생물학적 시스템에서 확률적 연결 모델 추론

박동석 · 송선희 · 나하선* · 김문환* · 배철수** · 나상동
조선대학교 · *한국전파기지국 · **관동대학교

Probabilistic Connection Models Representation of Systems Genetic

Dong-Suk Park · Sun-Hee Song, Ha-Sun Na* · Moon-Hwan Kim*,
Chul-Soo Bae** · Sang-Dong Ra
Dept. of Computer Engineering, Chosun University
*KRTnet Corporation Network Planning
sdna@mail.chosun.ac..kr

요 약

생물학적 유전자 배열에서 다양한 레벨로 분자 세포 간 네트워크를 입증하여 고 처리를 응용하여 수치학적인 표현 모델 분석으로 정보공학 네트워크를 연구한다. 확률적 그래프 모델을 사용하여 네트워크의 계층적 구성 특성을 이용하여 생물학적 통찰력을 확률함수를 응용해 복잡한 세포 간 네트워크에 대한 고 대역 처리 데이터의 근원인 DNA 마이크로 배열을 응용하여 유전자 베이스네트워크 논리를 유전자 표현 레벨로 나타낸다. 유전자 데이터로부터 확률적 그래프 모델들을 추정 및 분석하고 논리적으로 예측하여 확률적 그래프 모델이 정보공학 네트워크로 확장 추론 한다.

1. 서론

유전자 생물학에서 많은 변화가 일어나고 있는 계기는 IT와 BT를 접목하면서 크게 발전하고 있다. 생물학적 유전자 배열은 유전 스케일에서 세포를 입증할 수 있는 고 처리 분석의 개발로 가능하며, 이 분석법은 다양한 레벨에서 분자 네트워크와 그 요소들을 검출하여 DNA로부터 전령에 전사(轉寫)된 유전정보와 단백질-단백질과 단백질-DNA 상호관계, 염색체 구조, 단백질양, 위치측정, 환경의 영향에 의한 비유전성의 일시적 변이를 포함한다[1]. 이런 데이터를 통해 세포 처리 과정, 생물학적 네트워크 등 다양한 각도에서 정보공학 네트워크인 WWW에 응용하고자 한다.

생물학 유전자 배열에서 수치 해석적 도전은 기저의 메카니즘에 대한 생물학적 통찰력으로 고 처리된 다른 데이터를 전송하기 위한 방법을 제공한다. 세포 시스템을 분석으로 얻은 데이터의 통합은 유전자 발현과 단백질 상호작용보다 통일성 있는 재구성을 가능하게 하여 잡음 효과를 감소시킬 것이다. 그러나 이러한 통합을 얻기 위해서는 검측법과 짝을 이루는 생화학적 원리를 이해해야만 되기 때문에 분석된 결론은 단순히 데이터를 묘사하는데 그쳐서는 안되므로, 관련된 생물학적 객체와 절차에 대해 새로운 지식을 제공할 수 있다.

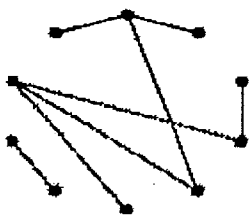
확률적 그래프 모델은 추상적인 것을 단순화한 형태에 다양한 조건에서 시스템의 행동과정을 발생시켜 이 행동들에서 시스템 요소의 역할을 추론하고, 확률적 모델 검측 잡음과 생물학적 시스템의 측면을 설명하기 위한 확률함수를 사용한다. 데이터를 분석하는 모델접근법에서 모델

의 공간을 정의하고 모델링하기 위해 논리적 추론을 통한 예측으로 데이터를 발견하는 학습절차법을 이용한다.

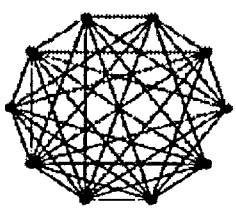
본 연구는 세포 간 네트워크의 모델분석을 위해 확률적 그래프 모델로 알려진 수학적 모델에서[2] 상호작용을 통계분야로 개발하고 확률적 결과와 모델을 연속적인 관측하여 확률적 그래프 모델을 추론하는 수치 해석적 절차로 연구한다.

