핵산과 아미노산의 결합 경향성을 발견하기 위한 알고리즘

한남식° 한경숙
인하대학교 전자계산공학과
han_3567@hotmail.com°, khan@inha.ac.kr

An algorithm for finding binding propensities of nucleic acids and amino acids

Namshik Han° Kyungsook Han
School of Computer Science and Engineering, Inha University

요 약

오늘날 핵산과 단백질의 결합체에 관한 자료가 PDB(Protein Data Bank)와 같은 공공 데이터베이스에 급속도로 증가되고 있고 하나하나의 자료 자체도 많은 양의 데이터를 가지고 있기 때문에 더 이상 수작업으로 이를 분석하기란 거의 불가능할 뿐 아니라 정확도에 많은 문제가 있다. 그래서 본 연구에서는 방대한 생물학 자료를 효율적으로 분석하기 위해 자동화된 알고리즘을 개발하여 수작업에 의존하던 기존방식을 개선하였다. 이 알고리즘으로 51개의 RNA와 단백질간의 결합구조로 구성된 Dataset과 129개의 DNA와 단백질 간의 결합구조로 구성된 Dataset 분석하여 각각의 경우에 있어서의 결합성향과 결합유형을 찾아내었다. 이러한 본 연구의 결과가 아직 구조가 밝혀지지 않은 단백질-핵산간의 결합부위를 예측하는 알고리즘 개발에 기초 자료로 이용될 수 있다. 신약을 개발하는 과정에서 표적단백질의 결합부위를 예측하는데 활용될 수 있을 것이다.

1. 서 론

바이러스에 의해 발병되는 질병들을 완치하기 위해서는 바이러스에 대한 보다 많은 연구가 요구됨과 동시에 감염된 단백질에 잘 결합하여 더 이상의 변이를 중단시킬 수 있는 신약개발이 있어야 한다. 이를 위해 핵산과 단백질의 결합부위를 예측할 수 있어야 하는데 아직 정확한 예측은 이루어지지 않고 있다. 그래서 실제 결합구조를 분석해 결합정보를 밝히는 연구가 요구되고 있다. 그런데 이러한 연구가 이루어지기 위해서는 많은 양의 서열과 수치 데이터 등으로 구성된 생물학 자료를 분석해야 하는 어려움이 따른다. 게다가 X-Ray Crystallography와 NMR (Nuclear Magnetic Resonance) 기법의 발달로 실제 결합구조를 밝히는 작업이 가속화되어 인터넷상의 공공 데이터베이스에 보고되는 자료가 급속도로 증가함으로 인해 더 이상 수동적인 방식에 의존하기에는 큰 어려움이 있다.

본 연구에서는 자동화된 분석기법을 개발하여 방대한 양의 생물학 자료를 분석하는데 이용함으로써 기존의 문제를 해결하였다. 그 결과 핵산과 단백질이 결합할 때 어떠한 결합성향을 보이며 결합형태는 어떠한 지를 발견하였다. 또한, 핵산의 두 종류인 DNA와 RNA가 단백질과 결합할 때 다른 결합성향과 결합형태를 보인다는 사실도 확인하였다.

2. Dataset 구성

본 연구를 수행하기 위해 2002년 9월까지 PDB에 보고된 단백질-RNA 결합구조 중 188개를 찾은 후, 그 중 해상도가 3.0Å 보다 좋지 않은 것을 제외시켜 139개의 결합구조를 선별했다. 그리고 Homologous Search를 하여 상동성이 있는 결합구조를 제외시켜 64개를 선별했으며, 최종적으로 인공적으로 합성되었거나 물이 낀 결합이 누락된 13개의 결합구조를 제외시킴으로써 51개의 단백질-RNA 결합구조로 Dataset을 만들었다. 단백질-RNA 결합구조는 이 51개 Dataset을 대상으로 분석하였으며, 단백질-DNA 결합구조의 Dataset의 경우는 직접 구성하지 않고 단백질-DNA 결합에 관한 선행연구에서 사용된 129개의 결합구조로 이루어진 Dataset을 사용하였다 [1].

3. 알 고 리 즘

단백질-핵산간의 결합구조를 자동화된 기법으로 분석하기 위해 분석 알고리즘을 구상하여 효율성과 정확성이 뛰어난 프로그램을 개발하였다. 이 프로그램이 수행되는 일련의 과정 4 단계로 구성되며 다음과 같다.

