시뮬레이티드 어닐링을 이용한 질환 관련 SNP 조합 선정 및 질환 예측

김동회*, 엄상용*, 김진*

요약

본 논문에서는 질환과의 연관성이 높은 SNP를 선정하고 질환 예측율 향상을 위해 확률 모델인 시뮬레이티드 어닐링 기술과 의사결정트리를 이용하였다. SA의 문제점인 시간 복잡도를 해결하기 위하여 빠른 휴리스틱 알고리즘에서 제공하는 SNP 조합에서 시작하여, 새로운 SNP 조합을 만들기 위한 효율적인 변환 규칙을 적용하였다. 연구 결과 효율적인 SA를 이용했을 때 기존의 특징선택 알고리즘에서 제시하는 SNP 조합과 다른 더 향상된 예측율을 보이는 새로운 SNP 조합을 얻을 수 있었다.

Finding Relevant SNP Sets and Predicting Risk to Disease Using Simulated Annealing

Dong-Hoi Kim*, Ssang-Yong Uhmn*, Jin Kim*

ABSTRACT

We applied simulated annealing algorithm and decision tree to find SNP sets relevant to a disease and predict the disease risk. For time complexity problem of SA, we construct an initial SNP set by fast heuristic algorithm and applied efficient transition rules to obtain new SNP sets. The experiment results show that we can obtain new SNP sets with the improved prediction performance compared to others by traditional feature selection algorithms.

Key Words : SNP, Risk Prediction, simulated annealing, decision tree, feature selection, machine learning

* 한림대학교 컴퓨터공학과
· 제1저자(First Author) : 김동회 · 교신저자(Correspondent Author) : 김 진
· 접수일(2010년 4월 30일), 수정일(1차 : 2010년 5월 28일), 게재확정일(2010년 6월 4일)

Ⅰ. 서 론

인간의 유전자 정보 중 개인 또는 집단 간의 차이를 연구하고 질환과의 연관성을 밝혀내어 질환예측 및 맞춤의학 토대를 마련하고자 최근 유전 정보의 차이를 나타내는 단일염기변이인 Single Nucleotide Polymorphism(SNP)[1]나 유전자 복제 수 변이인 Copy Number Variation(CNV)에 대한 많은 활발한 연구가 진행되고 있다. SNP는 CNV에 비해 먼저 연구가 활성화 되었고 방대한 량의 데이터가 축적되어 있어 컴퓨터를 이용한 연구에 용이한 이점이 있다.

현재 SNP를 이용해 질환을 예측하기 위해 SVM이나 의사결정트리[2][3][4][5], 뉴럴넷과 같은 다양한 기계학습기법들이 적용되고 있으며, 질환 관련 유전자의 수가 많아질수록 처리를 위한 시간복잡도와 공간복잡도가 증가함에 따라 최적의 SNP 조합을 선정하기 위해 변수 증가법(Forward Selection),변수 감소법(Backward Elimination)[6][7]등 다양한 특징선택(Feature Selection) 알고리즘들이 적용되고 있다. 하지만 SNP의 수가 많아질수록 지역 최소점(Local Minima)에 빠지기 쉬우므로 최적의 예측율을 제공하지 못할 수 있다.

본 논문에서는 질환과의 연관성이 높은 SNP를 선정하고 질환 예측율 향상을 위해 확률 모델인 시뮬레이티드 어닐링(Simulated Annealing, SA)[8] 기술과 의사결정 트리를 이용하였다.

SA는 가열 냉각의 개념을 도입한 알고리즘으로 전체 최적점(Global Optimum)을 제공해 줄 수 있다. 그러나 SA는 초기 상태에 따라 많은 실행 시간이 소요되는 문제점을 가진다. 이러한 초기 상태의 문제점을 해결하기 위해 변수 증가법을 이용하여 초기 상태를 선정하고 효율적인 변환규칙을 적용한 SA를 사용하였다. 실험 결과 효율적인 SA를 적용했을 때 향상된 질환 예측율을 얻을 수 있었으며, 기존의 특징선택 알고리즘에서 제시하는 예측율보다 더 향상된 예측율을 보이는 새로운 SNP 조합을 얻을 수 있었다. 이 새로운 SNP 조합들은 생명과학 연구 분야에 기여할 수 있을 것이다.

