유사 전사체 모델 탐색을 위한 위치 기반 블록 간의 유사도 비교 기법

김소라, 박태원, 황혜련, 조환규
부산대학교 컴퓨터공학과
e-mail : srkim_11@pusan.ac.kr

A Position-Based Block Similarity Computing Method for Similar Transcript Model Search

Sora Kim, TaeWon Park, HyeRyeon Hwang, Hwan-Gue Cho
Dept. of Computer Engineering, Pusan National University

요 약

전사체(transcript)는 유전자로부터 전사된 DNA 시퀀스 코드를 말한다. 전사체(transcript)의 발현된 형태에 따라 생성되는 단백질의 형태 역시 달라지므로 전사체 모델의 형태는 중요한 의미를 가지며 특정 위치의 전사체가 정상과 다르게 모델이 변할 경우 심각한 경우에는 유전자 질병에 노출될 수 있다. 현재 실험체에 대한 전사체 모형은 SpliceGrapher, Cufflinks 와 같은 상용화된 도구들을 사용하여 얻을 수 있다. 하지만 이런 도구 간의 결과 값 및 어노테이션 정보와 결과 값 간의 유사도 비교를 위한 방법론은 현재 알려진 바 없다. 대신 전사체 비교를 위해 모형 간의 차이를 눈으로 하나씩 비교하거나 전사체 위치를 이용한 산수 값을 이용한다. 본 논문에서는 전사체 모형 간의 유사도를 비교하기 위한 방법론을 제시하고 Homo sapiens grch37 어노테이션 파일과 SRR387514 실험 데이터 간의 유사도를 제시한 방법론을 이용하여 측정한 결과 값을 분석하였다.

1. 서론

모든 종(species)은 DNA 라 불리는 유전 물질을 가지며 해당 유전 물질은 유전자(gene)라 불리는 단위로 묶여서 하나의 형질을 나타낸다. 하나의 유전자는 엑손과 인트론이라 불리는 단위들로 이루어지는데 하나의 유전자 내에 존재하는 전체 엑손 중 전체 혹은 그 중 몇 개의 엑손이 묶이어 발현되었을 때의 유전 물질을 RNA 혹은 전사체(transcript)라 부른다.

전사체는 유전자 내에서 발현된 엑손의 뉴클레오타이드 시퀀스 정보에 따라 각기 다른 단백질을 생성한다. 따라서 심각할 경우 뉴클레오타이드 시퀀스 중 하나의 변형으로도 원래의 단백질이 아닌 다른 종류의 단백질을 생성하게 되고 이러한 단백질 종류의 변경이 신체에 어떠한 영향을 끼칠지는 알 수 없다. 따라서 이러한 현상 분석 및 특정 유전 질병, 암 혹은 신약 개발을 위하여 전사체를 분석하는 일의 중요도가 점차 높아지고 있다[1].

현재는 전사체를 분석하기 위해서 전사체 조립 프로그램을 이용하여 실험체의 RNA-Seq 데이터의 전사체 재조립 과정을 수행한 후 조립된 전사체 값을 텍스트 형태 그대로 분석에 사용하거나 시각화 도구를 이용하여 모형의 형태로 시각화하여 비교한다. 모형의 형태로 비교하는 것은 유전자가 엑손과 인트론으로 구성되어 있고 실제 전사체는 유전자의 엑손만이 발현되는 형태를 이용한 것이다. 이를 이용하면 발현된 전사체를 모형으로 나타냈을 경우 유전자의 특정 엑손이 발현되었는지에 대한 이해도가 텍스트 형태로 비교할 때보다 빠르다. 하지만 여러 개의 전사체를 비교해서 봐야 할 경우에는 하나씩 눈으로 확인해야 하며 서로 유사함의 정도를 값으로 표현할 수 없기 때문에 결과에 대한 비교가 힘들다.

본 논문에서는 텍스트 형태의 단순 위치 비교에서 발생하는 이해도 저하의 문제점과 시각화 형태의 위치 비교에서 발생하는 2 개 이상의 전사체 비교 시 확인 속도 저하 및 유사도를 표현할 수 없는 문제점을 해결하기 위하여 위치 정보를 가지는 블록 간의 유사함을 비교할 수 있는 기법을 새롭게 제안한다.

