ISSN 1598-0170 (Print)
ISSN 2287-1136 (Online)
http://www.jksii.or.kr

지식기반 유전자알고리즘을 이용한 한국인 빈발 HLA 대립유전자에 대한 결합 펩타이드 예측☆

Knowledge based Genetic Algorithm for the Prediction of Peptides binding to HLA alleles common in Koreans

조 연 진* 오 흥 범** 김 현 철***
Yeon-Jin Cho Heung-Bum Oh Hyeoncheol Kim

요 약

감염된 미생물에서 유래한 단백질 펩타이드가 HLA에 결합하여 숙주의 세포표면에 제시되면, T 세포가 이를 인식하여 면역반응을 유발함으로써 감염원을 제거하게 된다. HLA와 펩타이드간의 결합이 안정적일수록 T 세포반응이 강하게 일어나 효율적으로 감염원을 제거할 수 있다고 알려져 있다. 따라서 특정 HLA에 안정적으로 결합할 수 있는 펩타이드(HLA binder)를 찾아낼 수 있다면 감염질환이나 암의 예방을 위한 펩타이드 백신의 개발에 활용될 수 있다. 그런데 HLA는 매우 다형하기 때문에 하나의 집단 내에서도 어느 정도의 빈도를 가지는 대립유전자의 수가 매우 많다. 따라서 이들 모든 대립유전자들에 대해 가능한 펩타이드조합을 제작한 후 직접 실험을 통해 안정적으로 결합하는 펩타이드를 찾아내는 것은 매우 비효율적이다. 이를 극복하기 위하여 특정 HLA에 안정적으로 결합하는 펩타이드를 예측하는 정보전산적인 방법이 최근 개발되어 왔다. 이들 방법을 통해 제시된 펩타이드에 대해서만 직접 생물학적 실험을 시행함으로써 연구자는 검증해야 할 후보 펩타이드의 수를 현격히 감소시킬 수 있게 된다. 본 논문에서는 HLA 결합 펩타이드 예측을 위해 기계학습을 이용한 방법을 소개할 뿐만 아니라, 지금까지 HLA 결합 펩타이드 예측에 시도된 적이 없는 '지식기반 유전자알고리즘(knowledge-based genetic algorithm)'이라는 새로운 모델을 제시하고자 한다. 이것은 유전자알고리즘(GA)에 기반한 것이었지만 전문가 지식을 접목함으로써 GA보다 더 향상된 성능으로 한국인에 흔한 HLA에 결합하는 펩타이드를 예측하였다. 뿐만 아니라 이것은 결합하는 펩타이드의 규칙을 한국인에 흔한 HLA 대립유전자에 대하여 추출해 줄 수 있는 새로운 방법이었다.

ABSTRACT

T cells induce immune responses and thereby eliminate infected micro-organisms when peptides from the microbial proteins are bound to HLAs in the host cell surfaces, It is known that the more stable the binding of peptide to HLA is, the stronger the T cell response gets to remove more effectively the source of infection. Accordingly, if peptides (HLA binder) which can be bound stably to a certain HLA are found, those peptieds are utilized to the development of peptide vaccine to prevent infectious diseases or even to cancer. However, HLA is highly polymorphic so that HLA has a large number of alleles with some frequencies even in one population. Therefore, it is very inefficient to find the peptides stably bound to a number of HLAs by testing random possible peptides for all the various alleles frequent in the population. In order to solve this problem, computational methods have recently been developed to predict peptides which are stably bound to a certain HLA. These methods could markedly decrease the number of candidate peptides to be examined by biological experiments. Accordingly, this paper not only introduces a method of machine learning to predict peptides binding to an HLA, but also suggests a new prediction model so called 'knowledge-based genetic algorithm' that has never been tried for HLA binding peptide prediction. Although based on genetic algorithm (GA). it showed more enhanced performance than GA by incorporating expert knowledge in the process of the algorithm. Furthermore, it could extract rules predicting the binding peptide of the HLA alleles common in Koreans.