Ⅱ. 관련연구

2.1 SNP

인간의 유전자는 A, T, G, C 네 개의 염기의 서열이다. 단일 염기 변이인 SNP는 인간의 유전자 염기서열 중 개인 또는 집단 간 차이를 나타내는 단일 염기 변이로 유전 질환이나 질환에 대한 민감도에 영향을 미치며 또한 약물에 대한 반응에도 차이를 나타낸다. 그림 1은 SNP의 개념을 묘사한 것이다.

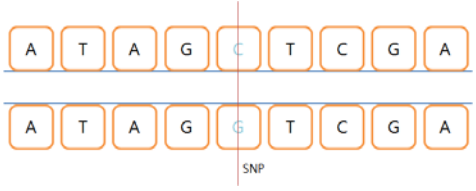

그림 1. SNP 개념
Fig. 1. SNP Concept

SNP를 분석하고 이해함으로서 질환을 미리 예측하거나 맞춤의학의 토대를 마련할 수 있다. 유전질환에 있어서 단일 염기 또는 소수의 염기 차이에 따라 질환 유무의 차이가 나타나는 경우가 다수 존재하며, 이 경우 유전자 실험을 통해 밝혀질 수도 있다. 유전질환의 또 다른 부류인 복잡한 질환이나 여러 개의 SNP 조합이 연관되어 질환에 영향을 미치는 경우 시간과 비용의 문제로 인하여 최적의 질환 관련 SNP 조합을 찾기 위한 컴퓨터 기술이 초기 분석 과정에 반드시 필요하게 된다. 그러나 관련 SNP 수가 증가할수록 예측율을 오히려 떨어트리는 다수의 SNP가 포함될 가능성이 발생할 수 있고, 이러한 SNP들이 포함됨으로써 컴퓨터

를 이용한 분석의 시간복잡도와 공간복잡도가 증가하게 된다. 이로 인하여 주어진 SNP의 모든 가능한 조합을 시도해 볼 수 없다. 따라서 최적의 예측율을 제공하는 SNP의 조합만을 추출하기 위한 특징선택 알고리즘이 적용되어야 한다.

2.2 시뮬레이티드 어닐링

SA 기법은 확률에 근거한 기법으로 combinatorial optimization 문제들에서 최적 비용을 찾아내는데 사용된다. SA는 높은 온도의 상태(state)에서 출발하여 Metropolis[9]에 의해서 제안된 상태전환규칙과 수용규칙을 적용함으로써, 계속적으로 현재의 상태에서 새로운 상태를 만들어낸다. 수용규칙의 기준은

1. 만일 $\Delta C \leq 0$, 새로운 상태를 수용한다.

2. 만일 $\Delta C > 0$, 수용확률 $P(\Delta C) = e^{-\frac{\Delta C}{T}}$ 를 가지고 새로운 상태를 수용한다.

이때 T는 온도이며 ΔC는 새로운 상태와 현재 상태의 비용 차이이다. 수용확률 $P(\Delta C)$는 시스템이 지역 최소점에 고정되어버리는 것을 방지한다. 온도 T는 새로운 상태를 수용하는 확률에 영향을 미친다. SA는 높은 온도 T에서 시작하여 annealing schedule에 따라 매번 온도를 점차 내린다. 높은 비용을 가진 상태를 새로운 상태로 수용하는 확률도, 온도가 내려감에 따라 작아진다.

SA과정은 조심스러운 annealing schedule과 많은 반복횟수가 허용되면 전역 최적점에 수렴하게 된다. 그림 2는 표준 SA의 흐름을 보인다.

SA의 가장 큰 문제점은 많은 컴퓨터시간이 걸린다는 점에 있는데 이는 기본적으로 SA은 monte-carlo 기법이기 때문이다. 본 논문에서는 이 컴퓨터시간을 줄이기 위해 몇 가지의 방법을 적용하였다.

