Journal of Life Science 2007 Vol. 17. No. 5. 714~718 ©JLS / ISSN 1225-9918

유전자 프로모터 예측을 위한 Support Vector Machine의 응용 방법에 대한 연구

김 기 봉*

상명대학교 공과대학 생명정보공학과

Received April 16, 2007 / Accepted May 8, 2007

A Study On the Application Methods of a Support Vector Machine for Gene Promoter Prediction. Ki Bong Kim*. *Department of Bioinformatics Engineering, Sangmyung University, Chunan 330-720, Korea* – The high-throughput sequencing of a lot of genomes has resulted in the relatively rapid accumulation of an enormous amount of genomic sequence data. In this context, the problem posed by the detection of promoters in genomic DNA sequences via computational methods has attracted considerable attention in recent years since exact promoter prediction can give a clue to the elucidation of overall genetic networks. In this study, applications of support vector machine(SVM) to promoter prediction are explored to show a right approaches to discriminate between promoter and non-promoter regions by means of SVM. The results of various experiments show that encoding method, encoding region and learning data constitution can play an important role in the performance of SVM.

Key words – high-throughput sequencing, promoter prediction, genetic networks, support vector machine, encoding method

서 론

다양한 유기체를 대상으로 한 활발한 유전체 프로젝트의 수행결과로 말미암아 밝혀진 유전정보의 양은 기하급수적으로 증가하고 있으며, 실험연구 기법의 자동화 및 대용량화의 가속으로 바이오데이터의 홍수 속에서 살아가고 있다고 해도 과언이 아니다. 2003년 4월에 인간 유전체 프로젝트의 종결을 공식적으로 선언한 이래로 이미 서열결정이 완료되어 인터넷상에 공개되어 있는 것만 하더라도 850여 종에 이른다[12]. 게다가 긴 서열단편들 간의 갭 채우기가 조만간 완성되어 200여 종의 유전체 완결본이 머지않아 인터넷상에 공개될 전망이다[12]. 이러한 유전체 프로젝트들은 엄청난 핵산 염기서열 데이터를 양산하고 있으며, 생성된 막대한 양의 유전분자 정보들을 해석하고 이해하는데 초점이 맞추어져 있는 전산생물학은 눈부신 발전을 거듭하고 있다. 비록 전산생물학은 추정 유전자와 그에 상응하는 기능들을 규명하는데 많은 기여를 하고 있으나, 현재 이용 가능한 유전체 정보들을 해독하기 위해서 해야 할 일들이 여전히 많이 산적해 있다[6,7]. 특히 유전자가 자신의 최종 산물을 생성하는 유전자 발현과정과 그러한 과정에서 이루어지는 복잡한 유전자 제어 및 조절에 있어서 더욱 더 그러하다. 이러한 문맥에서 전산기법에 의해 유전체 DNA 염기서열 상에서 특정 전사조절인자 결합부위 및 핵심 프로모터 영역을 예측하는 문제는 상당히 중요한 연구 관심사로 대두되고 있으며, 활발한 연구 및 투자가 절실히 요구된다. 전사조절인자 결합부위나 프로모터 영역을 예측하는 것은 전사와 번역의 정확한 핵산 서열 결정자들을 규정하는 생화학적 문제와 본질적으로 상당히 밀접한 관련이 있을 뿐만 아니라, 유전자의 발현양상을 예측하거나 유전자의 기능을 규명 하는데 결정적인 단서를 제공한다. 더 나아가서는 그러한 유전자 제어 모듈 및 계층구조를 파악함으로써 살아있는 생명체의 유기적 네트워크 현상들을 규명할 수 있을 뿐만 아니라, 총체적 생명현상을 이해하는데 결정적인 단서를 얻을 수 있기 때문에 더욱 더 이 분야의 연구 및 개발이 필요하다[7].

