector Machine(SVM)과 만성 간염환자와 정상인의 SNP 데이타를 사용하여 만성 간염에 대한 감수성(susceptibility)을 예측하였다. SVM [2]은 범용 목적의 지도된 패턴 인식(supervised pattern recognition)방법으로, 각종 암에 대한 감수성 예측 [3]과 DNA chip 데이타 분석[4], 단백질구조 예측[5]을 포함한 여러 생물학 파생 데이타를 분류하는데 성공적으로 사용되고 있다.

본 논문에서는 SNP 데이타를 사용하여 SVM을 학습시켜 만성 간염에 대한 감수성을 예측하는 방법에 대하여 논한다. 2장에서는 보다 구체적인 배경 연구에 대하여 논하며, 3장에서는 실험 및 결과에 대하여 논하며, 마지막으로 결론 및 향후 연구과제에 대하여 논한다.

2. 배경연구

2.1 SNP

인간 유전자의 서열은 Σ={A, T, G, C}의 알파벳으로 이루어진 23쌍의 유한한 길이의 문자열로 정의할 수 있다. 이 23쌍의 문자열의 길이를 합하면 약 30억 개 정도가 된다. 서열 정보는 변이(mutation)에 의해 그 내용이 달라질 수 있다. 개개인의 서열을 비교하였을 때 0.1%에 해당하는 부분은 다른 것으로 알려져 있으며, 이러한 차이에 의해 인종 간 혹은 개개인에 나타나는 다양성, 특정질병유무, 특정약물에 대한 반응성이 다르게 나타날 수 있다. 이러한 변이 가운데 가장 빈번하게 나타나는 SNP는 약 1000bp의 염기 당 1개의 빈도로 나타난다[6].

일반적으로 실험실에서 추출되는 지노타입은 배수염색체의 정보이기 때문에 염색체 각 위치에서 대립 형질이 동질접합(homozygous)일 경우에는 명확한 형태의 정보를 알 수 있지만, 이질접합체(heterozygous)일 경우에는 형질이 섞여있는 모호한 형태로 추출된다.

그림 1은 만성간염과 관련한 세 환자의 SNP 패턴을 나타낸 것이며, 표 1은 이 SNP 패턴을 관련 데이타로 표현한 것이다.

이때 SNP1을 예로 들면, 환자1은 이질접합인 GA를 가지고 있으며, 환자2, 환자3은 동질접합 GG, AA를 가지고 있다. 표 1에서 알 수 있는 것처럼 한 SNP에서는 최대 세 가지의 값이 나타날 수 있다.

