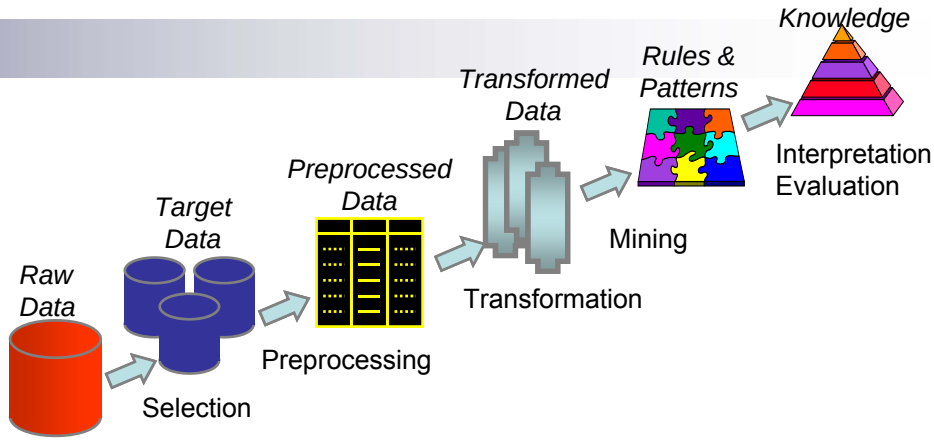
グラフ系列マイニング

猪口 明博
大阪大学 産業科学研究所
科学技術振興機構 さきがけ

研究の背景



- データマイニング
  - インフラ技術の高度化
    - 多様で大規模な情報やデータへのアクセス，蓄積が容易．
  - 多様で大規模なデータから有用な知識を発掘することは重要な課題．
- 頻出アイテム集合マイニング [Agrawal 94]
  - 頻出アイテム集合列挙問題
    - 一般に多くの事例を説明する知識は有用である．
  - バスケット分析
    - 例) スーパーマーケットのデータベースからよく売れる商品の組み合わせを高速に抽出する．
    - データベースはアイテム(商品)の集合からなる．

顧客1={食料品a,食料品b,日用品b,...}

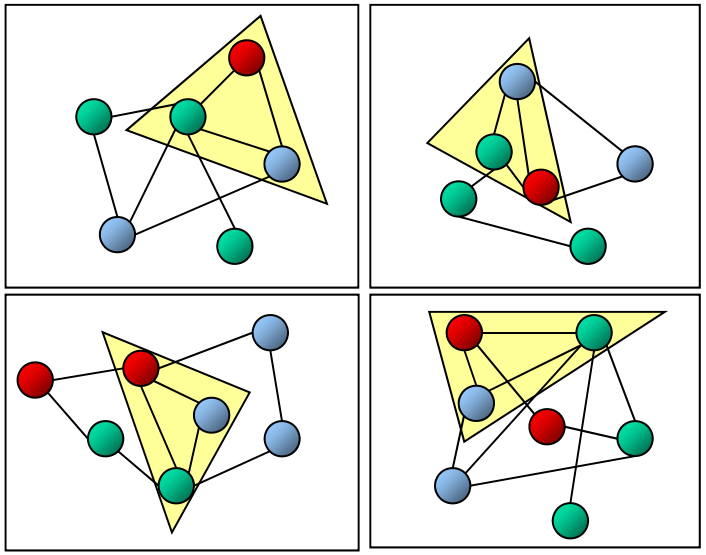
●●●

顧客n={食料品a,飲料水a,日用品b,...}

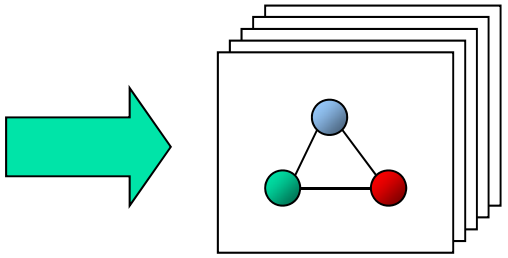
σ人以上の人が購入した
商品の組み合わせを全て列挙

{食料品a,飲料水a},
{食料品a,飲料水a,日用品b},
...

