http://dx.doi.org/10.3745/KIPSTD.2012.19D.1.021

SVM 모델을 이용한 3차원 패치 기반 단백질 상호작용 사이트 예측기법

박 성 희[†] · Björn Hansen[††]

요 약

모노머 단백질의 상호작용 사이트 예측은 기능을 알지 못하는 단백질에 대해서 이것과 상호작용하는 단백질로부터 기능을 예측하거나 단백질 도킹을 위한 검색 공간의 감소에 중요한 역할을 한다. 그러나 상호작용사이트 예측은 대부분 단백질 상호작용이 세포 내에서 순간적 반응에 일어나는 약한 상호작용으로 실험에 의한 3차원 결정 구조 식별의 어려움이 따르며 이로 인해 3차원의 복합체 데이터가 제한적으로 양산된다.

이 논문에서는 모노머 단백질의 3차원 패치 계산을 통하여 구조가 알려진 복합체의 상호작용사이트와 비상호작용사이트에 대한 패치 속성을 추출하고 이를 기반으로 Support Vector Machine (SVM) 분류기법을 이용한 예측 모델 개발을 제시한다. 타겟 클래스의 데이터 불균형 문제 해결을 위해 under-sampling 기법을 이용한다. 사용된 패치속성은 2차 구조 요소와 아미노산 구성으로부터 총 9개가 추출된다. 147개의 단백질 복합체에 대해서 10 fold cross validation을 통해서 다양한 분류모델의 성능 평가를 하였다. 평가한 분류 모델 중 SVM은 92.7%의 높은 정확성을 보이고 이를 이용하여 분류 모델을 개발하였다.

키워드 : 단백질 상호작용, 3차원 패치, 인터페이스, 바인딩 사이트 예측, SVM, 분류, 데이터 불균형

Prediction of Protein-Protein Interaction Sites Based on 3D Surface Patches Using SVM

Sung Hee Park[†] · Björn Hansen[††]

ABSTRACT

Predication of protein interaction sites for monomer structures can reduce the search space for protein docking and has been regarded as very significant for predicting unknown functions of proteins from their interacting proteins whose functions are known. In the other hand, the prediction of interaction sites has been limited in crystallizing weakly interacting complexes which are transient and do not form the complexes stable enough for obtaining experimental structures by crystallization or even NMR for the most important protein-protein interactions.

This work reports the calculation of 3D surface patches of complex structures and their properties and a machine learning approach to build a predictive model for the 3D surface patches in interaction and non-interaction sites using support vector machine. To overcome classification problems for class imbalanced data, we employed an under-sampling technique. 9 properties of the patches were calculated from amino acid compositions and secondary structure elements. With 10 fold cross validation, the predictive model built from SVM achieved an accuracy of 92.7% for classification of 3D patches in interaction and non-interaction sites from 147 complexes.

Keywords : Protein Interaction, 3D Patches, Interface, Prediction of Binding Sites, SVM, Classification, Imbalanced Data

1. 서 론

생체 내에서 단백질은 다른 단백질과 결합을 통해 기능을 수행하고 유전자 및 단백질 발현, 단백질 위치의 변화를 통한 단백질 상호작용의 변화를 초래한다. 궁극적으로 이러한 단백질 상호작용은 유전자 조절 및 신호전달체계와 같은 생물학적 프로세스에 관여하는 필수적 반응이다. 단백질 3차 구조의 원자 수준에서 단백질 상호작용에 대한 예측은 상호작용의 결합 친화성과 상호작용 파트너를 결정하는 특이성을 제공한다는 장점이 있다[1]. 전통적인 3차원 구조 분석에 이용되어온 상동성 모델링, 폴드 인식 및 단백질 도킹 기법들은 유전자 발현 및 상호진화와 같은 유전체 정보를 기반

※ 이 논문은 2009년도 한독대학원생교류지원사업과 한국학술진흥재단(KRF-2005-214-E00050)의 지원을 받아 수행되었음.
† 정 회 원 : 숭실대학교 생명정보학과 연구교수(교신저자)
†† 비 회 원 : University of Hamburg, Bioinformatics Center 박사과정
논문접수 : 2011년 11월 8일
수 정 일 : 1차 2012년 1월 11일, 2차 2012년 2월 27일
심사완료 : 2012년 2월 27일

한 예측 모델보다 더 정확한 물리적 상호작용을 예측 할 수 있다. 특히, 단백질 도킹 기법들은 3차원 구조에 대한 바인딩 사이트를 정확하게 식별 할 수 있는 계산 방법의 토대를 마련하였지만 바인딩 가능한 구조 공간을 검색해야 하는 높은 계산비용을 요구한다. 이러한 문제의 해결을 위해 바인딩 사이트 예측을 시도하여 도킹의 입력으로 사용될 단백질의 검색 공간을 줄이고 이미 개발된 정밀한 단백질 도킹 소프트웨어를 활용하는 연구가 시도되고 있다[2, 3]. 이러한 연구 경향을 반영하여 이 연구에서는 모노머 단백질 구조에서 생성될 수 있는 많은 3차원 패치 중 type I 에러의 임계값 미만을 허락하는 상호작용 패치의 후보 집합을 예측할 수 있는 기법을 제시한다. 이렇게 함으로써 제시하는 기법을 도킹을 위한 검색 공간의 감소와 새롭게 구조가 결정된 단백질에 대하여 기능적인 주석 추가 파이프라인 구축에 활용할 수 있다.