2.2 SVM

본 논문에서는 SNP를 사용한 감수성 예측을 위해 SVM을 사용하였다. SVM은 최근 생물정보학 연구분야

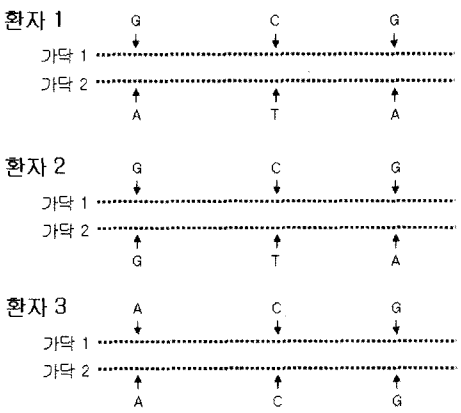

그림 1 환자 1, 2, 3에 대한 SNP 패턴

표 1 환자 1, 2, 3의 SNP 패턴의 데이타화

	SNP1	SNP2	SNP3	질환 유무
환자 1	G/A	C/T	G/A	유
환자 2	G/G	C/T	G/A	무
환자 3	A/A	C/C	G/G	유

에서 아주 많은 관심을 보이고 있는 기법이다. 이는 뉴럴 넷, case based reasoning 등과 같은 데이타 마이닝 기법과 비교하였을 때 상대적으로 정확도가 높고, 효과적인 결과를 제공해 주기 때문이다. SVM은 Vapnik와 Cortes에 의해 개발된 학습방법이다. SVM은 비선형적인 분류, 함수평가, 혹은 밀도 평가 등의 문제에 사용되는 강력한 기법이다. SVM의 기본 원리는 서로 다른 클래스 사이를 구분하는 다 평면의 최적의 최대의 마진(margin)을 찾아내는 데 있다. 이 기법의 목표는 가능한 한 작은 오류를 허용하며, 두 클래스 사이를 구별하는 확실한 평면들의 경계를 최대화하는데 있다. 또한 SVM은 데이타가 선형으로 구별되지 않을 경우에도 사용될 수 있다. 이 경우, 데이타는 비선형 함수를 사용하여 다차원의 자질 공간(feature space)에 사상된다.

학습을 위한 데이타 집합이 SVM에 주어지면, SVM은 하나의 모델을 형성한다. 이 모델은 추후 테스트를 위한 데이타 집합이 주어졌을 때, 출력으로 +1 혹은 -1의 값을 부여하게 된다. 본 논문에서 SVM은 주어진 SNP 집합에 의해 학습되고, 이후 미지의 SNP가 테스트 데이타로 주어졌을 때, 해당되는 클래스를 예측하게 된다.

2.3 문제정의

우리에게 특정 질환과 관련한 L개의 후보 SNP 정보가 있다고 가정한다. 또한 n명의 환자의 질환 여부와 이들에 대한 L개의 SNP 부위에 대한 실험정보가 있다고 가정한다. L개의 SNP정보를 사용하여 얻을 수 있는 가

능한 부분집합의 개수는 2^L-1개이다. 우리에게 예측률을 얻기 위한 하나의 기계학습 기법이 주어졌다 가정하고, 이 기계 학습 기법을 특정 부분 집합 $S \subseteq \{1...L\}$에 적용하여 얻은 예측률을 $p(S, n)$이라 하자. 이때 n명에 대한 L개의 SNP 정보를 사용하여, 최고의 예측률 p를 제공하는 부분집합 $S \subseteq \{1...L\}$을 구하려 하며, 이때의 부분집합을 S_{max}라 하자. 우리가 해결하려는 문제는 $p(S, n)$을 최대화하는 $S_{max} \subseteq \{1...L\}$을 찾는 것이라 정의할 수 있다. 이때 예측률 $p(S, n)$을 구하는데 사용되는 기법은 SVM을 포함한 다양한 여러 인공 지능적 예측기법이 될 수 있다.

이 문제를 해결하는 가장 단순한 방법은 모든 가능한 부분집합에 대하여 $p(S, n)$을 계산한 후 최대값을 제공하는 부분집합 S를 취하는 무차별적(brute-force)방법이 있다. 그러나 이 방법은 L개의 SNP가 있을 때, 예측률을 계산해야 하는 부분집합 S의 개수는 2^L-1이 되므로, L이 작은 경우에만 실용적이다. 실제 이 문제는 *NP-hard* 문제라 생각할 수 있다.

이 문제를 실용적으로 해결하기 위해서, 우리는 SNP들을 유전자 그룹별로 나누었다. 각각의 SNP와, 유전자별 SNP의 그룹, 전체 SNP에 대하여 SVM을 사용하여 그 가운데 최고의 예측률을 제공하는 것을 해답으로 취하였다.

3. 실험 및 결과

3.1 실험 데이타

예측률 실험에 사용된 데이타는 아주대학교 간 소화기 질환 유전체 센터[7]에서 얻어진 간질환 환자 가운데 SNP가 확보된 만성 간염환자와 정상인의 데이타를 이용하였으며, SNP데이타는 간 질환과 관련성을 가진다고 알려진 28개 SNP을 사용하였다. 이 28개의 SNP는 상호 연관성이 있을 것으로 추측되는 4개의 그룹으로 나누어진다. 표 2는 실험에 사용된 유전자의 SNP들과 해당 SNP가 가질 수 있는 염기와 환자그룹 및 환자수를 나타낸다. 모든 28개의 SNP 정보를 가지는 환자의 수는 만성간염 173, 정상 155명이다.