☞ keyword : Machine learning(기계학습), Genetic algorithm(유전자알고리즘), Knowledge based genetic algorithm (지식기반 유전자알고리즘), Rule generation(규칙생성), Prediction model(예측모델), HLA binding peptide prediction(HLA 결합 펩타이드 예측), Korean common HLA allele(한국인 빈발 HLA 대립유전자)

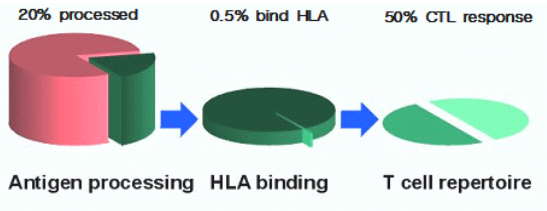
* 정 회 원 : (주)원일테크 연구원
jx@hanmail.net
** 정 회 원 : 울산대학교 의과대학 진단검사의학과 교수
hboh@amc.seoul.kr
*** 정 회 원 : 고려대학교 컴퓨터교육과 교수
harrykim@korea.ac.kr

[2012/07/17 투고 - 2012/07/20 심사 - 2012/08/16 심사완료]
☆ 이 논문은 2009년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임.[NRF-2009-351-H00018]

1. 서 론

HLA (Human leukocyte antigen)는 사람이 가지고 있는 유전자 중에서 가장 심한 다형성(polymorphism)을 보이는 유전자이다. 감염 미생물에서 유래한 펩타이드를 끼워 T 세포에 항원제시(presentation)함으로써 면역반응을 유도하는 것으로 알려져 있다. 이때 HLA에 의해 T 세포에 제시되어 면역반응을 유발하는 펩타이드를 T 세포 에피토프(epitope)라 한다[1].

HLA와 강하게 결합(binding)하는 T 세포 에피토프는 강한 T 세포 반응을 유도(selection hypothesis)한다. 따라서 감염의 예방을 위하여 미생물 고유의 단백질 중에서도 강한 면역반응을 유발하는 펩타이드를 발굴하거나, 암의 치료에 있어서 특정 암에서만 주로 발현하는 단백질 중 강한 면역반응을 유발하는 펩타이드를 발굴할 수 있다면 이들을 이용한 펩타이드 백신 혹은 치료제를 개발할 수 있다[2, 3]. 그러나 HLA에 결합되는 펩타이드를 발견하는 것은 매우 어려운 문제이다. 왜냐하면 (그림 1)과 같이 항원 단백질의 20%만 processing되고, 그 중 0.5%가 HLA 틈새(cleft)에 결합되며, 또 그 중 절반만이 CTL (Cytotoxic T cell)에 반응하는 것으로 알려져 있기 때문이다[1, 3]. 결국 1/2000의 펩타이드가 면역학적 반응성을 나타낸다(immunogenic)고 할 수 있는 것이다. 또한 immunogenic 펩타이드가 HLA 대립유전자에 따라 다르기 때문에 여러 종류의 대립유전자에서 면역성이 높은 펩타이드를 알아내는 과정은 매우 어려운 일이다. 그러므로 너무 많은 종류의 HLA에 대하여 결합 가능한 후보 펩타이드를 모두 실험실에서 직접 실험하여 T 세포 에피토프 여부를 가려낸다는 것은 현실적으로 매우 비효율적이라 할 수 있다[4, 5].



(그림 1) From proteins to immunogens

따라서 최근에는 기계학습 방법을 이용하여 특정 HLA 대립유전자에 결합 가능한 펩타이드를 예측하고 예측된 펩타이드를 기반으로 실험하게 함으로서 연구자가 펩타이드 합성(synthesise)과 실험해야 할 후보 binders의 수를 감소시켜 연구실험 비용을 효율적으로 줄이고자 하는 노력이 시도되고 있다[3, 4, 5].

본 논문에서는 HLA 결합 펩타이드 예측을 위해 기계학습 알고리즘을 이용한 방법들을 소개하고, 지금까지 HLA 결합 펩타이드 예측에 시도된 적이 없는 유전자알고리즘으로 HLA 대립유전자에 대한 결합 펩타이드 규칙을 추출한다. 이는 지금까지 여러 연구자들에 의해 시도되었던 인공신경망을 이용하여 예측 모델을 만들고, 이로부터 규칙 추출 알고리즘(Ordered Attribute Search, OAS)[6]을 이용하여 규칙을 추출한 후 다시 의·생물학 분야에서 높은 성능을 보여주고 있는 유전자알고리즘의 초기해 생성에 활용함으로써 더 다양하고 정확한 binder를 생성하는 예측 모델을 구축하는 것이다. 이를 본 논문에서는 지식기반 유전자알고리즘(Knowledge Based Genetic Algorithm, KBGA)이라고 제안한다. 그리고 지식기반 유전자알고리즘을 이용하여 한국인 빈발 HLA 대립유전자[7] 중 HLA class Ⅰ의 HLA - A, HLA-B에서 5% 이상의 높은 빈도를 보이는 A*2402 (22.5%), A*0201 (15.7%), A*3303 (14.4%), A*1101 (11.0%), A*0206 (8.9%), A*2601 (5.2%), B*5101 (12.1%), B*1501 (8.7%), B*4403 (7.4%), B*3501 (6.6%), B*4601 (6.2%), B*5801 (5.8%), B*5401 (5.0%)의 결합 펩타이드를 예측하는 규칙을 추출하였다. 추출된 분류 규칙은 다른 기계학습 알고리즘에서 발견되지 않은 다양하고 높은 성능의 새로운 규칙을 생성하였다.

