社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2005－ARC－162 (12)
2005－HPC－101 (12)
2005／3／8

ネットワーク並列構築を適用したBOAによる蛋白質構造予測

村尾直哉[†]　棟朝雅晴[††]　赤間　清[††]

Bayesian optimization algorithm (BOA) は現在の集団分布を推定したベイジアンネットワークによる確率モデルを構築し，得られたモデルに基づいて次世代集団を形成することで探索を行う最適化アルゴリズムであり，適切な符号化が保証されない GA 困難な問題においても解を求めることができる．しかし，この分布推定にかかる計算コストは問題サイズに依存しており，問題サイズの増加に対して計算コストが大きく増加することが知られている．これまでの研究では，生成される確率ネットワークの品質を下げずに並列構築する手法を提案してきた．本研究では，このネットワーク並列構築に基づく並列 BOA を用いて構造エネルギー最小化問題を解くことで，蛋白質の立体構造予測問を解決し，提案手法の有効性について検証する．

Structure Prediction of Protein by BOA with parallel network construction

Naoya Murao ,[†] Masaharu Munetomo [††] and Kiyoshi Akama[††]

Bayesian optimization algorithm (BOA) is an advanced optimization algorithm that effectively solves GA-difficult problems in which it is difficult to ensure tight encoding. BOA constructs Bayesian networks based on probabilistic distributions of current promising solutions, and generates next population based on the obtained networks. However, computational cost of the estimation depends on the problem size, and is known the calculation cost increases rapidly along increasing the problem size. In our previous study, we proposed a BOA with parallelized Bayesian network construction. In this paper, we try to solve proteins structure prediction problems that minimize the structual energy by employing BOA with the parallel network construction, and investigate the effectiveness of our approach.

1. はじめに

遺伝的アルゴリズムなどに代表される進化計算は多くの最適化問題に適用されているが，単純な遺伝的アルゴリズムでは解くことが困難な問題が存在することが知られている．それらは GA-difficult な問題と呼ばれ，解の候補である個体の符号化が不適切である場合などが含まれる．このような GA-difficult な問題を解くための手法の研究が進められており，集団中の個体の分布を推定して次世代の集団を生成する手法として分布推定アルゴリズム (Esitimation of Distribution Algorithms, EDAs) がある．この EDAs の中でも特に優れた手法として Bayesian Optimization Algorithm(BOA) [12] が提案されている．BOA は集団の分布推定モデルに条件付確率に基づくベイジアンネットワークを用い，生成される確率ネットワークに基づいて次世代の集団を生成していくことで解を探索する手法である．この集団の分布を推定して生成される確率ネットワークは個体の遺伝子座間における依存関係を表現していると言える．そのため確率ネットワークから次世代の集団を生成し，解を探索していくことで，ビルディングブロックの交換が適切に行われ，GA-difficult な問題においても最適解を得ることができる．また，BOA は最適解を得られるまでに必要となる適応度評価回数に関する計算コストが比較的小さく，問題サイズの増加に対して評価計算コストの増加がほぼ線形であることが示されている．しかし，実際の問題において確率ネットワーク構築に要する時間は無視できない場合がほとんどであり，問題サイズの増加に対して計算コストが大きく増加することが知られている．これまでの研究では，分布推定により構築される確率ネットワークの品質を下げずに並列構築する手法を提案し，計算コストを下げることに成功している [8]．

また，蛋白質の立体構造予測問題には X 線結晶構造解析は核磁気共鳴方 (NMR) などが使用されてきたが，

† 北海道大学工学研究科　システム情報工学専攻
Division of Systems and Information Engineering, Graduate School of Engineering, Hokkaido University
†† 北海道大学情報基盤センター　大規模計算システム研究部門
Division of Large-Scale Computational Systems, Information Initiative Center, Hokkaido University

1 つの蛋白質の解析に数ヶ月の期間が必要であることも多いことが知られている．蛋白質の立体構造はアミノ酸配列の情報とそれぞれの原子の間の角度によって決定されるエネルギーを最小にするような構造であることが知られており，SA や GA などを用いて最適化する研究も行われている．

本稿では，実際の問題の応用例としてこの蛋白質の立体構造予測問題を取り上げ，確率ネットワーク並列構築を適用した BOA を用いて最適化を行い，その性能について検証する．