2. 네트워크 그래프 탐색

확률적 그래프에서 네트워크의 계층적 구성 특성을 이용하여 생물학적 신호경로를 입증한 확률적 신호경로 그래프는 그림 1과 같이 네트워크 탐색 입증 그래프 등 분자생물학 분야에서 확률적 인식을 각 노드와 같은 수가 들어오는 에지와 나가는 에지를 가지기 때문에 실제 네트워크에서는 탐색입증이 5배 정도로 나타난다.



(a)



(b)

그림 1 네트워크 그래프 탐색
Fig 1 Network search graph

그림 1에서 확률적 그래프는 객체의 확률적인 면과 구조적인 면을 모두 함께 다룰 수 있기 때문에 복잡한 객체를 다룰 수 있으며, 다중 계층으로 확장되면서 네트워크 확률 탐색입증으로 객체 표현이 가능하다. 하부 계층에는 객체의 기본이 되는 정보를 표현하고 상위 계층에는 하부 계층의 조합으로 얻어질 수 있는 네트워크 확률적 정보를 추론함으로써 상위 계층으로 올라 갈수록 학습된 모델을 이용하여 탐색입증이 가능하여 진다. 여기서 확률적 탐색입증 그래프의 특성과 정보 표현 탐색 인식 등이 매칭 방법에 의해 이루어진다. 생물학적 시스템을 모델링 할 때 확률적 그래프에서 출력되는 결과를 네트워크 노드는 유도된 에지의 집합으로 이루어지므로[3] 확률적 네트워크 그래프는 다음과 같이 정의된다.

Definition 1 (network Graph) A network graph over $R_V \cup R_E$ is a 4-tuple $R=(V(R),E(R),\mu,\delta)$ such that

- R_V, representing a network vertex set, is an n-tuple($\alpha_1,\alpha_2,...,\alpha_n$) where each α_i, called a network vertex, is a network variable;
- R_E, called a network edge family, is an m-tuple($\beta_1,\beta_2,...,\beta_m$) where each β_j, called a network edge, is also a network variable;
- $V(R)$ is a finite and nonempty set of vertices;
- $E(R)\subset V(R)\times V(R)$ s a set of ordered pairs of distinct elements in $V(R)$
- $\mu: V(R)\rightarrow R_V$ and $\delta: E(R)\rightarrow R_E$ are functions.

$G=<N,E>$을 확률 모델 그래프 R에서 확률 함수 $M=(\mu,\delta)$에 의해 출력되는 그래프라고 하면, R에서 G가 출력될 확률은 G의 정점과 에지의 출력 확률의 곱으로 표현된다.

3. 유전자 결합 확률 분포 인식

통계적 모델에서의 인식이란 주어진 데이터와 모델의 관계에서 최대 확률을 얻어내는 모델을 추출하는 과정이라고 할 수 있다. 본 논문에서 사용한 랜덤 그래프 모델의 노드와 에지 모두 확률분포로 표현되는 모델이기 때문에 입력과 모델간의 최대 확률을 구하는 것이 각 단계의 모델

매칭 방법이다[4]. 모델을 M 이라하고 입력 그래프를 X 라고 했을 때 인식이란 주어진 관측된 그래프 X 에서 사후확률(posteriori probability)를 최대화 시키는 모델 $\hat{M}$ 을 찾는 것이다.

$$\hat{U} = \underset{U_i}{\operatorname{argmax}} P(U_i | X)$$

(3-9)

베이지안 규칙에 의해서 사후확률 $P(M_i | X)$ 는 다음 식 (3-10)과 같이 표현된다.

$$P(U_i | X) = \frac{P(X | U_i) P(U_i)}{P(X)}$$

(3-10)

위 식에서 $P(X)$ 가 M_i 에 독립이라면, 식 (3-11)과 같이 쓸 수 있다.

$$\hat{U} = \underset{U_i}{\operatorname{argmax}} P(U_i | X) P(U_i)$$

(3-11)

위 식에서 $P(X | U_i)$는 모델U_i에서 X가 관측될 확률이고 $P(U_i)$는 사전확률로서 전체 모델에서 모델 U_i의 빈도수이다. $P(U_i)$ 이미 알려져 있다면 $P(X | U_i)$만 얻어내면 되는데 $P(X | U_i)$는 바로 모델 U과 X의 매칭과정에서 계산되어 진다.