2. HLA 결합 펩타이드의 규칙추출 방법

현재 알려져 있는 한국인의 HLA 대립유전자 분포는 건강한 한국인 309명을 대상으로 분석하여 HLA-A, HLA-B 대립유전자 빈도를 구분하였다[7]. HLA-A 대립유전자는 22종, HLA-B는 41종이 동정되었는데, 이들의 혈청학적 표현형은 각각 HLA-A 11종 HLA-B 29종이었다.

HLA 결합 펩타이드 예측 모델을 만들기 위해 실험에 사용한 학습 데이터는 SYFPEITHI 데이터베이스[8]에서 753개 펩타이드와 MHCPEP 데이터베이스[9]에서 4,539개 펩타이드가 사용되었다.

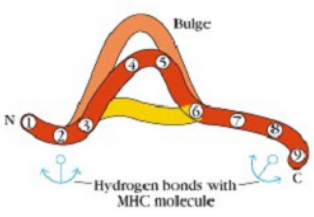
SYFPEITHI 데이터베이스는 논문으로 발표된 것들 중에서도 natural ligand에 대한 펩타이드 자료만으로 만들어진 것이므로 좀 더 양질의 것이라 할 수 있다.

실험에서는 SYFPEITHI 데이터베이스와 MHCPEP 데이터베이스의 binding 펩타이드 데이터를 사용하여 학습시키고, 각 데이터베이스에 포함되는 대립유전자에 대해

비교 평가하였다. Non-binder는 ENSEMBL 데이터베이스로부터 단백질을 무작위로 추출하여 일정한 크기로 자른 다음에 HLA 펩타이드 데이터베이스에 들어있는 서열을 모두 제거한 후 사용하였고, binder와 non-binder의 1:2 비율로 1764개 binder와 3528개 non-binder인 총 5292개의 펩타이드가 학습에 이용되었다.

2.1 HLA 결합 펩타이드의 규칙-motif 표현

일반적으로 HLA class Ⅰ분자의 경우에는 평균 9개의 아미노산이 틈새(cleft)에 들어간다. 이때 펩타이드는 HLA 틈새에 "똑딱 단추"처럼 결합(anchorage)이 이루어지는데, 그 위치는 거의 일정하게 2번째 아미노산(P2)과 carboxy-terminal(PC)인 것으로 알려져 있다. 여기서 P2는 class Ⅰ틈새에 결합하는 펩타이드 중에서 두 번째 위치의 아미노산 잔기를 의미한다. 결합은 우선 펩타이드의 carboxy-terminal에서 먼저 일어나고, 이후 P2가 틈새 내에서 적절한 위치를 찾는 방식으로 일어난다(그림 2).



(그림 2) HLA class Ⅰ에 결합된 펩타이드 모식도

HLA 결합 펩타이드에서 사용되는 규칙 즉, motif는 유사한 기능을 수행하거나 비슷한 구조를 이루기 위해서 서열의 부분적인 보존 영역 또는 서열 집합이 공유하는 짧은 서열 패턴을 의미한다.

x-[RK]-x-x-[RH]-x-x-x-[L]

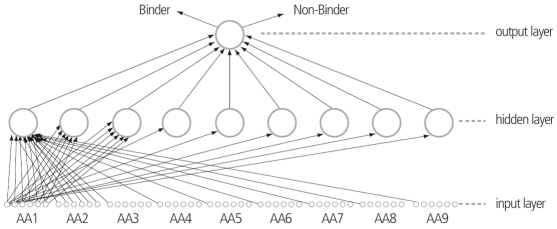
위의 정규식은 HLA-B*14 서열의 motif를 나타낸 것으로, 표현 의미는 아래와 같다.

x : 임의의 아미노산 잔기(residue)
[RK] : Arginine이나 Lysine 아미노산 잔기 중 하나
[RH] : Arginine이나 Histidine 아미노산 잔기 중 하나
L : Leucine 아미노산 잔기

이러한 HLA 결합 펩타이드의 motif 표현은 유전자알고리즘의 염색체를 표현하는 스키마 형식과 동일하다. 즉 유전자알고리즘의 규칙은 스키마로 표현되는데, 특정기호(아미노산 잔기)와 무관기호(*: 임의의 아미노산 잔기)로 정의하여 HLA binding 펩타이드 패턴을 표현한다.