이러한 문맥에서 본 논문에서는 최근들어 생물정보학 분야에서 널리 사용되고 있는 기계학습 알고리즘의 일종인 SVM (Support Vector Machine)[5,10]을 활용한 다양한 실험을 통해 진핵생물의 핵심 프로모터 영역을 예측하는데 있어서 여러 가지 고려사항 및 문제점들을 짚어보고 올바른 적용방법을 제시하고자 한다. 생물정보학 분야에서 *ab initio* 예측방법들, 즉 서열정보를 기반으로 하는 예측 방법들의 핵심 알고리즘으로 널리 사용되어 왔던 것들은 HMM (Hidden Markov Model)을 비롯한 확률적 모델들[2,8], 신경망 (Neural Network)을 비롯한 기계학습 기법들[3] 등이 과거에는 주류를 이루었으나 최근 들어서는 SVM (Support Vector Machine)이 널리 응용되고 있는 상황이다[3,4,11]. 특히 단백질 상호작용 예측 분야에서는 널리 사용되고 있다. 그러나 아직까지 유전자 프로모터 영역 예측 분야에 적용된 예는 그리 많지 않으며 초기 응용단계에 있는 실정이다[4]. SVM (Support Vector Machine)은 선형 학습기법이 가지는 한계점을 극복하기 위해 새로운 가설 공간을 도입하여 데이터들을 분류하는 방법이다. 이것은 두 부류의 데이터 사이의 거리가 가장 크게 되는 분할평면(hyperplane)을 찾고 그 분할 평면을 기준으로 입력서열을 분류하는 방법이다. SVM에

*Corresponding author
Tel : +82-41-550-5377, Fax : +82-41-550-5184
E-mail : kbkim@smu.ac.kr

서 제공하는 커널함수(kernel function)는 선형(linear), 다항식(polynomial), 반경기반(radial basis function : RBF) 및 시그모이드(sigmoid) 함수 등이 있다. 본 연구에서는 공개되어 널리 사용되고 있는 SVMlight[10]를 이용하였으며, 커널함수(kernel function)는 SVMlight의 도트 커널(dot kernel)과 다항식 커널(polynomial kernel)을 그대로 사용하였다. 본 논문에서는 프로모터 데이터로 EPD 데이터베이스[9]의 프로모터 영역 데이터를 사용하였고, 비(非)프로모터 데이터(non-promoter data)로는 GENIE[8] 프로그램에서 사용했던 CDS (coding sequence) 서열을 사용하였다.

재료 및 방법

앞에서 언급한 바와 같이 유전자 프로모터 영역에 대한 학습 및 검정 데이터로는 EPD (Eukaryotic Promoter Database) [9] 데이터를 사용하였다. EPD는 ISREC (Swiss Institute for Experimental Cancer Research)의 생물정보학 그룹에서 제작, 운영하는 데이터베이스이다. EPD는 실험적으로 전사의 시작위치가 규명된 진핵생물 POL II 계열의 프로모터 데이터를 집적해 놓은 것으로, 데이터베이스 내의 엔트리 프로모터 서열들 간의 중복을 없애고 최대한의 상세 주석정보를 담고 있는 잘 정리된 양질의 데이터베이스이다. 또한 TSS (Transcription Start Site)를 기준으로 상위부위(upstream)의 499 bp와 하위부위(downstream)의 100 bp (즉, -499 ~ +100 영역에 해당하는 프로모터 서열. TSS는 0로 표기)를 포함하여 전체 길이가 600 bp인 프로모터 서열 데이터들이 EPD 데이터베이스에 저장되어 있다. 이러한 이유에서 본 논문의 실험 데이터로 EPD를 활용하였다. 그리고 본 연구의 다양한 실험에서 비(非)프로모터 데이터로는 진핵생물의 유전자 구조 예측 프로그램인 GENIE[8]를 개발하기 위해 사용되었던 양질의 CDS 서열 데이터를 활용하였다.