본 논문에서 제안한 모델을 이용하여 최종적으로 글자를 인식 확률을 구하는 방법은 식 (3-12)와 같다.

$$P(U | X) = \prod_{i=1, x \subset X}^{n} P(U_{Gi} | x) \times P(U_C | X)$$

(3-12)

위 식에서 $P(U_{Gi} | X)$는 입력 획 집합 x에 대한 자소 모델 G_i의 매칭 확률이고, $P(U_C | X)$는 해당하는 글자 모델과의 매칭 확률이다. 각각의 자소가 매칭된 확률 $P(U_{Gi} | X)$을 모두 곱하고, 대상 글자 모델과의 매칭 확률 $P(U_C | X)$를 곱함으로써 글자 인식 확률을 얻을 수 있다. 위와 같이 글자인식 확률을 구해서 최고 높은 확률을 출력하는 글자가 매칭 글자가 된다.

학률 그래프 모델은 유전자 표현에서 하향식 계층분해에서 가장 하위부분에 있는 학률적 유전자 모델링하기 위해서 제안하였다. 그림 3-6과 같이 확률 그래프에서 주로 사용되는 유전자 획모델로 이루어진다. 본 논문에서는 유전자 단일 접속을 사용하지 않고, 두 개의 유전자 획을 표현할 수 있는 확률적 획모델을 제안하는데, 가장 기본이 되는 4가지의 유전자 획을 이용하여, 두 개의 유전자 획 사이의 결합을 정의 하고 유전자 획 모양에서 동그라미 획은 제외를 시켰는데, 그 이유는 제안된 유전자 획들을 결합하면 동그라미를 충분히 표현할 수 있기 때문이다.

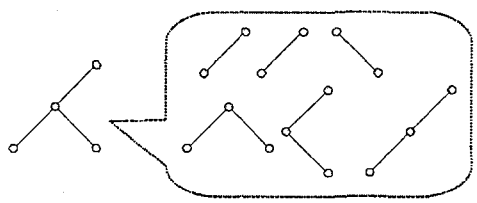


그림 2. 유전자 결합에서 확률분포 획 모델
Fig 2. Probability destribution dash model to gene combination

이와 같은 확률적 그래프로 정의된 유전자 획모델에서 확률적 모델의 에지와 노드의 파라미터를 결정한다. 본 연구에서는 유전자 획의 확률적 모델 각각의 매칭에서 세포간 네트워크가 형성하는 확률분포를 갖도록 정의 하였다.

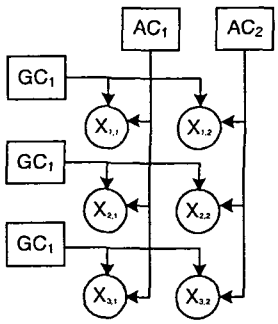
확률적 그래프 모델에서 각 에지에는 에지의 방향에 대한 확률분포가 존재하고, 두 개의 에지가 연결된 연결부분의 노드에는 두 에지간의 결합각에 대한 확률 분포 모델 매칭에서 세포간 인식 과정에서 상호 경로 연결 관계를 나타내는 확률함수를 응용하여 확률적 그래프 모델에서 두 개의 확률을 가질수 있는 인식 추론한다.

4. 유전자 표현 프로파일 모티프

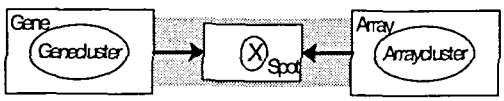
유전자 그래프 모델링을 확장시킬 수 있는 규칙을 포함해서 세포 간 네트워크에 대해 고대역처리 데이터의 주된 근원은 DNA의 아주적은 마이크로 배열(micro array)을 사용하여 얻은 유전 표현 프로파일 모티프[8]이다. 유전표현에서 임의 변수 $X_{g,a}$ 값에서 g는 유전자에 대한 지표이고 a도 분석에 대한 지표이다.

