연구논문

생물정보학적 접근을 통한 *Caenorhabditis elegans* 모델시스템의 생체내 RNAi 기능예측 및 비특이적 공동발현억제 현상 분석

김태호¹, 김의용², 주현^{2*}

Bioinformatics Approach to Direct Target Prediction for RNAi Function and Non-specific Cosuppression in *Caenorhabditis elegans*

Taeho Kim¹, Euiyong Kim², and Hyun Joo^{2*}

접수: 2010년 12월 22일 / 게재승인: 2011년 3월 21일
© 2011 The Korean Society for Biotechnology and Bioengineering

Abstract: Some computational approaches are needed for clarifying RNAi sequences, because it takes much time and endeavor that almost of RNAi sequences are verified by experimental data. Incorrectness of RNAi mechanism and other unaware factors in organism system are frequently faced with questions regarding potential use of RNAi as therapeutic applications. Our massive parallelized pair alignment scoring between dsRNA in Genebank and expressed sequence tags (ESTs) in *Caenorhabditis elegans* Genome Sequencing Projects revealed that this provides a useful tool for the prediction of RNAi induced cosuppression details for practical use. This pair alignment scoring method using high performance computing exhibited some possibility that numerous unwanted gene silencing and cosuppression exist even at high matching scores each other. The classifying the relative higher matching score of them based on GO (Gene Ontology) system could present mapping dsRNA of *C. elegans* and functional roles in an applied system. Our prediction also exhibited that more than 78% of the predicted

co-suppressible genes are located in the ribosomal spot of *C. elegans*.

Keywords: RNA interference (RNAi), Expressed Sequence Tag (EST), *in silico* Biology, Multiple Sequence Alignment, Cosuppression, dsRNA, *Caenorhabditis elegans*

1. 서론

생체 내 단백질 발현은 다양한 방법으로 억제 가능하다. 발현억제 양상은 전사과정중 발현억제 (transcriptional gene silencing: TGS)와 전사후 발현억제 (post-transcriptional gene silencing: PTGS)로 크게 나뉘어진다 [1]. TGS는 DNA로부터 mRNA가 전사되지 않음으로써 silencing을 일으키는 현상으로 유전자의 메틸화 반응이나 크로마틴의 구조변화 등이 관여한다. PTGS경우 전사과정에는 특이적 현상이 일어나지 않지만, mRNA가 특정원인에 의해 분해되거나 또는 번역이 저해됨으로써 유전자 발현이 억제되는 현상이다. RNA 간섭효과 (RNA interference, RNAi)는 PTGS현상으로 구분되며 dsRNA에 의해 서열 특이적으로 mRNA가 분해되어 단백질 발현이 억제하는 현상이다 [2,3]. 1998년 Fire등의 실험에서 이쁜꼬마선충 (*Caenorhabditis elegans*)에 이중나선 RNA (double-stranded RNA 혹은 dsRNA)들을 주입하여 서열특이적인 유전자 발현억제가 유발되는 유전자 침묵 (silencing)효과가 처음 발견된 이후 다양한 방법을 통하여 수많은 RNAi관련 실험이 진행되었다 [4-7]. 본 연구에서 모델 시스템으로 선충 *C. elegans* RNAi를 선택한 이유도 바로 이

¹일본 오사카대학교 WPI 면역선도연구센터

¹Laboratory of Systems Immunology, World Premier International Immunology Frontier Research Center, Osaka University, Suita, Osaka 565-0871, Japan.

²인제의대 생리학교실

²Department of Physiology and Integrated Biosystems, College of Medicine, Inje University, Busan 614-735, Republic of Korea.
Tel: +82-51-890-6452, Fax: +82-51-894-5714
e-mail: phyjoo@inje.ac.kr

와 같은 선행연구와 후속연구들에 의하여 RNAi현상에 대한 풍부한 실험데이터가 축적되어 있기 때문이다.

RNAi는 몇 가지 특이적인 분자생물학적 특징을 갖는다. RNAi는 서열에 매우 의존적이며 그 길이에 매우 민감하게 작용한다 [8,9]. 또한 RNAi현상은 mRNA의 프로모터나 인트론 서열에 silencing을 일으키지 않는 점도 중요한 특징이다. 21-23개 염기의 RNA가 타겟 서열 인식을 위한 최소한의 RNA 서열크기로 기능하고 있는 것이 밝혀졌고, 화학적으로 합성된 21-23서열의 dsRNA또한 타겟 mRNA를 절단 가능한 것이 증명되었다 [10]. 이 최소한의 염기서열을 지니는 dsRNA는 RNAi현상에 매우 중요한 요소 중 하나인 siRNA (short interfering RNA)라 명명되었다. siRNA는 서열특이성이 매우 높게 나타났고, 이 현상은 RNAi에서 가장 중요한 특징 중의 하나이다. 그러나 문제는 RNAi현상이 서열특이적으로 알려져 있음에도 불구하고, 목표 유전자와 서열이 매우 유사할 경우 내재 유전자들과의 원치 않는 공동발현억제 (cosuppression)를 보여주는 현상을 꼽을 수 있다 [11,12]. 따라서 비특이적 대상체와의 결합이 유발되기에 어느 서열정보를 이용하는 것이 보다 효율적인 발현억제가 될 수 있는지는 여전히 미지수다. 또한 서열이 목표유전자와 100% 일치되게 설계된 siRNA라 할지라도 대부분이 실험을 통하여 확인해 보지 않는 한 그 효과를 장담할 수가 없다.