2. 전사체 모형화

전사체를 모형화하기 위해서는 NGS 기계를 통하여 실험체의 RNA-Seq 데이터를 얻고 이를 이용하여 재조립하는 과정을 거쳐야 한다. 전사체를 조립하는 과정은 프로그램마다 조금씩 다르나 크게 2 개의 단계를 거쳐 조립된다. 첫 번째 단계에서는 TopHat[2], SpliceMap[3], MapSplice[4]와 같은 이어맞춤접합(splice junction)을 찾아주는 도구들을 사용하고 두 번째 단계에서 Cufflinks[5], Isolasso[6] 와 같은 도구들에 앞선

단계의 결과 값을 이용하여 전사체를 조립한다.

SpliceGrapher[7]와 같은 도구들은 내부에 두 개의 단계를 모두 가짐으로써 다른 도구를 사용하지 않고 도구 하나만을 사용하여 전사체 조립이 가능하다.

$$transcript_n =< chr, start, end, CDSs_n >$$
$$CDSs_n = \{CDS_1, CDS_2, \dots, CDS_m\}$$
$$CDS_m =< chr, start, end >$$

위의 표현식은 전사체를 텍스트로 나타낼때의 양식을 나타낸 것이다. 모든 전사체 조립 프로그램들의 경우 출력 값으로 텍스트 형태의 위치 정보를 위의 $transcript_n$ 표현식과 같이 제공한다. $transcript_n$ 표현식 내의 CDSs 는 해당 전사체의 CDS 집합을 나타낸 것이다. 유전자로부터 발현된 엑손 영역을 전사체에서는 코딩 시퀀스(coding sequence, CDS)로 명명한다. 실제로는 표현식보다 더 많은 정보를 제공하지만 위의 표현식은 모형화를 위해 필요한 정보만 간략히 표현하였다. 전사체 모형을 시각화하여 나타낼 때에는 전사체의 위치 정보를 이용하여 블록의 형태로 그림 1 과 같이 나타낸다.

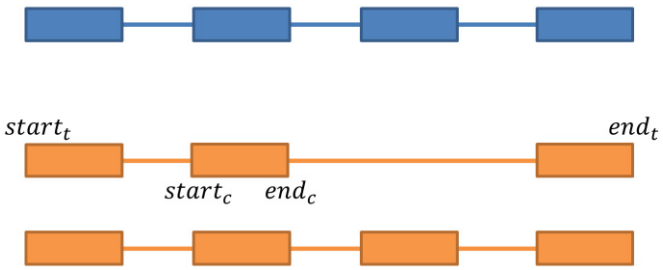

그림 1 표현식의 전사체를 구조화하여 나타낸 모형이다. 파란색 블록들은 유전자를 모형으로 나타낸 것이고 주황색의 블록들은 위의 유전자로부터 발현된 개별의 전사체를 구조화하여 나타낸 것이다. Start 와 end 표시는 위의 표현식에서 나타낸 위치 값들이 어떻게 사용되었는지 나타내준 것이다. $t = transcript$, $c = CDS$

그림 1 의 파란색 블록으로 구성된 모형은 $transcript_n$ 와 같은 유전자 정보를 이용하여 나타낸 유전자 모델이다. 유전자 내의 엑손, 인트론의 집합 중 실제 발현되어 단백질의 생성에 관여하는 엑손 영역을 블록으로 나타냈다. 주황색 블록은 표현식 $transcript_n$의 전사체 정보를 이용하여 전사체 모델을 나타낸 것이다. CDS 를 위치 정보를 가진 블록 단위로 표현하여 유사도를 측정하는데 사용하였다.

3. 전사체 모델의 종류

전사체를 조립하는 과정의 첫 번째 단계에서 어떤 종류의 선택적 이어맞추기(alternative splicing)가 나타나는냐에 따라 전사체 모델의 모양이 결정된다.

선택적 이어맞추기(alternative splicing)란 유전자로부터 전사체가 생성될 때에 그림 1 의 두 번째 전사체 모델과 같이 유전자 내의 엑손이 순서대로 붙는 것뿐 아니라 그림 1 의 첫 번째 전사체 모델과 같이 유전자 내의 엑손이 그림 2 와 같은 다양한 조합으로 이루어지는 것을 말한다.