최근에는 단백질 구조의 바인딩 사이트 예측을 위한 다양한 계산학적 방법들이 개발되었다[8-12]. 이러한 연구들은 단백질 상호작용 사이트가 물리화학적인 속성[1-5], 위상적 속성[6], 보존되는 아미노산[7]들에 의해서 특징지어짐을 보여주고 있다. 관련 연구들은 SVM(Support Vector Machine)[8-11], Random Forests[3,12] 등을 포함하여 다양한 분류 기법을 이용하고 있다. 제안 논문의 선행연구[13]에서 연관규칙 마이닝을 이용하여 4가지 상호작용 유형에 따라 상호작용 사이트의 패턴을 구별할 수 있는 연관규칙을 제시하였다. 발굴된 연관규칙 패턴을 이용하여 상호작용 사이트를 예측할 수 있음을 보였다. 이 논문에서는 선행연구를 기반으로 상호작용 사이트에 대한 3차원 패치를 정의하고 이 패치의 속성을 이용하여 상호작용 사이트를 예측할 수 있는 기법을 새로이 제시한다. 이 논문을 효율적으로 기술하기 위해 다음과 같이 구성된다. 2장은 상호작용 인터페이스 및 3차원 패치 정의, 3장은 패치 속성 추출 및 이를 통한 상호 및 비상호작용 패치의 식별, 4장과 5장에서는 상호작용 패치에 대한 분류 모델 구축 및 평가로 구성된다.

2. 3차원 패치 및 상호작용 인터페이스 정의

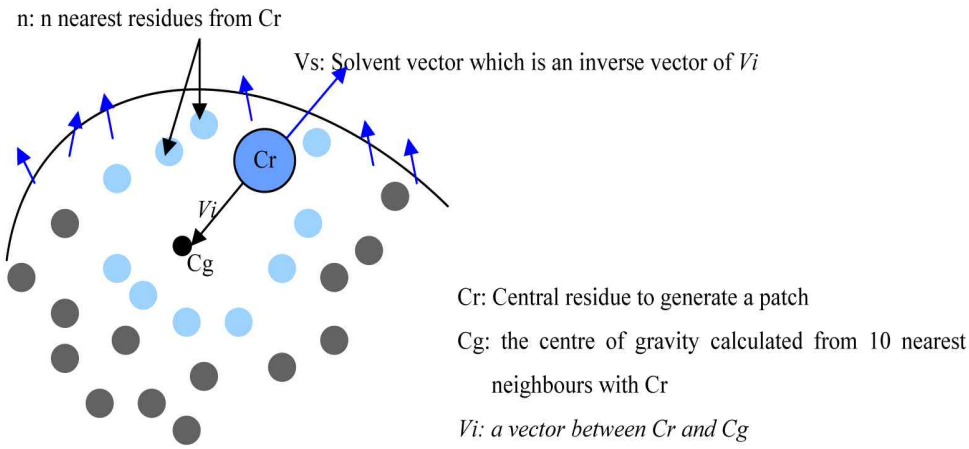
2.1 인터페이스

인터페이스(interface)는 상호작용에 참여하는 두 개의 단백질 분자가 결합하고 있는 양쪽 결합사이트를 의미한다. 인터페이스에 위치하는 원자의 식별은 원자나 아미노산 수준에서 접촉 거리를 이용하거나 *solvent accessible surface area (SASA)*-분자가 용매와 접촉할 수 있는 표면적-을 이용한다. 이 논문에서는 SASA를 이용하여 인터페이스를 정의한다. 인터페이스는 두 개의 단백질이 3차원 복합체를 형성할 때 SASA가 $SASA > 1\ A^{\circ 2}$ 로 감소하는 원자들 집합으로 정의할 수 있다. 단백질 복합체를 형성하지 않은 모노머 상태에 존재하는 두 단백질 분자가 상호작용하여 결합이 될 때 바인딩 사이트에 위치하는 원자들이 결합하여 표면에 노출되지 않게 되므로 SASA 값이 감소된다. 단백질 분자에 있는 모든 원자에 대한 SASA 값은 NACESS [14] 소프트웨어를 이용하여 계산되었다. 이 때 probe sphere를 1.5 A° 설정하여 계산한다.