2. BOA

BOA(Bayesian Optimization Algorithm) [12] は集団中の個体の分布推定にベイジアンネットワークに基づく確率ネットワークを用いており，遺伝子座間の依存関係を条件付確率を用いて定量的に表現している．分布推定で得られる確率ネットワークを用いて次世代の集団を生成することにより，遺伝子座間の依存関係を考慮した次世代集団の生成を行いながら探索が進められる．この BOA において重要な処理は，この分布推定による確率ネットワークを構築する部分である．現在の集団を適切に反映するような確率ネットワークを構築するために BOA では以下の手順を行う．

(1) 辺のつながれていないノードのみからなる初期ネットワーク B を生成する．
(2) 追加可能なノードのペアをつなぐ辺を確率ネットワーク B に追加したものを B' とする．
(3) 確率ネットワーク B' における確率ネットワーク評価メトリックを計算する．
(4) すべての追加可能な辺に対し，2, 3 を繰り返す．
(5) 得られた評価メトリックの中から最良のメトリックを返す辺を確率ネットワークに追加する．
(6) 終了条件が満たさていなければ 2 へ戻る．

ここで，メトリックとは構築される確率ネットワークを評価する基準となるものであり，現在のネットワークの状態と集団の分布から計算される．このように，BOA では現在の確率ネットワークに追加すると良いと思われる辺を探し，それを追加した際の情報を更新しながら確率ネットワークを構築していく．また，ベイジアンネットワークでは条件付き確率を用いるため，構築されるネットワークは非循環有向ネットワークである必要がある (図 1)．

この BOA では確率ネットワークの構築中に個体の適応度評価を行わないため，個体長 l の増加によって生じる適応度評価にかかる計算コストの増加が，線形もしくはそれに近い結果が示されており，$O(l^{1.55})$ であることが実験的にわかっている．しかし，上記のコストは適応度評価回数にもとづく計算コストであり，確率モデルの構築に関する計算コストは含まれていない．BOA における確率ネットワーク構築のコストは $O(kl^3 + kl^2 2^k N)$ であることがわかっており (ここで，k は確率ネットワークにおいて，各ノードに入る辺の最大数，l は個体長，N は集団サイズ)，実際にはこの計算コストを無視できない問題のほうが多いと考えられる．そのため，並列化に関する研究が行われている．

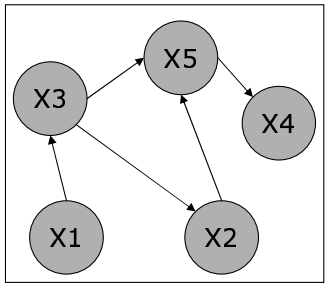

図 1 bayesian network の例

3. 並列確率ネットワーク構築

前節で見たように BOA の分布推定にかかる計算コストは問題サイズの増加に対して大きいことが知られている．そのため，並列化に関する研究が行われているが，メトリック計算の際に現在のネットワーク情報を利用するため，単純に分割し各ノードをプロセッサに割り当てて独立に更新していくと，構築されるネットワーク品質が下がる結果になる．また，構成される確率ネットワークはベイジアンネットワークを用いているために非循環ネットワークを構成している必要がある．一方で，各プロセッサがそれぞれのプロセッサで追加されたネットワーク情報を毎回通信しながら確率ネットワークを生成することは通信コストの増加につながり，そのコストは個体長の増加とともに大きくなることは明白である．本章では，まず従来手法である Ocenasek による並列化方法 [9] を述べた後で，これまでの研究で提案した並列化方法 [8] について説明し，それらの手法の比較を行う．

3.1 既存の並列 BOA

Ocenasek [9] により提案された確率ネットワーク並列構築方法は順列を用いて確率ネットワークの方向付けを行う．例えば 4 ノード $X = (X_1, X_2, X_3, X_4)$ に対して順列 $O = (3, 2, 1, 4)$ としたとき，X_2 は X_1 と X_4 に辺を追加可能でありメトリック評価計算などが必要があるが，X_3 の方向には辺を追加できないためメトリック評価計算が必要ない．同様に，X_1 は X_2 と