III. 본 문

3.1 문제정의

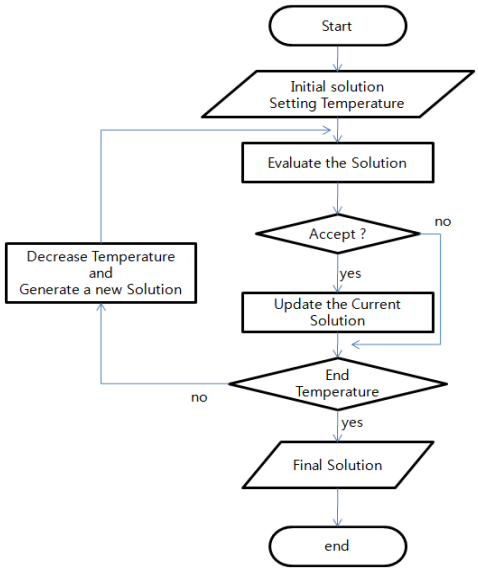

그림 2. 시뮬레이티드 어닐링 흐름도
Fig. 2. Simulated Annealing Flow

질환과 관련성이 있을 것으로 생각되는 L개의 후보 SNP 에 대해 n명의 실험군과 대조군 정보가 있다고 가정한다면 L개의 SNP 정보를 사용하여

얻을 수 있는 가능한 부분집합의 개수는 $2^l - 1$개이다. 예측율을 얻기 위한 하나의 기계학습 기법이 주어지며, 이 기계 학습 기법을 특정 부분 집합 $S \subseteq \{1...L\}$에 적용하여 얻은 예측율 $P(S,n)$ 이라 하자. 이때 n명에 대한 L개의 SNP 정보를 사용하여, 최고의 예측율 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 예측율 $P(S,n)$을 구하는데 사용되는 기법은 의사결정 트리와 같은 다양한 기계학습 기법이 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 SNP이 있을 때, 예측율을 계산해야 하는 부분집합 S의 개수는 2^l-1이 되므로, L이 작은 경우에만 실용적이다. 실제 이 문제는 NP-hard 문제라 생각할 수 있다. 따라서 최적의 SNP 조합을 선정하기 위해 휴리스틱 방법인 변수 증가법 ,변수 감소법과 같은 다양한 특징선택 알고리즘들이 적용되고 있다. 그러나 이러한 휴리스틱 방법들은 지역 최소점에 빠지기 쉬워 최적의 예측율을 제공하는 SNP 조합을 얻을 수 없을 수도 있다. 따라서 본 논문에서는 변수 증가법으로 얻어진 SNP 조합이 지역 최소점에 의해 얻어졌을 가능성이 있다고 보고 이를 초기 해로 보고 효율적인 SA 방법을 이용 최적의 해를 구하고자 한다.

3.2 데이터

실험을 위한 데이터는 plink[10]의 SNP 생성 기능을 이용하여 p-value<=0.001 의 SNP 데이터를 환자군(case)과 대조군(control)을 동일한 비율로 하여 생성하였다. 첫 번째 실험군의 경우 50개의 SNP정보에 환자군과 대조군 각각 45명의 데이터로 구성하였다. 두 번째 실험군의 경우 200개의 SNP정보에 대하여 환자군과 대조군 각각 100명의 데이터고 구성하였으며, 마지막 실험군의 경우 200개의 SNP 정보에 대하여 환자군과 대조군 각각 200명의 데이터로 구성하였다.

SNP의 데이터 표현의 경우, 예측율 제공을 위해 본 논문에서 사용한 기계학습방법인 의사결정 트리는 기호 데이터(AA , AT , TT , CC , GG , GC)를 사용할 수도 있으나, 차후 다른 기계학습 방법에서의 평가를 감안하여 AA , TT 또는 CC , GG와 같은 두 가지 동질접합형태와 AT 또는 CG와 같은 하나의 이질접합형태를 각각 1, 2, 3으로 표기하는 방법을 사용하였다.

3.3 비용함수를 최적화하기 위한 SA

3.3.1 높은 온도상태를 대신한 빠른 휴리스틱 알고리즘

SA의 높은 온도상태는 random search와 비슷하며, 낮은 온도 상태는 greedy local search와 비슷하다. 본 논문에서는 휴리스틱 알고리즘인 변수 증가법으로부터 얻어진 SNP 조합을 SA의 초기 상태로 하였다. 이와 같은 방법은 SA의 현재 상태를 최적에 가까운 (near-optimal) 상태에서 시작하게 함으로써 어닐링 시간을 단축시킬 수 있도록 한다.