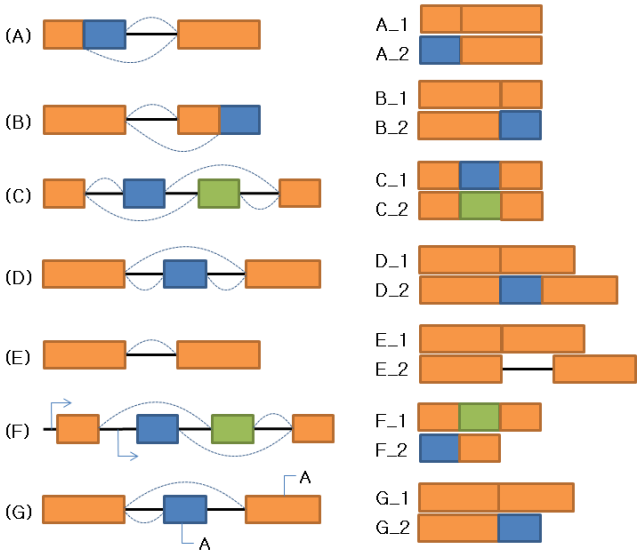

그림 2 선택적 이어맞추기(Alternative splicing)의 여러 가지 유형을 모형화하여 나타낸 것이다. 각 유형의 명칭은 (A) alternative 5' SS (B) alternative 3' SS (C) mutually exclusive exons (D) cassette exon (E) intron retention (F) alternative promoters (G) alternative polyadenylation 이다. 각각의 유형별로 알파벳_1, 알파벳_2 의 그림과 같이 같은 곳에서 전사되더라도 선택적 이어맞추기에 의해 발현되는 RNA 의 형태는 달라지게 된다.

현재 논문으로 발표된 선택적 이어맞추기 유형은 그림 2 와 같이 약 7 가지 경우가 있다[8]. 그림 2 와 같이 선택적 이어맞추기의 유형에 따라 같은 유전자에서 발현된 전사체의 경우에도 여러 가지 모형을 가지게 됨을 유추할 수 있다.

또한 똑같이 발현된 전사체의 경우에도 NGS 기계 및 여러 프로그램을 거치면서 데이터가 손상되어 어떤 기계에서 만들어진 데이터인지 혹은 어떤 프로그램을 통하여 생성된 결과 값인지에 따라 위치 정보가 정확하게 일치하지 않고 에러율에 따라 미세하거나 큰 차이가 있을 수 있다.

이러한 점을 고려하여 서로 완전히 일치하는 전사체 모델부터 서로 비슷하지 않고 극단적으로 위치가 벗어나는 전사체 모델 간의 비교가 가능하고 전사체 A 와 전사체 B 가 존재할 때 A 에 대한 B 의 유사도와 B 에 대한 A 의 유사도를 비교할 수 있도록 위치 정보를 가진 블록 간의 유사도를 측정할 수 있는 새로운 방법을 고안하였다.

4. 위치 정보를 가진 블록 간 유사도 비교 기법

4.1 $Cover_\alpha$를 이용한 유사도 측정 방법

실제 전사체 모델은 다양한 정보를 가지고 있지만

그 중 위치 정보만을 사용하여 전사체의 코딩 시퀀스(coding sequence, CDS) 간의 유사도를 측정하기 위하여 위치 정보를 가진 블록 간의 유사성을 측정할 수 있는 새로운 $Cover_\alpha$라는 방법을 고안하였다.

$$(\mathbf{4.1})\ Cover_\alpha(T_i|T_j) = \frac{\sum_{(s_n,e_n)\in CDSs_i}\sum_{(s_m,e_m)\in CDSs_j} f(s_n,s_m,e_n,e_m)}{n(CDSs_i)}$$

$$f(s_n,s_m,e_n,e_m) = Over_\alpha\left(\frac{L(max(s_n,s_m),\min(e_n,e_m))}{L(s_n,e_n)}\right)$$

$$(\mathbf{4.2})\ L(x,y) = y-x+1 \qquad if\ x<y, 0\ otherwise$$

$$(\mathbf{4.3})\ Over_\alpha(x) = 1 \qquad if\ x \geq \alpha, 0\ otherwise$$

식 (4.1)에서 s_n, e_n는 $CDSs_i$에 대한 시그마 계산 동안 CDS_n의 시작 위치와 끝 위치를 나타낸다.