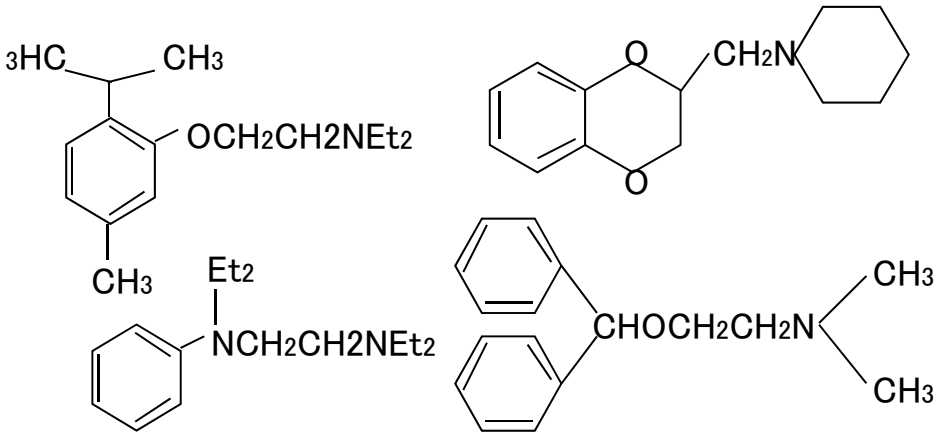
頻出部分グラフマイニング

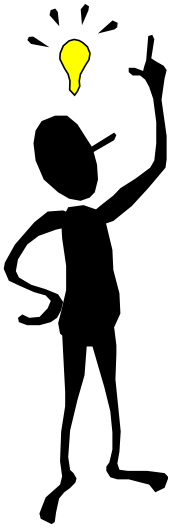


問題：
σ個以上のグラフに含まれる
全ての部分グラフを全て列挙

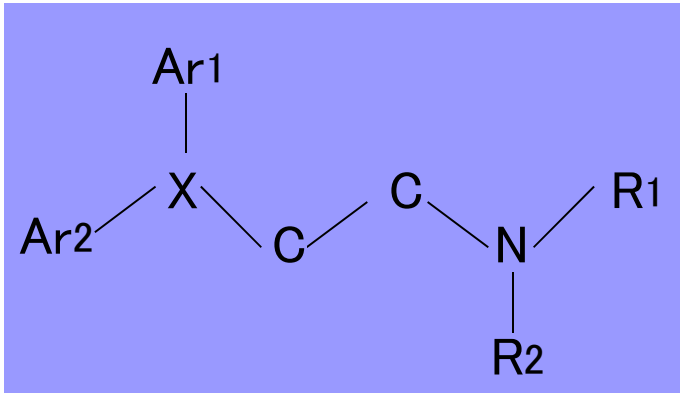


- 応用例
  - 抗ヒスタミン薬の共通パターン発見





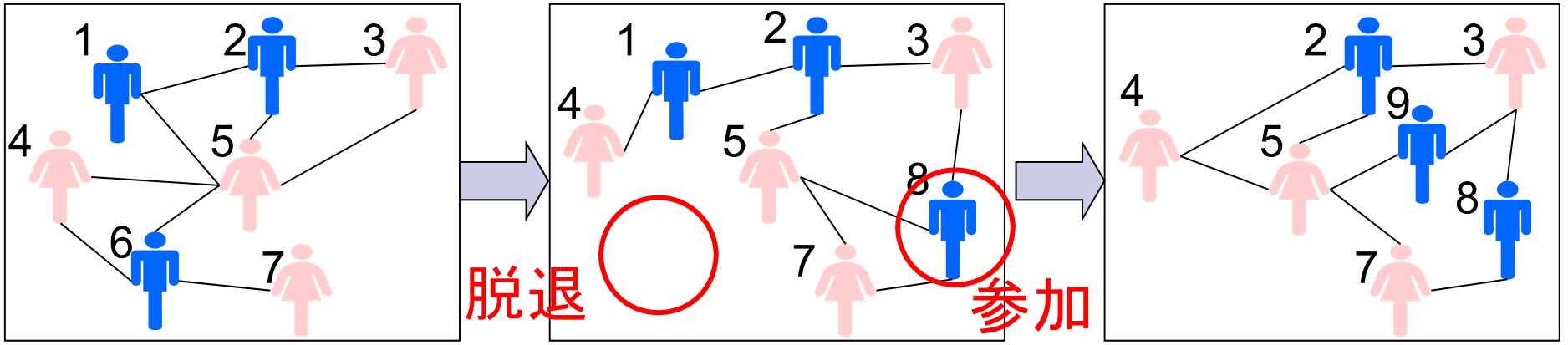
共通パターン



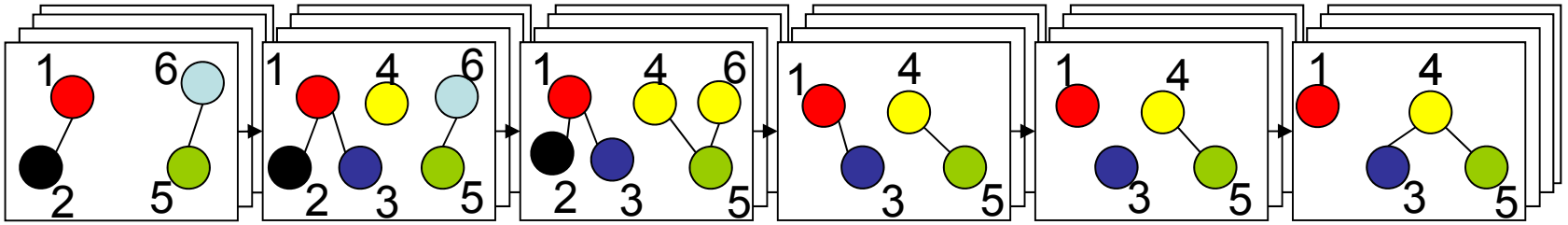
抗ヒスタミン薬の一般的な構造(母核)

グラフ系列の例

- ホームページのリンク構造の変化
  - HTML文章：頂点，ハイパーリンク：辺
- 人間関係ネットワークの変化
  - 人：頂点，人間関係：辺
- 遺伝子ネットワークの変化(進化)
  - 遺伝子：頂点，相互作用：辺
- 機械の組み立て
  - 部品：頂点，隣接する部品間：辺
- その他．．．

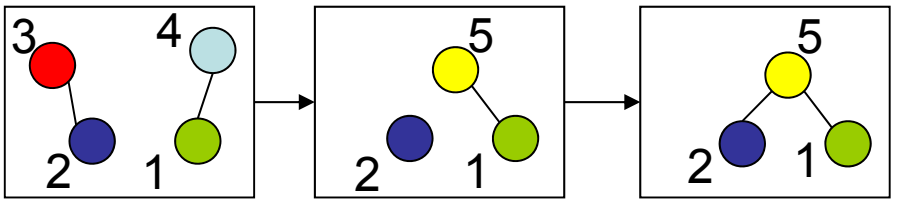


グラフ系列のマイニング



頻出する
部分系列を
マイニング

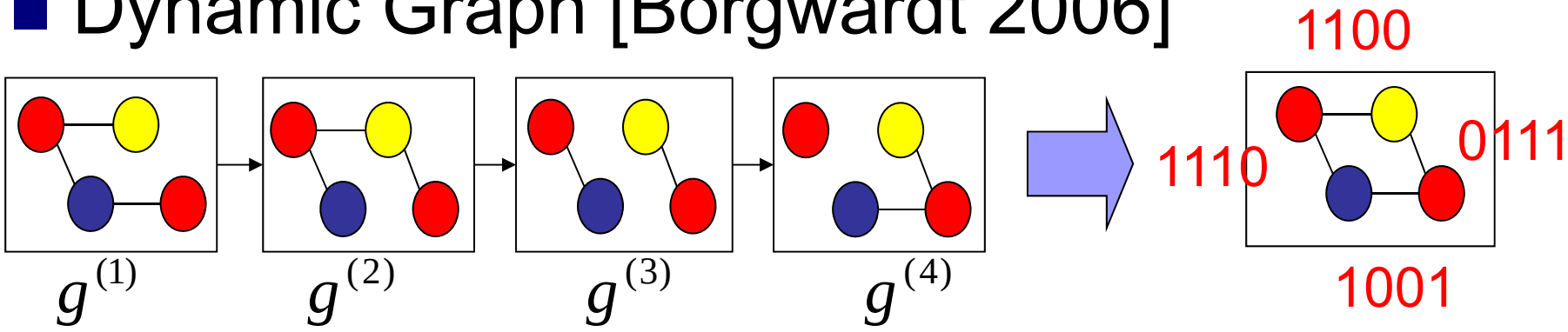
頻出変換部分系列
FTS (Frequent Transformation Subsequence)



