RNA 공통 구조 기술자 Committee Machine을 이용한 ncRNA 예측 성능 향상 기법

남진우[O,1,2] 김성규[1,2] 정제균[1,2] 장병탁[1,2,3]
서울대학교 대학원 생물정보학 협동과정[1]
서울대학교 바이오정보기술 연구센터(CBIT)[2]
서울대학교 컴퓨터공학부 바이오지능연구실[3]
{jwnam, skkim, jgjoung, btzhang}@bi.snu.ac.kr

Improving ncRNA Prediction using RNA common-structural Desciptor (RCSD) Committee Machine

Jin-Wu Nam[O,1,2] Sung-Kyu Kim[1,2] Je-Gun Joung[1,2] Byoung-Tak Zhang[1,2,3]
Graduate Program in Bioinformatics[1]
Center for Bioinformation Technology (CBIT)[2]
Biointelligence Laboratory, School of Computer Science and Engineering[3]
Seoul National University, Seoul 151-742, Korea

요 약

최근 포유동물의 유전체에는 알려진 것보다 훨씬 많은 RNA 전사체가 발현되고 있음이 밝혀지고 있으며, 그 중에 많은 부분이 non-coding RNA로 알려지고 있다. 세포 내에서 non-coding RNA의 기능이 훨씬 다양해지고, 중요해지고 있는 상황에서 새로운 non-coding RNA를 정의하고, 탐색하는 것은 가장 시급한 과제이다. 본 연구에서는 이전 연구에서 RNA 공통 구조 학습을 위해 제안되었던, esRCSG (evolutionary search for RNA common-structural grammar) 알고리즘의 성능 향상을 위해, committee machine을 도입한다. Committee machine은 마지막 세대에서 최적화된 RNA 공통 구조 기술자 (RCSD)와 차상위로 최적화된 기술자들 중, 양성데이터와 음성데이터의 치역을 합쳤을 때 특이도는 거의 변화가 없으면서 민감도의 증가가 가장 큰 기술자들의 집합이다. Committee machine은 특히 family type의 서열의 가진 특정 ncRNA에서 좋은 성능 향상을 보인다. microRNA를 이용한 성능평가에서 특이도의 변화가 거의 없이 민감도의 성능이 약 1.5배 향상되는 결과를 보였다. 이러한 특이도와 민감도가 높은 기술자를 이용함으로써 새로운 non-coding RNA를 예측하는 것을 약속할 수 있을 것이다.

1. 서 론

최근 식물과 동물에서 발견된 대규모의 전사체는 유전체의 약 3%가 발현된다는 기존의 믿음을 넘어서 비전사지역의 상당부분이 발현되고 있음을 말해주고 있다[1,2]. 더구나, 유전자가 전사되는 반대 가닥에서도 상당량의 전사체가 전사되고 있다는 사실은, 기존의 많은 학설을 뒤 업는 획기적인 결과이다. 특히, 발견된 전사체의 상당 부분이 non-coding RNA (ncRNA)이며, 그 중에서도 기능이 정확히 밝혀지지 않은 microRNA (miRNA)와 같은 small ncRNA이 많은 부분을 차지하고 있다는 것은 RNA world의 가설을 뒷받침하는 증거로써 시사하는 바가 크다[2]. 이러한 연구 보고는, 많은 전사체중에서 기능을 하는 새로운 ncRNA를 탐색하고 동정하는 작업을 긴급히 요구하고 있는 것이다.

세포 내에서 ncRNA의 기능은 대부분 그 서열와 구조에 의해서 유추될 수 있다. ncRNA의 서열과 구조는 기능의 보존을 위해 진화상에서 보존되어 왔고, 그것을 동정하는 주요한 자질로 사용되어 왔다. 보존된 구조와 모티프를 동정하는 가장 간단한 방법은 multiple sequence alignment을 하는 것이다. Profile hidden Markov model은 정렬된 서열로부터 전이확률과 발산확률로 보전된 구조와 서열을 profiling하는 확률 예측 모델이다[3]. INFERNAL과 같은 covariance 모델은 structural multiple alignment을 이용하는 좀 더 발전된 확률 모델이다 [4]. 최근 RNA 구조를 표현할 수 있는 문법과 언어가 개발되면서, 해석 가능한 RNA 모델이 속속 소개되고 있다[5,6,7]. 특히 RNAmotif는 RNA의 구조적인 패턴을 context-free 문법으로 표현되는 '기술자'(descriptor)로 정의하고, 정의된 기술자를 이용하여 서열 데이터베이스로부터 새로운 RNA를 탐색할 수 있는 기능을 제공한다[7]. 하지만 구조 기술자나 공통 구조를 표현하는 공통 구조 기술자(common-structural descriptor, CSD)는 컴퓨터를 이용한 자동 생성과 학습이 불가능하게 여겨졌다. 이전 연구에서는 이것을 극복하고자 genetic programming을 이용하여 정렬되지 않은 RNA 서열로 부터 공통 구조를 표현하는 기술자를 자동 학습하는 evolutionary search for RNA common-structural grammar (esRCSG) 알고리즘