3.3.2 변환규칙

L개의 SNP들에 휴리스틱 알고리즘 중 변수 증가법을 적용하여 초기 SNP 조합을 획득하게 된다. 획득된 SNP 조합에 변환규칙을 적용하여 새로운 후보 SNP 조합을 얻게 된다.

변환규칙이 적용될 때 기존 SNP 조합에 포함되지 않은 다른 SNP를 조합에 포함할지, 기존 SNP 조합에서 특정 SNP을 제거할지를 무작위로 결정한다. 이때 추가 또는 제거될 SNP의 수는 최소 1개에서 최대 3개를 무작위로 결정되도록 한다. 이 경우에 대표적인 특징 순위 알고리즘으로 기계학습 분야에서 특징의 중요도를 측정하는데 널리 사용되는 정보획득량(Information Gain)[11]을 사용하여 정보획득량이 높은 SNP 일수록 후보 SNP에 선택되어질 확률을 높이고, 낮을수록 기존 SNP들에서 제거될 확률을 높일 수 있도록 한다. 만약 모든 SNP이 선택되어지거나 삭제되게 되면 초기 SNP 조합을 다시 현재 SNP 조합으로 하여 변환규칙을 적용해 나가고, 그렇지 않을 경우 새로 만들어진 SNP 조합을 현재 SNP 조합으로 사용한다. 변환규칙은 그림 3과 같다.

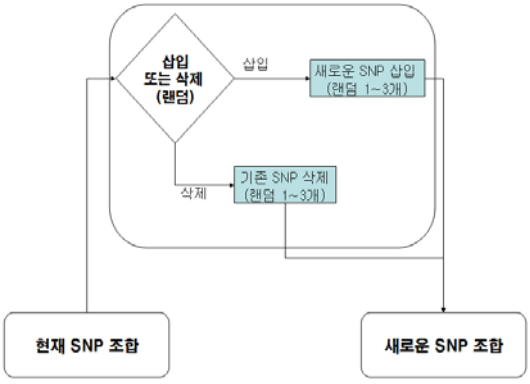

그림 3. 변환 규칙
Fig. 3. Conversion Rules

SA에서 현재 상태를 s , 비용함수를 f라 할 때, 상태 비용 C는 현재 상태에 비용함수 f를 적용하여 계산한다($C = f(s)$). 본 논문에서의 ΔC는 의사결정 트리에서 얻어진 예측율을 이용하여 새로운 예측율 C_{new}에서 현재의 예측율 $C_{current}$의 차이 값으로 다음과 같이 나타낼 수 있다.

$$\Delta C = C_{new} - C_{current} \quad (1)$$

3.3.3 annealing온도의 스케줄링

본 알고리즘의 annealing온도의 스케줄링은, $T = T_i \times e^i$이다. 이때 e는 상수이며, i는 반복횟수, T_i는 초기온도이다. e값은, k를 총 반복횟수, T_f를 마지막 온도라 할 때, 다음과 같이 정의된다.

$$e = (T_f / T_i)^{1/k} \quad (2)$$

그림 4는 본 논문에서 사용된 SA 알고리즘의 개관을 보인다.

```
begin SA
   current SNP set ← fast heuristic algorithm
   final SNP set ← current SNP set
  Cmin ← Cinit
  T ← Ti
  while(T>Ti)
    new SNP set ← current SNP set ± SNP set applying transition rule
    calculate Accuracy using decision tree
    Cnew ← Ccurrent
    if Metropolis conditions are satisfied then
      Ccurrent ← Cnew
      current SNP set ← new SNP set
      if(Cnew<Cmin) then
        Cmin ← Cnew
        final SNP set ← current SNP set
      end if
    end if
    T ← εT
  end while
 end SA
```

그림 4. 비용함수 최적화를 위한 SA
Fig. 4. SA for the optimization of the cost function

3.4 실험방법

본 논문에서는 예측율을 얻기 위한 기계학습 방법으로 C4.5 의사결정 트리를 사용하였다.

