계통유전체학과 COG를 이용한 유전자 기능예측

Gene Prediction Using Phylogenomics and COG

신창진[1], 강병철[2], 박준형[2], 신동훈[4], 김철민[1,2,3]

[1](주) 진인 유전체 의학 연구소

[2]부산대학교 대학원 생물정보협동과정

[3]부산대학교 의과대학 생화학교실

[4]국립암센터 암정보연구과

Chang-Jin Shin[1],Byeong-Chul Kang[2], Jun-Hyung Park[2], Dong-Hoon Shin[4], and Cheol-Min Kim[1,2,3]

[1]Institute for Genomic Medicine, GeneIn. Co., Ltd., Busan, Korea

[2]Interdisciplinary Program of Bioinformatics, Graduate School,Pusan National University

[3]Department of Biochemistry, College of National University, Pusan National University, Busan, Korea

[4]Cancer Information Branch, National Cancer Center Research Institute

E-mail : teragene@hotmail.com

요약

본 연구는 유전자 기능예측에 있어서 유사성 검색과 비교유전체학이 가진 한계를 극복하기 위하여 9종의 Human Herpesvirus를 대상으로 COG와 계통유전학적 방법을 적용하여 향상된 유전자 기능예측을 하고자 하였다. COG의 방법을 이용하여 114 HCOGs (Human Herpesvvirus COGs)를 구축하고, HCOGs를 바탕으로 유전자 컨텐츠트리를 제작하였다. 이 트리를 통하여 각 HCOG는 α-특이적 그룹, β-특이적 그룹, α, β, γ-특이적 그룹 중 하나에 속함을 보였다. 계통유전체학의 적용을 위하여 α, β, γ-특이 그룹에 속하는 ORF중 DNA polymerase를 이용하여 종트리를 제작하였다. SDI (Speciation and Duplication) 알고리즘을 통하여 148개의 당단백질에서 47개의 복제점을 예측하였고, 초기 HCOG의 제작에서 제외되었던 7 ORF는 당단백질과 관련된 5개의 HCOG로 재 정의 하였다. 이 연구를 통하여 COG는 ortholog 그룹을 클러스터링하는데 효과적인 방법이며, 이를 더욱 보완할 수 있는 방법으로 비교유전체학이 사용될 수 있음을 확인하였다. 이는 비교유전체학의 방법과 계통유전체학적 방법을 조화시켜 유전자 기능 예측을 보완할 수 있음을 보여 주었다.

1. 서론

게놈프로젝트 이후 포스트 게놈 시대로 접어들면서, 가장 큰 화두는 유전체 정보의 분석이다. 많은 유전자 서열들의 기능이 실험적으로 증명되지 않은 상태에서 제공되고 있기 때문에, 유전체의 기능 예측을 위한 생물정보학적 방법의 중요성이 고조 되고 있다.

생물정보학적 유전자 기능 예측은 기능이 알려진 유전자와 미지의 유전자가 염기서열의 유사성(similarity)이 있을 경우 기능적으로 유사한 역할을 할 것이라는 전제로, 알려진 서열 데이터베이스와 비교하여 얻어진 가장 유사성이 높은 결과로부터 그 유전자의 기능을 예측할 수 있다. 그러나, 유사한 서열을 가진 유전자들은 다양한 기능을 나타내기 때문에 서열 유사성만으로 동일한 기능을 예측하기는 어렵다.

생물은 종간에 진화적으로 보존적인 유전자를 가지며 유사한 기능을 나타낸다. 근연종 간에 유사한 기능을 하는 단백질은 그 유전자 서열간에도 유사성이 높다는 가정아래 종간 보존적인 유전자 즉, ortholog그룹을 모으는 COGs (Clusters of Orthologous Groups)가 제안되었다[1]. 그러나, COGs방법은 여러 종의 ortholog그룹을 탐색하는데 획기적인 방법이나 서열 유사성에 의한 클러스터링 방법에 의존함으로써 유전자들간의 ortholog와 paralog를 구

별하기 힘들다.