이에 대한 선 예측을 위하여 BLAST 검색과 같은 대상 생물체의 유전체 (genome) 정보를 이용하는 방법이 생물정보학적으로 일부 이용되고 있는 실정이지만 [13-15], 이의 직접적인 이용은 서열 분석기법의 한계로 인하여 극히 제한적이다. 가장 큰 문제점으로는, 일반적으로 사용되는 유전체 서열정보에는 비 정제된 발현부위와 인트론, 그리고 UTR (untranslated region) 영역들이 저장되어 있기에 BLAST 탐색 등을 이용하는 *in silico* 서열 유사도 분석 과정에서 많은 시행착오를 불러일으킨다. 진핵생물의 경우 빈번한 유전자 스플라이싱이 유발되어 유전체 수준에서의 탐색 효율성이 급격히 저하된다. 또한 전체 유전자들을 일일이 수작업을 통하여 발현 부위만 이용 하더라도, BLAST검색은 갭 (gap)을 허용하지 않아 탐색시 간과되는 서열들이 상당수 존재한다 [16]. 최근 갭을 허용하는 BLAST 2.0이 사용되는 경우도 원천적인 위드탐색 문제로 인하여 명확한 결과를 얻기가 어려운 실정이다. 이처럼 RNAi 설계작업중에 전체 유전자 규모의 서열 유사도 탐색을 우선시 하는 가장 주된 이유중의 하나는, 목표 유전자 외에 타 유전자와의 서열 유사성 문제가 발생될 경우 원치않는 공동발현억제가 발생되기 때문이다. 이러한 문제점을 해결하기 위하여 RNAi 설계시 목표 유전자 부위를 여러 부분으로 나누어 제작하고 있는 실정이다. 또한 동일한 유전정보를 지닌 siRNA에도 불구하고 대상체에 따라 달리 적용 될 수 있다.

RNAi의 작동 메커니즘은 Fig. 1에 나타내었다. 헤어핀 (hairpin)을 포함하는 긴 이중가닥인 dsRNA가 두 분자체로 구성된 dicer라고 하는 절단효소에 의해 절단되어 21-23개의 서열로 구성되는 siRNA를 형성하는 것으로부터 시작한다. 이는 긴 서열을 제외하고는 하등생물이나 포유동물 모두에 공통적으로 작용하는 것으로 알려져 있으며, mRNA분해와 발현억제 (translation suppression)는 공통적으로 동일 조절경

로상에 존재하는 것으로 알려져 있다. 이 절단효소는 RNaseIII family에 속하며 다섯 개의 도메인으로 구성되어 있다. Dicer의 기능은 현재까지 확실하게 알려진 것이 없지만, dsRNA를 절단하는데 매우 중요한 효소로 인식되고 있다. 이후 siRNA는 다중결합체인 RISC (RNA-induced silencing complex)와 결합하여 RNAi silencing복합체를 형성하고 RISC의 ATP에 의해 활성을 띠게 된다. RNAi silencing 복합체는 특정 mRNA의 상보성 (antisense)결합체로서 대상 유전자의 상보적 단백질 코딩영역에 작동하게 되고 결국 타겟 mRNA 기능을 정지시킴으로써 유전자 발현을 억제시킨다. 따라서, 이의 작동은 dsRNA와 mRNA간의 상보적 염기서열들의 유사성 정보에 의하여 결정된다. 서열 유사도 점수가 100% 동일한 경우는 mRNA가 분해되며 90% 이상 상보적인 경우 mRNA 분해는 없으나 발현을 억제한다고 알려져 있다 [17].

Fig. 1. An overview of RNA interference (RNAi) mechanism. Double-stranded RNA directs the cleavage of mRNA at 21 to 23 nt by a dicer and which also affects unwanted cosuppression of other proteins. This study suggests all-to-all pairwise alignments of desired RNA and EST sequences in order to predict *ab initio* gene cosuppression.

본 연구의 목적은 앞서 언급된 유전체 수준에서의 탐색이 지닌 한계점을 극복하기 위한 것이다. 수많은 실험들을 통하여 이미 알려진 *C. elegans* RNAi 서열들과 mRNA의 부분 단편들로 구성된 전사체 (EST, expression sequence tag)서열 사이에 다중서열정렬 (MSA) 연산도구를 이용하여 상보적 서열유사도 값을 계산하는 것이다. 이를 통하여 목표 유전자의 발현억제 서열정보가 다른 원치않는 mRNA들의 발현을 억제하는지에 대하여 조사하고, 전체 세포수준까지 확장된 공동발현억제 현상을 확인하는 것이다. 궁극적으로는, RNA 간섭효과가 원하는 목표유전자에만 선별적으로 가능하게끔 인위적 설계가 가능케 하고, 공동발현억제가 가장 빈번히 발생하는 유전자 서열들을 선별하여 유전자온톨로지 (GO, Gene Ontology) 기준에 맞추어 이들 분자적 기능에 따라 전체적인 작동 양상을 맵핑하는 것이다. 이를 통하여 기존의 유전체 수준에서의 BLAST검색에서 제시하지 못한 보다 직접적

인 RNAi현상을 구현 가능할 뿐만 아니라, 이들 빈번한 공동 발현억제 유전자군의 상세한 위치 판독, 세포 수준에서의 치명성, 분자적 수준까지 확장하여 목표 유전자의 기능 예측이 가능케 하는 것이다.

