Vol.2011-BIO-24 No.6
2011/3/10

生体分子ネットワークレイアウトのための 2 次元近似照合によるハイブリッドレイアウトアルゴリズム

井上健太郎† 下園真一†† 吉田英聡†† 倉田博之†

生体分子ネットワークマップを可視化するためには、生体分子ノードをレイアウトする計算方法が必要である。ネットワーク構造を詳細に理解するために、ノードのラベル領域を確保した高速なレイアウト法が望まれる。本研究では、従来の高速レイアウト法と 2 次元近似照合を組み合わせて、グリッド上にネットワークを配置する新規なハイブリッドレイアウトアルゴリズムを提案し、従来法との比較を行った。

Hybrid Layout Algorithm using Approximate Pattern Matching in Two Dimensional Space for Biochemical Network Layout

Kentaro Inoue†, Shinichi Shimozono††, Hideaki Yoshida††, Hiroyuki Kurata†

To visualize a biochemical network map, the coordinates of all the molecular nodes need to be fast calculated. To illustrate biological details of the network structures, it is critically important to ensure the space where the node labels are clearly attached. In this study, we propose the hybrid layout algorithm that combines traditional fast layout methods with the two dimensional approximate pattern matching, which fast places all the nodes on square grids, while capturing their topological features.

1. はじめに

分子生物学の発展により、遺伝子制御ネットワーク、シグナル伝達経路、代謝回路などの詳細なマップが明らかになっている[1]。大規模で複雑な生体分子ネットワークをコンピュータ上で視覚的に表現することは、生命科学研究者がネットワーク構造を理解するのに役立つ。コンピュータ上でネットワークマップを構築し、生命システムの解析を行うために、多くのソフトウェアが開発されている。私たちは生体分子ネットワークの構築からダイナミックシミュレーションまでを実行する CADLIVE (Computer-Aided Design of LIVing systEms)システムを開発している[2]。

生体分子ネットワークは大規模かつ複雑なため、手動でネットワーク構造を理解しやすくレイアウトするには時間と労力が必要である。そこで、生体分子をノード、反応をエッジとしてグラフ化し、自動レイアウトを行うアルゴリズムが開発されている。ノードのラベル情報を見やすくするため、ノードラベル間の重なりをなくす必要がある。この問題は、グリッド上にノードを配置してラベルスペースを確保することによって、ラベル間の重なりを回避することができる。私たちは、2005 年に生体分子ネットワークをノードの重複なくレイアウトを行うグリッドレイアウトアルゴリズムを開発した[3]。このアルゴリズムは、目的関数を定義して、それが最小になるようにレイアウトの探索を行う。その際、描画領域を設定し、その領域のすべてのグリッドにおいて、シミュレーテッド・アニーリングによって最適なレイアウトを探索する。レイアウトされた結果は、コンパクトかつクラスタ構造を捉えている。しかし、ネットワークが大規模になるほど、描画に必要なグリッド領域が増えるため、計算時間が膨大になるという問題点があった。この問題を解決するために、改良法として Lucid Draw が開発された[4]。Lucid Draw は、探索領域を狭くすることで計算速度は速くなったが、未だ大規模なネットワークには計算時間がかかる。

本研究では、ノードラベルスペースは考慮しないが、一般的に広く使われており、高速なレイアウト法であるスペクトル法[5]、バネモデルのレイアウト法[6,7]、自己組織化マップを利用したレイアウト法[8]を前処理として利用し、2 次元近似照合によって、グリッド上にノードを配置するハイブリッドレイアウトアルゴリズムを提案する。このアルゴリズムは、前処理アルゴリズムによるレイアウトの位置関係を保ったままグリッド上にレイアウトを行うため、ノードの重複がなく、ノードラベルスペースを確保したレイアウトが実現できる。

† 九州工業大学生命情報工学分野
Department of Bioscience and Bioinformatics, Kyushu Institute of Technology
†† 九州工業大学知能情報工学分野
Depart ment of Artificial Intelligence, Kyushu Institute of Technology

2. 方法

ハイブリッドレイアウトアルゴリズムは、２つのステップで実行される。まず、最初に従来の高速レイアウト法を使って、粗いレイアウト(前処理)を行う。ここでは、スペクトル法[5]、Kamada-Kawai[6]、Fructcharman-Reingold[7]、Gürsoy-Atun[8]を用いた。次に、前処理レイアウトされた各ノードから最も近い格子点に移動させる 2 次元近似照合を行う。