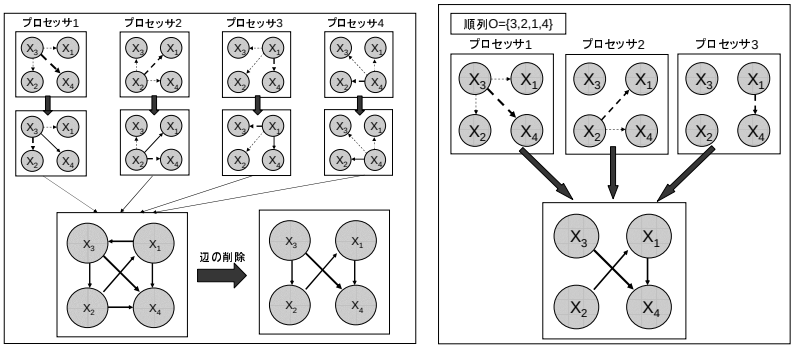

図 2 提案手法 (左) と従来手法 (右) における確率ネットワーク並列構築手法の比較

X_3 に対するメトリック評価計算が必要なくなる．また，辺の追加方向を限定することで非循環ネットワークを構成する可能性はなくなる (図 2 右)．

一般の問題では事象の因果関係が存在するため，ネットワークの方向があると考えられる．しかし，実際に解く問題の情報をあらかじめ得ていることは少なく，ランダムに順列を生成するしかないのが現実である．そのため探索空間の減少が起きる可能性が大きく，適切な順列を用いて探索されない場合には最適解を得ることができないと予測できる．

3.2 提 案 手 法

これまでの研究では，確率ネットワークの品質を下げずに並列構築する手法について検討してきた [8]．そのために各プロセッサが探索する範囲を限定せずに並列に確率ネットワークを構築する必要がある．しかし，そのようにすることで非循環ネットワークを構成する可能性があるので，枝切りという操作を加える．枝切りでは，循環サイクルを構成する部分の中からランダムに辺を削除する操作である．また，構成される確率ネットワークの品質を保つために，各プロセッサで追加された辺の情報などのネットワーク情報を共有する必要があるが，辺が追加されるごとに通信を行うのでは計算コストが高いことが考えられる．そこで，以下の T_i によって通信の間に各プロセッサで追加される辺の本数を制御する．

$$T_i = \begin{cases} a \times \frac{l}{P} & \text{if } i = 0 \\ c \times T_{i-1} & \text{otherwise} \end{cases} \tag{1}$$

ここで，l は個体長，P はプロセッサ数，$0 < a, c < 1$ である．T_i は上式のようにネットワークが出来上がるに従い値は小さくなる．確率ネットワークに追加されている辺の本数が多いとき，新たに追加した辺により循環サイクルを形成してしまう確率が高くなるため，辺の削除操作に伴うオーバーヘッドが増加する可能性がある．そこで，各プロセッサが行う辺の追加処理回数 T_i を確率ネットワークの大きさに従い制御することで，辺の削除操作によるオーバーヘッドの増加を抑えている．一方で，確率ネットワーク構築開始時など辺の本数が少ないときに，追加された辺により循環サイクルを形成する可能性は小さく，ある程度大きく独立に確率ネットワークを構築しても辺の削除操作によるオーバーヘッドは増加しないと考えることができる．図 2 左は，本手法の流れである．点線は評価されるメトリックのノードペアを示し，実線は追加された辺を示している．計算された T_i の分だけ確率ネットワークを生成し，マージした確率ネットワークに対し必要であれば枝切りを実行する．この後，T_i を再計算し，再び独立に確率ネットワークを生成することを繰り返す．

(1) 辺のつながれていないノードのみからなる初期ネットワークから開始する．
(2) 各プロセッサにおいて T_0 を計算する，$i \leftarrow 0$.
(3) 各プロセッサが割り当てられた領域に関して以下の操作を T_i 回繰り返す．
  (a) 各ノードペアに対し，辺を追加した場合のメトリックの計算を行う．
  (b) 得られた結果から，最良のメトリックの改善を与えるものに対し，辺の追加を行う．
(4) すべてのプロセッサの確率ネットワーク構造をまとめる．
(5) 循環サイクルを構成している部分が存在する場合には，循環サイクルを構成する辺の中からランダムに辺の削除を行う．
(6) 終了条件が満たされていなければ，結果として構成される確率ネットワーク構造をブロードキャストする．また，T_{i+1} を計算し $i \leftarrow i+1$ として，3 へ戻る．