2. 재료 및 방법

2.1. 다중서열정렬 연산도구를 이용한 RNAi 표적서열 검색
우선 세포내 *C. elegans*의 전사체인 EST와 dsRNA사이의 다중서열정렬을 수행하기 위해 검색하고자 하는 문자열 혹은 서열을 두 개씩 짝지어 쌍정렬을 수행하고 이들 사이에 가장 연관도가 높은 공통서열들을 뽑아내었다. 이를 위하여 동적 프로그래밍 기법중의 하나인 Smith-Waterman 알고리즘을 이용하였다 [18]. 이 알고리즘은 서열유사도 (sequence similarity) 점수화에 가장 높은 민감도를 보여준다. 두 비교서열들의 유사도 측정은 정확하게 수치화가 가능하지만 과도한 정확도로 인하여 연산지연이 발생하는 문제를 해결하기 위하여 task queue와 병렬 작업 로드밸런싱용 인터럽트가 가능도록 코딩된 ClustalXeed 버전을 사용하였다 [19]. 이는, 입력화일에 처리할 업무가 있다면 이를 FIFO 방식으로 나누어 작업량의 손실없이 실시간 처리가능한 인텔리전트 로드 밸런싱 알고리즘으로 기존 병렬연산에서의 단점을 보완한 프로그램이다.

2.2. 고속 대용량 연산 클러스터 구축

총 50개의 듀얼 Opteron 코어 (64 bit)와 기가비트 (Gigabit)로 연결된 클러스터 시스템, 병렬연산이 가능한 ClustalXeed 프로그램 코드를 일부 변경하여 구축하였다. 본 실험에 사용된 병렬연산 클러스터 장비는 서버단계와 클라이언트단계로 나누어 구현되었다. 여기에서 서버는 각 노드의 관리와 운영을 담당하는 마스터노드를 말하고, 클라이언트는 마스터노드의 관리에 따라 계산작업을 수행하게 되는 계산노드를 말한다. 마스터노드에는 리눅스 운영체제와 각 계산노드에 분산될 dhcpcd, tftpd, nfsd 등의 각 서비스 daemon을 설치하였다 [20-24]. 각 연산노드에서 사용하게 될 네트워크 파일 시스템, 클러스터의 관리와 로드밸런싱 등의 성능향상을 위한 openmosix [25]가 패치되어 있는 리눅스 커널과 동기중간 병렬처리를 지원하는 MPI (message passing interface)관련 응용프래임인 mpich [26]가 설치되었다. 각 클라이언트에서 사용될 운영체제 및 파일시스템은 실제 서버의 /tftpboot 디렉토리에 설치되었다.

2.3. dsRNA 서열과 발현단편 라이브러리

생체내 (in vivo) RNA 간섭현상을 예측하기 위해 RNA 간섭현상을 일으키는 근본이 되는 dsRNA의 서열과 ESTs간의 쌍정렬을 수행하였다. 이를 위하여 모델 시스템으로 현재까지 RNAi 관련 프로젝트 가운데 가장 먼저 데이터베이스화된 선충 *C. elegans* 내재 서열을 대상으로 발현단편을 통한 RNAi 분석 가능성을 예측해 보았다. *C.elegans*의 dsRNA 서열은 GeneBank (<http://www.ncbi.nlm.nih.gov>)에서 검색을

통하여 모두 475개의 dsRNA서열을 다운로드 받았다. 이 서열 중 가장 긴 서열의 길이가 1193931 bp이고, 가장 짧은 서열의 길이가 71 bp이다. 모든 서열의 평균길이는 4227.76 bp이다.

*C. elegans*의 전사체인 발현단편 (EST) 서열들은 *Caenorhabditis* Genome Sequencing Projects 데이터 베이스를 통하여 다운로드 받았다. Serolis (ver. 1.2.0.7)를 이용한 dot-plot기법을 통하여 EST 데이터베이스 내의 중복서열들을 제거하였다 [27]. 중복성이 제거된 발현 단편은 190247개의 서열이며, 가장 긴 서열의 길이가 985 bp, 가장 짧은 서열의 길이 12 bp, 평균서열의 길이가 438 bp이다. 총 쌍정렬 서열연산에 소요되는 시간 복잡도 $O(n)$ 은 다음과 같이 표현될 수 있다 [28].