위의 $Cover_\alpha$ 계산식을 이용하여 전사체 T_i와 전사체 T_j에 대하여 2 번에 걸쳐 서로에 대한 유사도를 계산할 수 있다. 전사체 T_i에 대한 전사체 T_j의 유사도와 전사체 T_j에 대한 전사체 T_i의 유사도는 그림 3 과 같이 다를 수 있다.

그림 3 위의 그림은 파란색 전사체 모형과 주황색 전사체 모형 간의 $Cover_\alpha$ 계산 값이 서로 다를 수 있음을 보여주는 모형도이다. α값을 100 이라고 하였을 때 주황색 전사체 모형은 $Cover_{100}(Orange|Blue) = \frac{2}{3}$가 되고, 파란색 전사체 모형의 경우에는 $Cover_{100}(Blue|Orange) = \frac{4}{4} = 1$ 로 서로 다른 값이 되는 것을 확인할 수 있다.

그림 3 의 파란색(Blue)과 주황색(Orange) 전사체 모형에 대한 $Cover_\alpha$ 는 $Cover_\alpha(Orange|Blue)$ 와 반대의 $Cover_\alpha(Blue|Orange)$의 2 개의 경우를 구할 수 있다. 우선 각 전사체를 구성하는 블록인 CDS 의 개수를 계산한다. 주황색 전사체 모델의 경우 $n(CDSs_O) = 3$ 이고 파란색 전사체 모델은 $n(CDSs_B) = 4$이다. α값은 사용자가 임의로 설정해 줄 수 있다. 계산의 편리성을 위해 100 으로 가정한다면 $Cover_{100}(O|B)$의 분자 값은 2, $Cover_{100}(B|O)$의 분자 값은 4 가 된다. 따라서 $Cover_{100}(O|B) = \frac{2}{3}$, $Cover_{100}(B|O) = \frac{4}{4} = 1$ 의 값을 가지게 된다.

$Cover_\alpha$는 서로 다른 전사체 모델간의 유사도를 모델을 구성하는 전체 블록 개수 중 상대 모델과 α값 이상으로 일치하는 블록의 개수를 보여줌으로써 사용자가 커버된 블록 개수 및 확률 값을 확인하여 대용량의 전사체 그룹에서도 비슷한 위치에 존재하는 전사체의 유무와 서로 간의 유사도를 확인할 수 있다.

하지만 이 방법의 단점은 단순화된 표현으로 인하여 CDS 길이와 전사체 간의 길이 차이에 대한 정보가 반영되지 않고 블록 단위의 커버 개수를 사용함으로 인하여 유사도의 확률 값이 일률적으로 나타나고 유사도가 실제 블록 간의 유사함에 비해 α 값 이상일 경우 획일화되는 과정으로 인하여 높거나 낮게 측정될 수 있다. 이를 해결하기 위하여 확장형 $Cover_\alpha$ 방법을 고안하였다.

4.2 $Cover^+{}_\alpha$를 이용한 유사도 측정 방법

$Cover_\alpha$ 계산법은 각각의 전사체를 구성하는 블록 간의 명확한 유사도보다 α 값 이상으로 일치하는 개수에 대한 정보만을 원할 경우 유용하나 명확한 유사도를 원할 경우 블록 간의 커버된 개수에 대한 확률 값이 되기 때문에 전사체의 길이나 블록의 길이에 대한 정보가 빠져서 유사도의 정확도가 떨어지게 된다. 예로 그림 3 의 경우에서 주황색 전사체의 경우 전체 3 개의 CDS 중 2 개의 CDS 가 일치함에 따라 $Cover_\alpha = \frac{2}{3}$라는 값이 되었지만 실제로는 일치하지 않는 CDS 의 길이가 일치하는 다른 2 개의 CDS 길이의 합보다 길기 때문에 2/3의 확률보다 더 낮은 확률이 되야 한다. 따라서 각각의 CDS 와 전체 전사체의 길이를 고려하는 확장형 $Cover_\alpha$, 즉 $Cover^+{}_\alpha$를 새롭게 고안하였다.