ハイブリッドレイアウトアルゴリズムの評価を行うために、計算速度、エッジの交差、描画領域に対するエッジの長さ、F 値について評価を行った。また、ハイブリッドレイアウトアルゴリズムを比較検証するために、ランダムレイアウトと私たちが以前に開発したグリッドレイアウトアルゴリズム[3]、Lucid Draw[4]を用いた。

2.1 ハイブリッドレイアウトアルゴリズム

2.1.1 前処理

2.1.1.1 スペクトル法

スペクトル法(SA)はグラフスペクトルを利用して、固有値が小さい２つの固有ベクトルを用いてレイアウトを行う。A を隣接行列、D をノードの次数の対角行列としたとき、グラフラプラシアン L を求め、L の固有関数を求める。

$$L = I - D^{-1/2} A D^{-1/2} \quad \cdots (1)$$

L の固有値 λ_i、固有ベクトル $\mathbf{v}_i$ としたとき、$0 = \lambda_1 \le \lambda_2 \le \ldots \le \lambda_n$ と並び替えたときの $\mathbf{v}_2$,$\mathbf{v}_3$ を座標軸として、各ノードに対応する $\mathbf{v}_2$,$\mathbf{v}_3$ の値をノードの座標とする。$\mathbf{v}_1$ は固有値 λ_1=0 となることから、固有ベクトルは不定となるので、ネットワーク構造の理解のためのレイアウトには役に立たない。SA はネットワークの密な領域のノード群が近くに配置されるようにレイアウトされる。

2.1.1.2 Kamada-Kawai

Kamada-Kawai (KK) はすべてのノード間がバネにより結合されるものと考え、バネに働く力と各バネの自然長を定義し、すべてのノードを円周上に並べた初期配置から各ノードを移動させることで全体のエネルギーの総計を小さくするアルゴリズムである。移動させるノードは Newton-Raphson 法を用いて、エネルギーが最小になるように移動させる。このアルゴリズムでは、グラフ理論的な距離が近いノードほど近くに配置されたレイアウトを得る。

2.1.1.3 Fructcharman-Reingold

Fructcharman-Reingold (FR) は KK と同様にバネモデルによる手法であるが、エッジで接続されたノード同士はノード間の距離の 2 乗に比例する引力を受け、すべてのノード同士はノード間の距離に反比例する斥力を受けるように定義する。さらに、温度の概念とその冷却過程を導入する (シミュレーテッド・アニーリング)。つまり、レイアウト計算の初期の段階では系全体の温度が高く、各要素が大きなエネルギーを持つ状態と考えて各ノードを大きく移動させ、計算が進むにつれて系全体の温度を低く、つまり各ノードの移動量を小さくする。これは一度の計算におけるノードの移動距離を制限し、計算を繰り返すごとにその範囲を減らしている。

2.1.1.4 Gürsoy-Atun

Gürsoy-Atun(GA)は自己組織化マップ(SOM)に基づいた手法で、レイアウトされた領域で次数の分布がほぼ等しくなるように、ノードを分散して配置する。

2.1.2 2 次元近似照合

2 次元近似照合のフローチャートを示す(図１)。最初に、前処理によって配置されたレイアウトを正の座標に平行移動させる。そして、ラベルに必要な領域を確保するため、グリッド領域を設定し、その領域に前処理で得られたレイアウトの相対的なノードの座標配置に拡大する。次に、各ノードに一番近い格子点に移動させる。このとき、もし格子点にノードが存在していれば、次に近い格子点に移動させる。この移動距離の総和が最小となる移動の組み合わせを動的計画法により探索する。動的計画法は、各ノードを移動させるパターンすべてを計算するため、膨大な計算量が必要となる。そこで、近似照合を行うノード群を分割して、そのノード群に対して動的計画法を適用する。分割法は Kd 木と 4 分木の２つの方法を用いた。ここでは、動的計画法を行うノード数が設定した数(カットサイズ)以下になるまで分割を繰り返す。

Kd 木は一次元における２分探索木の拡張であり、深さにより比較する成分を切り替える。本研究では深さで x 軸、y 軸の成分に交互に切り替える。まず、x 軸で座標の値を昇順に並び替えたとき、ノード数が半分となるノードを境に２分割する。次に、y 軸について考え、x 軸で２分割された各ノード群に対して、同様にして y 軸でのノード数が半分になるノードを境に２分割する。これを分割されたノード数がカットサイズ以下になるまで再帰的に分割を行う。