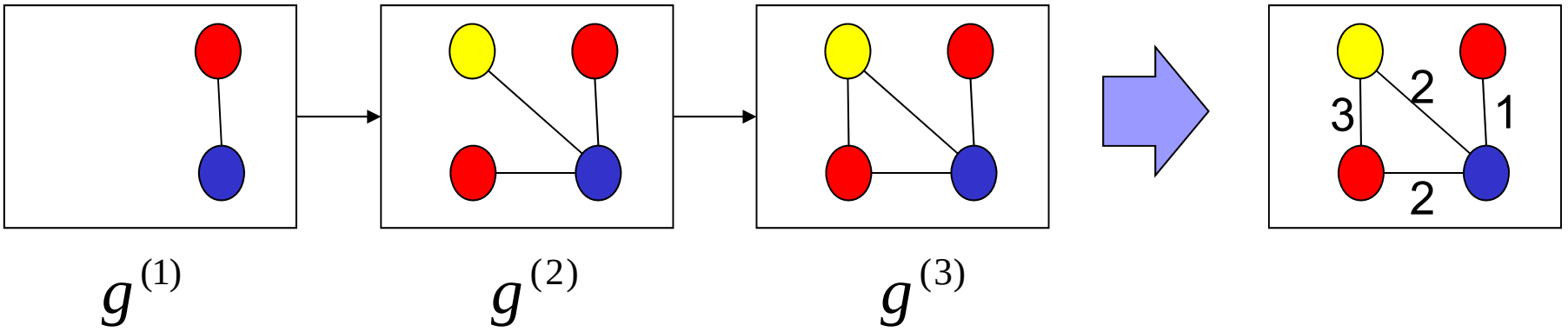
- グラフ系列
  - 頂点数，辺数が増減する．
  - 頂点ラベル，辺ラベルが変化する．
- 仮定
  - 各頂点は，頂点IDをもつ．
  - グラフ系列中の連続する2つのグラフの間では，構造が大きく変化することはなく，ごく一部の構造のみが変化する．
  - 系列中のグラフは，疎グラフである．

関連研究

- Dynamic Graph [Borgwardt 2006]

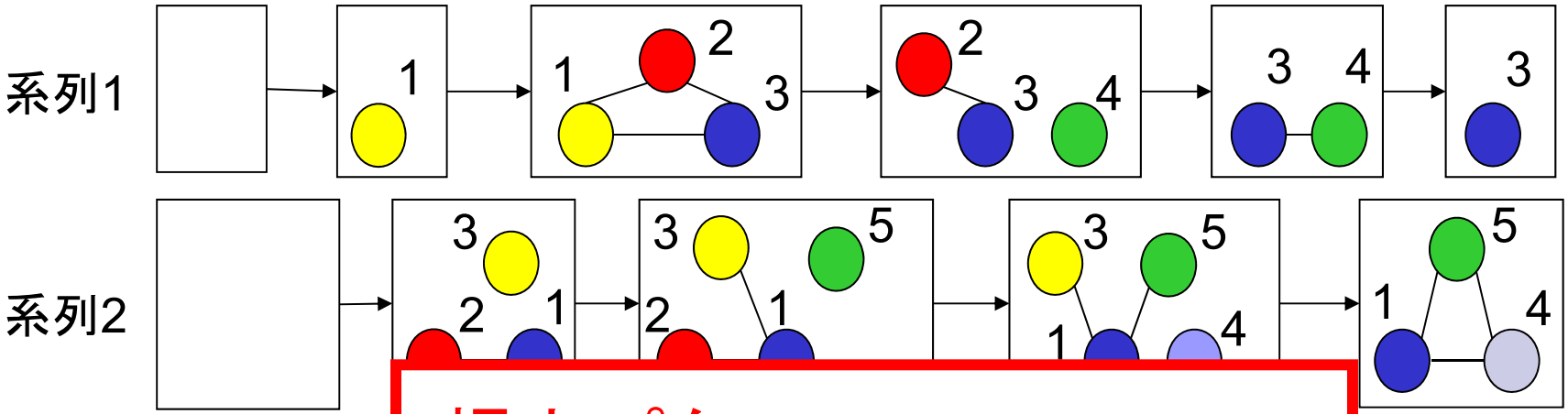


- Evolving Graph [Berlingerio 2009]



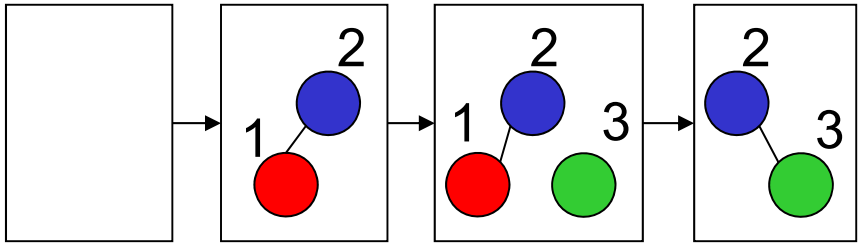
頂点数が増減するグラフやラベルが変化するグラフを扱うことができない.

GTRACEの基本アイデア [Inokuchi 2008]



〈((vi),(vi,vi,e
〈(vi,vi,vi
d),(ed,vd)〉
ed,vd)〉

頻出パターン



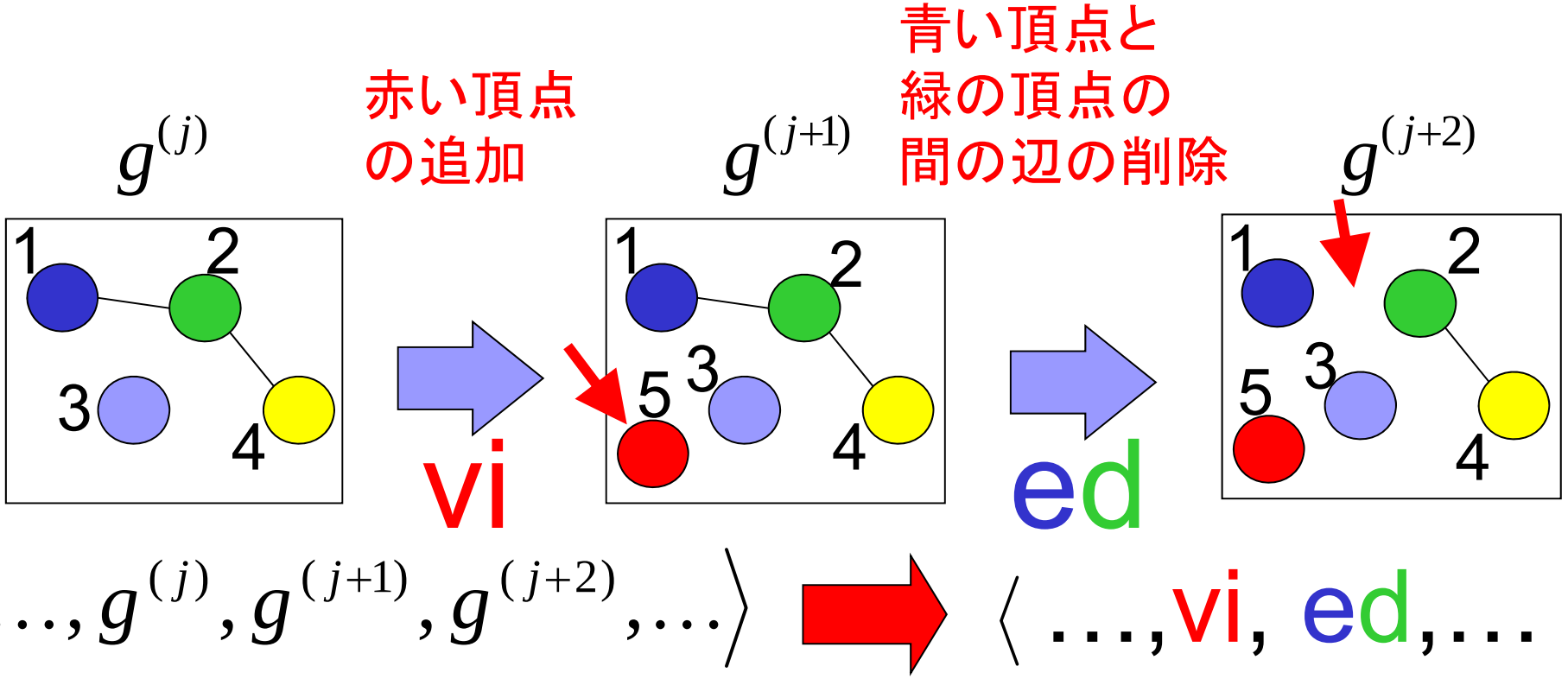
系列パターンマイニング

頻出部分系列 (FTS)