以下では，これまでの研究で行われた性能比較実験のまとめを行う．

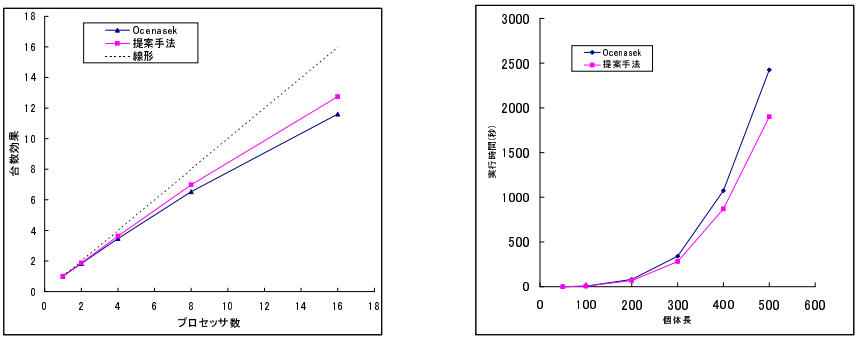

図 3 並列手法の比較実験．台数効果 (左)，問題サイズの増加に対する実行時間 (右)

3.3 性能比較結果

上記の 2 つの確率ネットワーク構築手法について，台数効果と問題サイズの増加に対する実行時間の測定を行ったのが図 3 である．対象関数は 5 ビットトラップ関数を用いている．また，実験環境は共有メモリ型並列計算機の SGI Onyx300(MIPS R14000/ 500MHz × 32CPUs / 16GB 共有メモリ) であり，10 回の試行の平均を記録してある．

結果から見てわかるように，従来手法よりも台数効果が高く，また問題サイズの増加に対しても実行時間の伸びは従来手法よりも小さいことがわかる．その要因として考えられることは構築される確率ネットワークの品質にある．実行 9 世代目に得られた確率ネットワークの一部 (図 4) について比較してみると，これまでの研究で提案した手法の方が先に収束していることがわかる (ここで，BOA では集団内で収束した遺伝子座に関しては辺を追加しない)．

これまでの研究で提案された本手法を蛋白質構造予測問題に適応し，その性能について検証する．

4. 実　　験

蛋白質は 20 種類のアミノ酸から構成しており，原子間における水素結合などのエネルギーの総和によりそのエネルギーが決定される．本稿では，岡本らの研究 [11] で使用された以下の蛋白質のエネルギー関数を用いてエネルギーを決定する．

$$E_{tot} = E_P + E_S$$

ここで，E_P は分子構造によるエネルギーで以下のように計算される．

$$E_P = E_C + E_{LJ} + E_{HB} + E_{tor}$$

$$E_C = \sum_{(i,j)} \frac{332 q_i q_j}{\epsilon r_i j}$$

$$E_{LJ} = \sum_{(i,j)} \left(\frac{A_{ij}}{\gamma_{ij}^{12}} - \frac{B_{ij}}{\gamma_{ij}^{6}}\right)$$

$$E_{HB} = \sum_{(i,j)} \left(\frac{C_{ij}}{\gamma_{ij}^{1}2} - \frac{D_{ij}}{\gamma_{ij}^{1}0}\right)$$

$$E_{tor} = \sum_{i} U_i(a \pm \cos(n_i \chi^i))$$

ここで，E_C は電気エネルギー項，E_LJ は Lennard-Jones 項，E_HB は水素結合項項，E_tor は回転エネルギー項である．また，E_S は溶媒によるエネルギーである．これらの計算は ECEPP/3 [3, 4] のエネルギー計算で用いられており，本稿における蛋白質構造予測に関しても，このエネルギー評価関数を用いて解を求める．

これまでに提案した本手法を用いて蛋白質の立体構造予測を行う．本稿では Met-enkephalin と C-peptide の 2 種類の蛋白質の構造予測を行う．Met-enkephalin は 5 個のアミノ残基と 19 個の二面角を持つ蛋白質であり，C-peptide は 13 個のアミノ残基と 64 個の二面角を持つ．それぞれの蛋白質のアミノ酸配列については表 1 である．本稿では各二面角を設計変数として扱い，$[-180, 180]$ である各二面角を 12 ビットで符号化して扱うものとし，実験環境は共有メモリ型並列計算機の SGI Onyx300(MIPS R14000/ 500MHz × 32CPUs / 16GB 共有メモリ) である．

