서열 및 상호작용 정보를 활용한 이종간 유사 기능 단백질 추출

설영주[1O] 김민경[2] 유성준[1] 박선희[3]
[1]세종대학교 컴퓨터 공학과
[2]이화여자대학교 컴퓨터 공학과
[3] 한국전자통신연구원
neutrian@macrogen.com[O], minkykim@ewha.ac.kr, sjyoo@sejong.ac.kr , shp@etri.re.kr

Ortholog protein finding System based on protein sequence and interaction information.

Young-Joo Seol[1O], Min Kyung Kim[2], Seong-Joon Yoo[1] , Seon Hee Park[3]
[1]School of Computer Engineering, Sejong University
[2]Department of Computer Science and Engineering, Ewha University
[3]Bioinformatics Team, Electronics and Telecommunications Research Institute

요 약

단백질 간 상호작용은 생물체 내에서 발생하는 모든 생명 현상을 이루는 기본 단위로써, 이를 종 수준에서 밝히고자 하는 시도가 yeast와 초파리,worm 등에서 보고되었다. 대량으로 존재하는 상호작용 데이터들은 종래에 서열로 시도되던 유연관계 비교 및 기능 유추 등에 기본 정보로 활용되고 있다. 본 연구에서는 다른 종에 속하는 동일 기능 단백질 즉, ortholog를 찾음에 있어, 기존의 서열 접근 방식 이외에 상호작용 정보를 추가로 사용하는 시스템을 고안하여 서열방식만을 활용하던 이전의 방식이 지니는 문제점을 극복하고자 하였다.

1. 서 론

생물학 정보들은 서로 이질적인 데이터를 참조하여 데이터간의 신뢰도를 측정하거나, 혹은 개선된 성능을 보이는 도구들을 개발하는 노력을 하고자 하는 추세에 있다. 특히 상호작용 정보의 경우에 있어서는 기능 예측, 새로운 컴플렉스(complex)의 발견, 생물학적 프로세스의 예측 등에 사용되기 시작한 정보로 유전자 발현 데이터등과 같이 사용되기도 한다. 이에 ortholog를 찾음에 있어서도 기존의 서열 데이터만을 사용하던 방식에서 벗어나 상호작용, 발현, 구조 등의 정보 등이 총체적으로 사용 되어질 것으로 예상된다.

본 연구에서는 다른 종에 속하는 동일 기능 단백질 즉, ortholog를 찾음에 있어, 기존의 서열 접근 방식이외에 상호작용 정보를 활용하여 찾아주는 시스템을 고안하고자 한다. 이는 서열 기반으로 찾아진 ortholog의 신뢰도를 나타내주는 척도가 될 수도 있으며, 서열의 유사도가 낮더라도 상호작용 패턴이 유사하면 *ortholog* 후보가 될 수 있다.

Ortholog 단백질은 이미 많은 연구가 진행된 종의 연구를 새로운 종에 적용 가능한 장점이 있다. 즉 기능 추정, 상호작용 추정 혹은 보존된 생체정보 경로 찾기 등에 있어 가장 기본적인 정보가 되기 때문에 정확한 ortholog를 찾는 것은 서열정보와 종수준에서의 상호작용 정보가 증가하는 현 시점에서 중요한 작업이 될 것으로 보인다.[1][2].

2. InterBlast 시스템

InterBlast 시스템은 다른 종에 속하는 유사 단백질(homolog) 즉, ortholog를 찾음에 있어 기존의 서열 기반 접근 방식이외에 상호작용 정보를 활용하여 서열 정보가 가지는 단점을 보완하면서 ortholog를 찾아주는 시스템이다.

2.1 InterBlast 시스템의 구성

InterBlast 시스템은 크게 세 개의 모듈로 나눌 수 있다. 전체 시스템 구성도는 <그림 1> 과 같다.

(1) Sequence Matching score Module 에서는 두 종의 Genome 서열을 블라스트로 돌린 결과를 가지고 매칭 스코어를 이용해서 두 종간의 유사도가 있는 단백질 쌍을 찾는다. 이때 적정한 threshold값을 가지고 서열 유사도가 아주 낮은 단백질들은 버리게 된다. 이 단계에서 블라스트는 비교 방향에 따라 결과 값이 틀려 짐으로 Reciprocal Best Hit 방법으로 블라스트의 결과를 처리한다.