4 分木は 2 次元空間を再帰的に 4 等分して分割する。まず、レイアウトされた領域に対して、４等分する。この分割された領域内のノード数がカットサイズ以下であれば、分割を終了し、カットサイズより多いノードが存在すれば、その領域を再度４等分する。

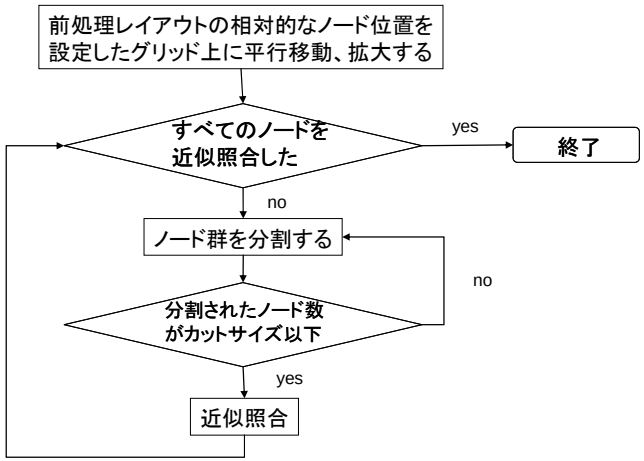

図１：２次元近似照合のフローチャート

2.2 グリッドレイアウト

私たちが開発したグリッドレイアウト(GL)アルゴリズム[2]は、ネットワークトポロジーとレイアウト結果の関係を評価関数で定義して、その評価値を最小化するようにレイアウトする。評価関数は各ノードの評価値の総和で計算される。各ノードの評価値は、ネットワークのノード間の最短経路長を重みに、グリッド上のマンハッタン距離をかけたものと定義される。この評価関数が最小となるレイアウトをシミュレーテッド・アニーリングによって探索する。この手法の特徴は、設定したグリッド上ですべての領域を探索することにより、最適化を図るため、広い範囲で評価値が低いレイアウトを探索することができる。

2.3 Lucid Draw

Lucid Draw (LD)は GL アルゴリズムを改良した方法で、最適化を行う過程が GL と異なる。LD は最適化の過程でノードを移動させる際、移動させるノードをネットワーク的な隣接ノードの隣のグリッドに限定してレイアウトを行う。GL に比べ、探索範囲が狭くなることから、計算量を大幅に削減できる。

2.4 評価法

CPU Core i7 2.8GHz, Memory 4Gbyte の計算機を用いて、計算速度、エッジの交差の割合、描画領域に対するエッジの長さ、F 値について評価を行った。各プログラムの開発は、前処理の SA は Matlab で作成し、KK、FR、GA、ランダムレイアウト(R)については MatlabBGL のプログラムで実行した。2 次元近似照合アルゴリズムは C 言語で作成した。

エッジの交差の割合はエッジの交差数を全エッジの２つの組み合わせで割ったものである。一般的に、エッジの交差が少ないほうが見やすいレイアウトである。

描画領域に対するエッジの長さ(Edge length)は、x 軸と y 軸のそれぞれで最小値、最大値の座標からなる長方形の面積に対して、全エッジの長さの割合を計算したものである。この値が低いことは、隣接するノードが密に配置されたことを示す。

F 値は，隣接するノードを隣接しないノードより相対的に近く配置するという考えに基づいた評価値である[8]。F 値は適合率(Precision)と再現率(Recall)との調和平均で定義される。適合率と再現率は、各ノードを中心に半径を広げたとき、その円内に存在するノードの中における隣接ノードの割合を示したものを適合率とし、隣接ノードのうちその円内にある隣接ノードの割合を示したものを再現率とする。これらの調和平均を計算して、ノード全体の平均を取ったものを F 値とする。

$$F_i(r_i) = 1 \Big/ \left\{ a\frac{1}{P_i(r_i)} + (1-a)\frac{1}{R_i(r_i)} \right\} \quad , \quad F = \sum_{i=1}^{N} \frac{F(r_i)}{N} \qquad \cdots (2)$$

P_i はノード i における Precision の値を示し、R_i はノード i における Recall の値を示す。r_i はノード i における P_i、R_i が最大となる半径を示す。N は総ノード数である。