〈(vi,vi,ei),vi,(ei,ed,vd)〉

グラフの変換

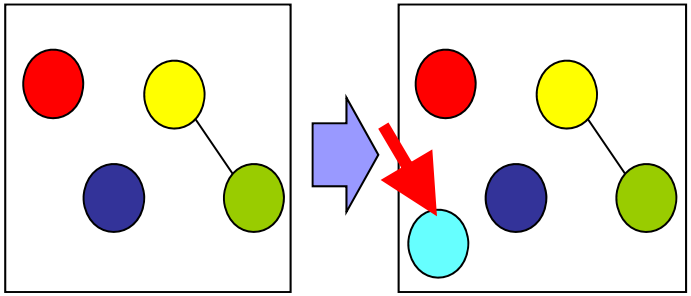
- 頂点や辺の追加，削除，ラベル変更をグラフの変化．



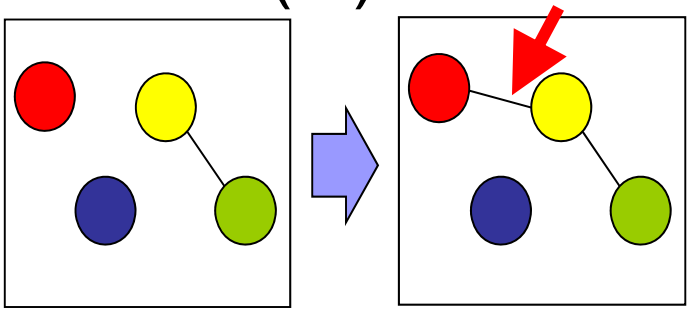
グラフの変化をアイテム集合（変換規則の集合）の系列に変換後，系列パターンマイニングアルゴリズムを適用し，FTSを列挙する．

6種の変換規則

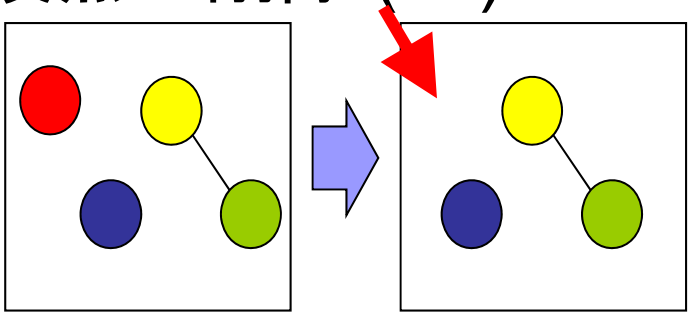
頂点の追加 (vi)



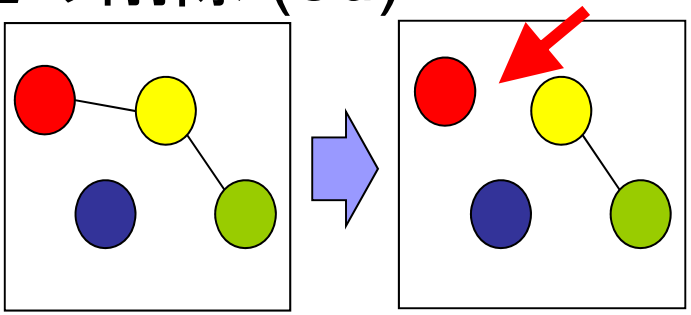
辺の追加 (ei)



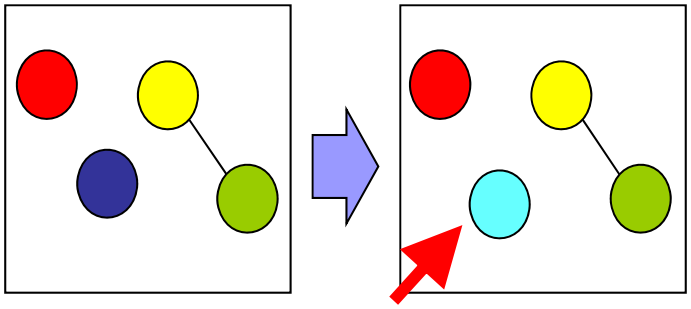
頂点の削除 (vd)



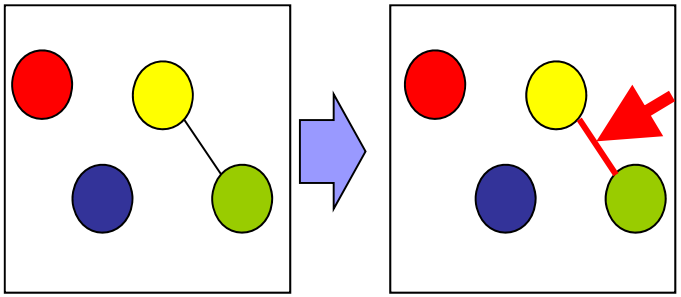
辺の削除 (ed)



頂点ラベルの変更 (vr)

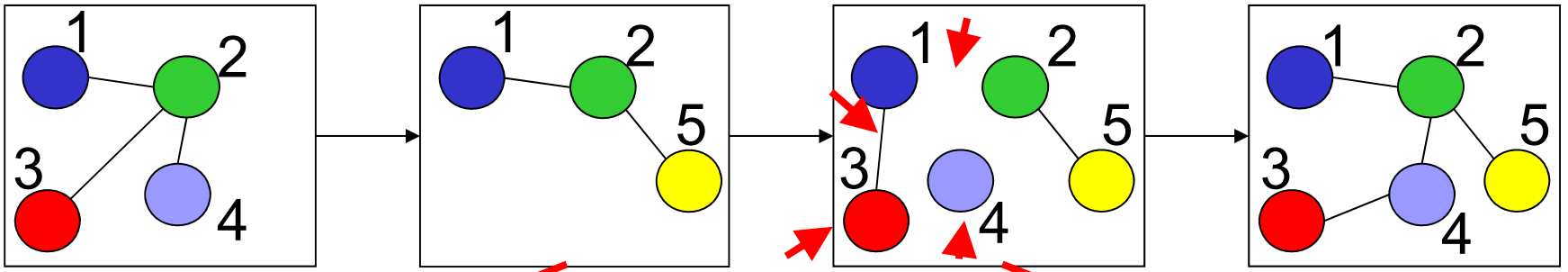


辺ラベルの変更 (er)