표 2 28개의 SNP와 환자 수

Group	No	SNPs	Sequence	환자 수
Group1	1	*CCR5(-2459)*	*G/A*	간염 환자:173 정상:155
	2	*RANTES(-403)*	*G/A*	
	3	*MCP1(-2518)*	*G/A*	
	4	*CCRR2-V64I*	*G/A*	
	5	*CXCR1-S276T*	*C/G*	
	6	*CXCR4-I138I*	*G/A*	
Group2	7	*IL1B-31*	*C/T*	
	8	*IL1B-511*	*C/T*	
	9	*IL1RN-S130S*	*C/T*	
	10	*IL1RN-3UTR*	*C/G*	
	11	*MBP-G54D*	*A/G*	
Group3	12	*IRF1(-410)*	*G/A*	
	13	*IFNGR2-Q64R*	*G/A*	
	14	*IRF1(-388)*	*C/T*	
	15	*IL-10(-592)*	*A/C*	
	16	*IL-10(-1082)*	*G/A*	
	17	*IFNGR1(-56)*	*C/T*	
	18	*IFNGR1(+95)*	*C/T*	
	19	*IFNG(+874)*	*A/T*	
	20	*TNF-238*	*G/A*	
	21	*TNF-308*	*G/A*	
Group4	22	*IL18-S35S*	*C/A*	
	23	*MMP3-E45K*	*G/A*	
	24	*MMP3-D96D*	*C/T*	
	25	*MMP3-A362A*	*C/T*	
	26	*MMP9-R279Q*	*G/A*	
	27	*MMP9-Q688R*	*G/A*	
	28	*MMP9-G607G*	*C/A*	

3.2 실험방법

SVM은 두 개의 클래스를 판별하는 SVM Light와 두 개 혹은 그 이상의 클래스를 판별하는 SVM Multi-class가 있다. 최초에 우리가 시도하였던 실험은 세분화된 클래스(정상, 간염보균, 간경화, 만성간염환자)들을 판별하는 것이었다. 이를 위해 복수개의 클래스를 판별하기 위해 사용되는 SVM Multiclass[8]를 사용하였으나, 의미 있는 결과 값을 얻을 수 없었다. 우리는 실험 범위를 축소하여 정상 혹은 만성 간염의 두 가지 클래스를 구별하기로 하였다. 두 개의 클래스를 구별하는데 SVM Multiclass를 사용하여도 되지만, SVM light[9]를 사용하는 것이 보다 일반적이다.

SNP 데이타는 두 개의 동질접합과 하나의 이질접합의 3개의 값을 가진다. 예를 들어 특정 SNP의 부위가 CC, CT, TT의 세 가지 값을 가진다고 가정하자. 이 값을 SVM을 사용하여 학습시키기 위해서는 숫자 값으로 이들을 변환해야한다. 일반적으로 생물학적으로 이질접합은 두 개의 서로 다른 동질접합에서 하나씩의 정보를 가져왔다고 생각하므로 이질접합에 대하여 가운데 값을 부여하는 경우가 대부분이다. 다른 한 방법으로는 각 SNP에 대해 두 자리 수의 숫자를 부여하는 방법이다. 예를 들어 A=1, G=2, C=3, T=4로 대치하게 되면, CC는 33, CT는 34, TT는 44가 된다. 그러나 이러한 두 자리 숫자부여방식은 숫자 값이 커지기 때문에, 사용하는 SNP의 수가 커질수록, SVM의 수행속도가 현저하게 느려진다. 우리는 이전 실험(pretest)에서 두 개의 실험방법을 모두 사용하였으나, 두 방법 모두 유사한 예측률을 보였다. 따라서 본 논문에서는 SNP 데이타를 이질접합의 경우 1로, 두 개의 서로 다른 동질 접합의

값을 0, 2로 랜덤하게 값을 부여한 결과를 논의한다.