2.2 인공신경망

인공신경망을 이용하여 HLA 결합 펩타이드를 예측하는 과정에서는 입력 데이터로 사용되는 binder와 non-binder의 아미노산 각각을 20bit로 인코딩하게 된다. 각 아미노산은 1bit 만 1이고 나머지 19bit는 0인 형태로 표현한다. 즉, Alanine (A)는 '10000000000000000000'으로 변환되어, 9 mer의 경우 총 180(=9×20)개의 입력(input) 노드를 갖는다[5]. 그리고 1~10개의 은닉 노드별 실험을 통해 가장 좋은 결과를 보인 4개의 은닉 노드와 HLA binding 클래스(binder: 1, non-binder: 0)를 구분하는 1개의 출력(output) 노드를 갖는다(그림 3).



(그림 3) HLA binder 예측을 위한 인공신경망의 구성

이들 binder와 non binder를 학습데이터 세트로 하여 인공신경망을 반복적으로 돌리게 되면 인공신경망의 출력값과 실제 분류 값을 비교하면서 값의 차이에 대한 평균제곱오차(mean square error, MSE)가 최소가 되도록 수정하면서 학습해 나간다. 인공신경망은 HLA 결합 펩타이드 예측을 위해 그동안 가장 많이 사용되어왔던 알고리즘으로서 결정트리와 은닉마르코프 모델을 능가하는 높은 성능을 보여주었다.

Honeyman[1] 등은 인공신경망 방법을 이용하여 80%의 예측 정확성을 보고하였다.

그러나 인공신경망은 학습된 가중치의 의미를 쉽게 이해하기 어렵다는 단점이 있기 때문에 이를 해결하기 위하여 인공신경망으로부터 규칙을 추출하려는 연구가 많이 진행되어 왔다. 즉 HLA binder여부를 예측하면서 왜 binder가 되는지도 알기위해 최근에는 인공신경망으로부

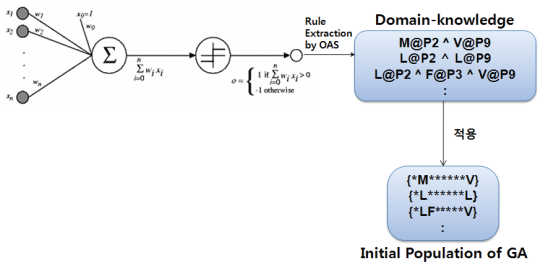
터 규칙을 추출하는 연구들이 보고되고 있으며 그 중의 하나가 OAS (ordered attribute search)알고리즘이다[6].

2.3 지식기반 유전자알고리즘

유전자알고리즘은 초기 개체집단(initial population)의 문자배당, 수치변환(coding) 등을 어떻게 설계하는지가 매우 중요한데, 아직까지 이론이 잘 정립되어 있지 않기 때문에 설계자의 경험에 의존하는 경향이 많다. 그리고 모든 연산들이 랜덤(random)하게 일어나기 때문에, 이미 검증된 유용한 지식들이 활용되지 않는 단점이 있다[10]. 따라서 유전자알고리즘을 의·생물학 자료에 효율적으로 활용하기 위해서는 유전자알고리즘에 생물학적 지식을 접목시키는 새로운 시도가 필요하다.

본 논문에서는 HLA 결합 펩타이드 정보를 OAS 알고리즘에 적용하여 초기 개체집단을 생성하고, PSSM을 이용하여 아미노산의 출현빈도에 따라 돌연변이 개체 선택율을 달리하였다. 이것을 본 논문에서는 지식기반 유전자알고리즘(knowledge-based genetic algorithm)이라 한다.

지식기반 유전자알고리즘에서의 개체표현은 예를 들어 OAS 알고리즘으로부터 "if M@P2 and V@P9 then binding"이라는 규칙을 추출하였다면, 이를 (그림 4)와 같이 '{*M******V}'로 인코딩하여 해당 아미노산과 '*' (don't care symbol)로 초기 개체집단을 설정한다는 점이 기존의 유전자알고리즘과 큰 차이점이다. 즉 개체를 구성할 때 어디에 어떤 아미노산과 '*' (don't care symbol)을 배당할지를 인공신경망에서 추출된 규칙인 domain knowledge를 이용한다는 것이다.