여기서 n 은 총 서열의 수, n_i 는 i 번째 서열의 길이를 의미한다. 순환관계 합은 $2n + 1$ 로 귀결이 되며, 시간의 복잡성만 고려할 때 2.5 GHz 이상의 클럭 주파수를 가지는 i86x급 싱글 PC 베이스로 최소한 364일 이상의 연산시간이 소요됨을 알 수 있다. 본연구에 사용된 ClustalXeed 분산 알고리즘과 $50 \times \text{CPU}$ 클러스터 시스템이 적용될 경우, 병렬화에 따른 스칼라화 (scalability) 지수가 1.4 정도로 측정되며 시간 복잡도에 의한 연산시간은 최종 14일 정도로 줄어들 수 있음을 알 수 있으며, 실제 연산시간도 비슷하게 소요되었다.

2.4. 연산 매개변수 추정 및 쌍정렬 점수 획득

대상으로는 dsRNA와 EST간의 서열일치도 결과값이 이용되었으며 전역정렬 (global alignment)에서 필요로 하는 SP (sum of pairs) 점수는 제외시켰다. 이유는 다중정렬을 통한 서열간 클래스 형성과는 달리, 단지 dsRNA와 ESTs간의 서열정렬 점수만을 필요로 하는 본 실험에서는 불필요한 과정이기 때문이다. 유전자 삽입 (insert), 삭제 (delete)에 의하여 발생하는 서열 매트릭스의 개별 점수화를 위하여 본 연구에서는 Smith-Waterman 알고리즘에 사용된 갭 벌점 (gap penalty) 값을 (-14, -4)로 적용하여 쌍정렬값 연산을 수행하였다. 첫번째 수치는 처음 발생하는 갭에 대하여, 두 번째는 연속적인 갭 발생에 대한 감점치를 부여한 것이다. 이는 EST와 같이 매우 짧은 단편화된 서열들에 대하여 합리적인 수준 하에서 결정된 값이다. 다중서열정렬 점수분석을 위하여 생성된 RNAi와 EST의 다대다 (m:n) 매트릭스는 총 1억3천7백만 쌍으로 전체 매트릭스의 서열정렬에 소요된 CPU시간은 약 1.92×10^4 sec 정도가 되었다.

3. 결과 및 고찰

RNAi와 전사체 발현단편군 사이의 표적서열 일치도 점수 결과를 살펴보면 전체 쌍정렬 매트릭스의 평균 서열유사도 스코어는 6.12%로 나타났고, 이중 85% 이상의 일치서열을 보여주는 매트릭스는 전체서열 조합속에서 불과 0.0046%인 것으로 나타났다. 이는 RNA간섭현상뿐만 아니라 비특이적 결합에 따른 발현 억제까지도 고려할 때 완전 서열일치도를 보여주는 21개의 연속된 염기서열이 다른 mRNA서열들에 대해서도 유의적인 수준으로 비특이적 단백질발현 방해 (혹은 억제)가 가능한 범주안에 속하는 값이다. 이보다 더

높은 90%에서 100% 범위에서는 dsRNA가 목표 mRNA에 대하여 발현을 억제할 수 있는 수준을 넘어 RNAi현상까지 가능하다. 서열유사도점수가 85% 이상을 보이는 EST서열들은 총 6348개이며, 90% 이상은 6308개이다. 이 결과를 종합하면 크게 두 가지의 상반된 의미로 해석될 수 있다. 우선 *in vivo*에서 RNAi현상이 발생할 수 있는 확률에 대하여 매우 긍정적으로 해석이 가능하다는 것이다. 일반적으로 dicer에 의해 21-23개의 염기서열이 siRNA로 생성되어 RNAi현상을 일으킨다는 관점에서 보면 dsRNA와 ESTs사이의 전체 평균 서열유사도 점수가 6.12%의 상동성을 가진다는 것은 이 실험에 사용된 전체 dsRNA의 평균서열길이가 약 4228 bp 이므로 이중 최소한 259 bp가 확률적으로 ESTs와 상동성을 보인다는 점이다. 따라서 서열일치도의 연속성을 고려하지 않는다는 전제하에서 본 결과는 생체내에서 RNAi현상이 일어날 수 있는 확률이 다소 존재한다고 예측할 수 있다. 그러나 dsRNA의 어느 부분이 dicer에 의해 siRNA로 만들어지는지 전혀 예측할 수 없는 상황에서 앞의 예측은 맞지 않다. 실험에 사용된 475개의 dsRNA와 190247개의 ESTs 사이에 1:1 쌍대칭정렬 서열유사도 점수가 평균 6.12%로 계산된 것과 대비되어 100%의 상동성을 보이는 서열들은 전체 EST중 단지 0.0015% (Fig. 2)만 존재한다. 이는 mRNA와 siRNA-RISC복합체간의 결합을 조절하는 작동기구만을 고려한다면 사용된 dsRNA가내재 유전자 서열들과의 중복도가 그리 높지 않음을 보여주는 것 같지만, 하나의 dsRNA라도 최소한 서너개의 내재된 EST들과 일치하고 있기에 목표 유전자 서열과 100% 일치된 유전자가 주입되더라도 매우 안정적인 RNAi현상을 얻을 수 있으리라는 보장이 될 수 없음을 시사한다.

Fig. 2. The calculated matching scores of all-to-all pairwise sequence comparisons between 475 dsRNA and 190247 *C.elegans* ESTs. The generated numbers of pair matrices are 136694858 and the averaged total pairalign similarity score is 6.12 %.