학습데이타와 테스트데이타는 전체 환자데이타를 이용하였다. 실험은 28개 SNP정보를 공통적으로 가지는 환자그룹(만성 간염:173명, 정상:155)과 환자그룹에 따라 분류한 4개의 그룹을 각각을 실험하였다. 예측률 평가를 위한 방법으로는 leave-one-out cross validation을 사용하였다[10]. 이 방법은 실험그룹의 전체 n 명의 환자 가운데 하나의 환자정보를 간염여부를 판정하기 위한 테스트 데이타로써 사용하며, 나머지 n-1개를 학습데이타로 사용한다. 이러한 방식을 모든 n명의 환자정보에 라운드 로빈 방식으로 적용하여, 이들을 테스트함으로써 어느 정도의 정확도를 가지고 질환과 정상을 구별할 수 있는가를 산출하였다. 이러한 방식에 의해 n개의 결과값을 얻을 수 있는데, 이러한 방식을 적용한 이유는 환자의 SNP 정보를 얻는 것이 매우 어려워, 기존 정보를 최대한 활용해야 하기 때문이다.

3.3 실험결과

SVM Light에서 선형 커널을 사용하여 실험한 결과는 다음과 같다. 진단결과의 정확도는 민감도(sensitivity)와 특이도(specificity) 및 정확도(accuracy)를 사용하여 나타낼 수 있다. 민감도는 실제 특정 질환을 가진 사람이 질환을 가졌다고 판정되는 확률을 의미하며, 특이도는 정상인 사람이 정상으로 판정되는 확률을 의미한다. 예를 들어 90%의 민감도의 실험결과는 실제 질환자중 90%가 질환으로 판정됨을 의미하며 실제 질환자의 10%는 정상으로 잘못 판정됨을 의미한다. 90%의 특이도는 실제 정상인 가운데 90%가 정상으로 판정됨을 의미하며, 10%는 질환자로 잘못 판정됨을 의미한다. 표 3은 특이도와 민감도 및 정확도를 계산하는데 사용되는 표이다.

$$Sensitivity = \frac{TP}{TP+FN}$$

$$Specificity = \frac{TN}{TN+FP}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

위의 표에 의거하여 민감도는 $TP/(TP+FN)$, 특이도는 $TN/(TN+FP)$로 표현할 수 있다. 또한 정확도는 $(TP+TN)/(FP+FN)$으로 표현된다.

주어진 28개의 SNP의 모든 부분집합에 대하여 정확도를 구하는 것은 비실용적이기 때문에, 각각의 SNP, 4개의 SNP그룹, SNP전체에 대하여 예측률을 구해보았다. SNP가 얻어진 유전자 그룹을 부분집합으로 한 이유는 해당 유전자가 특정 질환에 영향을 미친다고 가정하면, 해당 유전자에서 얻어진 SNP 그룹이 동일한 이유로 해당 질환에 영향을 미칠 수 있다고 생각할 수 있기 때문이다.

우리는 각각의 SNP와 상호간 관련성이 있을 것으로 생각되는 SNP들을 대상으로 유전체 분석을 위해 그룹화한 4개의 SNP 그룹, 또한 전체에 대하여 SVM을 이용하여 정확도를 계산하였다. 계산된 결과는 표 4와 같다.