비교유전체학적 연구방법을 보완하기 위해 Eisen[2]은 계통유전체학(phylogenomics)을 제안하였다. 계통유전체학적 방법은 새로운 유전자의 기능 예측을 위해 상동성이 있는 유전자들을 알려진 데이터베이스를 통해 검색하고, 검색된 유전자들과 미지의 유전자를 포함한 계통수에 알려진 표준 계통수를 적용해서 진화적 관계를 반영하여 새로운 유전자가 어떤 ortholog나 paralog그룹에 속하는지를 유추한다. 이로써 상동성 유전자 그룹 내에서 ortholog와 paralog를 구분할 수 있다[3].

본 연구에서는 유전자 기능 예측방법으로 사용되고 있는 유사성 검색 및 비교유전체학이 가진 한계를 극복하기 위하여 계통유전체학을 도입하여 상동성 유전자 그룹내에서 복제점을 예측하고, 그로부터 정확한 ortholog와 paralog를 구분함으로써 각 방법들이 가진 한계를 극복하여 보다 향상된 annotation결과를 얻고자 시도되었다.

· 연구대상으로 특히 다른 생명체에 비해 유전자 변이가 심한 바이러스에 대해서 진화적 유연관계를 기반으로 바이러스 유전자의 기능을 예측하고자 하였다. 이를 위해 유사성에 의한 기능 예측, 비교유전체학에 의한 기능 예측을 수행하였고 진화적 유연관계를 반영하기 위해 계통유전체학적 기법을 적용하여 유전자의 기능을 예측하고 앞의 두 방법의 결과와 비교하였다.

2. 실험방법

본 연구는 2단계의 과정(COG technology part, Phylogenomics part)을 거쳐 수행되었다 (그림 1).

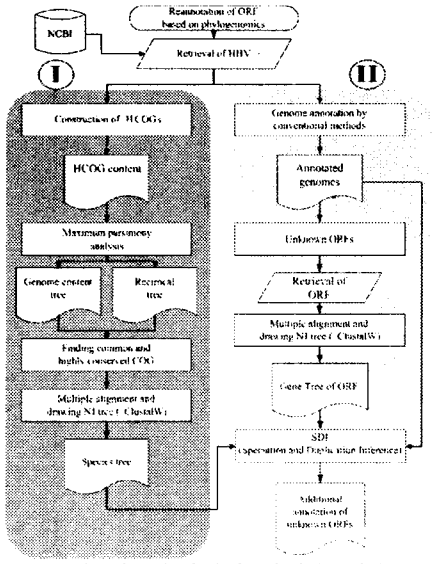

그림1. COG와 계통유전체학 방법을 이용한 유전자 예측 흐름도

2.1. COG technology part

HCOGs 구축

NCBI로부터 9 Human Herpesvirus (HHV)의 모든 ORF서열(942개)들을 다운로드하고, 이 서열들을 Tatusov[1]의 COG방법을 응용하여 114개의 HCOG를 구축하였다.

유전자 컨텐츠 트리 및 reciprocal 트리 작성

모든 HHV에 대하여, HCOG가 각 HHV에 존재하는지, 존재하지 않는지를 1과 0으로 할당 하여 HCOG membership matrix를 만들었다. 각 행은 0과 1의 이진 이산 서열을 나타내고, 이를 maximum parsimony방법을 이용하여 유전자 콘텐츠트리와 reciprocal 트리를 제작하였다. 특히, reciprocal 트리는 HHV의 서브패밀리에 대한 보존적인 HCOG를 알 수 있게 하였다.

종 트리를 제작하기 위한 대표적 ORF 선정

9 HHV의 ORF를 모두 포함하는 HCOG 중 DNA polymerase를 포함하는 HCOG의 ORF를 종트리 구성재료로 선정하고, clusterW[4]를 이용하여 종트리를 구성하였고, 이는 SDI 알고리즘[5]에 적용하였다.