2.5 検証モデル

KEGG [1]から大腸菌の代謝ネットワークを CADLIVE Converter[10]を用いて、CADLIVE 形式に変換し、解糖系や TCA 回路などを組み合わせて、９つのネットワークを作成した。

3. 結果・考察

2 次元近似照合のカットサイズに対する動的計画法の計算時間を測定した(図２)。近似照合するノード数の増加に伴い、計算時間が膨大になる。今回、近似照合の計算時間が 1 秒以内に収まるように、近似照合を行うカットサイズを 10 とした。また、２次元近似照合を行う際の初期配置のグリッド領域はネットワークのノード数で設定した。

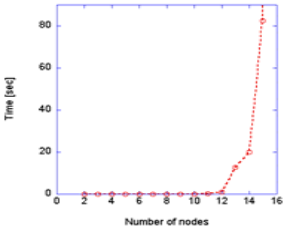

図２：動的計画法による計算時間

Vol.2011-BIO-24 No.6
2011/3/10

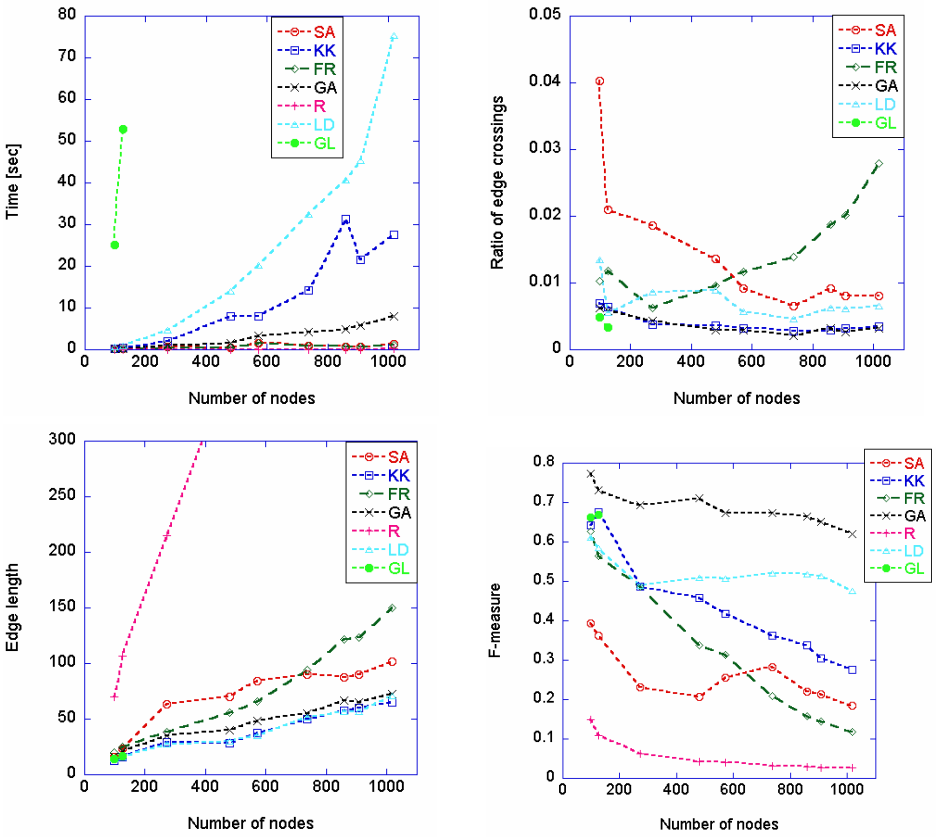

図３：Kd 木を用いたハイブリッドレイアウトとランダムレイアウト、従来のグリッドレイアウトの評価
SA:スペクトル法、KK:Kamada-Kawai、FR:Fructcharman-Reingold、
GA: Gürsoy-Atun、R:ランダムレイアウト、LD:Lucid Draw、GL:グリッドレイアウト。
各評価値は計算速度(左上)、エッジの交差の割合(右上)、描画領域に対するエッジの長さ(左下)、F 値(右下)である。R のエッジの交差の割合は約 0.23 である。