SVM은 통계학적 이원 분류 기법의 하나인데, 본 논문에서 수행할 실험과 같이 유전자 프로모터 영역이 '맞다' 혹은 '아니다'로 분류되는 예에 적합하다. SVM은 인코딩(encoding)에 의해 특성 벡터(feature vector)의 형태로 표현된 데이터를 기계 학습을 통해 분류 모델을 만들고 다른 데이터를 그 모델에 적용해 얼마나 잘 분류하는지 여부를 살펴본다. 본 논문의 실험에서는 서열의 특성들을 잘 반영하는 인코딩 방법을 찾기 위해 다양한 인코딩 방법들을 여러 실험을 통해 검정해 보았다. 이러한 인코딩 방법뿐만 아니라 실제 식별에 유익한 프로모터 영역을 찾기 위해 프로모터 영역 중에서 다양한 부분을 positive 데이터로 설정해서 실험해 보았다. 본 논문에서는 앞에서 언급했듯이 SVMlight (http:// svmlight.joachims.org/)[10]를 사용하였고 커널함수(kernel function)는 SVMlight의 도트 커널(dot kernel)과 다항식 커널(Polynomial kernel)을 그대로 사용하였다.

앞에서 언급한 것처럼 학습 데이터는 EPD 데이터베이스 사이트(http://www.epd.isb-sib.ch/)의 분류범주 중에서 "homo sapiens"범주의 서열을 추출하여 사용하였다. 실험 당시 EPD에는 1,871개의 homo sapiens 서열이 있었는데, 극히 일부는 서열결정에 있어서 확실성이 결여된 서열이 존재했다. EPD에 있는 서열에서 핵심 프로모터 영역을 정확히 알지는 못하므로 일반적으로 프로모터 영역이라 예상되는 영역을 추출하였다. 그러나 종마다 프로모터의 영역이 유동적이므로 영역을 변경하면서 4번의 실험을 거쳤다. 그리고 핵산 잔기들(residues) 간의 의존성(dependency)을 고려하여 1개, 2개, 3개 및 6개의 핵산 잔기들을 단일체로 간주하여 인코딩하여 실험 및 검정을 해 보았다. 프로모터로 예측되는 positive 데이터와 프로모터가 될 수 없는 영역을 negative 데이터로 추출하여 각각을 학습(training) 데이터와 검정(test) 데이터로 나누었다. 개별 실험에서 학습 성능을 평가하기 위해서 일반적으로 널리 사용되는 특이성(Specificity : S_p)과 민감성(Sensitivity : S_n)을 사용하였으며, 이들은 다음과 같이 정의된다.

$$S_n = \frac{TP}{TP+FN}, \quad S_p = \frac{TP}{TP+FP} \qquad \text{(수식 1)}$$

위에서 TP는 True Positive, FN은 False Negative, FP는 False Positive를 의미한다.

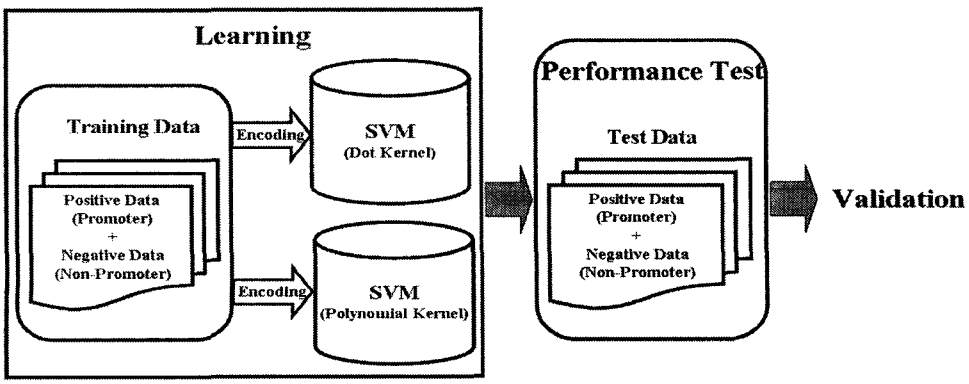

Fig. 1. Schematic diagram outlining the experimental procedure.