2.2 3차원 패치(3D surface patch)

3차원 패치는 Jones과 Thornton[1]이 정의를 이용하며 알고리즘 1에 의해서 생성한다. (그림 1)에서 표면벡터(Vs)는 두 개의 점(Cr (패치의 중심 레지듀)와 Cg(구조 중심 레지듀) 에 의해 생긴 벡터 (Vi)에 대해서 단백질 표면을 향하여 역벡터를 구하면 표면벡터(Sv)가 된다. 패치의 중심 레지듀 Cr은 3차원 패치를 생성시 패치의 중심에 해당하며 이 레지듀를 중심으로 벡터 제약사항을 만족하는 이웃한 레지듀가 패치에 해당된다. Cg는 Cr에서 가장 가까운 10개의 레지듀들의 중심이 된다.

좀더 상세히 3차원 패치를 정의하면 다음과 같다. 3차원 패치는 패치의 중심 레지듀 (Cr)와 표면벡터 solvent vector (Sv) 사이에 110°의 각 안에 존재하는 n개의 이웃한 표면 레지듀로 구성된다. N (디폴트, n=40)은 분자 표면에 위치한



(그림 1) 3차원 패치 정의 및 표면벡터 제약사항

[알고리즘 1] 3차원 패치 계산

```
Algo1: Calculation of a 3D surface patch
Input:
S: a set of C-alpha atoms of surface residues C={}
K: number of surface residue in a patch
Output:
P: a set of n-patches P={}
Pi :a set of C atoms of surface residues in a patch
// 패치 초기화
// i 번째 패치 속한 중심탄소 집합의 초기화
Pi ={}
// Pi 패치의 중심
Cgi: a centre point of 10 nearest neighbors for a patch Pi
// Pi 패치의 중심 원자로부터 K 개의 가장 가까운 중심탄소 원자 집합 (KNNi)
KNNi: K-nearest neighbors for Pi

// 단백질의 표면 레지듀 Cri 에 대해서 K 개의 가장 가까운 중심탄소 원자를 구함(KNNi))
For each surface residue (Cri) in S
 KNNi ← find K nearest surface residues from Cri
//KNNi 로부터 패치의 중심 Cgi 를 계산함
 Cgi ←calculate centre of 10 nearest surface residues from KNNi
 // Cri 와 Cgi 의 vector 를 구함
 Vi ← calculate a vector (Vi) between Cri and Cgi
 // Vi 에 대한 inverse vector 를 구하여 Solvent Vector 로 설정
 solvent vector(Vsi)← inverse the vector of Vi
 // KNNi 에 해당하는 원자가 Cri 와 Vsi 사이에 110 도의 각을 유지하는지 평가
    For each nearest surface residues Rj in KNNi
      angle(Vsi, Rj) ← calculate an angle between Vsi and Rj
      IF ( angle(Vsi, Rj) < 110 ) then
          Pi ← Rj
      IF
    End for
P ←Pi
End for
```

레지듀로 패치의 크기를 결정하며 사용자에 의해서 주어진다. 표면 레지듀 (surface residue)는 상대적 분자 표면적(RSA)이 5% 이상(RSA > 5%)인 레지듀가 해당된다. 상대적 분자 표면적은 주어진 아미노산(X)을 ALA-X-ALA의 3개의 펩타이드로 치환할 때 표면적과 비교한 퍼센트지로 다음 <식 1>과 같이 계산된다[14].

RSA = (absolute SASA / SASA(ALA-X-ALA)) * 100 <식 1>

3차원 패치에 존재하는 원자들은 서열상에서 연속적이지 않을 수 있다. 패치에 속한 레지듀와 Vs 사이에 110° 의 각을 유지하는 벡터제약 사항 (그림 1)은 패치가 평면이 되지 않게 한다.

3. 3차원 패치의 속성 추출

각 3차원 패치를 특징 지을 수 있는 11개의 물리화학적 속성을 계산하였다. 절대 표면적(absASA: absolute ASA), 상대 표면적(relASA: relative ASA), 아미노산 갯수(nAA:number of amino acids), 원자 갯수(nAtom: number of atoms, 2차구조 개수 (nSSE: number of Secondary Structure Elements), 수소성 (HH: hydrophobicity), 레지듀 경향성 (inPro: residue propensity), Secondary Structure Elements (SSEs) content (Helix, Strand, Non-Regular), overlapping ratio with Interaction Sites (ORIS). 각 속성에 대한 상세한 설명은 아래와 같다.

절대 표면적(absASA)

바인딩 사이트에 대한 SASA는 복합체를 형성할 때 감소되는 SASA 값의 총합으로 계산된다 (<식 2> 참조). A와 B가 결합하여 복합체 AB를 만드는 두 단백질이라고 가정하면, $SASA_A$, $SASA_B$와 $SASA_{AB}$는 각각 단백질 A, B, AB의 SASA 값이다. N은 인터페이스에 존재하는 총 원자의 수를 나타낸다. 절대표면적(absASA)는 다음 수식에 의해 계산된다.

$$absASA_{A,B} = \sum_{i=1}^{n} (SASA_A(i), SASA_B(i)) - SASA_{AB}(i)) \quad <식\ 2>$$

상대 표면적(relativeASA)

상대적 분자 표면적은 주어진 아미노산(X)을 ALA-X-ALA의 3개의 펩타이드로 치환할 때 표면적과 비교한 퍼센트지로 <식 1> 같이 계산된다 [14]. 상대표면적의 값이 클수록 레지듀가 단백질 내부 보다는 표면에 위치함을 나타낸다.