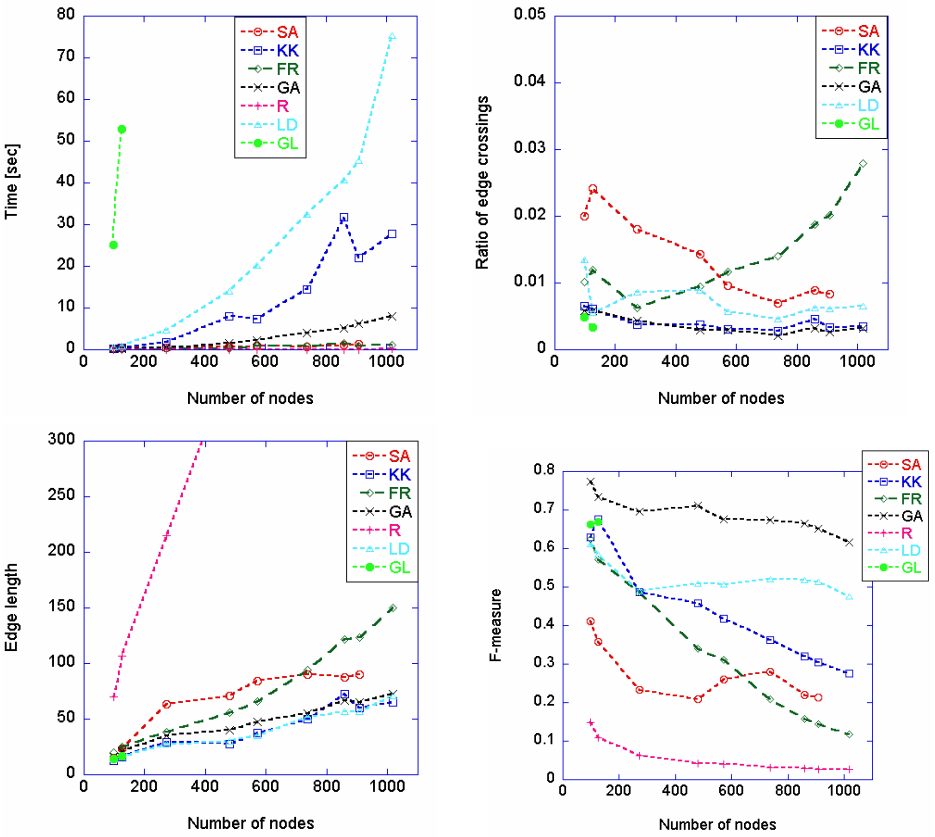

図４：４分木を用いたハイブリッドレイアウトとランダムレイアウト、従来のグリッドレイアウトの評価
ラベルの詳細は図３と同様。

Vol.2011-BIO-24 No.6
2011/3/10

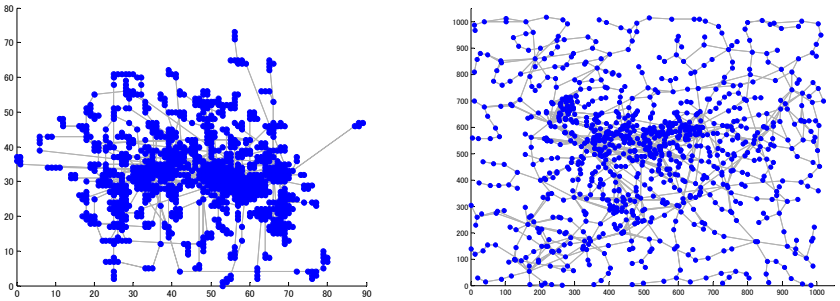

図５:LD(左)と Kd 木を用いた GA のハイブリッドレイアウト(右)のマップ
ノード数 1019、エッジ数 1463 の代謝ネットワーク。

Kd 木と４分木を用いた､それぞれのハイブリッドレイアウト､ランダムレイアウト、従来のグリッドレイアウトの評価値の比較を行った(図３、図４)。

４分木によるノード群の分割は、面積に応じて分割を繰り返すため、ある座標付近にノードが多く配置されると、分割ができなくなる。今回、SA ではノード数が 1000 程度のとき､ある座標付近にノードが密集したため､ノード群の分割ができなかった。これはグリッド領域を広くすることで分割が可能になる。