Met-enkephalin は X 線結晶構造解析や数値実験などにおけるこれまでの研究結果から，そのエネルギー値が $E \leq -11 kcal/mol$ となることが知られている．本稿でも実験結果がこの値よりも小さくなることが期

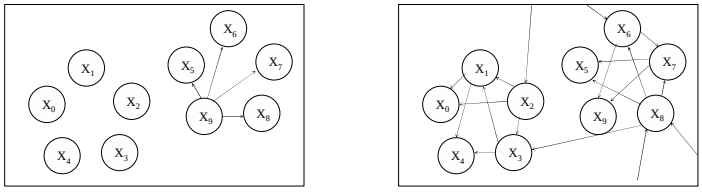

図 4 9 世代目に得られた確率ネットワーク．提案手法 (左) と Ocenasek による従来手法 (右)

表 1 各蛋白質のアミノ酸配列

蛋白質	アミノ酸配列	二面角数
Met-enkephalin	Tyr-Gly-Gly-Phe-Met	19
C-peptide	Lys-Glu-Thr-Ala-Ala-Ala-Lys-Phe-Glu-Arg-Gln-His-Met	64

表 2 実験によって得られたエネルギー値

蛋白質名	得られたエネルギー値	最適とされる領域
Met-enkephalin	-12.229kcal/mol	$\leq$ -11kcal/mol
C-peptide	-45.127kcal/mol	$\leq$ -42kcal/mol

待される．また，C-peptide については，岡本らによる数値実験 [10] においてエネルギー値が $-42kcal/mol$ 以下となることが示されており，岡本らによって得られた構造は X 線結晶構造解析による結果と類似していることが示されている．そのため，C-peptide に関する本実験の結果がそれよりも小さくなることが目標となる．

これまでに説明してきた確率ネットワークに並列構築を適用した BOA をそれぞれの蛋白質に適用した結果のエネルギー値は表 2 である．ここで，集団サイズは Met-enkephalin に対して 6000，C-peptide に対して 12000 を用いて実行した．また，BOA における各ノードに対する辺の最大入力数を $k = 8$ と設定し，用いたプロセッサ数は 4 台として実行した．

上述したように，これまでの研究において示されている最適領域に本稿の結果も含まれていることがわかる．また，実験はそれぞれ 5 回ずつ行ったが，すべての実験で最適領域に含まれる結果となった．また，評価回数は Met-enkephalin では 30 万回程度，C-peptide では 70 万回程度となった．廣安らによる研究では，2 個体分散遺伝的アルゴリズムによる蛋白質立体構造予測を行っており，Met-enkephalin における計算コストは 190 万回である [2]．このことから，本研究における手法が評価計算コストの面で優れていると言える．しかし，廣安らの手法では分布推定は行わないため，実時間の観点間から考察することは今後の課題となるだろう．

また，この結果を可視化ソフト Rasmol で示したものが図 5 である．それぞれの実験によって得られた構造は X 線結晶構造解析による結果と類似していることが言える．

5. おわりに

本稿では，これまでに提案してきた確率ネットワーク並列構築を適用した BOA について，実際の問題として蛋白質構造予測問題を扱い，その性能について検証してきた．結果から言えるように，このような問題においても適切に解を得ることができており，また問題サイズを大きくした場合においても解は最適な領域に含まれていることがわかる．今後の課題としては，より大きな問題サイズの蛋白質についての検証と他の最適化手法との比較をより詳細に行うことが挙げられる．

参 考 文 献

1) 三木光範，廣安知之，小椋信弥，青井桂子，吉田武史，岡本祐幸. 遺伝的交叉を用いた並列 SA によるタンパク質立体構造のエネルギー最小化, 情報処理学会 SACSIS‘03 講演論文集, pp.341-348/
2) 廣安 知之，三木光範，岩橋 崇史，岡本祐幸. 2 個体分散遺伝的アルゴリズムによるタンパク質立体構造のエネルギー最小化, 計測自動制御学会 SICE Annual Conference 2003.
3) F. Eisenmenger and U. H. E. Hansmann, “ Global Minimum Configuration of a Small Peptide for the ECEPP/2 and ECEPP/3 Force Field, ” Chem. Phys. Lett. 286, 86-92 (1997).
4) F. Eisenmenger and U. H. E. Hansmann, “ Variation of the Energy Landscape of a Small Peptide under a Change from the ECEPP/2