을 제안하였다[8].

본 연구에서는 esRCSG의 정확도(accuracy)를 높이기 위한 방법으로, CSD의 committee machine을 탐색하는 알고리즘을 제안하고, microRNA에 적용하여 향상된 예측 성능을 보인다.

2. Evolutionary search for CSD

2.1 RNA 구조 기술자

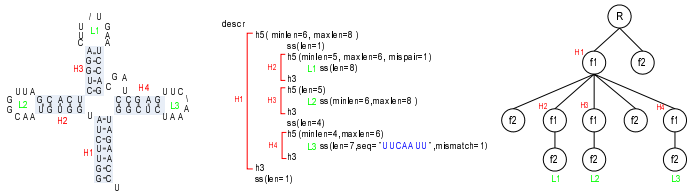

그림 1. tRNA 구조 기술자의 예와 트리표현

RNA CSD는 2차구조의 stem-loop 구조를 기본으로, bulge, internal loop, flanking sequence 등을 표현한다. 그림 1과 같이 tRNA는 3개의 stem-loop(H2, H3, H4)와 한 개의 stem 구조(H1)를 가지고 있으며, stem-loop사이의 bulge구조로 표현된다[8]. 기술자에서 구조는 'mispair', 'minlen/maxlen', 'len'로 서열은 'seq', 'mismatch' 의 파라미터들로 자세히 표현하게 된다. 이러한 기술자는 함수 f1 과 f2로 정의된 함수트리구조로 변환 가능하며, 이 트리는 유전자 프로그래밍(genetic programming)의 하나의 염색체(chromosome)로 사용된다[8].

2.2 알고리즘

```
begin                                   /* Structural Learning Step */   Step 1
    t = 0                               /* t = generation number */
    initialize P(t)                     /* random initialization */
    evaluate P(t)
    while (not termination-condition) do
    begin
      A = A + Top P(t)                  /* Move Top 5 to Archive(A) */
      t = t+1
      select P(t) from P(t-1)           /* Ranking selection */
      crossover-mutate P(t) except Best /* Elitism*/
      evaluate P(t)
end
```

```
w = wordwise(training data)                                              Step 2
begin                                   /* Sequential Learning Step */
        u = 0
        initialize P(u) from A with w   /* Incorporate words */
        evaluate P(u)
        while (not termination-condition) do
        begin
          A' = A' + Top P(u)
          u = u+1
          select P(u) from P(u-1)       /* Ranking selection */
          mutate P(u) for only seq. except Best
          evaluate P(u)
        end
end
```

그림 2. esRCSG 알고리즘

esRCSG는 그림 2에서처럼 구조 학습단계, 서열학습 단계의 두 단계로 CSD를 학습한다. 구조 학습단계에서는 간단한 유전알고리즘을 루틴을 따르지만, elitism을 사용하고, 각 세대의 top5는 archive를 만들어 따로 저장하게 된다. 두 번째 단계에서 archive에 저장된 CSD를 복제하여 wordwise에 의한 서열 조각들을 포함시켜 학습한다. 이때 crossover 연산자는 사용치 않으며 mutation 연산자만을 사용한다.