$$(\mathbf{5.1})\ Cover^+{}_\alpha(T_i|T_j) = \frac{\sum_{(s_n,e_n)\in CDSs_i}\sum_{(s_m,e_m)\in CDSs_j} f(s_n,s_m,e_n,e_m)}{\sum_{(s_n,e_n)\in CDSs_i} L(s_n,e_n)}$$

$$f(s_n,s_m,e_n,e_m) = Over^+{}_\alpha\left(L(max(s_n,s_m),min(e_n,e_m)),L(s_n,e_n)\right)$$

$$(\mathbf{5.2})\ Over^+{}_\alpha(x,y) = x \qquad if\ \frac{x}{y} > \alpha, 0\ otherwise$$

$$(\mathbf{5.3})\ \mathrm{wCover}_\alpha(T_i,T_j) = (Cover^+{}_\alpha(T_i|T_j) + Cover^+{}_\alpha(T_j|T_i)) \cdot \frac{\sum_{(s_n,e_n)\in CDSs_i} L(s_n,e_n)\cdot\sum_{(s_m,e_m)\in CDSs_j} L(s_m,e_m)}{\sum_{(s_n,e_n)\in CDSs_i} L(s_n,e_n) + \sum_{(s_m,e_m)\in CDSs_j} L(s_m,e_m)}$$

확장된 버전은 기존 $Cover_\alpha$ 측정 방식에서 전사체 길이에 대한 계산이 보강된 측정 방법으로 $Cover_\alpha$ 와 같이 전체 구간 중 일치하는 구간의 길이 비율까지 구한 후, α 값 이상이 될 경우 일치하는 구간 길이의 합을 전체 전사체 길이와 나누어 주는 방식을 사용함으로써 그림 3 과 같은 경우에 $Cover_\alpha$ 방식보다 더 정확한 유사도를 계산할 수 있다.

또한 $Cover^+{}_\alpha$를 이용하여 전사체 $T_i|T_j$혹은 $T_j|T_i$에 대한 것만 아니라 T_i, T_j의 유사도를 식(5.3)와 같은 방식으로 계산할 수 있다. wCover_α 계산식을 이용함으로써 서로 간의 유사도를 높은 순, 낮은 순으로 정렬이 가능해진다. 이와 같은 실수 값을 사용하여 사용 자가 원하는 유사도 정도를 즉각적으로 선택하여 확인할 수 있다.

표 1 Grch37 와 SRR387514 데이터 전사체 간의 유사도를 α에 대하여 측정한 값

α	Grch37		SRR387514		$Cover_\alpha$		$Cover^+{}_\alpha$		$wCover_\alpha$
	start	End	Start	end	$SRR387514 \mid Grch37$	$Grch37 \mid SRR387514$	$SRR387514 \mid Grch37$	$Grch37 \mid SRR387514$	
50	880077	894620	887377	887983	1.00	0.11	0.98	0.15	0.25
	880077	894620	883504	884041	1.00	0.11	0.77	0.10	0.17
	2518235	2520863	2518254	2519292	0.67	0.43	0.75	0.57	0.64
90	880077	894620	887377	887983	1.00	0.11	0.98	0.15	0.25
	880077	894620	883504	884041	0.50	0.11	0.36	0.10	0.13
	2518235	2520863	2518254	2519292	0.67	0.29	0.75	0.38	0.52

5. 실험 및 결과

5.1 실험 데이터

실험 데이터로는 Ensembl[9]로부터 제공되는 homo sapiens grch37 의 유전자 모델 어노테이션 파일의 전사체 모형 데이터와 NCBI[10]에서 제공하는 human embryonic kidney cell 타입의 RNA-Seq 데이터인 SRR387514 를 tophat, isolasso 프로그램을 이용하여 조립한 전사체 결과 파일 중 1 번 염색체의 일부 데이터를 이용하였다.

실험은 실제 데이터를 $Cover_\alpha$, $Cover^+{}_\alpha$와 $wCover_\alpha$를 이용하여 전사체 간의 유사도를 측정한 값이 α값이나 사용한 측정 방식에 따라 어떻게 다르게 나타나는지 확인하는 방향으로 진행되었다.

5.2 결과

표 1 은 Grch37 어노테이션 파일의 7661 개의 전사체와 SRR387514 데이터의 7652 개의 전사체 간의 유사도를 비교한 결과 값 중 일부이다.