다음에 수행된 실험으로는 dsRNA서열들과 ESTs 사이에 서열유사도가 90% 이상인 6308개의 쌍정렬 매트릭스만을 추출 및 분리하여, 공동발현억제 유발 후보군에 대한 정보 획득 및 이들의 생체내 주요 분자적 기능들에 대한 분석이다. 비 특이적 공동발현억제가 유발되는 분자적 기능을 가진

dsRNA들에 대하여 유전자온톨로지 (GO) 분류기준에 따라 기준하여 클러스터 매핑을 시도해 보았다. GOC (The Gene Ontology Consortium)는 functional biology 판독에 필수적인 요소인 controlled vocabulary를 체계적으로 정리하였으며 유전자 혹은 단백질의 기능적 분류기준을 판독할 수 있는 근간의 제시한다. GO에는 분자적 기능 (molecular function), 세포구성체 (cellular component), 생물학적 과정 (biological processing)과 같은 세 가지의 최상위 분류기준이 존재하며, 본 연구에서는 우선 분자적 기능에 맞추어 매핑을 수행하고자 하였다. 다른 카테고리 중첩이 되는 dsRNA는 GO에서 제시하는 'molecular function'의 기준에 따라 재 분류를 하였다. GO mapping은 GO데이터베이스에서 BLAST 검색을 통해 상동성별로 GO를 예측할 수 있도록 만들어진 인터넷 사이트인 GoFigure의 GO 분류 도구를 사용하여 수행하였다 [29,30].

비특이적 공동억제유발이 예상되는 6308개의 *C. elegans* EST서열들에 대한 GO 매핑 결과는 Table 1에 나타나 있다. 구조적분자활성 (structural molecule activity, 36%), 결합관련 (binding, 25%), 촉매활성 (catalytic activity, 17%), 운송체활성 (transporter activity, 5%), 사용하지 않거나 퇴화된 분자기능 (obsolete molecular function, 3%) 순으로 분류되었으며 유전자온톨로지 분자기능의 어느 분류범주에도 포함되지 않는 서열들이 11%로 분류되었다. 본 결과를 통하여 매우 중요한 사실을 확인 가능하였는데 구조적분자활성, 결합, 촉매활성 기능과 관련된 서열이 전체 비특이적 공동발현억제의 78%를 차지하고 있으며, *in vivo*에서 RNAi에 의한 비특이적 공동발현억제가 이 세 기능에 집중되어 있는 것을 알 수 있다. 특히 구조적분자활성 관련 유전자중 리보솜 (ribosome) 구성체에 관한 발현단편 유전자들이 전체 6803개의 매칭 유전자중 무려 2050개나 차지하였다. 즉, 가장 정교하게 특정서열만을 지니는 dsRNA 설계를 통하더라도 많은 수의 서열들이 바로 이 리보솜 구성체 유전자와 원치않는 서열 특이적 공동발현억제를 유발할 가능성이 매우 높다는 점이다. 이는 과거 RNAi 실험결과들에서 보여주듯이 대다수의 원치않는 공동발현억제가 치명적인 세포사멸 (death)을 유도하는 이유에 대한 중요한 해답이 될 수 있다. 리보솜 구성체 유전자중에서도 알집 (ovary)에 연관된 유전자는 전체 70% 이상을 차지하였다. 이러한 결과들은 RNAi현상이 생물의 발생 (development)에도 매우 밀접한 관계를 갖고 있다는 것을 보여줄 뿐만 아니라 [31], Tijsterman과 그 동료들에 의하여 이쁜꼬마선충을 통하여 연구되었던 배아치사율 (lethality of *C. elegans* embryo) 결과에도 이 현상이 매우 밀접한 연관이 있음을 시사한다 [32,33]. 그들은 결핍시 초기배아상태에 치명상을 주는 pos-1 유전자를 long dsRNA (800 bp), siRNA (25 bp), asRNA (25 nt), 혹은 sRNA (25 nt) 순으로 절단하여 이쁜꼬마선충에 주입한 결과, 800 bp 길이의 dsRNA에서는 100%의 작동성을, dicer에 의하여 작은 조각들로 편집된 siRNA들에서는 10-15%의 작동성을 확인하였다. 안티센스 RNA (asRNA)는 그 길이에 따라 절반 이상의 초기배아들에서 50% 정도의 작동성을 보여주었다. 이는 RNA간섭현상이 그 주입 유전자의 길이에 따라 상이하게 작용될 수 있으며,

Table1. Cluster mappings of used 475 dsRNAs in *C.elegans* ESTs based on gene ontology system