실험 Ⅰ

Table 1에서 보듯이 EPD 프로모터 서열 데이터에서 -15 ~ 50 영역의 서열을 positive 데이터(1,762개)로 15 ~ 80 영역의 서열을 negative 데이터(1,734개)로 하였고, 인코딩할 때에는 개별 위치의 핵산 잔기가 독립적인 존재로 인식할 수 있도록 인코딩하였다. 도트 커널과 다항식 커널 각각에 대해 학습시키고 검정 데이터를 통해 학습효과를 검증하였다(Table 1. 참조).

실험 Ⅱ

Table 2에서 보듯이 EPD 프로모터 서열 데이터에서 -65 ~ 0 영역의 서열을 positive 데이터(1,768개)로 35 ~ 100 영역의 서열을 negative 데이터(1,729개)로 하고 인코딩할 때에는 개별 위치의 핵산 잔기가 독립적인 존재로 인식할 수 있도록 인코딩하고 도트 커널과 다항식 커널 각각에 대해 학습시키고 검정 데이터를 통해 학습효과를 검증하였다(Table 2. 참조).

실험 Ⅲ

Table 3에서 보듯이 EPD 프로모터 서열 데이터에서 -150 ~ -85 영역의 서열을 positive 데이터(1,758개)로 35 ~ 100 영역의 서열을 negative 데이터(1,730개)로 하고 인코딩할 때에는 개별 위치의 핵산 잔기가 독립적인 존재로 인식할 수 있도록 인코딩하고 도트 커널과 다항식 커널 각각에 대해 학습시키고 검정 데이터를 통해 학습효과를 검증하였다(Table 3. 참조).

실험 Ⅳ

Table 4에서 보듯이 EPD 프로모터 서열 데이터에서 -50 ~ 50 영역의 서열을 positive 데이터(1,749개)로 -200 ~ -100 영역의 서열을 negative 데이터(1,754개)로 하고 인코딩할 때에는 개별 위치의 핵산 잔기가 독립적인 존재로 인식할 수 있도록 인코딩하고 도트 커널과 다항식 커널 각각에 대해 학습시키고 검정 데이터를 통해 학습효과를 검증하였다(Table 4. 참조).

실험 Ⅴ

Table 5에서 보듯이 EPD 프로모터 서열 데이터에서 -50 ~ 50 영역의 서열을 positive 데이터(1,750개)로 -200 ~ -100 영역의 서열을 negative 데이터(1,754개)로 구성하였다. 그리고 이전 실험에서와는 달리 인코딩할 때에는 개별 위치의 핵산 잔기를 독립적으로(즉, 네 가지의 핵산에 대해 개별적으로 간주하여 인코딩함) 간주하지 않고 인접한 두개의 핵산들 간의 상호의존성을 반영할 수 있도록 인코딩하였다. 즉, 16개(즉, 4 x 4)의 신호로 인코딩하였다. 이전의 실험에서 마찬가지로 도트 커널과 다항식 커널 각각에 대해 학습시키고 검정 데이터를 통해 학습효과를 검증하였다(Table 5. 참조).

실험 Ⅵ

Table 6에서 보듯이 EPD 프로모터 서열 데이터에서 -50 ~

Table 1. The performance result of the first experiment in which each nucleotide was encoded as an independent one. "+" and "-" symbols represent positive and negative data respectively.

		학습 데이터		검정 데이터	
데이터 종류		+	-	+	-
데이터 수		1,762	1,734	100	100
도트 커널	민감성(S_n)	53.0 %			
	특이성(S_p)	50.0 %			
다항식 커널	민감성(S_n)	59.0 %			
	특이성(S_p)	53.2 %			

Table 2. The performance result of the second experiment in which each nucleotide was encoded as an independent one. "+" and "-" symbols represent positive and negative data respectively.