(2) Interaction path alignment Module은 두 종의 상호작용 정보를 이용해서 두 개의 상호작용 네트웍을 만든다. 두 개의 상호작용 네트웍을 다시 하나의 Global 네트웍으로 만들게 되는데 이때 서열 유사도를 가지고 이러한 작업

을 수행하게 된다. 이 Global 네트웍은 블라스트에서처럼 gap과 mismatch의 정보를 각 에지에 가지고 있게 된다. 이는 전체 서열 유사도와 두 개의 단일한 네트웍으로 가능한 모든 노드와 에지들로 만들어 진다. 결국 이를 통해 상호작용 정보를 점수화 할 수 있게 된다.

(3) Resove Module에서는 이 두 정보를 가지고 두 종 사이에 가장 유사한 단백질 쌍을 찾게 된다. 또한 이를 중심으로 paralog를 찾게 된다.

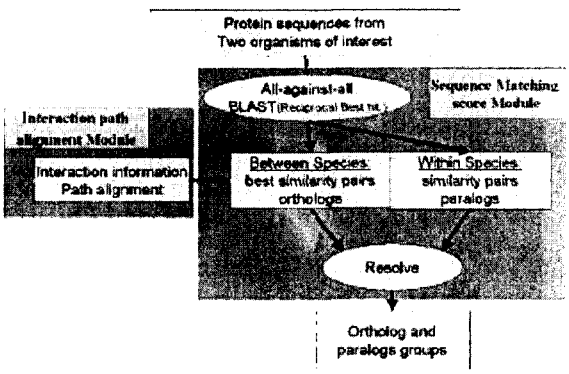

그림 1 InterBlast 시스템 구성도

2.2 InterBlast 시스템의 기능

첫 번째 단계에서는 두 종을 A-A,B-B,A-B,B-A로 방향을 바꿔가면서 Blast를 실행시킨다. 그 결과에서 50비트 이하의 값들로 매칭이 된 쌍은 고려대상에서 제외한다.. 두 번째 단계에서는 각 종의 상호작용 정보를 가지고 각각의 상호작용 네트웍을 만든다. 세 번째 단계에서는 서열 유사도를 보고 Global 네트웍을 만들게 된다. 이 Global 네트웍은 매칭된 두 종의 모든 단백질쌍을 노드로 가지고 있다. 즉, A:B,A:C 와 같은 노드가 존재한다는 것이다. 그리고 이 노드들 사이에는 에지가 존재하는데 에지에 가중치를 줄때 direct, mismatch, gap 에 따라 가중치가 틀리다. 각 노드는 서열 유사도를 갖는 것끼리 묶고 이에 상호작용 정보로 합쳐진 노드들을 에지로 연결할 때 유사도가 있는 노드들끼리 상호작용 정보도 존재하면 Direct로 연결하고, 하나 건너 즉 gap이 존재할 때는 Gap으로 설정하고, 두 개의 노드가 모두 서열 유사도가 있는 노드 사이에 서열 유사도가 없는 노드가 존재할 때는 Missmatch로 설정해서 각 에지에 각각의 점수를 할당하게 된다. 이는 Blast에서 시퀀스 유사도를 점수화할 때의 방법과 비슷한 방법이다. 이 글로벌 네트웍을 가지고 상호작용 점수를 시퀀스 유사도와 합해서 매치된 쌍에 대한 점수를 매기게 된다.

<수식 1>에서 보면 sum_score는 서열 유사도와 상호작용 정보의 점수의 합계이고 seq_score_percent와 interaction_percent는 서열정보를 전체 점수에 차지하는 비율을 조정하는 것이 된다 (현재 디폴트는 서열 유사도 정보와 상호작용 정보를 각각 50%씩 할당하고 있다). 또한 seq_match_score는 블라스트를 실행했을 때 나오는 서열 유사도 정보이고 lengthQ와 lengthT는 쿼리 단백질의 서열 길이와 데이터베이스 단백질의 서열 길이를 나타낸다. sum_interaction_score는 글로벌 네트웍 상에서 노드에 있는 에지들에 할당된 점수들을 합한 점수이고 single_edge_score는 각 종의 상호작용 네트웍에서 노드에 있는 에지 하나의 점수로 여기서는 10으로 하고 있다. graph_a_edges_size 와 graph_a_edges_size 는 각 종의 상호작용 네트웍의 각각의 노드에 있는 에지수이다.