2.2. Phylogenomics part

9 HHVs의 ORF 서열 annotation

전통적인 유전자 명명 방법으로 EC number, PIR, BLOCK, Prosite, Pfam, PDB의 데이터베이스를 통해서 942개의 ORF를 명명하였다.

Glycoprotein을 가진 HCOG 선정

HCOG 중 glycoprotein을 가진 HCOG들을 추출하고, 이를 상동성 유전자 그룹으로 하였다.

SDI 알고리즘 적용 및 ortholog ORF 찾기

상동성 유전자 그룹 및 HCOG 제작에서 빠진 싱글 ORF와 함께 유전자트리를 제작하고, 파트 1에서 제작된 종트리를 SDI 알고리즘에 적용하여 복제점을 예측하고, HCOG를 재 정의하였다.

3. 결과 및 고찰

HCOGs 구축

9종의 HHV에 대해서 blastp를 수행한 결과 942개의 BLAST 결과를 얻었으며, 이는 NCBI에 annotation 되어있는 9종의 모든 ORF 수와 일치함을 확인하였다. BLAST 결과를 파싱하여 종간 상호 best hit를 가지는

14,347개의 best hit쌍을 얻었으며, 모든 best hit쌍을 그래프로 표현하여 3종간의 best hit을 보이는 최소 COG를 얻었다. 그 결과 6,884개의 최소 COG가 생성되었으며, 이 최소 COG를 merging하여 최종적으로 114개의 HCOGs를 구성하였다.

모든 HCOG들을 확인한 결과 하나의 HCOG를 구성하는 ORF수는 최대 10개이며, 최소 3개였다. 9종의 HHV를 대상으로 하였으므로 이상적인 최대 ORF의 수인 9를 초과한 HCOG는 어느 한 종에서 paralog가 있음을 뜻하는 것으로 간주할 수 있다. ORF수 10을 가지는 HCOG는 HCOG010 (tegument protein)와 HCOG039 (capsid protein)으로 2개였으며 각각 HHV8과 HHV3에서 유전자 복제가 있었음이 예상된다.

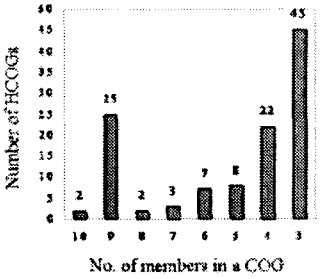

그림2. HCOG를 구성하는 ORF수에 따른 HCOG 분포도

유전자 컨텐츠트리 및 reciprocal 트리 작성

0또는 1로 구성되는 HCOG와 HHV간의 membership matrix를 만들고, 이 데이터로부터 Phylip에서 제공하는 Camin-Sokal의 maximum parsimony 알고리즘을 적용하여 유전자 컨텐츠 트리 및 reciprocal 트리를 작성하였다. 그림3은 reciprocal tree를 구한 결과를 보인 것이다.