소수성(Hydrophobicity): 바인딩 사이트의 평균 소수성 정도를 측정하기 위해서 Fauchere and Pliska [15]가 제시한 아미노산의 소수성 인덱스 기준을 사용하였다. 평균 소수성 정도(Hydrophobicity)는 아미노산의 소수성인덱스(HIAA)와 총 아미노산의 수(NAA)를 이용하여 <식 3>같이 계산된다.

$$HH = \frac{\sum_{I=1}^{I} HI_{AA}}{N_{AA}} \qquad <식\ 3>$$

레지듀 경향성: 레지듀 경향성은 바인딩사이트에서 다른 아미노산과 비교하여 특정 아미노산의 상대적인 출현 빈도를 나타내며 이를 통해 아미노산의 바인딩 사이트 선호 경향성을 나타낸다. 레지듀 경향성은 <식 4>에 의해서 계산된다.

$$\frac{\sum_{I=1}^{n} AAP_I}{N_R} \qquad <식\ 4>$$

AAp_i: 20개의 각 아미노산의 바인딩사이트 출현 경향성(AA Propensity)에 대한 자연로그 값

N_R: 바인딩사이트의 레지듀 총 수

20아미노산 바이딩 사이트 경향성(AA propensity)은 선행연구[13]에서 계산된 값을 사용하였으며 이것은 147개의 복합체를 구성하는 바인딩 사이트로부터 계산되었다.

2차 구조 구성(SSE content): 패치를 구성하고 모든 원자에 대해 2차 구조에 위치하고 있는 원자의 수를 퍼센트로 계산한다. 2차 구조의 타입은 DSSP[16] 소프트웨어 프로그램의 주석에 의해 정의하였으며 helix, strand와 비정형 구조(non regular region)로 나눈다. 비정형구조는 turn, bend와 loop 타입을 포함하고 helix와 strand를 제외한 타입을 포함한다. 위에서 설명한 9가지 속성은 상호작용사이트를 기술하기 위해 이용된 속성으로 패치의 속성 기술에 이용되며 선행연구[13]를 참조하여 계산되었다.

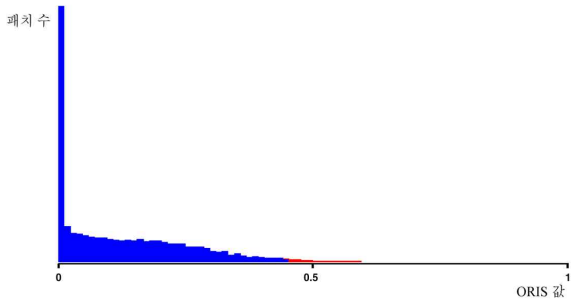
교차비율 (ORIS: Overlapping Ratio with Interaction Site): ORIS은 상호작용 사이트와 3차원 패치 사이 교차되는 비율로이다. ORIS의 계산은 상호작용사이트와 3차원 패치 양쪽에 모두 나타나는 원자의 비율을 나타낸다. 147개의 복합체에 대한 상호작용 사이트를 식별하고 각 상호작용 사이트에 포함되는 3차원패치에 위치하는 표면 레지듀를 식별하여 ORIS를 계산하였다(< 식 5> 참조).

$$ORIS = \frac{Interaction\ site \cap\ 3D\ patch}{Size\ of\ a\ 3D\ Patch} \qquad <식\ 5>$$

1: interaction patches where ORIS is ≥0.45

0: non-interaction patches where ORIS is 0

ORIS의 값에 따라서 3차원 패치가 상호작용 패치 또는 비 상호작용패치로 분류된다. 즉, (그림 2)와 같이 ORIS 값이 0.45보다 큰 3차원 패치는 상호작용패치 (ORIS ≥ 0.45)로 정의되고 상호작용 사이트와 중첩되는 원자를 하나도 포함하지 않은 3차원패치는 비상호작용 패치로 정의된다.