2.3 적합도 함수

$$F_{i,j} = k \times a_{i,j} + (1-k) \times r_{i,j} - C_{i,j} \quad ,0 \le k \le 1$$

$$TS_{i,j} = TD_{i,j} \times 10 + NN_{i,j}$$

$$C_{i,j} = \frac{1}{(NS+PS)^2} \times \frac{TS_{i,j}}{TS_{best,j-1}}$$

적합도 $F_{i,j}$는 위 식에서처럼 j 세대 i번째 개체의 민감도(a)와 정확도(precision, r)로 표현되며, k에 의해 민감도와 정확도의 trade-off를 결정한다. 또한 C라는 복잡도 수치를 포함하여 학습되는 트리의 크기를 조절하게 된다. 복잡도는 노드의 개수(NN)와 트리의 깊이(TD), 음성데이터 크기(NS)와 양성데이터 크기(PS)에 의해서 결정된다[8].

3. Committee machine 구축

Committee machine은 최종 학습된 CSD들 중 양성데이터에 대해서 가장 배타적(exclusive)이고 음성데이터에 대해서 가장 내포적(inclusive)인 관계를 나타내는 CSD set으로 구성된다. 이러한 관계는 아래 식처럼 B로 정의된 관측값을 이용하여 수치화 할 수 있고, set은 B 값이 가장 크면서 threshold를 넘는 CSD pair (CSD_{pair*})를 반복해서 탐색하여 얻을 수 있다. 이때 paired CSD는 다음 탐색 시 하나의 CSD로 간주된다.

$$CSD_{pair*} = \arg\max_{pair}(B)$$

$$B = \left(\frac{\Delta SP_i}{(\Delta SN_i + \alpha)} - \frac{\Delta SN_i}{(\Delta SP_i + \alpha)} \right)$$

여기서 α는 pseudo-value이고, ΔSN_i와 ΔSP_i는 두 개의 CSD를 데이터에 대해 합집합 연산을 시행했을 때 증가되는 민감도와 특이도를 나타낸다. 즉 특이도는 거의 변함이 없으면서, 민감도가 높아지는 CSD pair는 높은 B값을 갖게 되어 committee member로 등록된다. 이러한 committee machine 구축 과정의 알고리즘을 그림 3에서 확인할 수 있다.

```
begin                                   /* Boosting step*/          Step 3
    d = 0                               /* recursive depth */
    for(i = 0; i<n; i++)                /* n number of Archive A' members */
      E_d = search_best_exclusive(best, A'(i))
    if (evaluate(E_d) – evaluate (E_d-1) > threshold)
      d = d + 1
      recursive_call(E_d, A')           /* recursive search */
    else
      return E_d                        /* base line */
end
```

그림 3. Committee machine 구축 알고리즘

4. 실험 데이터 및 결과

4.1 실험 데이터 및 설정

알고리즘의 적용을 위해 152개의 human pre-miRNA 양성데이터 서열과 random 서열과 타 ncRNA로 이루어진 500개의 음성데이터를 사용한다. 양성데이터는 miRBase (http://microrna.sanger.ac.uk/sequences/)에서 다운로드 받을 수 있다. esRCSG에서 population 크기는 300 세대수는 30으로 고정했으며, k값을 0부터 1까지 0.1씩

증가해가면서 최적화하였다. Committee machine 구축 과정에서 threshold 값은 3으로 정하였고, CSD의 정확도를 측정하기위해 *F*-measure(민감도와 특이도의 조화평균)를 사용한다.

4.2 committee machine 도입 효과

esRCSG에 의해 학습되는 pre-miRNA의 CSD는 유전체상의 예측을 위해 높은 특이도를 요구받는다. 그리하여 *k*를 0.9로 설정한 뒤 학습을 진행하였다. 그림 4의 회색 막대는 committee machine을 사용하기 전 최적 pre-miRNA CSD의 민감도와 특이도, 그리고 정확도를 보여주고 있다. 반면 흰색 막대는 committee machine을 이용하여 예측한 결과를 보여주고 있다. 결과에서 committee machine을 사용한 후 특이도에는 거의 변화가 없었으나, 민감도에서 약 1.5배 정도의 향상이 있었고, 이것은 정확도의 향상에 영향을 미쳤다.

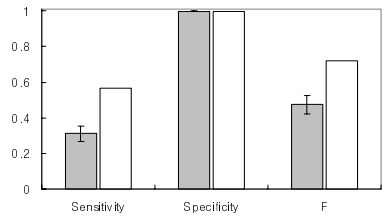

그림 4. Committee machine 효과

그림 5는 *k*의 변화에 따른 committee machine에 의한 민감도 향상(막대)과 committee 멤버의 평균수(line)를 보여주고 있다. 민감도의 향상은 0.8에서 가장 많이 이루어졌지만, member의 평균수가 많게 나타나고 있어 적절치 않아 보인다. *k*가 0.6 이하에서는 높은 민감도의 CSD가 학습되어 committee machine의 효과는 나타나지 않는다.