		학습 데이터		검정 데이터	
데이터 종류		+	-	+	-
데이터 수		1,768	1,729	100	100
도트 커널	민감성(S_n)	61.0 %			
	특이성(S_p)	51.7 %			
다항식 커널	민감성(S_n)	57.0 %			
	특이성(S_p)	55.3 %			

Table 3. The performance result of the third experiment in which each nucleotide was encoded as an independent one. "+" and "-" symbols represent positive and negative data respectively.

		학습 데이터		검정 데이터	
데이터 종류		+	-	+	-
데이터 수		1,758	1,730	100	100
도트 커널	민감성(S_n)	42.0 %			
	특이성(S_p)	44.7 %			
다항식 커널	민감성(S_n)	58.0 %			
	특이성(S_p)	52.7 %			

Table 4. The performance result of the fourth experiment in which each nucleotide was encoded as an independent one. "+" and "-" symbols represent positive and negative data respectively.

		학습 데이터		검정 데이터	
데이터 종류		+	-	+	-
데이터 수		1,749	1,754	100	100
도트 커널	민감성(S_n)	63.0 %			
	특이성(S_p)	57.3 %			
다항식 커널	민감성(S_n)	49.0 %			
	특이성(S_p)	61.2 %			

Table 5. The performance result of the fifth experiment in which nucleotides were encoded by dinucleotide, reflecting the dependency between two. "+" and "-" symbols represent positive and negative data respectively.

		학습 데이터		검정 데이터	
데이터 종류		+	-	+	-
데이터 수		1,750	1,754	100	100
도트 커널	민감성(S_n)	52.0 %			
	특이성(S_p)	53.1 %			
다항식 커널	민감성(S_n)	52.0 %			
	특이성(S_p)	60.1 %			

Table 6. The performance result of the sixth experiment in which nucleotides were encoded by trinucleotide, reflecting the dependency among them. "+" and "-" symbols represent positive and negative data respectively.

		학습 데이터		검정 데이터	
데이터 종류		+	-	+	-
데이터 수		1,749	1,703	100	100
도트 커널	민감성(S_n)	49.0 %			
	특이성(S_p)	51.0 %			
다항식 커널	민감성(S_n)	50.0 %			
	특이성(S_p)	57.5 %			

50 영역의 서열을 positive 데이터(1,749개)로 -200 ~ -100 영역의 서열을 negative 데이터(1,703개)로 구성하였다. 앞의 실험에서 보다는 더 확장해서 인접한 세 개의 핵산들 간의 상호의존성을 반영할 수 있도록 인코딩하였다. 즉, 64개(즉, 4 x 4 x 4)의 신호로 인코딩하였다. 이전의 실험에서 마찬가지로 도트 커널과 다항식 커널 각각에 대해 학습시키고 검정 데이터를 통해 학습효과를 검증하였다(Table 6. 참조).

실험 Ⅶ

Table 7에서 보듯이 EPD 프로모터 서열 데이터에서 -50 ~ 50 영역의 서열을 positive 데이터(1,747개)로 -200 ~ -100 영역의 서열을 negative 데이터(1,754개)로 구성하였다. 앞의 실험들에서 사용했던 인코딩 방법 보다는 더 확장된 개념에서 인접한 여섯 개의 핵산들 간의 상호의존성을 반영할 수 있도록 인코딩하였다. 즉, Hexamer를 단일체로 인코딩하였다(즉, 가능한 모든 경우의 입력신호의 수는 4^6개). 이전의 실험에서와 마찬가지로 도트 커널과 다항식 커널 각각에 대해 학습시키고 검정 데이터를 통해 학습효과를 검증하였다(Table 7. 참조).