표 4 실험집단 및 정확도, 민감도, 특이도

실험집단	정확도	민감도	특이도
각각의 SNP	40.5%~53.6%	76.7~98.2%	1.3~2.6%
그룹1	50.9%	98.1%	2.6%
그룹2	66.5%	93.6%	34.6%
그룹3	39.9%	73.8%	2.6%
그룹4	53.7%	98.6%	5.5%
전체	67.1%	77.5%	55.5%

28개의 SNP에 대하여 각각의 정확도를 계산해본 결과 각각의 SNP와 그룹에 대하여 40.5%부터 53.6%사이의 정확도를 나타냈다. 이 경우 특이도가 상당히 낮게 나타나는 것을 볼 수 있다. 이는 감수성을 예측하는데 있어서 한쪽 클래스로(실험의 경우 질환클래스) 편파 판정 된다고 볼 수 있다. 백혈병과 같은 유전질환은 하나의 SNP의 차이에 의한 것으로 하나의 SNP가 질환과 정상을 정확히 구분할 수 있으나, 본 논문 대상인 만성 간염의 경우 많은 SNP가 질환 감수성에 영향을 미치는 질환으로 한 개의 SNP를 사용하거나 SNP 그룹이 예측율을 높일 수 있는 패턴의 조합이 아닌 경우 클래스에 대한 판별력이 떨어진다고 볼 수 있다. 위의 표에서 알 수 있는 것처럼, 특정 SNP에 의한 정확도보다, 그룹별 정확도가 더 높았다. 각 그룹의 정확도를 비교한 결과 전체 SNP를 사용한 결과의 정확도가 67.1%로 가장 높았다. 이 정확도는 우연한 확률보다는 현저하게 높은 수치이다. 따라서 만성 간염 예측을 위해서는 28개의 SNP집합 전체를 사용하는 것이 가장 효율적이라 판단된다. 일반적으로 학습시키는 SNP의 개수가 많을 경우 더 높은 정확도를 제공할 것으로 예상할 수 있다. 그러나 위의 표는 학습시키는 SNP의 개수가 많다고 하더라

표 3 특이도, 민감도 및 정확도의 계산

		질환여부	
		+(질환)	-(정상)
실험결과	+(질환판정)	True Positive(TP)	False Positive(FP)
	-(정상판정)	False Negative(FN)	True Negative(TN)

도 더 좋은 정확도를 얻을 수 있는 것을 보장할 수 없다는 것을 보여준다. 그룹 3의 경우 개개의 SNP의 정확도보다 그룹으로 학습시킨 경우 오히려 더 낮은 예측률을 보였다.

또한 SNP에 건강관련 인자들을 추가하였을 때의 효과를 알아보기 위해 가장 간단한 인자인 성별과 나이를 추가하여 테스트하여 보았다. 전체그룹에 환자의 성별과 나이를 추가하여 SVM을 사용하여 민감도, 특이도, 및 정확도를 계산한 결과 모든 값이 7% 정도 추가 향상된 결과를 가져왔다. 전체그룹과 관련된 정확도에 대한 결과가 표5와 표6에 나타나있다. 전체 SNP데이타와 건강관련인자인 성별과 나이를 추가한 결과 74.7%의 예측률을 얻을 수 있었다. 이 결과에서 알 수 있는 것처럼 몇 가지의 환경인자들을 포함하여 SVM을 사용한다면 보다 향상된 정확도를 얻을 수 있을 것이다. 예를 들어 흡연여부, 음주여부, 식생활습관, 주거환경 등의 간염과 관련된 인자들과 SNP를 결합하여 예측해본다면 만성 간염 감수성의 예측률을 보다 높일 수 있을 것이다.