$$sum\_score = seq\_score\_percent*(seq\_match\_score/(lengthQ+lengthT)) + interaction\_percent*((sum\_interaction\_score)/(sigle\_edge\_score*0.5*(graph\_a\_edges\_size+graph\_b\_edges\_size)))$$

수식 1 서열 유사도와 상호작용 정보의 수식

네 번째 단계에서는 중심 ortholog를 찾는다. 각 클러스터 별로 이전 단계에서 계산된 점수를 비교해서 가장 큰 점수를 갖는 단백질이 각 종의 중심 ortholog가 된다. 다섯 번째 단계에서는 중심 ortholog의 paralog를 찾는다. 이는 각 종의 중심 ortholog와 서열 유사도가 있는 것들을 후보로 해서 각각 클러스터를 만든다.이 단계에서는 동일한 단백질이 다른 클러스터에 존재할 수 있다.

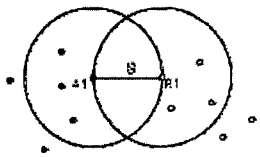

그림 2 Cluster of orthologs and paralogs

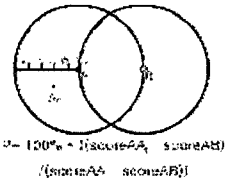

그림 3 Paralog의 confidence value

<그림 2>에서 검은색과 회색 점은 각각의 종의 서열 유사도 점수를 나타낸다. 여기서 중심 Ortholog는 A1과 B1이 된다. 이 둘의 서열 유사도 점수는 S로 나타낸다. 여기서 같은 원안에 있는 점들은 각각의 paralog를 나타내는데 이 들은 다른 종과의 ortholog보다 같은 종의 중심 ortholog와 거리상으로 더 가까이 있다고 가정하고 있다. 따라서 점수 S보다 얼리 있는 점들에 대해서는 paralog로 고려하지 않게 된다. Paralog를 고려할 때는 거리를 나타내는 S 점수를 서열 유사도만을 고려했다. 5번째 단계에서는 paralog 후보들에 대해 중심 ortholog 와 paralog사이의 거리를 정수로 하는 신뢰 값를 계산한다. <그림 3>을 보면 그 신뢰 값(confidence value)을 계산하는 수식과 그림이 나와 있다. 신뢰 값은 중심 ortholog와의 유사도에 따라 0%에서부터 100%까지 나올 수 있다. 중심 ortholog의 신뢰 값은 따라서 항상 100%가 된다. InterBlast는 서열 유사도와 상호 작용 정보를 이용해서 ortholog를 찾아내기 때문에 중심 ortholog 보다 더 높은 서열 유사도를 갖는 paralog들이 존재 할 수도 있다. 6번째 단계에서는 paralog와 ortholog의 클러스터들이 다른 클러스터들과 겹치는 부분을 해결하는 작업을 몇 가지 정해진 규칙에

따라 하게 된다.

규칙1은 중심 Otholog A2와 B2가 이미 더 점수가 높은 그룹 A1과 B2에 속해 있다면 같은 클러스터로 합치게 된다. 규칙2는 중심 Ortholog A와 두개의 같은 점수를 갖는 B1 과 B2 가 있을 때 이들을 합친다. 규칙3은 같은 클러스터 안의 중심 Otholog A1이 A2보다 더 높은 점수를 가질 때 A2와 Ortholog관계인 B2를 제외한다. 이때 (Score(A2-B2) -Score(A1-B1) > 0.5 이면 제외하게 된다. 규칙4는 새로운 Otholog 후보가 이미 다른 그룹에서의 신뢰 값이 50% 이상일 때 합친다. 규칙5는 이전 규칙들에서 남은 paralog 그룹의 겹치는 부분 중심 ortholog의 점수에 따라 분리 한다. 이 작업이 끝나면 하나의 클러스터에 신뢰 값이 100%인 ortholog가 여러 개 존재할 수도 있게 된다. 규칙4는 새로운 Otholog 후보가 이미 다른 그룹에서의 신뢰 값이 50% 이상일 때 합친다. 규칙5는 이전 규칙들에서 남은 paralog 그룹의 겹치는 부분 중심 ortholog의 점수에 따라 분리한다. 위의 Resolve Module은 Inparanoid System이 사용하는 방식을 따랐다[3].