(그림 4) 학습된 OAS에서 추출된 domain knowledge를 이용한 KBGA

또한 돌연변이를 일으킬 때 position specific scoring matrix (PSSM)를 이용하여 아미노산의 위치와 종류에 따라 돌연변이 개체 선택율을 달리함으로써 HLA binding motif에 더 높은 비율의 해를 얻을 수 있도록 하였다. 예를 들어 (표 3)은 HLA-A*0201에 대한 PSSM인데 matrix의 값 중에는 음수의 경우도 있기 때문에 약간의 변형을 통해 돌연변이 개체 선택율로 이용해야 한다. 예를 들어 1번 위치에 alanine (A)이 오는 경우 -1.495의 값을 가지므로 이를 양수로 변형하기 위하여 exponential 함수를 적용하면 0.224로 변형할 수 있다 [e=0.024]. 이렇게 (표 3)을 모두 변환한 다음에 룰렛 알고리즘을 이용하여 돌연변이 개체로 사용 할 수 있다. 그렇게 되면 20개의 아미노산 중에서 랜덤하게 임의의 아미노산으로 변형하는 것이 아니라, 중요 binder 부위가 더 높은 비율로 돌연변이를 일으키면서 binder를 찾을 확률이 각 위치의 scoring matrix 만큼 높아지게 된다.

(표 3) HLA-A*0201의 PSSM

a.a.	A	R	N	D	C
1	-1.495	-1.083	-3.268	-6.993	-5.672
2	-5.118	-6.546	-7.537	-7.9	-5.94
3	1.653	-6.084	-0.625	-0.932	-4.561
4	0.233	-2.55	0.223	3.68	-2.691
5	0.456	-2.353	-1.165	-0.328	-2.24
6	-0.183	-4.935	-1.439	-2.314	-2.294
7	0.842	-5.104	-0.745	-1.178	-3.143
8	0.527	-2.046	0.15	-4.452	-3.716
9	0.067	-7.04	-7.545	-7.803	-5.58
a.a.	Q	E	G	H	I
1	-5.302	-6.296	-1.494	-3.323	-1.891
2	-0.148	-6.933	-7.957	-6.958	-0.14
3	-1.012	-5.995	-2.802	-2.136	1.114
4	-0.468	3.53	-0.495	-0.082	-6.32
5	-1.171	-0.066	0.422	1.443	0.083
6	-0.718	-2.99	-2.508	-0.364	1.876
7	0.341	-0.974	-5.031	0.762	-0.665
8	0.256	1	-4.76	-0.42	-1.666
9	-6.693	-7.1	2.093	2.656	-6.852
a.a.	L	K	M	F	P
1	-1.261	0.458	2.146	5.205	-7.122
2	5.337	-6.603	5.163	-4.431	-7.398
3	0.827	-5.313	4.237	2.488	-1.366
4	-5.758	-1.677	-3.271	-2.634	2.448
5	-0.455	-1.191	0.034	0.566	-3.076
6	2.225	-4.765	0.738	1.324	-0.059
7	-1.278	-4.021	1.226	3.125	2.873
8	-0.42	-1.666	-1.172	1.606	2.971
9	2.656	-6.852	-2.582	-5.071	-7.101

a.a.	S	T	W	Y	V
1	0.029	-2.169	2.477	4.955	-2.162
2	-6.417	-3.863	-6.405	-5.692	-1.767
3	0.796	-2.155	1.901	1.864	-1.729
4	1.232	-1.244	0.078	-0.909	-4.081
5	0.661	0.093	2.659	2.797	-0.274
6	-0.178	0.564	-0.482	0.723	0.823
7	-0.628	0.769	3.524	2.441	-1.481
8	0.404	-0.204	0.926	3.378	-3.245
9	-5.748	-3.842	-6.9	-5.901	5.257

a.a.: amino acid

이러한 지식기반 유전자알고리즘을 기반으로 한 예측 모델은 기존의 인공신경망을 기반으로 개발된 Net MHCpan과 NetMHC 보다 평균 20% 더 높은 예측율을 나타내었다[11].

3. 실험결과 및 생성된 규칙의 성능

본 논문에서는 HLA 결합 펩타이드 예측 규칙을 생성하기 위해 처음으로 유전자알고리즘 사용을 시도하였다. 그리고 유전자알고리즘 초기 개체집단의 생성과 돌연변이율을 결정하는데 이미 알려져 있는 생물학적 지식(domain-knowledge)을 이용하였다. 이러한 지식기반 유전자알고리즘을 사용한 경우, 다른 기계학습들 보다 더 다양하고 정확한 규칙을 추출하여 우수한 예측 성능을 나타내었다. 이것은 예측율의 성능 평가를 통해 확인할 수 있었다. HLA 결합 규칙의 예측율(PPV : positive predictive value)은 다음과 같이 정의하였다.

$$PPV = \frac{num\ of\ true\ positive\ examples}{num\ of\ examples\ matched\ by\ the\ condition\ part} = \frac{TP}{(TP+FP)} \times 100$$

HLA binder 예측 규칙과 같은 경우는, binder와 non-binder 정보를 이용했다 하더라도, 생성된 규칙이 '왜 binder가 아닌지'에 대한 negative (non-binder) 규칙이 아닌, '왜 binder인지' 에 대한 positive (binder) 규칙이기 때문에, 규칙 성능은 TP (true positive)와 FP (false positive)로만 사용하여 성능을 평가한다. 따라서 이 경우에는 성능 평가 척도를 P/FP로 표시하거나 혹은 PPV (positive predictive value)로 나타낸다[12].