3.1. Phase 1

첫 번째 단계에서는 효율적 분석을 위한 기초 작업 단계로서 다음 단계들에서 사용될 기초 자료를 정리하여 분류해 배열에 저장하는 작업이 수행된다. 즉, 이 단계에서는 PDB 파일과 수소결합을 표시해 주는 프로그램인 HBPlus의 결과 파일에서 아미노산과 핵산의 서열, 위치좌표 및 결합관계 등의 자료를 뽑아 각각의 배열에 저장하게 된다.

3.2. Phase 2

두 번째 단계는 핵산 내부의 결합 구조를 검토하여 핵산들간 내부결합관계를 저장하는 작업이 수행되는 단계이다. Phase1에서 작성된 배열을 이용하여 핵산의 경우 다른 핵산의 Base와 결합을 이룬 Pair인지 아닌지의 여부를 확인하여 핵산 자체의 결합구조를 표시해 Linked-List에 저장한다. 즉, Linked-List를 보게 되면 결합한 핵산간의 관계를 알 수 있으므로 추후 이중결합과 다중결합을 분류할 때 유용하게 사용된다.

3.3. Phase 3

표 1. RNA의 각 Base 원자와 아미노산간의 수소결합 분포

	A				G					C			U			TOTAL
	N1	N3	N6	N7	N1	N2	N3	O6	N7	O2	N3	N4	O2	N3	O4	
NH	41(59)							16(34)	4(15)	7(12)	11(41)		10(17)		6(15)	95
C==O			69(69)		3(4)	62(41)						27(61)		15(47)		176
ARG	7(10)	1(17)						14(30)	8(31)	23(40)	12(44)		17(28)		6(15)	88
LYS		1(17)		2(10)				16(34)	7(27)	3(5)	2(7)		5(8)		13(33)	49
THR			19(19)	18(90)		1(1)		1(2)	3(12)						4(10)	46
SER	19(27)	4(67)	5(5)			2(1)				6(11)	2(7)		11(8)			49
TYR						1(1)								1(3)		2
ASN	2(3)		1(1)		4(5)				4(15)	13(23)		5(11)	10(17)	4(13)	6(15)	49
GLN			3(3)			8(5)	4(100)			5(9)		4(9)	5(8)	6(19)	4(10)	39
GLU			1(1)		43(56)	46(30)						2(5)		4(13)		96
HIS													2(3)			2
ASP					27(35)	31(21)						6(14)		2(6)		66
MET			2(2)													2
CYS	1(1)															1
TOTAL	70	6	100	20	77	151	4	47	26	57	27	44	60	32	39	760

세 번째 단계에서는 단일결합, 이중결합, 다중결합의 3가지 결합유형으로 분류하는 작업이 수행된다. 3가지 결합유형을 나누는 것은 각 유형별 결합성향이 다르기도 할 뿐 아니라 단순한 단일결합 보다 둘 이상의 수소결합을 취하는 잔기간의 결합형태를 쉽게 찾아주기 위함이다.

Phase 1, 2에서 작성된 배열과 Linked-List를 검토하여 핵산 중 Pair를 이룬 것들을 찾아 결합관계를 확인하여 2개 이상의 수소결합을 취하고 있는 경우만 분류함으로써 단일결합을 이중결합과 다중결합으로부터 구별해 낸다. 분류된 집합에서 두개 이상의 Base pair와 결합하고 있는 아미노산을 찾아 다중결합으로 분류한다.

이러한 분류를 위해 핵산의 Base간의 거리를 구해주는 함수로 핵산의 서열과 위치정보가 저장된 배열과 핵산의 내부구조를 저장한 Linked-List를 비교하여 해당 핵산간의 최대 거리를 구한다. 그리고 그 거리가 0인 것은 Base pair를 이루지 않은 단순 Base인 것으로, 1인 것은 연속된 Base pair인 것으로, 그 이상인 것은 Base pair가 연속되어 Stem구조를 이룬 것으로 구분한다. 즉, 0인 경우는 Base pair를 이루지 않아 다중결합이 될 여지가 없기 때문에 곧바로 이중결합으로 분류가 가능함으로 불필요한 연산을 줄여주는 장점이 있다.

3.4. Phase 4

네 번째 단계는 파싱과정을 통해 각 아미노산과 핵산별 결합성향과 결합형태를 찾아낸다. 파싱을 하면 각 아미노산과 핵산의 총 수소결합회수 뿐 아니라 원자수준과 잔기수준에서 각 원자와 잔기별 수소결합 회수, 결합대상 및 동시에 몇 개의 결합을 하는지 등에 대한 분석을 한다. 이러한 자료를 다시 파싱하여 결합성향과 결합유형을 찾아낸다. 결합유형을 찾을 때, 단일결합의 경우는 단순히 일대일의 결합을 취하는 것이기 때문에 제외되고 이중결합과 다중결합만을 대상으로 함으로 효율적으로 처리할 수 있다.