표 5 전체그룹과 관련된 민감도, 특이도 및 정확도

		질환여부		
		+	-	
실험결과	+	134	69	민감도=77.5% 특이도=55.5% 정확도=67.1%
	-	39	86	

표 6 전체그룹과 나이 및 성별을 포함하였을 때의 민감도, 특이도, 정확도

		질환여부		
		+	-	
실험결과	+	146	56	민감도=84.4% 특이도=63.9% 정확도=74.7%
	-	27	99	

4. 결론 및 향후 연구

본 논문에서는 환자의 SNP 데이타를 이용하여 환자의 만성 간염감수성의 정도를 예측하기 위해서 SVM을 사용한 결과를 보였다. 또한 SVM을 사용하여 성공적으로 만성 간염 감수성을 예측할 수 있음을 보였다. 이는 간 질환이 유전적인 요인 뿐만 아니라 다양한 건강요소 및 생활환경도 중요한 요인임을 감안할 때 상대적으로 높은 예측률 이라고 볼 수 있다. 따라서 SVM은 SNP와 환자와 관련된 다양한 특징들을 사용할 경우 만성 간염감수성을 예측할 수 있는 강력한 도구가 될 수 있을 것이다. 우리는 생활습관, 환경인자 및 더 많은 수의 SNP 정보를 사용하여, 보다 높은 정확도를 가지며, 만성간염뿐만이 아닌 여러 간 질환에 대한 세분화된 판정을 얻을 수 있는 방법을 추후 연구하고자 한다.

참 고 문 헌

[1] A. M. Glazier, J. H. Nadeau, and T. J. Aitman, "Finding genes that underlie complex traits. Science," 298(5602):2345-2349, Dec 2002.

[2] V. N. Vapnik, "The Nature of Statistical Learning Theory," Springer, 1995.

[3] J. Listgarten, S. Damaraju, B. Poulin, L. Cook, J. Dufour, A. Driga, J. Mackey, A. Wishart, R. Greiner, and B. Zanke, "Predictive models for Breast Cancer susceptibility from multiple single nucleotide polymorphism," Clinical Cancel Research vol. 10 2725-2737, April 15, 2004.

[4] T. S. Furey, N. Duffy, N. Cristianina, D. Bednarski, M. Schummer, D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," Bioinformatics, 6(10):906-914. 2000.

[5] Y. D. Cai, X. J. Liu, X. b. Xu and G. P. Zhou, "Support Vector Machines for predicting protein structural class," BMC Bioinformatics 2:3 2001.

[6] J. I. Bell, "Single Nucleotide Polymorphism Disease Gene Mapping," Arthritis Research, Vol.4, pp.s273-s278, 2002.

[7] http://www.agcg.re.kr

[8] T. Joachims, "Making lagre-scale SVM learning practical. In B. Schoelkopf, C. Burges, and A. Smola, editors, Advances in Kernel Methods - Support Vector Learning," MIT Press, 1999.

[9] http://svmlight.joachims.org

[10] B. Efron, "Bootstrap Methods: Another Look at the Jackknife," The Annals of Statistics 7, 1-26 1979.

김 동 회

2001년 건국대학교 컴퓨터과학과(학사)
2003년 한림대학교 컴퓨터공학과(석사)
2007년 한림대학교 컴퓨터공학과(박사)
2007년~현재 한림대학교 정보전자공과대학 정보통신공학부 전임강사. 관심분야는 생물정보학, 데이타마이닝

엄 상 용

1987년 한림대학교 전자계산학과(학사)
1997년 한림대학교 컴퓨터공학과(석사)
1999년 한림대학교 컴퓨터공학과(박사수료). 관심분야는 분산/병렬처리, 프로그래밍 언어, 생물정보학

함 기 백

연세대학교 의과대학 졸업, 전문의, 의학박사. 연세대학교 의과대학 소화기내과 전임강사. 미국 국립암연구소 연수. 아주대학교 의과대학 소화기내과 교수. 아주대학교의료원 간 및 소화기질환 유전체 연구센터 소장. 현 분당제생병원소화기내과. 대한암예방학회 이사. 대한소화기학회 섭외이사. 헬리코박터학회 학술위원. 프리라디컬학회 부회장. 관심분야는 소화기내시경학, 위장학, 종양학, 생물정보학

김 진

1984년 Korea University, Physics. 1990년 Michigan State University, Computer Science, MS. 1996년 Michigan State University, Computer Science, Ph. D. 2007년 현재 한림대학교 정보전자공과대학 정보통신공학부 교수. 관심분야는 생물정보학, 알고리즘, 인공지능