(그림 2) ORIS 분포

4. 분류 모델 구축

모노머 단백질 구조가 주어질 때, 상호작용 사이트의 예측은 단백질 구조로부터 생성된 3차원 패치를 이진 클래스로 (상호작용 또는 비상호작용) 나누는 이진 클래스 분류 문제로 정형화된다. 3차원 패치의 이진 분류를 위하여 ORIS를 제외한 10개 패치 속성을 어트리뷰트로 사용하였으며 ORIS는 타겟 클래스를 결정하는 속성으로 분류모델의 트레이닝시 제외되었다. 상호작용패치(ORIS ≥ 0.45)는 positive 클래스로 정의하고 비상호작용 패치 (ORIS=0)는 negative 클래스로 정의하였다. 3차원 패치는 비상호작용 패치에 비하여 상호작용패치의 수가 극히 적어서 각 클래스의 데이터 분포가 불균형을 이룬다. 이러한 불균형 클래스 분류 문제를 최소화하기 위하여 일반적으로 under-sampling과 over-sampling 방법을 사용하고 있으나 이 연구에서는 under-sampling을 사용하였다. Under-sampling 기법은 소수 범주의 수만큼 다수 범주에서 데이터를 샘플링하여 사용하는 기법으로 3차원 패치의 이진분류에 over-sampling보다 적합하다. Negative 데이터인 비상호작용 패치에 대한 데이터가 positive 데이터인 상호작용 패치 데이터보다 극히 많은 경우로 over-sampling를 할 경우 전체 트레이닝 데이터의 증가로 실행 시간의 증가를 초래하는 단점이 있다. 또한

positive 데이터가 반복적으로 샘플링되므로 분류모델이 트레이닝 데이터에 오버피팅 (overfitting)될 가능성이 높다.

의사결정트리 (DT: Decision Trees)[17], Random Forest(RF)[18], Support Vector Machines (SVM)[19]와 같은 분류 모델을 이용하여 분류기를 구축하였다. WEKA 머신 러닝 라이브러리[20] 버전 v 3.6.6을 이용하여 10 fold cross validation기법으로 분류모델의 성능을 테스트 하였다.

5. 실험 및 평가

5.1 데이터 셋

선행연구에서 사용된 상호작용 복합체 데이터와 동일한 데이터를 사용하였다. 147개 단백질 복합체 (<표 1> 참조)에 해당하는 PDB 단백질 구조 데이터로부터 총 54,407개의 3차원 패치가 생성되었다. 이 중 1085개의 패치는 상호작용 패치에 포함되며 13,310의 패치는 비상호작용 패치에 해당된다. 상호작용 패치는 전체 패치 중 1.99 %에 해당하는 극히 적은 분포를 보이며 under-sampling을 통하여 상호작용 패치 클래스와 비상호작용 클래스의 패치의 수가 균형을 이루도록 한다. Under-sampling 후에도 평균 원자 수와 표면적의 값은 under-sampling 이전의 값들과 통계적으로 유의한 차이가 없음을 <표 2>에서 알 수 있다.

〈표 1〉 단백질 복합체 데이터

상호작용 타입	단백질 복합체 수
Enzyme-inhibitors	25
Non Enzyme-inhibitors	21
Hetero-obligomers	14
Homo-obligomer	87

5.2 분류 모델의 평가

세 가지 분류 모델에 대한 파라메터는 WEKA 머신러닝 라이브러리의 기본 설정을 사용하였으며 모델 구축 시 변경된 파라메터 값은 <표 3>에 표시하였다. 원본데이터에 대한 세 가지 분류 모델에 대한 true positive (TP) 비율은 positive 클래스와 negative 클래스 각각에 대하여 0.99와 0.701을 보인다 (<표 3> 참조). 데이터의 수가 극히 적은 positive 클래스에 대한 TP 값은 negative 클래스에 비해 작지만 false positive(FP)의 비율은 의사결정트리나 SVM 모델에서 낮은 에러율을 보인다. 데이터의 불균형을 고려할 때 낮은 FP 비율과 높은 정확성은 아니지만 분류 모델로 사용 가능한 정확성을 보임을 확인할 수 있다. under-sampling후에 positive 클래스에 대한 TP 비율은 0.89~0.91까지 향상되는 높은 정확성을 보인다. SVM이나 random

〈표 2〉 데이터의 특성

1) 원본 데이터		
	interaction patch (ORIS ≥ 0.45) class=1	non-interaction patch (ORIS=0) class=0
#. of patchs	1085	13310
avg. #. of atoms	305.71	312.45
avg. #.of ASA(Å^2)	2827.3	2232.46
2) under-sampling 후의 데이터		
	interaction patch class=1	non-interaction patch class=0
#. of patchs	1085	1064
avg. #. of atoms	305.71	312.83
avg. #. of ASA(Å^2)	2827.3	2224.32

〈표 3〉 분류 모델의 정확성

Data Set	Decision Tree				Random Forest (Tree[a] =10)				SVM (RBF Kernel gamma[b] =50)			
	Class=0		Class=1		Class=0		Class=1		Class=0		Class=1	
	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
Raw Data	0.988	0.30	0.691	0.012	0.996	0.294	0.706	0.004	0.998	0.294	0.706	0.002
Undersampling	0.890	0.110	0.890	0.110	0.924	0.099	0.900	0.076	0.943	0.088	0.912	0.057

[a]: Random Forest 에서 random sample에 학습되는 트리의 수

[b]: 비선형커널인 RBK 커널의 gamma 값

forest 모델에서는 positive 클래스의 TP 값이 차이가 나지 않는 높은 정확성을 보인다. <표 4>에서 positive 클래스의 TP 값이 가장 높은 SVM 모델에 대해서 상세한 정확성을 보인다.