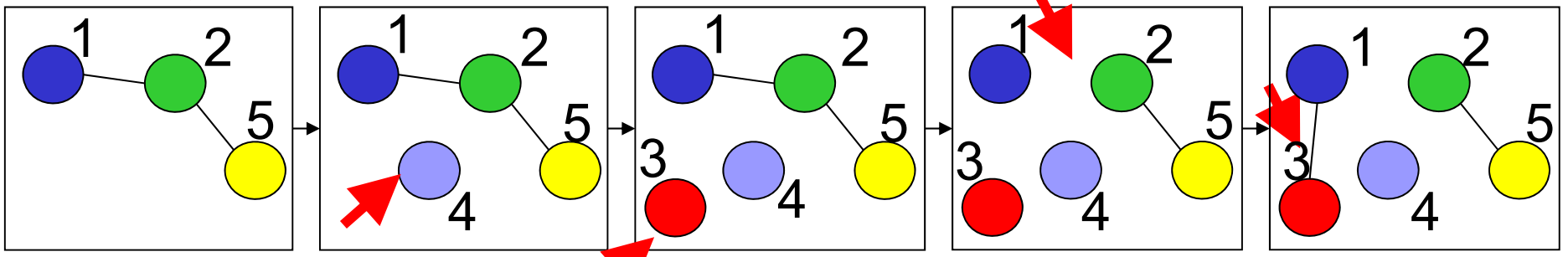


グラフ系列の補間

観測されたグラフ系列



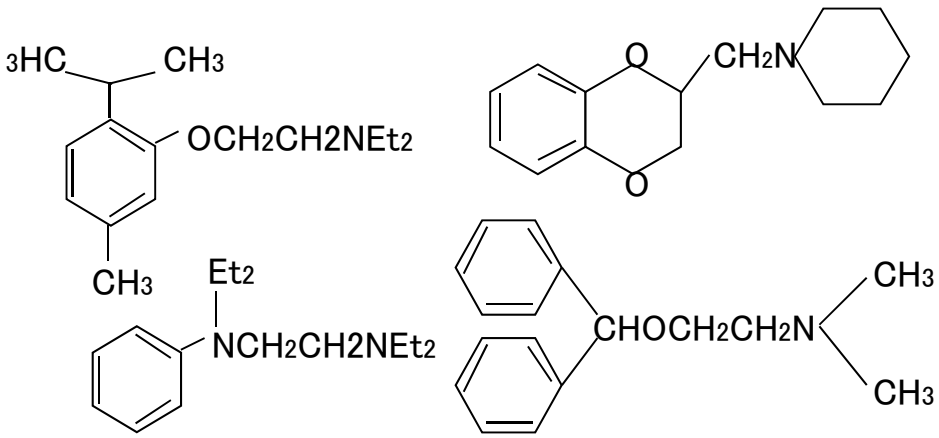
補間されたグラフ系列



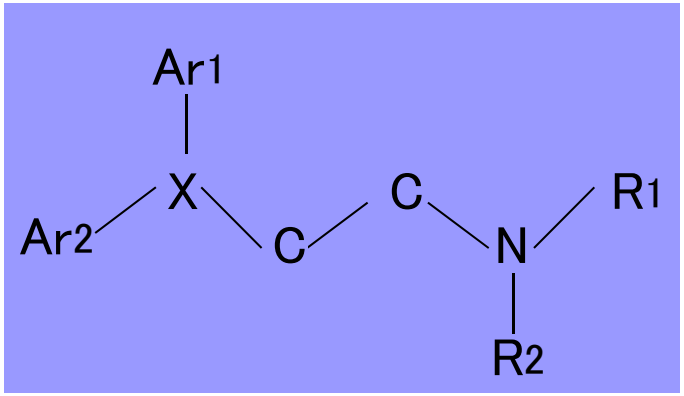
〈…,(vi, vi, ed, ei),…〉

応用例

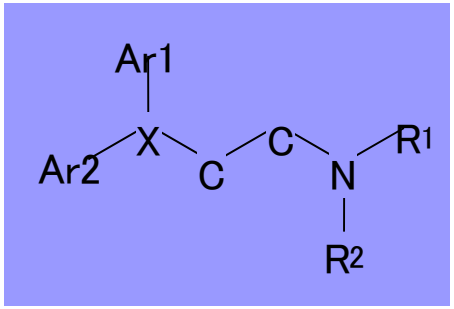
抗ヒスタミン薬の共通パターン発見



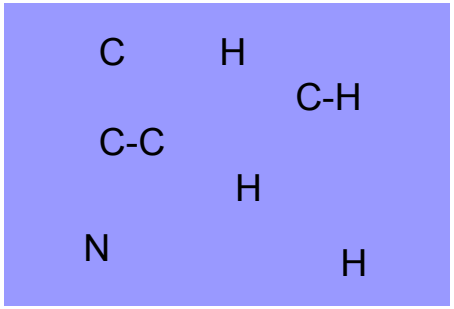
共通パターン



抗ヒスタミン薬の一般的な構造(母核)