모델링 가설은 유전자가 동일하게 표현된 유전자 클러스터로 분할된다는 것과 각 클러스터

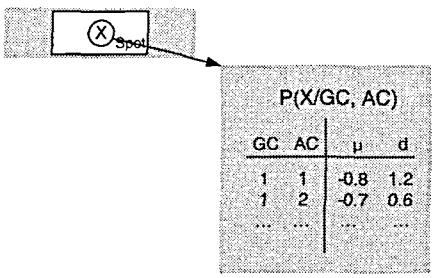
내의 유전자가 각 배열에서 전형적인 표현 레벨을 가진다. 여기서 대열집단으로 분할된다고 가정할 때 유전자 표현은 같은 배열집단에 속한 마이크로 배열과 비슷할 것이라고 가정할 수 있다. 여기에 임의의 변수를 더하여 모델을 제시하고 이 모델에서 Gene$Cluster_g$ 를 유전자 g의 집단으로 할당량을 나타내며, Array$Cluster_g$는 대열 a의 집단 할당량을 나타낸다.



(a) 유전자 네트워크 구조



(b) 유전자 집단 모델



(c) 유전자 할당량

그림 3. 3개의 유전자와 2개의 배열을 단일로 한 베이스 네트워크 집단

Fig 3. 3 gene and arrangement of 2 things alternative one way one base network group

그림 3에서 대열집단의 유전자 g를 유전표현 Gene$Cluster_g$와 Array$Cluster_g$에 달려있다 그러므로 이 모델은 특정 유전자 집단과 대열집단에 대응하는 모든 검측을 같은 조건 분류에 의해 통제된다고 가정할 때 모델을 베이스의 네트워크로 각 표현 특성 $X_{g,a}$가 Gene$Cluster_g$와 Array$Cluster_g$의 근본이 된다. 실제의 네트워크 구조는 데이터에 있는 유전자의 수와 대열에 따라 달라지기 때문에 베이스의 네트워크로는 두 가지 중요한 측면을 명확하게 나타낼 수 있다. 첫째, 임의 변수는 유전자와 같은 각 실체의 특성을 나타낸다. 둘째 템플릿은 같은 타입의 모든 엔티티에 의해 공유되므로 P($X_{g,a}$|Gene $Cluster_g$, Array$Cluster_g$)의 조건적 확률은 g와 a의 선택이 비슷하므로 규칙성을 찾아냄으로써 더 간결한 모델을 구현한다. 유전자집단 문제에 대한 템플릿 모델을 보여주고 유전자와 대열만 주어지면 베이스의 네트워크를 발생시킨다.

유전자 할당량이 원래의 검측에서 동일한 유전자 표현을 가진 각 블록으로 분할 할 가능성이 매우 높다[5]. 여기서 E-step과 M-step 사이에서 반복되는 기대의 최대값을 사용하여 분배를 찾을 수 있다. 여기서 E-step은 유전자 대열의 확률적인 집단배분을 찾기 위해 현 파라미터를 사용하며 M-step은 이 할당량을 토대로 각 유전자/배열집단 조합 내의 분포를 재 추정한다. 이 모델에서 생물학적 메카니즘에 대한 통찰력을 얻는 데에도 확장될 수 있으며, 유전자의 동일표현이 동일규칙에 반영되므로 핵심적인 조절 메카니즘에는 유전자 촉매 인자 부분에 전사 인자가 결합되는 것이 포함된다. 따라서 유전자의 촉매 인자 지역에 있는 전사인자 결합 지점을 찾아낸다.