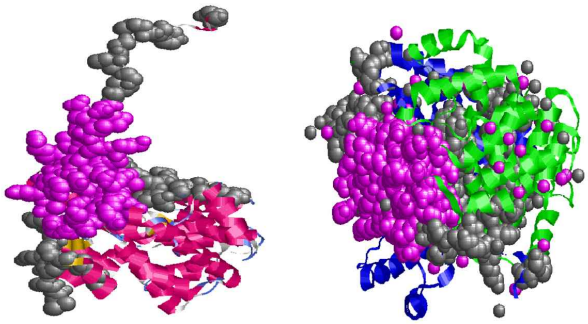
〈표 4〉 SVM 분류 모델의 상세한 정확성

Class	TP rate	FP rate	Precision	ROC area	F-Measure
0	0.943	0.088	0.913	0.928	0.928
1	0.912	0.072	0.928	0.928	0.927

5.3 생물학적 관점에서 분류 결과 해석

(그림 3) 3차원 패치를 분류하는데 사용한 9개의 속성의 타겟 클래스에 따른 분포를 보여준다. 주목할 만한 것은 상호작용 패치나 비상호작용 패치 클래스에 따라 원자의 개수는 차이가 없지만 패치의 표면적(absASA)과 상대 표면적(relASA)은 상호작용 패치에서 더 넓게 관찰된다. 또한 레지듀 경향성은 소수성에 비해 영향이 없게 보이며 소수성 정도를 나타내는 hydrophobicity가 상호작용 패치에서 더 강하게 보인다. 즉, 3차원 상호작용 패치는 넓은 표면적을 가지며 소수성이 강한 레지듀로 구성됨을 알 수 있다. 또한 3차원 상호작용 패치는 더 적은 수의 2차 구조로 구성되며 주로 β-strand의 비율이 α-helix보다 더 적으며 주로 비정형 2차 구조의 비율이 높음을 알 수 있다.

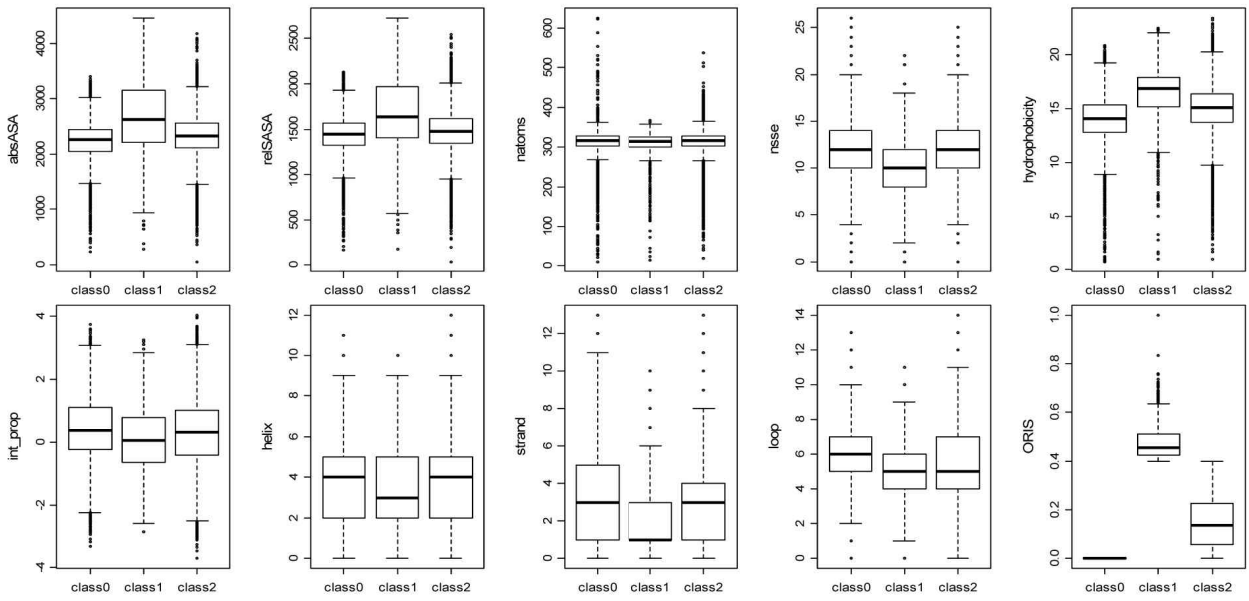
3D 패치가 바인딩사이트와 가장 높은 비율(ORIS=0.675)로 겹쳐진 패치는 1aj8 단백질의 B체인에 13번 레지듀를 중심으로 생성한 패치(1aj8B_13)이다. 1aj8단백질 복합체는 100C° 가까운 고열에서 산소가 없는 환경에서 생존할 수 있는 고세균류에서 채취한 구연산합성효소(citrate synthase)로써 호모다이머 (homodimer)를 형성한 3차원 구조다. (그림 4)의 (a)는 모노모상태에서 1aj8B_13 패치의 모양을 살펴본 것이고 (b)는 동일한 바인딩 사이트를 이용하여 다이머를 형성한 구조를 보인다. 실제로 자연계에서는 이 단백질은 (그림 4)의 (a)와 같이 monomer상태로 존재하지 않고 항상 (그림 4)의 (b)처럼 다이머를 형성하는 obligomer이다. <표 5>는 상호작용 사이트에 대한 속성과 패치의 속성을 나타낸다. 각 상호작용 사이트와 패치 사이의 원자의 수는 차이가 있지만 절대 표면적 (absSASA)은 차이가 거의 없다는 것이다. 패치가 대부분의 단백질의 표면에 위치하고 있는 레지듀로 구성 됨을 시사하고 있다.