여기서 TP는 MHCPEP, SYFPEITHI DB에서 구한 binder를 기준으로 정의하였고, FP는 ENSEMBL DB에서 binder를 제외하고 랜덤하게 구한 non-binder를 기준으로 정의하였다.

인공신경망은 데이터 셋의 size가 20개 이상일 때 유의미한 결과를 보인다는 Pierre[4]의 실험결과에 따라, size가 20개를 유지하는 3-fold cross validation방법을 사용하였다. 입력노드 180개, 은닉노드 2개, 출력노드 1개의 구조로, 각 DB의 alleles은 평균 90%의 일반화 성능과 100%의 학습 정확률을 나타내었다.

지식기반 유전자알고리즘과 기계학습 알고리즘에서 추출된 HLA 결합 규칙이 실제로 유용한지에 대한 타당성은 기존에 실험적으로 알려져 있는 HLA Facts Book*과 비교하여 검증하였다. 그 결과 지식기반 유전자알고리즘에서 생성한 결합 규칙은 HLA FactsBook의 motif를 모두 포함하고 있어(표 4), KBGA의 HLA 결합 규칙이 실제 사용가능하다는 신뢰성을 확인하였다. 또한 KBGA를 통해 새롭게 발견된 규칙은 음영색으로 표시하였다(표 5).

(표 4) HLA-A*2402에 대한 결합 규칙과 PPV

Classifier	Positive Rule	Motif	TP/FP (PPV)
HLA FactsBook	F@P2	*F*******	16/5 (76.2)
	I@P9	********I	20/5 (80.0)
	F@P9	********F	14/3 (82.4)
	Y@P2	*Y*******	46/5 (90.2)
	Y@P2 ^ L@P9	*Y******L*	28/2 (93.3)
OAS알고리즘	Y@P2 ^ L@P9	*Y******L*	28/2 (93.3)
지식기반 유전자 알고리즘	F@P2	*F*******	16/5 (76.2)
	I@P9	********I	20/5 (80.0)
	F@P9	********F	14/3 (82.4)
	Y@P2	*Y*******	46/5 (90.2)
	Y@P2 ^ L@P9	*Y******L*	28/2 (93.3)

(표 5) HLA-B*5101에 대한 결합 규칙과 PPV

Classifier	Positive Rule	Motif	TP/FP (PPV)
HLA FactsBook	I@P9	********I	20/2 (90.9)
OAS알고리즘	P@P2 ^ I@P9	*P******I	13/0 (100.0)
지식기반 유전자 알고리즘	P@P2	*P*******	24/3 (88.9)
	I@P9	********I	20/2 (90.9)
	P@P2 ^ I@P9	*P******I	13/0 (100.0)

* HLA FactsBook은 생물학 연구실 실험을 통해 저널에 발표된 결과들을 수집한 것으로 HLA 유전자에 대한 그동안의 지식을 총괄하여 HLA분야에서는 백과사전과 같은 교재이다.

지식기반 유전자알고리즘을 이용하여 HLA 결합 펩타이드의 'if-then' 규칙을 추출한 결과 HLA A*2402에서 예측율 93.3%의 "If Y@P2 ^ L@P9 then binding"인 규칙을 추출하였다. 이 규칙의 의미는 펩타이드의 2번째 아미노산이 'Y'이고, 9번째 아미노산이 'L'이면 HLA A*2402 틈새에 결합한다는 것이다. 이 결합 규칙은 motif 형태로도 표현할 수 있는데, 예를 들어 Y@P2 ^ L@P9을 motif 형식으로 표현하면 '*Y****** L'이 된다(표 4). 그런데 기계학습으로 생성된 규칙들 중에서 domain knowledge를 이용하지 않은, 일반적인 유전자알고리즘의 규칙들은 모두 예측율(PPV) 75% 이하여서 (표 4)와 (표 5)에 포함되지 못하였다.