수학적 모델로 전환시키는 방법에서 하나는 특징적인 결합지점을 가지고 있는 촉매에 주석을 하고 그 후 이것을 유전자 실체의 새로운 특성으로 사용하는 것이다. 임의적인 이원변수 $R_{g,j}$는 유전자 g가 전사인자 j와 결합지점을 가지고 있고 각 유전자의 집단 할당량이 관련 결합지점과 표현 특성(2,5)에 직접 영향을 주는 모델을 구상할 수 있다. 유전자 실체에 적용해 볼 때 유전자 집단에 초점을 맞춘다. 이 결합지점은 표현을 예측하기 때문에 다른 결합지점을 발생하려고 하지 않고 또 다른 결합지점 표현 데이터에도 나타나 있지 않은 조건이 된다. 실제 촉매 시퀀스가 주어질 때 결합지점의 확률을 모델링함으로써 모델을 증대시킬 수 있으므로 촉매 지역과 모델 촉매 시퀀스 Seq_g에 따라서 $R_{g,j}$를 나타내는 새 객체를 도입할 수 있다. 이 조건적 확률 파라미터는 전사인자가 인식하는 특정한 모티프를 규정한다.

수학적 모델 구성은 Gene$Cluster_g$와 연관된

조건적 분포의 표현으로 촉매에 포함하고 결합지점이 존재하는 것이 유전자가 어떤 집단에 속하는지를 결정하는데 예측할 수 있다. 지금까지 연구된 조건적 확률은 의사결정트리(11)에서 포괄적으로 보였다. 모델 집단에 할당된 유전자를 설명하는 새로운 결합지점을 찾는 단계가 있고, 또 그들의 표현 프로파일과 촉매 지역을 토대로 한 집단에 유전자를 재배분하는 단계로 본다. 각 집단 내 표현의 배분을 재평가하는 단계는 학습의 두 종류의 데이터 사이 정보의 흐름을 포함하고 약간의 결함도 허용할 것이다. 이러한 정보는 유전자 집단 변수에 의해 전해지기 때문에 유전자 집단 변수는 통일성 있는 유전 표현 프로파일과 유사한 촉매들을 다 가진 유전자 집단을 나타내어 결합지점 모티프로 확인하였다.

5. 결론

세포 간 네트워크의 모델 분석을 확률적 그래프 모델로 접근법에 의해 유전자 표현 레벨을 논의하여 생물학 지식을 정보공학적 네트워크로 분석 추정하는 연구이다.

생물학적 통찰력을 제공하는 모델을 얻기 위해 세포간 시스템 분석으로 얻은 유전자 발현과 단백질 상호작용보다 동일성 있는 재구성이 가능했고, 잡음 제거를 감소시켰다. 또 모델에 초점을 두어 그래프 모델을 명시하는 능력과 데이터를 통합시킨 것이다. 유전자 모델을 개선하는 요소는 생물학적 조절 메카니즘에 의해 간접 조절을 추론하는 능력과 같은 데이터 본질에 대한 통찰력이 있었고 생물학적 원리를 모델 디자인 통합을 학습하는데 제한 할 수 있었으며 실제 세부사항들을 포착하는 모델들을 분석하여 고 처리 데이터의 양과 다양성이 디자인을 포함한 단일 세포, 합성 기관, 전체 유기체, 그리고 사회 전체 수준에서의 시스템 등을 수치적 모델 구성에서 조건적 분포의 표현과 결합지점에 존재하는 유전자가 어떤 집단에 속하는지를 결정하는데 예측할 수 있어서 생물학적 통찰력을 얻는데 충족시키는 연구가 필요함을 지적할 수 있었다. 앞으로 생물학적 세포 간 요소의 모델링 선택 확장에 더 연구한다.

참고문헌

1. N. Friedman, L. Getoor, D. Koller, A. Pfeffer, Relational Data Mining, S. Dvzeroski, N. Lavrac, Eds. (Springer-Verlag, Berlin, 2001), pp. 307.337.
2. Y. Barash, N. Friedman, J. Comp. Biol. 9, 169 (2002).
3. I. Holmes, W. Bruno, Proc. Int. Conf. Intell. Syst. Mol. Biol. 8, 202 (2000).
4. M. Deng, T. Chen, F. Sun, Proceedings ofthe 7th annual International Conference on Computational Molecular Biology, M. Vingron et al., Eds. (ACM Press, New York, 2003), pp. 95.103.
5. A. Hartemink, D. Gifford, T. S. Jaakkola, R. A. Young, Pac. Symp. Biocomput. 6, 422 (2001).