α값에 따라 2 개의 그룹으로 나뉘고 각 그룹의 실험 데이터는 같다. SRR|Grch37 에 대한 계산 값들의 경우 커버된 블록의 개수를 이용한 $Cover_\alpha$ 값과 달리 길이 정보가 포함된 $Cover^+{}_\alpha$ 값이 다르게 나타나는 것을 확인할 수 있다. 특히 두 번째 데이터의 유사도는 $Cover_\alpha$에 비해 더 낮게 나타나고 세 번째 값의 경우 더 높게 나타나는 것을 확인할 수 있다. 실제 CDS 데이터를 이용한 검증 단계에서 $Cover^+{}_\alpha$ 방식의 유사도가 모형 간의 실제 유사 정도와 비슷하게 표현되었음을 확인하였다.

또한 $wCover_\alpha$ 값을 이용하여 서로에 대한 유사도가 아닌 둘 간의 유사도를 계산하였다. $Cover^+{}_\alpha$ 방식의 SRR|Grch37 에 대한 결과 값과 Grch37|SRR 에 대한 결과 값은 차이가 많이 나는데 이를 단순히 양 쪽의 결과 값을 더한 후 $1/2$을 하는 계산보다 $wCover_\alpha$를 이용한 방식이 더 정확하게 두 개의 전사체 간의 유사도를 나타냄을 확인하였다. $Cover_\alpha$의 경우 α값이 50 일 때, 첫 번째와 두 번째 데이터의 경우 서로 간의 유사도가 같을 것으로 예상되나 $wCover_\alpha$의 값과 같이 실제로는 다르게 나타남을 확인할 수 있다.

6. 결론

어노테이션 데이터와 실험데이터 간의 비교나 서로 다른 실험군 간의 비교가 필요할 때 단순한 위치 값의 비교나 모델의 모형 형태에 따른 시각화 비교가 아니라 $Cover_\alpha$ 와 $Cover^+{}_\alpha$ 방법을 사용하여 전사체 모델 간의 유사도를 확률 값으로 측정할 수 있게 되었다. 이러한 수치 비교를 통하여 사용자의 전사체 모델 간의 유사도에 대한 이해도를 높이고 유사도 값에 따른 데이터 정렬 및 특정 전사체 모델의 탐색이 가능할 것으로 예상된다.

Acknowledge

본 연구는 질병관리본부 학술연구용역과제(2012-E72006-00) 연구비를 지원받아 수행되었습니다.

참고문헌

[1] Y. Takei, K. Kadomatsu and et.al., “A Small Interfering RNA Targeting Vascular Endothelial Growth Factor as Cancer Therapeutics”, Cancer Research, Vol. 64, No. 10, pp. 3365-3370, 2004

[2] C. Trapnell, L. Pachter and S. L. Salzberg, “TopHat: discovering splice junctions with RNA-Seq”, Bioinformatics, Vol. 25, No. 9, pp. 1105-1111, 2009

[3] K. Au, H. Jiang and et al., “Detection of splice junctions from paired-end RAN-seq data by SpliceMap”, Nucleic Acids Research, Vol. 38, No. 14, pp. 4570-4578, 2010

[4] K. Wang, D. Singh and et al., “MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery”, Nucleic Acids Research, Vol. 38, No. 18, pp. e178, 2010

[5] C. Trapnell, B. A. Williams and et al., “Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation”, Nature Biotechnology, Vol. 28, No. 5, pp. 511-515, 2010

[6] W. Li, J. Feng and T. Jiang, “IsoLasso: A LASSO Regression Approach to RNA-Seq Based Transcriptome Assembly”, Journal of Computational Biology, Vol. 18, No. 11, pp. 1693-1707, 2011

[7] M. Rogers, J. Thomas and et al., “SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data”, Genome Biology, Vol. 13, No. 1, pp. 1-17, 2012

[8] P. Rajan, D. J. Elliott and et al., “Alternative splicing and biological heterogeneity in prostate cancer”, Nature Reviews Urology, Vol. 6, No. 8, pp. 454-460, 2009

[9] European Molecular Biology Laboratory, Ensembl http://asia.ensembl.org/index.html

[10] National Center for Biotechnology Information, NCBI, http://www.ncbi.nlm.nih.gov/