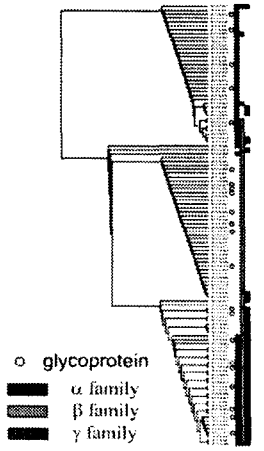

그림3. HHV의 reciprocal 트리

*각 노드는 HCOG를 의미한다

HCOG로부터 서브패밀리가 모여진 형태를 보기위해서, reciprocal tree를 제작하여 본 결과 α, β 및 γ의 모든 서브패밀리에서 분포하는 HCOG 32종, α, β 특이적인 HCOG 9종, β, γ 특이적인 HCOG 11종, γ, α 특이적인 HCOG 3종, α 특이적인 HCOG 23종, β 특이적인 HCOG 36종으로 나뉘어 졌다. 그림3의 reciprocal tree를 볼 때 α, β, γ 서브패밀리가 모두 모여있는 것들은 매우 보존적인 유전자들인 반면에 γ에만 속하는 유전자들이 따로 존재하지 않는 것은 COG알고리즘이 최소 COG로 3종이 기본이나 γ 서브패밀리의 경우 2 종만이 존재하므로 γ특이적인 HCOG가 나타나지 않는 것으로 판단된다. 본 연구자는 세 가지의 서브패밀리가 모두 존재하는 HCOG 그룹이 보존성이 높은 유전자로 구성된 ortholog 그룹으로 판단하였고, α 서브패밀리에만 존재하는 HCOG 23종과 β 서브패밀리에만 존재하는 HCOG 36종이 있음을 확인하였다.

유전자복제의 예측과 ortholog 재정의

계통유전체학적 기법인 SDI 알고리즘에 당단백질의 유전자계통수와 DNA polymerase로부터 얻은 종트리를 입력하고 유전자 복제점을 검색하였다. 당단백질의 경우는 47개의 복제점을 발견하였고, ATV 뷰어[6]를 통해 확인하였다. 그림 4는 예측된 복제점을 분석하기 위하여 복제점으로 예측된 부분을 확대한 그림이다. 그림 4에서 계통수에 표기되어있는 ⓐ점은 예측된 복제점을 의미하며 각 유전자의 기능을 확인하기 위한 annotation분석 결과를 함께 도시하였다. M은 membrane으로 예측된 결과이며, P와 I는 각각 BLOCK, Pfam에서의 개별 히트 결과를 보인 것이다. 본 예제에서는 HCOG023와 HCOG082는 ⓐ점을 기준으로 두 그룹은 마지막 공통 조상으로부터 복제되어 나누어진 것으로 사료되었다.

HCOG023은 HHV의 서브패밀리인 α, β, γ 모두에서 존재하는 보존적인 HCOG이며, HCOG082는 β 서브패밀리 특이적인 HCOG이다. 그림 4에서 복제점 ⓐ에서 α, β, γ 공통 HCOG(HCOG023)와 β 서브 패밀리 특이적인 HCOG(HCOG082)가 나뉘어 졌음을 알 수 있다. 각 데이터베이스 검색결과를 살펴 보면, protein classification에서 HCOG023은 단백질의 분류가 HCOG23:np040152와 HCOG023:np043795는 membrane 구조를 포함하고 있는 것으로 나타나며, HCOG082에서는 모두 membrane 구조를 포함하였다. 또한 Pfam과 BLOCK 검색 결과, HCOG023은 ssDNA bining protein과 가장 높은 유사성을 보였고, HCOG082는 Transmembrane 4 family와 높은 유사성이 있음을 확인하였다. ⓐ점을 기준으로 HCOG023과 HCOG082로

두 개의 그룹으로 나눌 수 있었고, 기존의 HCOG구성에서 제외되었던 HCOG999:np572100과 HCOG999:np039876이 ⓐ점 아래 서브트리에 존재하므로 HCOG082와 동일한 기능을 하는 ORF로 판단할 수 있다. 이로부터 SDI에 의해서 예측된 복제점이 서로 다른 기능의 HCOG들을 분류할 수 있으므로 계통유전체학의 적용이 유효함을 알 수 있다.