실험 Ⅷ

Table 8에서 보듯이 EPD 프로모터 서열 데이터에서 -250 ~ 50 영역의 서열을 positive 데이터(840개)로 하고, GENIE 시스템의 CDS 데이터를 negative 데이터(840개)로 구성하였다. 게다가 positive 데이터와 negative 데이터의 조성(composition)의 차이를 반영할 수 있도록 인코딩할 때 각 데이터들의 G+C%도 인코딩시켰다. 인코딩은 개별 핵산을 단일 독립체로 간주할 수 있도록 인코딩하였다. 이전의 실험에서와 마찬가지로 도트 커널과 다항식 커널 각각에 대해 학습시키고 검정 데이터를 통해 학습효과를 검증하였다(Table 8. 참조).

Table 7. The performance result of the seventh experiment in which nucleotides were encoded by hexamer, reflecting the dependency among them. "+" and "-" symbols represent positive and negative data respectively.

		학습 데이터		검정 데이터	
데이터 종류		+	-	+	-
데이터 수		1,747	1,754	100	100
도트 커널	민감성(S_n)	46.0 %			
	특이성(S_p)	49.5 %			
다항식 커널	민감성(S_n)	51.0 %			
	특이성(S_p)	60.7 %			

Table 8. The performance result of the eighth experiment in which nucleotides were encoded by single unit and G+C% also was encoded to reflect the composition difference between positive and negative data. "+" and "-" symbols represent positive and negative data respectively.

		학습 데이터		검정 데이터	
데이터 종류		+	-	+	-
데이터 수		840	840	50	50
도트 커널	민감성(S_n)	56.0 %			
	특이성(S_p)	54.9 %			
다항식 커널	민감성(S_n)	66.0 %			
	특이성(S_p)	63.5 %			

결과 및 고찰

본 논문에서는 다양한 인코딩 방법 및 인코딩 영역선택을 통해서 유전자 프로모터 예측을 위한 SVM 활용 방안을 나름대로 제시하고자 하였다. 앞의 다양한 실험들의 결과를 통해서 알 수 있듯이 예상대로 커널함수의 측면에서 본다면 다항식 커널이 도트 커널보다 대체로 더 나은 성능을 보였다. 그러나 인코딩 대상영역과 인코딩 방식에 따라 어떤 경우에는 오히려 도트 커널이 더 좋은 성능 결과를 보였다(실험 IV 및 Table 4 참조). 유전자 예측이나 다양한 신호부위 예측 등에서 인접한 핵산들 간의 상호의존성을 반영하기 위해 k-mer를 주로 사용하고 있고, 또한 일반적으로 k값이 높을수록 좋은 것으로 인식되고 있으나, 본 논문의 실험 IV, V, VI

및 VII (각 실험에서 k값을 각각 1, 2, 3, 6 으로 사용하였음) 등의 결과를 비교해보면 알 수 있듯이 개별 핵산을 독립적 개체로 인코딩하였을 경우 예상과는 반대로 오히려 더 높은 성능결과를 보였다. 그리고 인접한 핵산간의 의존성을 반영하기 위해 2개, 3개, 및 6개를 단일체로 보고 인코딩하였을 경우에는 2개, 즉 dimer 기준으로 인코딩하였을 경우의 다항식 커널이 trimer와 hexamer 단위로 인코딩 했을 때의 도트 커널과 다항식 커널 보다는 더 좋은 성능을 나타냈다. 그리고 positive와 negative의 경계가 뚜렷한 데이터를 사용하였을 뿐만 아니라 새로운 특성, 즉 G+C% 값까지 인코딩한 8번째 실험에는 도트 커널과 다항식 커널 둘 다 다른 실험결과에 비해 상대적으로 좋은 성능을 보였고, 특히 다항식 커널의 경우 앞의 모든 실험결과 보다도 훨씬 좋은 성능을 보여주었다. 게다가 Fickett와 Hatzigeorgiou이 기존의 다른 프로그램들의 성능을 비교분석한 결과와 비교해 보더라도 본 연구에서 행해진 실험결과들은 중상위 이상의 성능을 보여준 것으로 여겨진다[1].

