2ZB-4

確率ネットワークを用いた相同タンパク質検索ツール

吉原 久雄* 賀屋 秀隆§ 松井 藤五郎† 朽津 和幸‡ 大和田 勇人†

東京理科大学大学院理工学研究科経営工学専攻* 同 理工学部 経営工学科†
同 応用生物科学科‡ ゲノム創薬研究センター§

1 序論

新たに発見されたたんぱく質の機能予測は，バイオインフォマティクスの分野では最も重要な計算問題の一つである。現在のバイオインフォマティクスの分野では、HMMER，FASTA 等様々なツールが用意されている．生物学者全員がその全てのツールを使用するわけではないが、たんぱく質の機能はそのようなツールを幾つか使用し、それから得られたそれぞれの結果を比較して予測する。ここでの判断は数値的なものではなく、各々の生物学者が持つ知識や経験を考慮して行われる。そのようなことから，様々な要素を考慮して予測を行うシステムが以前から研究されている [1] [2]．又、知識を基にモデルを作成できるものとしてベイジアンネットワークというモデルがある [3]．ベイジアンネットワークは、データからモデルのパラメータを学習するのと同等に統計的な推論を行うことが出来る柔軟かつ力強いフレームワークでるため、最近様々な要因を考慮するために使用されている．
それゆえ、この論文で我々は，ベイジアンネットワークを基とし，生物学者の持つ知識や経験といったものを考慮するたんぱく質の機能予測を行う手法を提案する．

2 提案手法

我々が提案する確率ネットワークは，ベイジアンネットワークを基にしている．生物学者の独自の知識や経験則を全て反映させるために，ゼロからネットワークを全て構築する．考慮すべきものはある程度決まっているので，使用するノードの種類は最初に用意しておき，それを使用して自由にネットワークを構築できるようにする．自由に構築すると言っても，基本は”これを持っていればあるタンパク質である”という形である．

我々が提案する確率ネットワークを用いて作成した例を図 1 に示す．これは，BH1，2，3 と呼ばれる 3 つのドメイン領域と膜貫通領域を持つ Bcl-2 family をモデル化した例である．基本的に，モデルは「Tool Node」「Function Node:Require Node, Optional Node」「Optional Node」「Result Node」の４つのノードを使用して構築される．Required と Optional は扱いは異なるが，Function Node としてまとめている．又，モデルは上下に分けることが出来，それぞれ Subjective block と Objective block と呼ぶことにする．前者は生物学者自らが作成するもので，後者はその作られたモデルに対応して逆向きのノードが自動的に作成される．

2.1 Subjective block

Subjective block では全てのパラメータを生物学者がインプットする

2.1.1 Tool Node

Tool Node とは，確率ネットワークのモデル内で予測を行う際に使用されるツールのノードである．このノードは特に複雑な入力ではなく，信頼度を入れることによって条件付確率表 (以下 CPT) が作成される．

2.1.2 Function Node

Require Node と Optional Node とは，あるタンパク質がある要因を持っているということを表すノードである．このノードのパラメータは Tool Node の信頼度を使い，以下の重付き平均で求める．この式で値を求めることにより，直感的な確率にすることができる．

$$Pr(D_j = T|\pi(T_j)) = \frac{\sum_{t\in T} Pr_{T_i} \prod(1-Pr_{T_i})}{\sum_{t\in T} Pr_{T_i} \prod(1-Pr_{T_i}) + \sum_{t\in T} Pr_{F_i} \prod(1-Pr_{F_i})}$$

ここで Pr_{T_i} は，ツール i が “ ある要因を持っている ” と判断する確率であり，Pr_{F_i} はその逆で，ツール i が “ ある要因を持っていない ” と判断する確率である．

2.1.3 Result Node

結果ノードとは，あるタンパク質が最終的にそのタンパク質である，ということを確率ネットワーク内で表すノードである．このノードの CPT は Function Node から得られた値を基に，同時確率で求めることができる．最終的に”あるタンパク質である”という確率は，基本的にはベイジアンネットワークと同じように同時確率で求めることができるが，”ある要因は無くても良いが，あったほうがより良い”

Development of motif search system that enable burial of knowledge of biology

Hisao Yoshihara*, Hidetaka KAYA§,Tohgoroh MATSUI†, Hayato OHWADA†, and Kazuyuki KUCHITSU‡

*Department of Industrial Administration, Graduate school of Science and Technology, Tokyo University of Science, †Department of Industrial Administration, Faculty of Science and Technology, ‡Department of Applied Biological Science, Faculty of Science and Technology,§Genome and Drug Research Center

といった場合，異なる式で求めることができる．

$$Pr(T_1,...,T_l,D_1,...,D_m,F,R) = 1-\{(1-Pr(T_1,...,T_l,D_1,...,D_m,R))Pr(F|\pi(T_n))\}$$