計算時間について、GL はノード数が 100 程度までは 1 分以内で計算できるが、ノード数が増えるにつれ、膨大な時間がかかる。LD は GL と比べると高速であるが、ノード数が 1000 を超えると 1 分以上かかる｡ハイブリッドレイアウトは 2 次元近似照合が 1 秒以下で計算され､ほとんどが前処理の計算時間であった｡KK の場合を除いて、ハイブリッドレイアウトはノード数が 1000 程度であっても、数秒で計算が終わる。

エッジの交差の割合について、GL は他の手法と比べ、最もエッジの交差の割合が少なかった。KK と GA は LD と比べるとエッジの交差の割合が少ない。

描画領域に対するエッジの長さについて、GL、LD、KK、GA はほぼ同じ値であった。これらの手法は描画領域に対して、隣接ノードがコンパクトに配置されている。

F 値について、GL は LD と比べると高い値であるが、GA は GL や LD より高い値を示した。GA は、隣接ノードと隣接していないノードの関係が明確に分けられてレイアウトされている。

今回、用いた前処理の中では、GA が LD より高速に計算され、またエッジの交差は少なく、隣接ノードの関係を示す Edge length や F 値についてよい結果となった。このことから、ハイブリッドレイアウトアルゴリズムは GA を前処理として用いることにより、従来のグリッドレイアウト法より良い評価結果が得られる。

4. おわりに

現在、さまざまな生体分子ネットワークを可視化するためのソフトウェアが開発され、これらは生命システムの理解に役立っている。膨大な規模の生体分子ネットワークの理解を助けるために、大規模なネットワークのレイアウト技術は重要である。本研究で提案したハイブリッドレイアウトアルゴリズムは、従来の高速なレイアウト法に 2 次元近似照合を用いたことに新規性がある。2 次元近似照合を用いることで、生体分子ノードのラベル領域を確保したグリッドレイアウトを行うことができ、ノードの重なりがないレイアウトを実現することができた。今回用いた前処理のレイアウト法の中では、GA が最も良い結果を示した。GA は従来のグリッドレイアウト法である GL や LD より評価がよく、また高速な計算法であった。

本研究は 2 次元近似照合との組み合わせにより、ノードのラベル領域を確保できるグリッドレイアウトが実現する。今回用いた評価値は、エッジの交差や隣接ノードの位置関係のみに注目した。ネットワーク構造を視覚的に理解するためには、ネットワーク中に存在する機能モジュールが一目でわかるようになることが理想的である。今後、こういったクラスタ構造の定義を行い、レイアウト評価の指標として提案し、ネットワークの性質による最適なグリッドレイアウト法が実現されることを試みる。

参考文献

1) M. Kanehisa, S. Goto, S. Kawashima, A. Nakaya, The KEGG databases at GenomeNet, *Nucleic Acids Res.*, 30(1), 42-46, 2002
2) H. Kurata, K. Inoue, K. Maeda, K. Masaki, Y. Shimokawa, Q. Zhao, Extended CADLIVE: a novel graphical notation for design of biochemical network maps and computational pathway analysis, *Nucleic Acids Res.*, 35(20):e134, 2007
3) S. He, J. Mei, G. Shi, Z. Wang, W. Li, LucidDraw: efficiently visualizing complex biochemical networks within MATLAB**,** *BMC Bioinformatics*, **11**:31, 2010
4) W. Li and H. Kurata, A grid layout algorithm for automatic drawing of biochemical networks, *Bioinformatics*, 21, 2036-2042, 2005
5) Y. Koren, On spectral graph drawing, *Lect. Notes Comput. Sci.*, 2697, 496-508, 2003
6) T. Kamada and S. Kawai, An algorithm for drawing general undirected graphs, *Information Processing Letters*, 31, 7-15, 1989

7) T. M. J. Fructcharman and E. M. Reingold, Graph drawing by force-directed placement, *Software-Practice & Experience*, 21 (11), 1129-1164, 1991
8) A. Gursoy and M. Atun, Neighborhood Preserving Load Balancing: A Self-Organizing Approach, *LNCS,* 1900, 324-341, 2000
9) T. Yamada, K. Saito, N. Ueda, Cross-Entropy Directed Embedding of Network Data, *Proceedings of the Twentieth International Conference on Machine Learning* (*ICML-2003*), 832-839, 2003
10) K. Inoue, K. Maeda, Y. Kato, H. Kurata, CADLIVE: A platform for network modeling and simulation of biological pathway, *The Proceeings of the 2010 Annual Conference of the Japanese Society for Bioinformatics*, P06, 2010