Gene Ontology Categories	Representation	% Representation of total
Biological process		
Physiological process		
Metabolism	17	<1%
biosynthesis	51	1%
electron transport	34	1%
protein metabolism	40	1%
catabolism	23	<1%
nucleobase, nucleoside, nucleotide and nucleic acid metabolism	23	<1%
Cell growth and/or maintenance	8	<1%
cell proliferation	59	1%
cell growth	1	<1%
Cellular process		
Cell communication		
cell-cell signaling	12	<1%
signal transduction	9	<1%
Obsolete biological process	7	<1%
Development	3	<1%
Cellular component		
Cell		
Membrane	79	1%
integral to membrane	25	<1%
Intracellular		
Cytoplasm	35	1%
Nucleus	12	<1%
Molecular function		
Binding	58	1%
nucleotide binding		
purine nucleotide binding	373	6%
nucleic acid binding	298	5%
DNA binding	316	5%
RNA binding	125	2%
nuclease activity	29	<1%
translation factor activity, nucleic acid binding	6	<1%
metal ion binding		
calcium ion binding	158	3%
transition metal ion binding	53	1%
Protein binding	47	1%
transcription factor binding	81	1%
cytoskeletal protein binding	15	<1%
lamin binding	9	<1%
cytoskeletal protein binding	1	<1%
Catalytic activity		
hydrolase activity	226	4%
peptidase activity	257	4%
hydrolase activity, acting on acid anhydrides	98	2%
hydrolase activity, acting on ester bonds	90	1%
hydrolase activity, acting on glycosyl bonds	25	<1%
ligase activity	56	1%
ligase activity, forming carbon-nitrogen bonds	4	<1%
transferase activity	48	1%
transferase activity, transferring phosphorus-containing groups	44	1%
transferase activity, transferring acyl groups	12	<1%
transferase activity, transferring glycosyl groups	4	<1%
oxidoreductase activity	41	1%
Oxidoreductase activity, acting on diphenols and related substances as donors	23	<1%
oxidoreductase activity, action on NADH or HADPH	22	<1%
oxidoreductase activity, action on CH-OH group of donors	19	<1%
disulfide oxidoreductase activity	9	<1%

lygase activity	3	<1%
lygase activity, forming carbon-nitrogen bonds	2	<1%
kinase activity		
protein kinase activity	47	1%
lipid kinase activity	25	<1%
helicase activity		
ATP dependent helicase activity	2	<1%
Chaperon activity	18	<1%
co-chaperone activity		
co-chaperonin activity	6	<1%
Enzyme regulator activity		
enzyme activator activity		
Phospholipase activator activity	15	<1%
enzyme inhibitor activity		
protease inhibitor activity	7	<1%
GTPase regulator activity		
small GTPase regulatory/interacting protein activity	30	<1%
Molecular functional unknown	1	<1%
Obsolete molecular function		
caspase activity	15	<1%
DNA-directed RNA polymerase III activity	1	<1%
proteasome endopeptidase activity	187	3%
Signal transducer activity	33	1%
receptor activity		
transmembrane receptor activity	12	<1%
ligand-dependent nuclear receptor activity	3	<1%
Structural molecule activity	39	1%
structural constituent of ribosome	2050*	32%
structural constituent of cytoskeleton	209	3%
extracellular matrix structural constituent	26	<1%
structural constituent of cuticle	7	<1%
Transcription regulator activity	9	<1%
transcriptional repressor activity	28	<1%
RNA polymerase II transcription factor activity	6	<1%
Transporter activity	12	<1%
protein transporter activity	117	2%
electron transporter activity	65	1%
channel/pore class transporter activity		
alpha-type channel activity	51	1%
porin activity	49	1%
carrier activity		
electrochemical potential-driven transporter activity	20	<1%
carbohydrate transporter activity		
sugar porter activity	2	<1%
Others	Representation	% Representation of total
uncategorized	226	4%
Total	6308	100%

유전자 조절 연구방법론에 매우 중요하게 이용될 수 있음을 보여준 결과이다. 지금까지 연구된 합성 RNAi들은 하등생물 체나 식물, 특히 포유동물류의 특이 유전자 발현 억제에 매우 쉽고도 높은 효과를 발휘할 수 있는 것으로 알려져 있지만, 가장 중요한 문제점으로 부각되었던 점은 인터페론반응이나 치명적인 생체활성에 영향을 줄 수 있는 타 유전자들과의 원치 않는 공동발현억제이다. 이쁜꼬마선충이나 초파리와 같이 하등생물체에서는 길이가 긴 dsRNA가 사용 가능하기에 높은 서열특이성을 인위적으로 부여 가능한 것에 반하여 포유동물의 경우 인터페론 면역계라 일컫어지는 비특이적 반응경로

활성화에 기인되어 30 nt 이상의 염기에는 효과가 없는 문제점이 있었다. 따라서 본 연구 결과를 활용하면 이러한 문제점 해결을 위하여 30 nt 이내로 제한된 dsRNA를 보다 정교하게 설계하고 생체내 공동발현억제와 같은 부작용을 선예측할 수 있다는 점에서 매우 유용하게 사용될 수 있다.