これにより，その要因を持っていなかったとしても最終的な確率を下げることなく計算することができる．

2.2 Objective block

Objective block では，全てのパラメータは自動的に計算される．データは自動的に PROSITE から取得し使用する．そこから統計量に基づき CPT の構築を行う．最終的に，Subjective block の結果ノードとオブジェクティブノードの Function Node を使用し Result Node の CPT を変更する．ここで，ベイジアンネットワークでも使用されいている信念伝播を用いることによって間のノードの CPT を更新することが出来，統計量も織り交ぜた値とすることが出来る．

3 実験

3.1 実験方法

複数のドメインを持つようなタンパク質を認識するように構築することが可能なため，我々はそのようなタンパク質が検出可能かどうか実験を行った．データベースは，ランダムな文字列で，長さ 1000 のアミノ酸配列が負事例として 10000 本，正事例として 1000 本含まれている．正事例の配列は，今回 Bcl-2 family をターゲットにしているのでそれに似るように作成した．確率的ネットワークで使用したツールは Blast，Fasta，HMMER，SOSUI である．モデルの構築やそれぞれのパラメータの入力は専門家に行ってもらった．又我々は，各々のドメインを分けて検索し，結果を統合することが出来るシステムである MDHMMER [4] とも比較を行った．

3.2 実験結果

図 2 は，確率的ネットワークで様々なシステムを使用した場合の結果を ROC 図で表したものである．このネットワークでは SOSUI を使用しているので，膜貫通領域を持つタンパク質は上位にくる．他の様々なシステムと比較すると，我々の提案手法はかなり性能がよいことが分かる．例えば，Blast と比較してみると，負事例が 300 のとき約 70 倍の正事例を発見している．又，今回データからパラメータを計算し構築したベイジアンネットワークとも比較を行っている．各々のノードの CPT の値は異なるが，ほぼ同じ結果となっていた．この二つの性能を比較するために，最小二乗法で値を比較した．目的変数の値を 1 とする．提案手法では，確率が一番高い正事例の値が 0.494 であるのに比べ，ベイジアンネットワークでは 0.243 であった．このことから，我々の提案手法の方が目的に対する誤差が少なく，性能がよいと言える．

4 結論

今回，様々な要因を考慮することにより，システムを単独で使用するよりも性能が上がり，ドメイン以外の要因も考慮することによって性能が大幅に上がった。これにより，

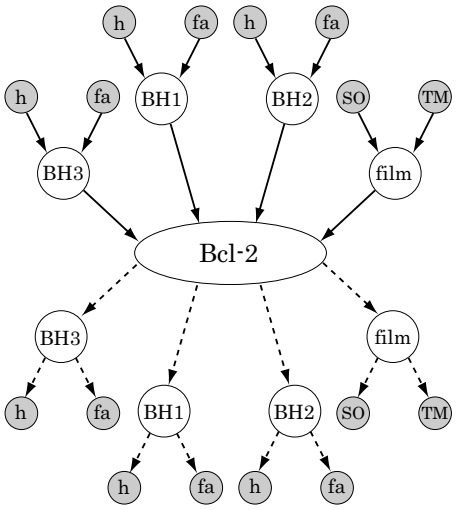

図 1　モデルの構築例

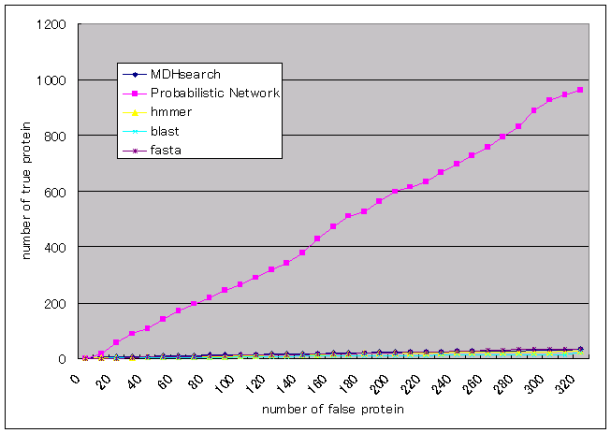

図 2　実験結果

各々の生物学者の持つ知識や経験則をを考慮することは，タンパク質の機能予測において有効な手段であることが分かった．

参考文献

[1] Ashutosh Garg Vladimir Pavlovic and Simon Kasif. A bayesian framework for combining gene predictions. *Bioinformatics*, Vol. 18, pp. 19–27, 2002.

[2] Z. Ghahramani A. Raval and D. L. Wild. A bayesian network model for protein fold and remote homologue recognition. *Bioinformatics*, Vol. 18, pp. 788–801, 2002.

[3] J. Pearl. Probabilistic reasoning in intelligent systems. *NetworksofPlausibleInference, Morgan Kaufmann*, 1988.

[4] 瀬下, 松井, 大和田. Multi-domain hmmsearch:マルチドメインを持つ遠縁なタンパク質のための相同性検索ツール. 第 5 回情報科学技術フォーラム (FIT-2006) 情報科学技術レターズ,, pp. 153–156, 2006.