최적의 질환 관련 SNP 조합은 가장 높은 질환 예측율을 제공하는 SNP 집합으로 예측율에 대한 평가는 목표 변수의 실제 범주와 모형에 의해 예측된 분류범주 사이의 관계를 나타내는 표 1과 같은 정오분류 행렬(confusion matrix)[12]를 이용하였다.

표 1. 예측율 계산
Table 1. Predictive value calculating

		Actual	
		+	-
Predict	+	True Positive (TP)	False Positive (FP)
	-	False Negative (FN)	True Negative (TN)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

TP는 실제 질환클래스를 질환으로 정확히 예측한 것을 의미하며, FP는 실제 정상클래스 이지만 질환클래스로 오분류한 것을 의미한다. FN은 실제 질환클래스를 정상클래스로 오분류한 것을 의미하며, TN은 실제 정상클래스를 정상클래스로 정확히 예측한 것을 의미한다. 정오분류행렬에 의해 계산된 예측율은 SA의 비용으로 사용하며 전 상태의 비용과의 차이 값을 최적의 SNP 조합을 선출하기 위해 사용되었다.

우선 SA의 초기 상태에 포함될 수 있는 후보 SNP들은 의사결정 트리에 휴리스틱 특징선택 방법인 변수 증가법을 이용하여 평가 하였으며, 이 때 Leave-One-Out Cross Validation (LOOCV)[13]방법을 사용하였다. LOOCV 는 환자군과 대조군의 자료 수가 적을 경우 모든 자료를 학습과 평가에 사용하기 위한 방법으로 N개의 자료에 대해 N-1 개의 데이터를 학습에 사용하고 한 개를 평가에 사용하는 것을 N번 반복 수행함으로서 예측율을 측정하는 방법이다. 이 평가에서 가장 높은 예측율을 제공하는 SNP들로 초기 상태를 구성하였다.

비용함수가 최적화된 SA는 초기 SNP 조합을 초기 상태로 하여 시작온도 T = 20부터 최종온도 T = 0.1 이 될 때 까지 100,000회의 변환규칙을 적용하면서 매 변환에서 얻어지는 새로운 SNP 조합에 대해 예측율을 평가하여 더 높은 예측율을 보이는 경우에는 이를 새로운 현재 상태로, 예측율이 현재 상태 보다 낮은 경우에는 SA에서의 확률에 따라 새로운 현재 상태로 사용할 지를 결정한다. 이 과정을 통해서 현재 까지 최고의 예측율을 보이는 SNP 조합을 최종적으로 선출하게 된다. 이 때 예측율을 평가하기 위한 방법으로 의사결정트리를 사용하였으며, 이 과정에서도 LOOCV 방법을 사용하였다.

3.5 결과

표 2는 본 논문에서 적용한 비용함수를 최적화한 SA를 통해 얻어진 결과이다.

표 2. 실험 결과
Table 2. Experimental results

No	총 SNP수	초기 SNP 조합 수	예측율	새로운 SNP 조합 수	예측율
1	50	5	89%	19	90.5%
2	200	91	89%	86	90%
3	200	103	88.75%	86	91.75%

첫 번째 실험에서는 50개의 SNP를 가진 환자군과 대조군 각각 45개의 데이터를 대상으로 변수 증가법으로 얻어진 5개의 SNP 조합을 초기 해로 하여 최종적으로 예측율 향상을 보이는 19개의 새로운 SNP 조합을 얻었다. 두 번째 실험에서는 200개의 SNP를 가진 환자군과 대조군 각각 100개의 데이터를 대상으로 변수 증가법으로 얻어진 91개의 SNP조합을 초기 해로 하여 최종적으로 예측율

향상을 보이는 86개의 새로운 SNP 조합을 얻었다. 세 번째 실험에서는 200개의 SNP를 가진 환자군과 대조군 각각 200개의 데이터를 대상으로 변수 증가법으로 얻어진 SNP 조합을 초기 해로 하여 예측율 향상을 보이는 86개의 새로운 SNP 조합을 얻었다.