본 논문의 실험결과를 토대로 알 수 있는 사실은 생물학적 신호부위의 특성을 잘 반영할 수 있도록 생물학적 문맥(Biological Context)을 잘 투영할 수 있게 인코딩하여야 하며, 단순히 프로모터 영역뿐만 아니라 인터론(Intron), 유전자간 영역(Intergenic region), 및 엑손(Exon) 영역까지도 학습 범주로 포함한다면 SVM의 성능을 훨씬 향상시킬 수 있다는 것이다. 게다가 일반적으로 다항식 커널이 더 강력하지만 경우에 따라서는 주어진 상황과 활용 방도에 따라서 단순한 도트 커널이 더 나을 수 도 있다는 것이다. 즉, 사용자의 용도와 데이터의 특성에 따라 적합한 커널함수를 사용해야 한다는 사실이다.

요 약

유전가의 구조 예측 및 발현 기작에 대한 연구는 매우 중요한 사안으로 대두되고 있다. 특히 유전자 발현 제어에 중요한 역할을 하는 프로모터 영역을 예측하는 것은 전체 생명체 네트워크 규명을 위한 단초를 제공하기 때문에 많은 연구가 이루어지고 있다. 본 논문에서는 이러한 진핵생물의 유전자 프로모터 예측을 위한 Support Vector Machine (SVM) 활용방안에 대한 연구내용을 다루고 있다. 특성 벡터 값 생성을 위한 인코딩 방법 및 학습 데이터들의 구성에 대한 다양한 실험을 통해 SVM 활용 방안에 대한 올바른 방향을 제시하고 있다.

참 고 문 헌

1. Fickett, J. W. and A. C. Hatzigeorgiou. 1997. Eukaryotic promoter recognition. *Genome Res.* **7**, 839-844.
2. Fofanov, Y., Y. Luo, C. Katili, J. Wang, Y. Belosludtsev, T. Powdrill, C. Belapurkar, V. Fofanov, T. Li, S. Chumakov and B. Pettitt. 2004. How independent are the appearance of n-mers in different genomes. *Bioinformatics* **20**, 2421-2428.
3. Gangal, R. and P. Sharma. 2005. Human pol II promoter prediction: time series descriptors and machine learning. *Nucleic Acids Res.* **33(4)**, 1332-1336.
4. Gordon, L., A. Chervonenkis, A. Gammerman, I. Shahmuradov and V. Solovyev. 2003. Sequence alignment kernel for recognition of promoter regions. *Bioinformatics* **19(15)**, 1964-1971.
5. Joachims, T. 1999. *Advances in Kernel Methods - Support Vector Learning*. pp. 169-184, MIT Press. Cambridge, MA USA.
6. Jung, M., W. Park and K. Kim. 2004. Development of integrated system for motif and domain search. *Journal of Life Science* **14(6)**, 991-996.
7. Jung, M., W. Park and K. Kim. 2004. Development of web-based assistant system for protein-protein interaction and function analysis. *Journal of Life Science* **14(6)**, 997-1002.
8. Kulp, D, D. Haussler, M.G. Reese and F. H. Eeckman. 1996. A generalized Hidden Markov Model for the recognition of human genes in DNA. *Proc Int Conf Intell Syst Mol Biol.* **4**, 134-42
9. Perier, R. C., V. Praz, T. Junier, C. Bonnard and P. Bucher. 2000. The Eukaryotic Promoter Database (EPD). *Nucleic Acids Research* **28**, 302-303.
10. SVMlight, http://svmlight.joachims.org/.
11. Zhang, Y., C. Chu, Y. Chen, H. Zha and X. Ji. 2006. Splice site prediction using support vector machines with a Bayes kernel. *Expert Systems with Applications* **30**, 73-81.
12. http://www.integratedgenomics.com