頻出部分グラフが連結

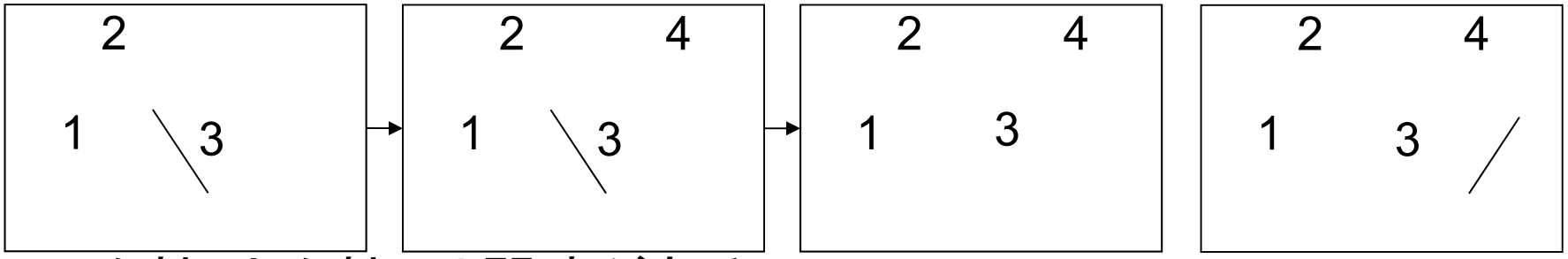


頻出部分グラフが非連結

理解が容易

理解が困難

関連のあるFTSのマイニング

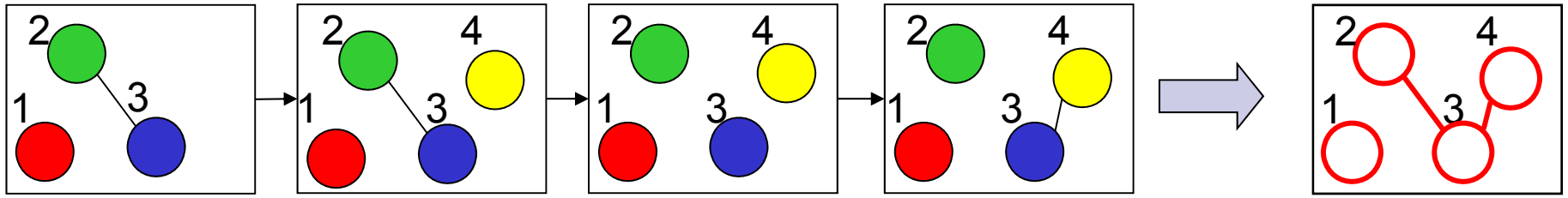


女性2と女性3は関連がある.
女性3と男性4は関連がある.

⇒ 女性2と男性4は女性2を介して関連があると考える.

男性1は他の人と関連がない.

和グラフ



FTSの和グラフが連結であるならば,「関連がある」と定義する.

他の頂点と関連のない頂点を除くことで，互いに関連のある頂点と辺からなるFTSのみをマイニングする.

グラフ系列マイニング問題

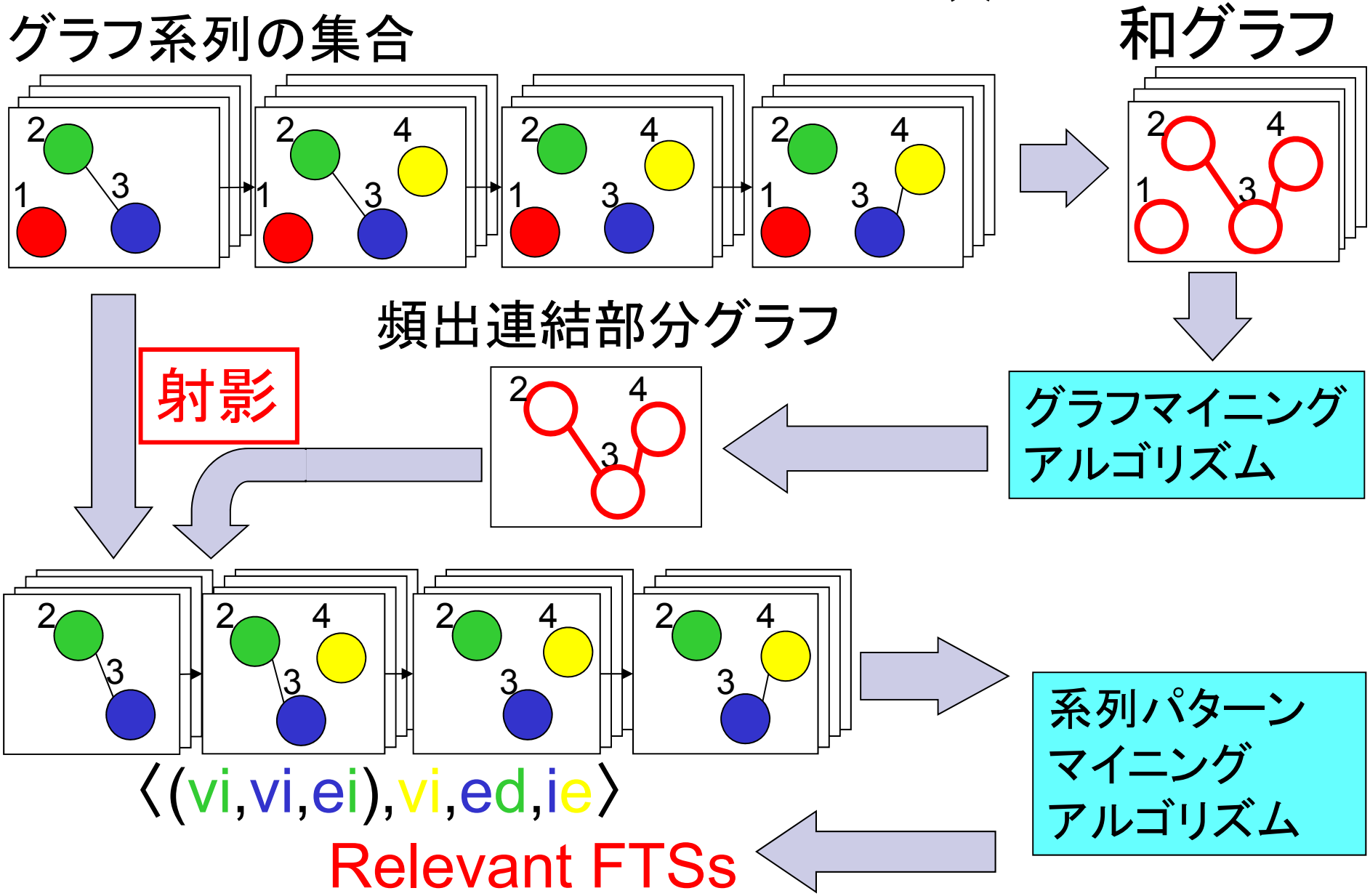
- 変換部分系列の支持度

$$\sup(seq(d')) = \frac{|\{d_i \mid d_i = \langle g_i^{(1)} ... g_i^{(n)} \rangle\}, seq(d') \subseteq seq(d_i), seq(d') : relevant|}{|\{d_i \mid d_i = \langle g_i^{(1)} g_i^{(2)} ... g_i^{(n)} \rangle\}|}$$

$seq(d)$: 変換規則の系列

- 頻出変換部分系列 (FTS：Frequent Transformation Subsequence)
  - 最小支持度以上の支持度を有する変換部分系列
- 支持度の逆単調性
  - $seq(d'_1) \subset seq(d'_2) \Rightarrow \sup(seq(d'_1)) \geq \sup(seq(d'_2))$
- グラフ系列マイニング問題
  - グラフ系列の集合が入力として与えられたとき，全てのFTSを列挙すること