3. Experiment

3.1 Data

상호작용 정보는 DIP에서 가져왔다. 여기서 우리는 ***Drosophila melanogaster***의 7052개의 단백질에 대한 20789개의 상호작용 정보와 ***Saccharomyces cerevisiae*** 의 4749개의 단백질에 대한 15131개의 상호작용 정보를 가져왔다. DIP에 있는 단백질 정보는 Inparanoid의 단백질 정보와 다소 차이가 있다. 단백질 ID 가 달라서 이들을 매칭 시키기 위해 단백질 서열 매칭을 했다. 따라서 원래의 상호작용 정보에서 fly는 18197개의 상호작용 정보와 yeast의 경우는 12868개의 상호작용 정보만이 유효하다.

Organism	Dip protein	Inparanoid Protein	Same protein
DM(fly)	7052	18931	6628
SC(yeast)	4749	6705	4398

표 1 DIP data와 Inparanoid Data 의 중복된 단백질 데이터

3.2 결과

<표 1>의 데이터를 가지고 InterBlast를 실행하면 1782개의 ortholog 클러스터가 나온다. 이 결과를 이전의 시스템인 Inparanoid와 비교하면 <표 2>와 같다.

	InterBlast	InParanoid	Identical	Different	
				Miss	Order
Source protein	2947	3792		353	126
Target protein	2143	2473		27	42
cluster group	1782	1963	1308		

표 2 InterBlast와 InParanoid의 결과 비교표

4. 결론

상호작용 정보의 비중을 낮출 수록 서열만으로 작동하는 Inparanoid와 비슷한 결과를 나타낸다. 이에 비해 상호작용 데이터의 비중을 높임에 따라서 결과는 Inparanoid와 달라진다.

본 연구에서 사용하는 상호작용 데이터의 경우 false positive 혹은 false negative 데이터 등이 존재함이 알려져 있다. false negative란 상호작용을 하는 단백질임에도 불구하고 아직은 알려지지 않은 것들로써 본 연구에서 사용된 초파리의 상호작용 개수는 yeast 보다 적은 상태이다 또한 false positive data 들은 상호작용을 하지 않음에도 상호작용 하는 것으로 알려진 것들로 본 연구에서 ortholog를 찾음에 있어 스코어에 영향을 줄 가능성이 있다. 이와 같이 상호작용 데이터 자체의 문제점으로 인하여 전체 시스템의 성능에 영향을 줄 가능성이 충분히 있을 것으로 보인다.

그럼에도 불구하고 상호작용의 숫자나 그 신뢰도는 앞으로도 더 개선될 것으로 생각되며, 이러한 상황에서는 상호작용 데이터를 사용하는 것이 서열 데이터만을 사용하는 것보다 더욱 강력한 수단이 될 것으로 보인다. 즉 서열로는 상동성이 떨어짐에도 불구하고 상호작용 파트너간에 동일 기능 단백질로 둘러 쌓인 것들을 찾아주는 데 기여할 수 있을 것으로 보인다.

상호작용 파트너들의 유사성을 검색하는 수단으로 현재는 서열 정보를 다시 사용한다. 따라서 서로 간에 실재적인 ortholog 관계가 있는 것이 아니라도 서열 유사성만 있는 경우도 상호작용 파트너가 유사한 것으로 보는 점은 앞으로 더 개선 해야 할 것으로 보인다.

5. Reference

[1] Tao-Wei Huang et al, POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome, Bioinformatics, doi:10.1093/bioinformatics/bth366., 2004

[2] Lisa R. Matthews et al, Identification of Potential Interaction Networks Using Sequence-Based Searches for Conserved Protein-Protein Interactions or "Interologs", Genome Res., 11,2120 - 2126, 2001

[3] Remm M, Storm CE et al, Automatic clustering of orthologs and in-paralogs from pairwise species comparisons., J Mol Biol. 314(5), 1041-52, 2001