(a) 모노모 상태의1aj8A　　(b) dimmer를 형성한 homo-obligomer의 복합체

단백질 1aj8A, 1aj8B는 각각 파랑과 초록 색으로 표현되었다. 회색의 spacefill은 상호작용 사이트이며 분홍색의 spacefill 은 3차원 패치 1aj8B_13이다.

(그림 4) 1aj8의 3차원 패치 및 상호작용 사이트



class0: negative 클래스, class1: positive 클래스, class2: negative와 positive 클래스에 포함되지 않은 패치집합

(그림 3) 타켓 클래스 별 3차원 패치 속성 분포

〈표 5〉 1aj8의 상호작용 사이트 및 패치 속성

ID	HH	absSASA	nAA	nAtom	nSSE	inPro	sRatio	Helix	Strand	NR	PPI type
1aj8A	0.412	3596.04	114	498	31	0.528	30.73	29.52	22.89	47.59	homoobligomer
1aj8B	0.413	3618.77	111	487	29	0.12	30	25.05	23.2	51.75	
1aj8B_13	16.8	3509	40	329	10	0.51	10.81	2	3	5	-

6. 결론 및 토의

제시한 기법은 기존연구[1]에 기반하여 3D패치 정의와 속성을 추출하였다. 상호작용 사이트의 예측모델을 구축하기 위해 의사결정트리, random forest와 SVM에 대한 분류 모델을 구축하고 성능을 평가하였다. 여러 가지 분류 모델 중 정확성이 가장 높은 SVM 분류 모델을 적용하여 예측 모델을 구축한다는 것이 기존연구[1]과 차별화 될 수 있다. SVM을 사용한 기존의 기법[9, 10, 11]과는 다르게 2차 구조 정보를 반영함으로써 상호작용사이트에 작용하는 단백질 폴드정보를 추가적으로 사용하여 모델의 정확성을 높였다.

2차 구조를 포함한 9개의 패치 속성은 원본 데이터의 극심한 불균형 분포에도 불구하고 3차원 패치를 상호작용 및 비상호작용 패치로 구분 짓는 높은 분류 파워를 가진다. 3차원 패치 중 상호작용 클래스에 해당하는 패치들은 패치의 모든 원자들이 상호작용 사이트에 해당하는 것이 아니고 일부분 45% 이상이 상호작용사이트에 해당된다. 그러므로 일부분의 속성은 비상호작용 사이트에 해당하는 속성이 분류 모델 구축에 반영되고 이러한 요인이 정확성을 낮추는데 작용했을 것이다. 이러한 현상을 반영하듯 원본 데이터에 대해 구축된 분류 모델에서 positive 클래스에 비해 negative 클래스의 정확성이 높다. 하지만 반대로 false positive의 값은 positive 클래스에서 더 낮은 에러율을 보인다.

Under-sampling을 이용한 트레이닝 데이터에 대해 구축된 분류 모델의 정확성은 3차원 패치를 상호작용과 비상호작용 패치로 예측에 활용할 수 있을 정도의 높은 정확성을 보인다. 예측 모델을 비상호작용 패치의 예측에 활용함으로써 전체 단백질 중 상호작용 사이트에 해당되지 않은 부분을 제거하고 남은 부분을 상호작용 사이트로 예측하고 이것을 도킹을 위한 입력으로 사용할 수 있다. 그러므로 도킹을 위한 검색 공감을 줄이는 도킹파이프라인의 전 단계에 활용가능 하다. 제안 연구에서는 단백질 서열에 대한 속성을 사용하지 않고 상호작용 사이트 예측 기법을 개발하였다. 향후 연구로서 단백질 서열의 진화적 관계 정보를 이용하여 구조가 존재하지 않은 단백질에 대해서도 상호작용 사이트를 예측할 수 있도록 단백질과 서열 정보를 추가하여 제안 기법을 확장할 수 있다. 또한 단백질 상호작용 사이트는 하나의 속성으로 완벽하게 기술되지 않고 여러 속성의 결합을 통해 기술되어진다. 현재까지 많은 연구를 통해서 사용되는 속성들은 한정되어 있으므로 새로운 속성의 발굴이 요구된다.