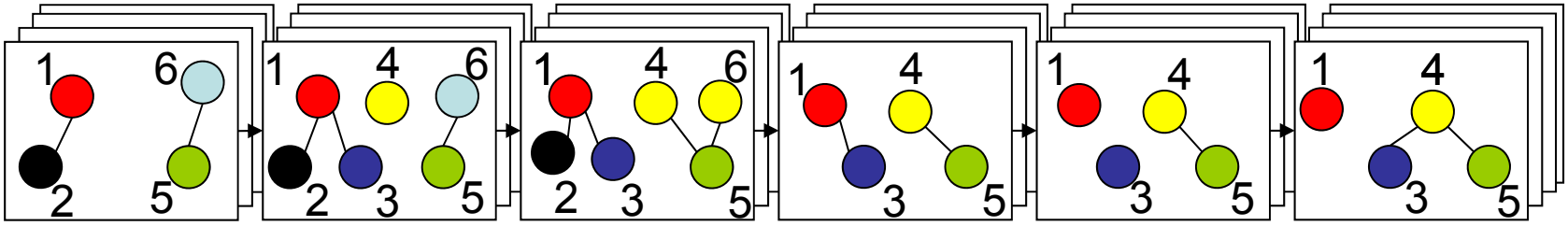
GTRACEのマイニング手順



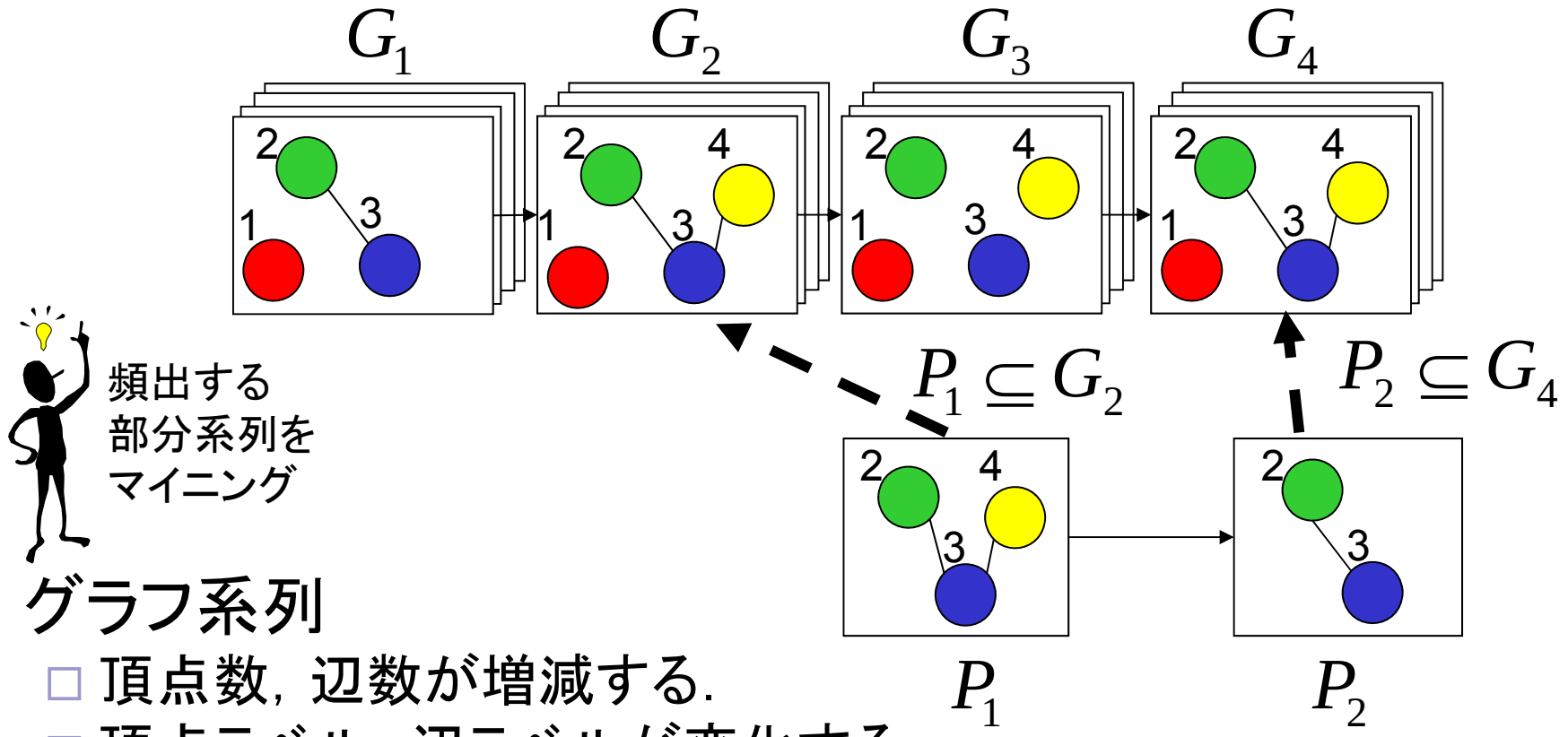
GTRACEの課題

- GTRACEは観測されたグラフ系列中の連続する2つのグラフで，その大部分は変化せず，ごく一部の構造が変化することを仮定

- 観測されたグラフ系列中の連続する2つのグラフが大きく変化する場合には，変換規則の系列が長くなり，膨大な計算時間を要する.