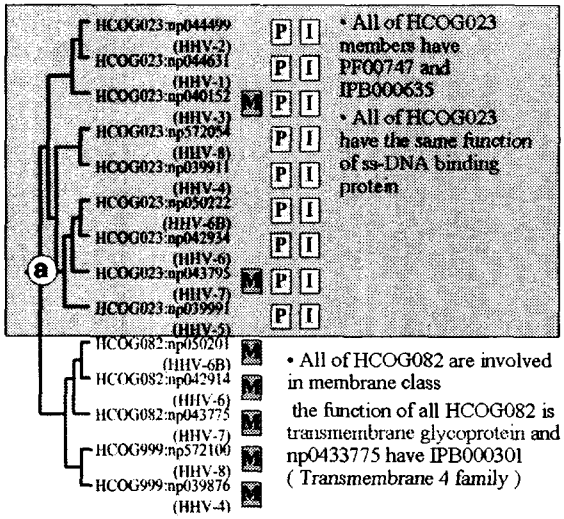

그림4. 복제점을 통해 예측된 ortholog 그룹

이 과정을 당단백질을 가지는 HCOGs에 대해서 반복하여 표1과 같이 HCOGs를 재구성하였다. 이를 통해 HCOGs에서 ortholog로 분류하지 못한 ORF들을 다시 재정의하여 향상된 ortholog 그룹을 만들 수 있었다.

표1. 재 정의된 HCOG

HCOG ID	added ORFs		Reassigned ID
HCOG042	np042993		HCOG042.1
HCOG061	np039952		HCOG061.1
HCOG080	np040053	np040056	HCOG080.1
HCOG082	np572100	np039876	HCOG082.1
HCOG087	np039971		HCOG087.1

4. 결론 및 향후계획

본 연구에서는 유전자 기능 예측방법으로 사용되고 있는 유사성검색 및 비교유전체학이 가진 한계를 극복하기 위하여 계통유전체학을 도입하여 정확한 ortholog와 paralog를 구분함으로서 각 방법들이 가진 한계를 극복하여 보다 향상된 유전자 기능예측 결과를 얻고자 시도되었다.

연구대상으로 특히 다른 생명체에 비해 유전자변이가 심한 바이러스 중 HHV의 진화적 유연관계를 기반으로 유전자의 기능을 예측하고자 하였다. 이를 위해 유사성에 의한 기능 예측, 비교유전체학에 의한 기능 예측의 방법으로 COG 방법을 도입하여 HHV 9종의 942 ORF로부터 114개의 HCOGs를 구축하였고, 이로부터 reciprocal tree를 제작함으로서 각 HHV의 서브패밀리가 가진 특이 단백질을 얻을 수 있다. SDI 알고리즘을 적용하여 HCOGs 구축시 ortholog 그룹에서 제외되었던 당단백질 ORF들을 재 배치하여 향상된 ortholog 그룹을 만들었다.

이로부터 다른 생물체에 비해 그 다양성이 많은 것으로 알려진 바이러스의 유전자 기능예측에 있어서 비교유전체학의 방법이 가진 한계를 계통유전체학의 방법을 적용함으로서 향상된 유전자 기능 예측을 하였다. 비교유전체학의 방법과 계통유전체학적 방법을 조화시켜 유전자 기능 예측을 보완할 수 있었다.

향후 본 연구에서 제한하였던 HHV의 당단백질 이외의 HHV ORF에 대하여 검증할 것이며, 본 연구방법을 사용하여 다른 인체 감염성 바이러스에서도 재해석이 가능할 것이다.

5. 참고문헌

[1]Tatusov RL, Koonin EV, Lipman DJ. "A genomic perspective on protein families", Science, Vol. 278, pp.631-637, 1997.

[2]Eisen JA. "Phylogenomics:Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis", Genome Res, Vol. 8, pp.163-167, 1998.

[3]Bouzat JL, McNeil LK, Robertson HM, Solter LF, Nixon JE, Beever JE,. Gaskins HR, Olsen G, et al. "Phylogenomic Anlaysis of the α Proteasome Gene Family from Early-Diverging Eukaryotes", J.Mol. Evol, Vol. 51, pp.532-543, 2000.

[4] Thompson, J.D., Higgins, D.G., Gibson, T.J., "CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.", Nucleic Acids Research., Vol. 22, pp.4673, 1994.

[5]Zmasek CM, Eddy SR. "A simple algorithm to infer gene duplication and speciation events on a gene tree. Bioinformatics", Vol. 17, pp.821-828, 2001.

[6]Zmasek CM, Eddy SR. "ATV:display and manipulation of annotated phylogenetic trees", Bioinofrmatics, Vol.17, pp.383-384, 2001.