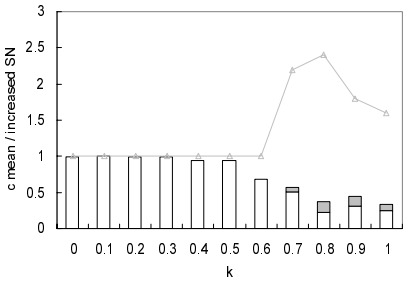

그림 5. 민감도의 향상과 committee 멤버의 평균수

4.3 구축된 CSD committee

그림 6은 최종 학습된 pre-miRNA CSD committee를 보여준다. Committee는 총 3개의 CSD로 이루어져 있으며 최적의 CSD는 1번이다. 1번과 2번의 *B* 값은 32.9로 특이도에서는 거의 변화 없이 민감도에 많은 향상을 보였으며, 다음 (1,2)번과 3번의 *B* 값은 3.3으로 나타나 약간의 민감도가 증가하였다. 이는 1번과 2번이 각기 다른 영역의 양성데이터를 대변하고 있음을 말해주며, pre-miRNA가 family 형태를 가지고 있다는 사실은 committee machine의 필요성을 대변한다.

```
descr1
h5 ( len=14, seq="ccaguuc", mismatch=4 )
        ss ( minlen=11, maxlen=25 )
        h5 ( minlen=1, maxlen=13 )
                h5 ( minlen=2, maxlen=16 )
                        ss ( minlen=10, maxlen=21 )
                h3
        h3
h3

descr2
h5 ( len=24, mispair=5, seq="cccugcu", mismatch=3 )
        ss ( minlen=8, maxlen=17 )
h3

descr3
h5 ( len=24, mispair=5, seq="caucgug", mismatch=3 )
        ss ( minlen=8, maxlen=17 )
h3
```

그림 6. pre-miRNA CSD committee

5. 결론

기존의 esRCSG의 알고리즘은 pre-miRNA와 같이 family 형태로 된 ncRNA에 대해 좋은 민감도를 갖는 CSD를 학습하지 못했다. 이러한 단점을 극복하기 위해 양성데이터에 대해 배타적이고, 음성데이터에 대해 내포적인 committee machine 탐색하였고, 정확도의 큰 향상을 가져왔다. 반대로 week runner의 교집합 연산을 통해 낮은 *k*에서 특이도의 향상을 가져오는 방법도 생각해 볼 수 있으며, 이는 boosting과 유사한 효과를 가져 올수 있을 것으로 보인다.

감사의 글

이 논문은 교육부 BK21 사업, 산업자원부 차세대 신기술 과제 및 과학기술부 국가지정연구실사업(NRL), 서울과학장학금에 의하여 지원되었음.

참고문헌

[1] Katayama,S., Tomaru,Y., Kasukawa,T., Waki,K., et al., Antisense trasncription in the mammalian transcriptome. *Science*, 309:1564-1566, 2005.

[2] Lu,C., Tej,S.S., Luo,S., Haudenschild,C.D., Meyers,B.C. and Green,P.J. Elucidation of the small RNA component of the transcriptome. *Science*, 309:1567-1569, 2005.

[3] Eddy, S.R. Profile hidden Markov models. *Bioinformatics*, 14(9):755-63, 1998.

[4] Eddy, S.R. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics*, 3(1):18, 2002.

[5] Sakakibara,Y. Pair hidden Markov models on tree structures, Pair hidden Markov models on three structures. *Bioinformatics*, 19:i232-240, 2003.

[6] Cai,L., Malmberg,R.L. and Wu,Y. Stochastic modeling of RNA pseudoknotted structures: a grammatical approach. *Bioinformatics*, 19:i66-73.

[7] Macke,T.J., Ecker,D.J., Gutell,R.R., Gautheret,R.R., Case,D.A., and Sampath,R. RNAmotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res*, 29(22):4724-4735, 2001.

[8] Nam,J.W., Joung,J.G., Ahn,Y.S. and Zhang,B.T. Two-step genetic programming for optimization of RNA common-structure. *LNCS*, 3005:73-83, 2004.