(표 6) 한국인 빈발 HLA-A, -B에 대한 결합 펩타이드 (예측율 90%이상)

Allele (빈도)	Positive Rule	Motif
A*2402 (22.5%)	Y@P2 ^ L@P9	*[Y]******[L]
A*0201 (15.7%)	L@P2 ^ V@P9	*[L]******[V]
A*3303 (14.4%)	R@P9	********[R]
A*1101 (11.0%)	K@P9	********[K]
A*0206 (8.9%)	V@P2 ^ V@P9	*[V]******[V]
A*2601 (5.2%)	V@P2 ^ Y@P9	*[V]******[Y]
B*5101 (12.1%)	P@P2 ^ I@P9	*[P]******[I]
B*1501 (8.7%)	Q@P2 ^ F@P9	*[Q]******[F]
B*4403 (7.4%)	E@P2 ^ F@P9	*[E]******[F]
B*3501 (6.6%)	P@P2	*[P]*******
B*4601 (6.2%)	F@P9	********[F]
B*5801 (5.8%)	S@P2 ^ W@P9	*[S]******[W]
B*5401 (5.0%)	P@P2	*[P]*******

(표 6)은 한국인 빈발 HLA 대립유전자에 대한 결합 펩타이드 중, 5% 이상의 높은 빈도를 보이는 HLA-A 6종, HLA -B 7종에서 결합 규칙을 생성한 것이다. 지식기반 유전자알고리즘의 motif 규칙은 의·생물학분야에 임상적 적용이 가능하도록 예측율 90% 이상의 기준을 두어 나타내었다.

4. 결 론

유전자알고리즘은 학습 자료에 과적합(over-fitting)되지 않기 때문에 새로운 문제에 대해서 잘못된 해답을 낼 우려가 적고, 수렴된 해집합은 직접 확인할 수 있기 때문에 규칙을 추출할 수 있는 장점이 있다.

유전자알고리즘을 이용한 결합 펩타이드 규칙은 학습된 가중치 이면의 기호화된 의미를 해석하기 어려운 인공신경망이나 SVM 보다 HLA binder 예측의 이해와 해석을 위한 기계학습 방법으로서 더 뛰어난 점이라 할 수 있다.

지식기반 유전자알고리즘에서 생성된 결합 규칙은 motif 패턴으로 표시되는데, 이 motif의 규칙형태는 사용자로 하여금 예측된 결과를 쉽게 이해 할 수 있도록 도와주며 따라서 의·생물학분야에 임상적 적용이 가능하다는 장점이 있다.

본 논문에서는 (1) HLA 결합 펩타이드 정보를 OAS 알고리즘과 유전자알고리즘에 적용시켜 HLA 펩타이드 간의 결합 규칙을 생성하였고 (2) PSSM (position specific scoring matrix)을 유전자알고리즘의 돌연변이 연산에 적용함으로서 정확한 해를 찾을 가능성이 높아지게 하였다. 또한 (3) 지식기반 유전자알고리즘을 이용하여 한국인 빈발 HLA 대립유전자에 대해 90% 이상의 결합 펩타이드 규칙을 생성하였다. 이것은 유전자알고리즘의 초기 집단 생성에서 어떤 우수형질을 확보했는지의 여부가 결과에 큰 영향을 준다는 것과 돌연변이 생성에 관해 기존의 생물학적 지식을 이용하는 것이 예측 신뢰성을 더 높인다는 것을 확인해 주는 것이다.

HLA 결합 펩타이드를 전산적(생물정보학적)인 기법으로 예측하는 것은 최근 중요하게 다루어지고 있는 'Immunoinformatics' 기술의 하나이다. 그러나 이러한 기법은 실험 데이터의 부족으로 인하여 연구실에서의 생물학적 실험을 완벽하게 대체할 수는 없지만 필수적인 실험의 수를 최소로 압축하여, 연구자로 하여금 좀 더 정밀한 실험에 집중시켜 연구실험 비용 최소화 및 시간 절약이라는 장점을 누리도록 도와준다. 최근 이용할 수 있는 생물학적 자료들이 체계적으로 누적되고 있으며, 아울러 이를 효율적으로 다룰 수 있는 생물정보학적 기법 또한 빠르게 발전하고 있는 점을 고려하면 HLA 결합 펩타이드를 예측하는 시스템은 향후 계속해서 발전 구축되어질 전망이다.

HLA가 민족마다 서로 다른 특징을 가지고 있기 때문에 한국인을 대상으로 한 펩타이드 결합 연구는 반드시 시행되어야 하는데, 아직까지 국내에서는 이런 연구가 시행된 사례가 없다. 그리고 HLA-결합 펩타이드를 예측하는 시스템[3, 5]은 소수 국외에서 개발된 바 있으나 HLA-결합 예측 규칙을 추출하는 시스템은 아직 보고된 바가

없다. 본 연구에서 제시한 지식기반 유전자알고리즘은 그러한 생물정보학적 발전에 일부 기여할 것으로 사료된다.