참 고 문 헌

[1] S. Jones and J. M. Thornton, “Analysis of protein-protein interaction sites using surface patches.”, J Mol Biol, Vol.272, pp.121-132, 1997.

[2] S.B. Qin and H.X. Zhou, “A holistic approach to protein docking.”, Proteins Vol.69, pp.743 - 74, 2007.

[3] Z. Qiu and X. Wang, “Prediction of protein-protein interaction sites using patch-based residue characterization.”, Journal of Theoretical Biology, Vol.293, pp.143-150, 2011.

[4] W. S. Valdar and J. M. William, “Protein-protein interfaces: Analysis of amino acid conservation in homodimers.”, Proteins Vol. 42, No.1, pp.108-124, 2001.

[5] H. Neuvirth, R. Raz and G. Schreiber, “ProMate: a structure based prediction program to identify the location of protein-protein binding sites.”, J Mol Biol, Vol.338, pp.181-199, 2004.

[6] F. P. Davis and A. Sali, “PIBASE: a comprehensive database of structurally defined protein interfaces.”, Bioinformatics, Vol.21, No.9, pp.1901-1907, 2005.

[7] C. D. Livingstone and G. J. Barton, “Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation.”, Computer Applications in the Biosciences, Vol. 9, No.6, pp.745-756, 1993.

[8] J. R. Bock and D. A. Gough, “Predicting protein-protein interactions from primary structure.”, Bioinformatics, Vol.17, No.5, pp.455-460, 2001.

[9] A. Koike and T. Takagi, “Prediction of protein-protein interaction sites using support vector machines.”, Protein Eng Des Sel, Vol.17, No.2, pp.165-173, 2004.

[10] Bradford J. R. and D. R. Westhead, “Improved prediction of protein-protein binding sites using a support vector machines approach.”, Bioinformatics, Vol.21, No.8, pp.1487-1494, 2005.

[11] H. Zhu, F. S. Domingues, I. Sommer and T. Lengauer, “NOXclass: prediction of protein-protein interaction types.”, BMC Bioinformatics, Vol.7, No.27, 2006.

[12] X. W. Chen and M. Liu, “Prediction of protein-protein interactions using random decision forest framework.”, Bioinformatics, Vol.21, No.24, pp.4394-4400, 2005.

[13] S. H. Park, J. A. Reyes, D. R. Gilbert DR, J. W. Kim and S. Kim, “Prediction of protein-protein interaction types using

association rule based classification.", BMC Bioifnromatics, Vol.10, No.36, 2009.
[14] S. J. Hubbard, S. F. Campbell and J. M. Thornton, "Molecular recognition: conformational analysis of limited proteolytic sites and serine proteinase inhibitors.", J. Mol. Biol., Vol.220, pp.507 - 530, 1991.
[15] J-L. Fauchere and V. E. Pliska, "Hydrophobic parameters p of amino acid side chains from partitioning of N-acetyl-amino-acid amides.", Eur J Med Chem., Vol.18, pp.369-375, 1983.
[16] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.", Biopolymers, Vol.22, No.12, pp.2577-637, 1983.
[17] R. Quinlan, 'C4.5: Programs for Machine Learning', Morgan Kaufmann Publishers, 1993.
[18] L. Breiman, "Random Forests.", Machine Learning, Vol.45, No.1, pp.5-32, 2001.
[19] J. Platt, "Using Analytic QP and Sparseness to Speed Training of Support Vector Machines.", NIPS 11, pp.557-563, 1999.
[20] I. H. Witten and E. Frank, 'Data Mining: Practical machine learning tools and techniques',2nd ED, San Francisco: Morgan Kaufmann, 2005.

박 성 희

e-mail : shpark@ssu.ac.kr
1996년 충북대학교 도시공학과(공학사)
2001년 충북대학교 전자계산학과 (이학석사)
2005년 충북대학교 전자계산학과 (이학박사)
2005년~2006년 Univ. of Glasgow, Bioinformatics Research Centre Post-Doc Fellows
2007년~현 재 숭실대학교 생명정보학과 연구교수
관심분야 : 생명정보학, 시스템생물학, 생물 & 의학 데이터 마이닝, 시공간데이터베이스

Björn Hansen

e-mail : hansen_bjoern@web.de
2007년 University of Luebeck, Molecular Life Science(이학사)
2011년 University of Hamburg, Bioinformatics Center(이학석사)
2011년~현 재 University of Hamburg, Bioinformatics Center 박사과정
관심분야 : 생명정보학, 시스템 생물학, 단백질구조 모델링