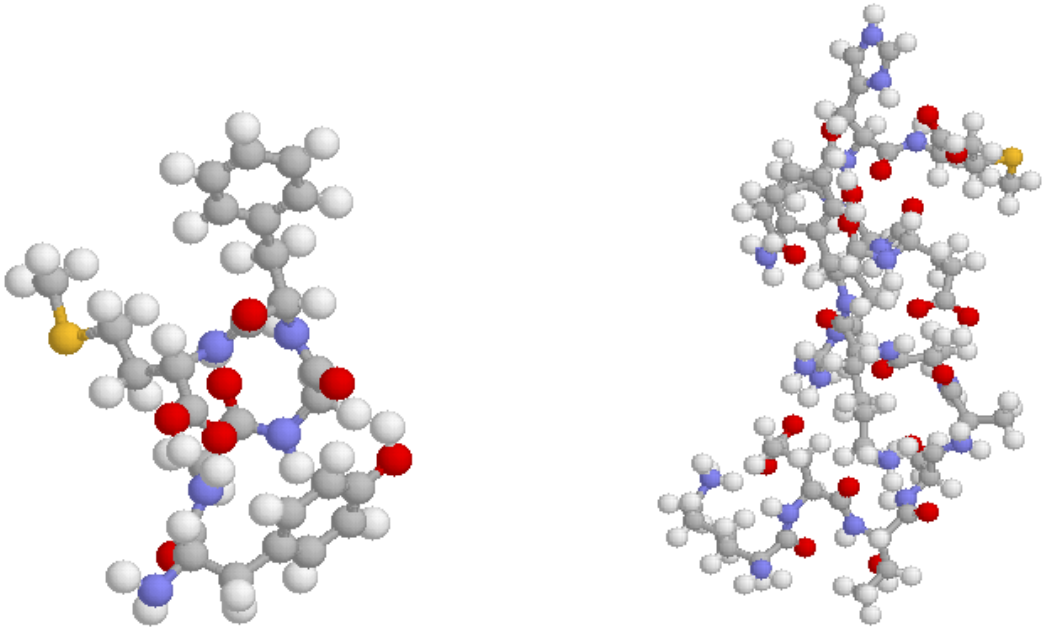

図 5 蛋白質構造予測結果. Met-enkephalin(左) と C-peptide(右)

Force Field to ECEPP/3, ” J. Phys. Chem. B 101, 3304-3310 (1997).

5) David Heckerman. A Tutorial on Learning With Bayesian Networks. Microsoft Research, MSR-TR-95-06, 1995.

6) David Heckerman and Christopher Meek. A Bayesian Approach to Learning Bayesian Networks. Microsoft Research, MSR-TR-97-07, 1997.

7) Erick Cantú-Paz. Designing Efficient and Accurate Parallel Genetic Algorithms. PhD thesis. University of Illinois at Urbana-Champaign, 1999.

8) 村尾直哉, 棟朝雅晴, 赤間清. Bayesian Optimization Algorithm におけるネットワーク並列構築手法の検討, 情報処理学会 MPS シンポジウム論文集（複雑系科学シンポジウム 2004）, pp.271-276 (2004)

9) Ocenasek Jiri. Parallel Estimation of Distribution Algorithms. PhD. Thesis, Faculty of Information Technology, Brno University of Technology, Brno, Czech Rep., 2002, pp. 1-154.

10) Yuko Okamoto, Takeshi Kikuchi, and Hikaru Kawai. Prediction of Low-Energy Structures of Met-Enkephalin by Monte Carlo Simulated Annealing. CHEMISTRY LETTERS, pp. 1275.1278, 1992.

11) Yuko Okamoto. Protein Folding Problem as Studied by Monte Carlo Simulated Annealing. Physical Properties Research, Vol. 70, No. 6, pp. 719.742,1998.

12) Martin Pelikan, David E. Goldberg and Erick Cantú-Paz : BOA:The Bayesian optimization algorithm. Technical Report IlliGAL Report No.99003, University of Illinois at Urbana-Champaign, Urbana, IL, 1999.

13) Yang Xiang and Tongsheng Chu. Parallel learning of belief networks in large and difficult domains. Data Mining and Knowledge Discovery, 3(3) : 315–339, 1999.