향후 연구 과제로는 유전자알고리즘의 초기 집단 생성과 돌연변이에 사용한 도메인 지식을 선택과 교차 등의 다양한 연산에 적용함으로써 보다 도메인에 적합한 (domain specific) 지식기반 유전자알고리즘을 검토할 필요가 있다.

참 고 문 헌

[1] Brusic, V. Bajic, V.B. Petrovsky, N., 'Computational methods for prediction of T-cell epitopes -a framework for modelling, testing, and applications.', Elsevier Inc. Science Direct, pp.436-443, 2004.

[2] Lafuente, EM. and Reche, PA., 'Prediction of MHC-peptide binding: a systematic and comprehensive overview.', Curr Pharm Des, pp.3209- 3220, 2009.

[3] Zhang, L. Udaka, K. Mamitsuka, H. Zhu, S., 'Toward more accurate pan-specific MHC- peptide binding prediction: a review of current methods and tools.', Brief Bioinform, pp.350-364, 2011.

[4] Donnes, P. Kohlbacher, O., 'SVMHC: a server for prediction of MHC-binding peptides.', Nucleic Acids Research, pp.194-197, 2006.

[5] Tong, JC. Tan, TW. Ranganathan, S., 'Methods and protocols for prediction of immunogenic epitopes.', Brief Bioinform, pp.96-108, 2006.

[6] Kim H., 'Computationally Efficient Heuristics for If-Then Rule Extraction from Feed-Forward Neural Networks.', Lecture Notes in Artificial Intelligence, pp.170-182, 2000.

[7] 황상현, 오흥범, 양진혁, 권오중, 한국인의 HLA-A, -B, -C 대립유전자와 일배체형 분포, 대한진단검사의학회지, 제24권, 제 6호, pp.396-404, 2004.

[8] Rammensee, H.G. Bachmann, J. Emmerich, N.P. Bachor, O.A. and Stevanovic, 'SYFPEITHI:data base for MHC ligands and peptide motifs.',Immunogenetics, pp.213-219, 1999.

[9] Brusic, V. Rudy, G. Harrsison, LC., 'MHCPEP, a database of MHC-binding peptides:update 1997.', Nucleic Acids Research, pp.368-371, 1998.

[10] Fernandez, M. Caballero, J. Fernandez, L. Sarai A., 'Genetic algorithm optimization in drug design QSAR: Bayesian-regularized genetic neural networks (BRGNN) and genetic algorithm optimized support vectors machines (GA-SVM).', Molecular diversity, pp.269-289, 2011.

[11] Cho, YJ. Kim, H. OH, HB., 'Generating Rules for Predicting MHC Class I Binding Peptide using ANN and Knowledge-based GA.', jdcta International Journal, pp.111-119, 2009.

[12] Loong, TW., 'Understanding sensitivity and specificity with the right side of the brain.', BMJ, pp.716-719, 2003.

◐ 저 자 소 개 ◑

조 연 진

1998년 한신대학교 정보통신학과 졸업(학사)
2001년 명지대학교 교육대학원 전자계산교육학과 졸업(석사)
2009년 고려대학교 대학원 컴퓨터교육학과 졸업(박사)
2009년~2010년 울산대학교 의학과 연구원
2012년~현재 (주)원일테크 연구원
관심분야 : 바이오인포매틱스, 기계학습, 데이터마이닝
E-mail : jx@hanmail.net

오 흥 범

1988년 서울대학교 의학과 졸업(학사)
1996년 서울대학교 대학원 의학과 졸업(석사)
1999년 서울대학교 대학원 의학과 졸업(박사)
1998년~현재 울산대학교 의과대학 진단검사의학과 교수
관심분야 : HLA, 바이오인포매틱스, 데이터마이닝, etc.
E-mail : hboh@amc.seoul.kr

김 현 철

1988년 고려대학교 전산과학과 졸업(학사)
1990년 Univ. of Missouri - Rolla 졸업(석사)
1998년 Univ. of Florida 졸업(박사)
1998년 GTE Data Services, Inc. 시스템 분석가
1998년~1999년 삼성 SDS 책임컨설턴트
2005년~2006년 Univ. of Florida 대우교수
2010년~2011년 일본 홋카이도대학 정보기반센터 특임교수
1999년~현재 고려대학교 컴퓨터교육과 교수
관심분야 : 컴퓨터교육, 스마트러닝, 기계학습알고리즘
E-mail : harrykim@korea.ac.kr