앞서 결과들을 종합하면, RNAi 실험을 통하여 생체내 목표 유전자만의 특이적 발현억제를 성공적으로 이끌기 위해서는 신뢰할 수 있는 siRNA 설계기술이 필수적이다. 본 연구는 생체내 RNAi 주입시 비특이적 공동발현억제를 막을 수 있는 새로운 방법을 제시한다. 본 연구에서는 방대한 서열정보들을 신속

히 처리가능한 병렬화 알고리즘 프로그램인 ClustalXeed를 이용하여 모델 시스템으로 선충 *Caenorhabditis elegans* ESTs와 RNAi현상의 기본이 되는 dsRNA간의 특이적 혹은 비특이적 결합 가능성을 예측하였다. 서열정렬결과 전체 1.37×10^9 개의 쌍정렬 가운데 불과 0.0015% 서열만이 100%의 상동성을 보였다. 비록 매우 적은 수치 같지만, 실제로는 본 분석에 사용된 전체 190247개의 EST 단편서열중에 2천여개 이상이 전체 475개의 RNAi들과 100% 서열일치를 보여 준다. 즉, 평균적으로 1개의 RNAi서열이 3-4개 이상의 타 유전자군과 공동발현억제를 유발할 수도 있다는 점이다. 이와 같은 결과를 통해 우리는 기본적으로 RNAi현상이 mRNA와 siRNA-RISC 복합체간의 서열유사도에 크게 영향을 받는다는 기존의 RNAi 메카니즘외에 또 다른 공동발현억제 유발 기구의 존재 혹은 유사한 연결고리가 존재할 수 있다는 것을 확인하였다. 따라서 본 연구 결과에 따르면 아직도 RNAi현상의 많은 메카니즘들이 미궁명 상태임에도 불구하고 siRNA와 mRNA의 결합을 보다 선택적이고 효율적으로 조절하기 위해서는 EST 발현단편 라이브러리로부터 선행적인 상동성 예측이 중요함을 알 수 있다.

4. 결론

RNA간섭효과 (RNA interference, 혹은 RNAi)는 특정 유전자의 dsRNA를 생체내에 주입하여 그 일치서열 유전자의 전사후 발현억제를 유도하는 기술이지만 성공적인 실험을 위하여 생체내 작동 신뢰도가 높은 siRNA 디자인이 필수적이다. 이와같은 혁신적인 유전자 다운레귤레이션 기술임에도 불구하고, 표적유전자만을 효율적으로 억제하기가 용이하지는 않다. 이는 표적서열외적으로 존재하는 높은 서열유사도를 가진 타 유전자군의 공동발현억제가 빈번히 유발되기 때문이다. 이에 본 연구에서는 전사체 라이브러리를 이용한 공동발현억제 선 예측 기법을 제시하였다. 발현이 억제된 유전자들의 대다수가 dsRNA서열들과 매우 유사도가 높다는 점에 착안하여, 이를 바이오인포매틱스 기법을 통하여 분석하고 공동발현억제뿐만 아니라 원치않는 유전자 침묵현상(gene silencing)을 미리 방지할 수 있는 방안으로써, 이는 기존 연구들에서 사용되어온 유전체 (genome) 정보를 그대로 받아들여 분석하는 방법과는 달리, 직접적인 단백질 발현 산물인 전사체 라이브러리를 사용한 차별화된 연구이다. 본 연구에서 얻어진 결과는 모델시스템으로 사용된 이쁜꼬마선충 (*C. elegans*)뿐만 아니라, 최근 식물, 초파리, 포유동물까지 확대되어 대상 유전자의 생체내 발현억제 기능을 연구하는데 많이 이용될 수 있을 것으로 사료된다. 결론적으로, dsRNA를 시작으로 발생하는 RNAi현상 메카니즘에는 현재 알려져 있는 것과는 또 다른 요소, 즉 특정구조 혹은 특이위치의 결합에 관련된 공통서열 도메인 (common sequence domain)이 타 유전자 그룹에도 널리 분포되어 있음을 예측할 수 있었다. 본 연구에서 제시한 분석기법은 RNA 간섭효과의 광범위한 이용을 제한하는 비특이적 공동발현억제를 미리 예측가능하게, 향후 RNAi 현상을 이용한 유전자치료 및 특이적 발현억제

연구에 크게 이용될 수 있을 것으로 생각된다. 특히, siRNA나 shRNA등의 대규모 선별작업시 손쉽게 이용될 수 있으며 목표로 하는 대상 유전자만의 발현조절에 이용이 가능하다. 유전자 수준에서 질병 발생의 직접적인 원인 해석 및 정확한 PTGS 대상 유전자만의 선별이 가능하다는 장점을 내포하고 있어 유전자를 직접 표적으로 치료할 수 있는 궁극적인 방법론이 될 수 있을 것으로 믿는다.

감사

본 연구는 2010년도 인제대학교 교내학술 연구비로 수행되었으며 이에 감사드립니다.