4. 결 과

4.1. DNA와 RNA의 결합성향 비교

단백질-DNA, 단백질-RNA 결합구조의 두 Dataset을 대상으로 분석하여 각 결합구조들에 대한 분석결과를 얻고 이를 비교하여 단백질-DNA, 단백질-RNA 각각의 결합성향과 결합유형을 발견하였다. 일반적으로 DNA는 Double strand 구조를 취하는 반면 RNA는 Single strand 구조를 취하고 있는데 이러한 구조적인 차이가 단백질과 결합할 때 각기 다른 결합성향과 형태를 띄게 한다는 것을 발견할 수 있었다. DNA Base가 단백질과 수소결합을 이루는 비율은 32%에 불과하지만 RNA Base의 수소결합 비율은 50%에 다다르는 것을 발견하였다. 즉, DNA Base는 DNA가 Double strand 구조를 이루는 과정에서 이미 다른 DNA Base와 결합을 하였기 때문에 단백질과 결합할 기회가 적은 반면 RNA Base는 RNA가 Single strand 구조이기 때문에 다른 RNA Base와 결합하지 않은 개방된 Base가 대부분 이기 때문에 pair를 이룬 DNA Base보다 단백질과 결합하기가 용이하다는 것이다.

DNA와 RNA 별 수소결합 시 선호하는 아미노산을 살펴보면, DNA는 GLY와 ALA를 RNA보다 선호하는 경향이 있으나 RNA는 대개의 경우 그것들과 잘 결합하지 않는다는 것을 발견하였다 [1, 2]. 반면, RNA는 GLU와 ASP와 다수의 결합을 하는 경향을 발견하였다. 더 자세히 분석해 보면 단백질과 RNA 결합구조에서 특이한 성향을 발견하였는데 이중결합의 경우 Guanine은GLU, ASP와 많은 결합을 하는 반면 Adenine은 THR, LYS와 잘 결합하는 경향성을 보인다. 이러한 특이 성향은 결합형태에도 영향을 미치게 되기 때문에 높은 경향성을 보이는 아미노산과 핵산들은 결합형태에서도 주요한 역할을 한다. 다음 절에서 소개될 가장 많은 결합형태인 GLU-G 결합의 경우 37번 발견되어 주요 결합성향은 결합형태에 영향을 준다는 것을 입증해 주고 있다. 표 1은 핵산 Base의 각 원자가 아미노산과 결합한

표 2. 단백질-RNA 결합의 유형

GLU		THR		ARG						ASP		GLN						ASN		SER	
OE1	OE2	OG1	OG1	NH1	NH2	NE	NH2	NH1	NH2	OD1	OD1	OE1	NE2	OE1	NE2	OE1	NE2	OD1	ND2	OG	OG
A	A	D	A	D	D	D	D	D	D	A	A	A	D	A	D	A	D	A	D	A	D
G		A		C		U		G		G		U		G		G		A		A	
N1	N2	N7	N6	N3	O2	O2	O2	O6	O6	N1	N2	N3	O4	N2	O2	N2	N3	N6	N1	N6	N1
D	D	A	D	A	A	A	A	A	A	D	D	D	A	D	A	D	A	D	A	D	A
37		18		11		2		1		12		2		1		1		1		1	

것을 분석하여 수소결합 회수를 보여주는데, 여기서 보아도 GLU와 ASP가 많은 결합을 함을 알 수 있다. 괄호안의 숫자는 RNA Base의 한 원자가 특정 아미노산에 결합하는 확률이다. 그리고 아미노산의 경우, Main chain은 핵산의 Backbone과 같이 구별성이 없기 때문에 Amide group (NH)과 Carbonyl group (C=O)의 두 부분으로 분류하여 표기하였기 때문에 나머지 아미노산은 해당 아미노산의 Side chain만을 의미한다.

4.2. RNA의 결합유형

DNA에 관한 선행연구는 많이 있었으나 RNA에 관한 연구는 아직 그에 비하면 적은 수준이다. 그래서 본 연구에서는 RNA의 결합성향을 밝혀내는데 주안점을 두었다. RNA의 결합성향은 4.1.절에서 언급하였으므로 RNA의 특징적인 결합유형과 선행연구와의 비교한 결과를 살펴보겠다 [3].