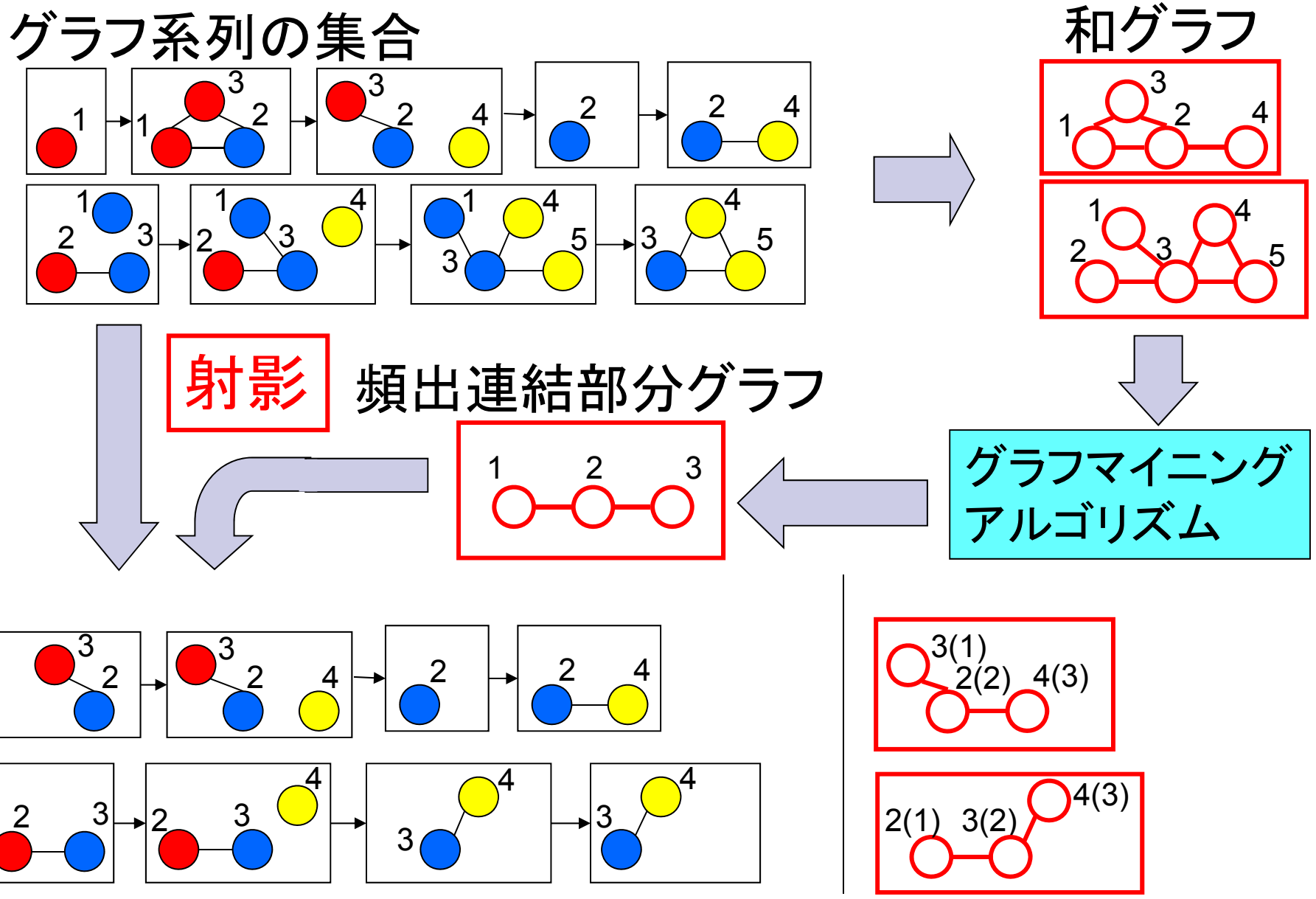


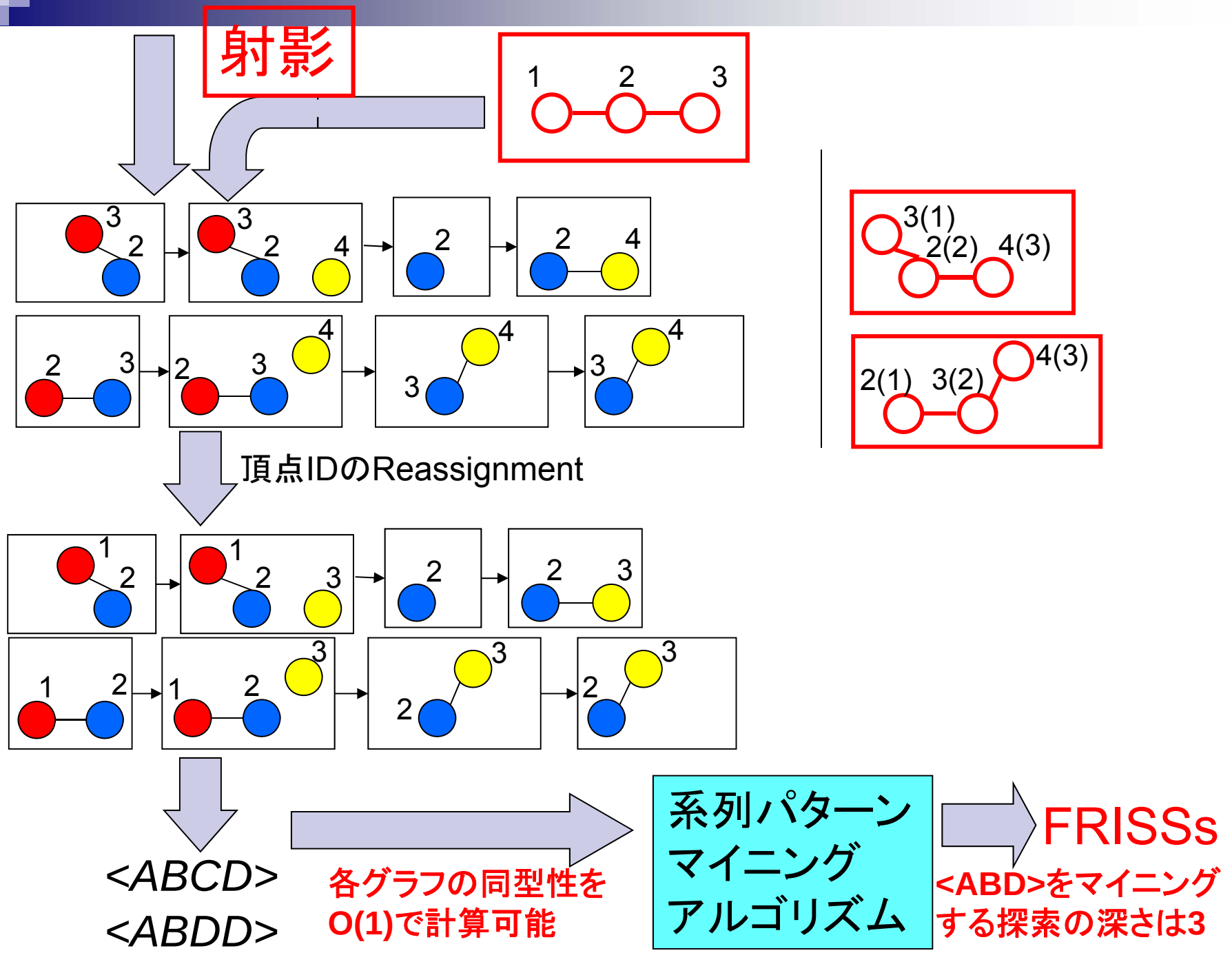
FRISSMiner [Inokuchi 2010]



- グラフ系列
  - 頂点数，辺数が増減する.
  - 頂点ラベル，辺ラベルが変化する.
  - 各頂点は，IDをもつ.
  - ~~グラフ系列中の連続する2つのグラフの間では，構造が大きく変化することはなく，ごく一部の構造のみが変化する.~~

FRISSMinerのマイニング手順




射影
1 2 3
3(1) 2(2) 4(3)
2(1) 3(2) 4(3)
頂点IDのReassignment
<ABCD>
<ABDD>
各グラフの同型性を
O(1)で計算可能
系列パターン
マイニング
アルゴリズム
FRISSs
<ABD>をマイニング
する探索の深さは3

GTRACEとFRISSMinerの比較

	GTRACE	FRISSMiner
前提	連続する2グラフ間で構造が大きく変化しない	なし
グラフ系列の表現形式	変換規則の系列	グラフの系列
取り出されるパターン	共通する変化	共通する構造

まとめ

- グラフ系列マイニング
  - GTRACE [Inokuchi 2008]
    - 変換規則の系列でグラフ系列を表現
    - グラフ系列の集合から共通する変化を列挙
  - FRISSMiner [Inokuchi 2010]
    - グラフ系列の集合から共通する構造を列挙
- 課題
  - グラフ構造の変化の予測
    - データの背後に隠された変動やそのパターンを検知