References

1. Sijen, T., I. Vijn, A. Rebocho, R. van Blokland, D. Roelofs, J. N. M. Mol, and J. M. Kooter (2001) Transcriptional and posttranscriptional gene silencing are mechanistically related. *Curr. Biol.* 11: 436-440.
2. Fire, A., S. Xu, M. Montgomery, S. Kostas, S. Driver, and C. Mello (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* 391: 806-811.
3. Hammond, S. M., A. A. Caudy, and G. J. Hannon (2001) Post-transcriptional Gene Silencing by Double-stranded RNA. *Nature Rev. Gen.* 2: 110-119.
4. Montgomery, M. K., S. Xu, and A. Fire (1998) RNA as a target of double-stranded RNA-mediated genetic interference in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* 95: 15502-15507.
5. Ashrafi, K., F. Y. Chang, J. L. Watts, A. G. Fraser, R. S. Kamath, J. Ahringer, and G. Ruvkun (2003) Genome-wide RNAi analysis of *Caenorhabditis elegans* fat regulatory genes. *Nature* 421: 268-272.
6. Kamath, R. S., M. Martinez-Campos, P. Zipperlen, A. G. Fraser, and J. Ahringer (2000) Effectiveness of specific RNA-mediated interference through ingested double-stranded RNA in *Caenorhabditis elegans*. *Genome Biol.* 2: 1-10.
7. Maeda, I., Y. Kohara, M. Yamamoto, and A. Sugimoto (2001) Large-scale analysis of gene function in *Caenorhabditis elegans* by high-throughput RNAi. *Curr. Biol.* 11: 171-176.
8. Sijen, T., J. Fleenor, F. Simmer, K. L. Thijssen, S. Parrish, L. Timmons, R. H. Plasterk, and A. Fire (2001) On the role of the RNA amplification in dsRNA-triggered gene silencing. *Cell* 107: 465-476.
9. Cerutti, L., N. Mian, and A. Bateman (2000) Domains in gene silencing and cell differentiation proteins: the novel PAZ domain and redefinition of the piwi domain. *Trends Biochem. Sci.* 25: 481-482.
10. Schwarz, D. S., G. Hutvagner, T. Du, Z. Xu, N. Aronin, and P. D. Zamore (2003) Asymmetry in the assembly of the RNAi enzyme complex. *Cell* 115: 199-208.
11. Ketting, R. F. and R. H. Plasterk (2000) A genetic link between co-suppression and RNA interference in *C. elegans*. *Nature* 404: 296-298.
12. Dernburg, A. F., J. Zalevsky, M. P. Colaiácovo, and A. M.

- Villeneuve (2000) Transgene-mediated cosuppression in the *C. elegans* germ line. *Genes Dev.* 14: 1578-1583.
13. Zhong, W. and P. W. Sternberg (2006) Genome-wide prediction of *C. elegans* genetic interactions. *Science* 311: 1481-1484.
 14. Arziman, Z., T. Horn, and M. Boutros (2005) E-RNAi: a web application to design optimized RNAi constructs. *Nucleic Acids Res.* 33: W582-W588.
 15. Kamath, R. S., A. G. Fraser, Y. Dong, G. Poulin, R. Durbin, M. Gotta, A. Kanapin, N. Le Bot, S. Moreno, M. Sohrmann, D. P. Welchman, P. Zipperlen, and J. Ahringer (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* 421: 231-237.
 16. Hutvagner, G. and P. D. Zamore (2002) RNAi: nature abhors a double-strand. *Curr. Opin. Genet. Dev.* 12: 225-232.
 17. Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.
 18. Smith, T. F. and M. S. Waterman (1981) Identification of common molecular subsequences. *J. Mol. Biol.* 147: 195-197.
 19. Kim, T. and H. Joo (2010) ClustalXeed: a GUI-based grid computation version for high performance and terabyte size multiple sequence alignment. *BMC Bioinformatics* 11:467.
 20. HPC (High Performance Computer) Linux Cluster How To. <http://wiki.kldp.org/KoreanDoc/html/HPC-KLDP/index.html>. (2003).
 21. Barak, A. and O. La'adan (1997) The MOSIX multicomputer operating system for high performance cluster computing, *Future Gen. Comp. Sys.* 13: 361-372.
 22. Building a diskless Linux Cluster for high performance computations from a standard Linux distribution. <http://www.uni-due.de/~bt0756/publications/2003-cluster/>. (2003).
 23. Dynamic Host Configuration Protocol. <http://www.ietf.org/rfc/rfc2131.txt>. (1997).
 24. NFS version 3 Protocol Specification. <http://www.ietf.org/rfc/rfc1813.txt>. (1995).
 25. OpenMosix Project. <http://openmosix.sourceforge.net>. (2008).
 26. MPICH2. <http://www.mcs.anl.gov/research/projects/mpich2/>. (2010).
 27. Serolis: Dot-plot software for literal and genetic sequences and DNA translation. <http://www.code10.org>. (2009)
 28. Chen, Y., A. Wan, and W. A. Liu (2006) Fast parallel algorithm for finding the longest common sequence of multiple biosequences. *BMC Bioinformatics* 4: S4.
 29. Khan, S., G. Situ, K. Decker, and C. J. Schmidt (2003) GoFigure: automated gene ontology annotation. *Bioinformatics* 19: 2484-2485.
 30. The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics* 25: 25-29.
 31. Piano, F, A. J. Schetter, D. G. Morton, K. C. Gunsalus, V. Reinke, S. K. Kim, and K. J. Kempfues (2002) Gene clustering based on RNAi phenotypes of ovary-enriched genes in *C. elegans*. *Curr. Biol.* 12: 1959-1965.
 32. Tijsterman, M., R. F. Ketting, O. L. Kristy, S. Titia, and R. H. Plasterk (2002) RNA helicase MUT-14-dependent gene silencing triggered in *C. elegans* by short antisense RNAs. *Science* 295: 694-697.
 33. Maeda, I., Y. Kohara, M. Yamamoto, and A. Sugimoto (2001) Large-scale analysis of gene function in *Caenorhabditis elegans* by high-throughput RNAi. *Curr. Biol.* 11: 171-176.