RNA의 특징적인 결합유형을 표 2 에서 보여주고 있다. 앞서 언급한 GLU-G 결합 및 기타 결합유형을 볼 수 있다. 도표의 윗 열은 결합에 참여하는 아미노산과 그의 원자들이고, 아래 열은 핵산과 그의 원자들이다. 각 원자들 밑에 표기된 A와 D는 수소결합 시 각 원소가 Acceptor의 역할을 하는 경우 A로, Donor의 역할을 하는 경우 D로 표기한 것이다. 그런데, 일반적으로 높은 결합성향을 보인 LYS의 경우는 특별한 결합유형을 취하지 않고 있는데 이것은 LYS의 Side chain이 이 질소로만 구성되어 있어 핵산의 Base 부분과 많은 수소결합을 하지 못하기 때문이라고 판단된다. 반면, 결합성향과 결합유형 모두에서 높은 비중을 차지하는 GLU와 ASP는 아미노산 자체의 분류에서도 Acidic side chain group에 속해 있듯이 Side chain이 산성을 띄기 때문에 핵산의 Base와 수소결합을 빈번히 이루기 때문에 LYS와는 달리 특정한 결합유형을 취할 수 있다고 본다.

본 연구의 신뢰도를 측정하기 위해 단백질-RNA 결합구조에 관한 선행연구 [3]와 비교하였다. 선행연구에서 사용된 Dataset을 본 연구에서 개발한 프로그램으로 분석한 결과 선행연구와 거의 비슷한 결과를 얻을 수 있었다. 또한, 본 연구에서 사용한 Dataset을 분석해 얻은 이미 언급한 결과들과 비교하여도 큰 차이점은 발견할 수 없었다.

5. 향 후 과 제

결합관계에 대한 분석 결과를 단백질-핵산 간의 결합부위 예측 프로그램의 기초자료로 사용하고 있다. 예측을 위해서는 실제 결합구조에서 발견한 결합특성과 결합유형을 이용하여 임의의 서열이 주어졌을 경우 해당 서열에서 각 잔기별 결합가능수치를 계산하는 과정을 거쳐 가장 결합할 확률이 높은 예측부위를 찾기 위함이다. 즉, 실제의 단백질-핵산 결합관계에서 결합성향이 높다고 나온 잔기들과 결합형태에 빈번히 참여하고 있는 잔기들은 결합부위가 알려지지 않은 단백질과 핵산이 결합할 때에도 수소결합에 참여하여 결합부위가 될 확률이 높기 때문에 예측할 때 이러한 잔기들은 다른 잔기보다 높은 가중치를 받게 된다.

6. 결 론

본 연구를 통해 자동화된 프로그램으로 방대한 양의 생물학 자료를 분석하여 단백질과 핵산의 결합성향과 결합유형을 발견하였다. DNA와 RNA간의 결합성향이 다르다는 것을 확인하였고 RNA만의 특징적인 결합유형 또한 찾아낼 수 있었다. 공용화된 웹 데이터베이스에 급속도로 증가하는 생물학적 자료의 추세를 감안한다면 대다수의 과정이 수동적이던 기존 방식으로는 매우 긴 시간이 요구되거나 많은 비용이 들어 연구할 수 없었던 부분에 대한 연구를 컴퓨터 프로그램을 이용하여 가능하게 함으로써 생물학적 문제 해결에 이바지 할 수 있다.

특히 본 연구는 표적 단백질에 잘 결합할 수 있는 핵산의 서열과 구조 예측에 사용되고 있으므로 감염된 단백질에 잘 결합할 수 있는 신약의 서열과 구조를 밝혀주는 신약개발연구에 기여할 것으로 기대한다.

후기

본 연구는 정보통신부 정보통신 선도기반기술개발사업 (과제 번호 01-PJ11-PG9-01BT00B-0012)의 지원에 의하여 이루어졌음.

참 고 문 헌

1. N.M. Luscombe, R.A. Laskowski, and J.M. Thornton, Amino acidbase interactions: a three-dimensional analysis of proteinDNA interactions at an atomic level, *Nucleic Acids Res., 29,* 2860-2874, 2001.
2. S. Jones, P. van Heyningen, H.M. Berman, and J.M. Thornton Protein-DNA Interactions: A Structural Analysis. *J. Mol. Biol., 287,* 877-896, 1999.
3. J. Allers, and Y. Shamoo Structure-based Analysis of Protein-RNA Interactions using the program ENTANGLE. *J. Mol. Biol., 311,* 75-86, 2001.