실험 결과 세 가지 경우 모두 예측율이 향상되었음을 볼 수 있으며, 이는 SA가 기존 휴리스틱 방법에 의해 구해진 결과 보다 향상된 질환 예측율을 보이는 SNP 조합을 얻을 수 있었음을 볼 수 있다. 특히 보다 적은 수의 SNP 조합으로 높은 예측율을 얻을 수 있었으며, 이는 SA가 SNP 조합의 수를 최소화 할 수 있는 방법이 될 수 있음을 나타낸다.

Ⅳ. 결론 및 향후 연구

본 논문은 SNP 분석을 통한 질환예측과 최적의 예측율을 제공하는 SNP 조합을 선별하기 위한 비용함수 최적화 SA의 적용을 제시하였다. 최근 SNP 분석에 있어서 대용량 SNP 데이터 분석을 위해 컴퓨터를 이용한 기초 분석은 꼭 필요하다. 또한 SNP 데이터의 증가에 따라 효율적인 컴퓨터 기술 또한 필수적이라고 할 수 있다. 이에 제시된 방법은 질환 예측에 있어서 기존 휴리스틱 방법에 비해 더 향상된 예측율을 가지는 SNP 조합을 선별 할 수 있음을 보였으며, SNP 조합의 수를 최소화 하는 방법을 제시함으로서 SNP을 이용한 진단 의학에 있어서 비용 절감 효과를 기대할 수 있다.

본 논문에서는 p-link의 SNP 자동 생성 기능을 이용한 데이터를 기반으로 하였으나, 향후 실제 질환 관련 SNP들을 대상으로 한 연구에 있어서도 유용할 것이다.

참고문헌

[1] Anthony J.Brookes "Review The essence of SNPs" GENE pp178 1999.

[2] J.R. Quinlan, “Induction of decision trees”, Machine Learning 1 81-106; reprinted in: J.W. 1986.

[3] Shavlik and T.G. Dietterich, eds., “Readings in Machine Learning” ,Morgan Kaufmann, San Mateo, CA, 1986.

[4] J.R. Quinlan, “C4.5: Programs for Machine Learning” , Morgan Kaufmann, San Mateo, CA, 1993.

[5] Tjen-Sien Lim, Wei-Yin Loh, and Yu-Shan Shih, An empirical comparison of decision trees and other classification methods, TR-979, Dept. of Stat., Univ. of Wisconsin, Madison, June 1997.

[6] R. Kohavi and G. John. “Wrappers for feature subset selection.” Artificial Intelligence, 97 (1-2), 273-324, 1996.

[7] A.J. Miller, “Subset Selection in Regression” , Chapman and Hall, London, 1990.

[8] Kirkpatrick, S., Gelatt, C. D., Jr. and Vecchi, M. P. . “Optimization by simulated annealing.” *Science*, **220**, 671-680, 1983

[9] metropolice, N., Rosenbluth, M., Rosenbluth, A., Teller, A. and Teller, E. “Equation of state calculations by fast computing machines.” *J. Chem. Phys*., **21**. 1087-1092, 1953.

[10] http://pngu.mgh.harvard.edu/~purcell/plink/

[11] S. Kullback , “Information theory and statistics” , John Wiley and Sons, NY , 1959.

[12] Geisser, Seymour , “Predictive Inference.” New York: Chapman and Hall. ISBN 0412034719 , 1993

[13] Kohavi, Ron , "A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2 (12): 1137 - 1143. 1995

Acknowledgement

본 연구는 2009년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구입니다. (2009-0077545)

김동희(Dong-Hoi Kim)

2002년 한림대학교 컴퓨터공학 석사
2007년 한림대학교 컴퓨터공학 박사
2007년~현재 한림대학교 컴퓨터공학과 전임강사
※ 관심분야 : 데이터마이닝, 생물정보학

엄상용(Ssany-Yong Uhm)

1997 한림대학교 컴퓨터공학과 (석사)
1999 한림대학교 컴퓨터공학과 (박사수료)

※ 관심분야 : 분산/병렬처리, 프로그래밍 언어, 생물정보학

김진(Jin Kim)

1990년 Michigan State University, Computer Science 졸업(석사)
1996년 Michigan State University, Computer Science 졸업(박사)
1997년~2000년 건국대학교 전산과학과 교수
2001년~현재 한림대학교 컴퓨터공학과 교수
※ 관심분야 : 생물정보학, 유전자 알고리즘, 의료정보학