化学構造マインニングの諸問題

刈田 学
関西学院大学
情報メディア教育センター
http://www.clab.kwansei.ac.jp/~okada/
okada@kwansei.ac.jp

目次
■ マインニングに何を求めるか？
■ 決定木のどう見るか？
■ 落ちこぼれの少なそうな探索法
■ ドーパミン受容体アンタゴニスト
活性の解析
■ おわりに

マイニングに何を求めるか？
■ かつて「マイニング」は悪口であった
■ 変に使うとまずい
■ Lead化合物構造の認識はどの種類？

<table>
<thead>
<tr>
<th>目的</th>
<th>内容</th>
<th>理解</th>
<th>例</th>
</tr>
</thead>
</table>
| 分類                  | すべての事例に確定的な回答  | 不要  | プラン化自動演繹
文字解釈 |
| 状況把握              | すべての問題に複雑的な回答  | 必要  | 医療診断        |
| 当てもの              | 一部の問題にもっとも望しい回答  | ？     | 作成の予測
価値の予測 |

多変量解析での色々な方法
■ 回帰
  ◆ PLS
  ◆ Back propagation NN
■ 判別
  ◆ Counter propagation
  ◆ RBF
■ 主成分, 因子
■ クラスタリング
  ◆ SOM

マイニングの色々な方法
■ 決定木
■ カスケード モデル
■ 例外性発見
■ 総合論理プログラミング
■ 法則を求める 程度

最初は決定木
■ すでに一般的
  ◆ Answer treeなど
■ 専門家が難しいことを書きすぎて
誤解がある
■ 易しく、簡単に使える
■ 限界もある
決定木

Gini index vs. Entropy

\[
\text{Gini-index} = \sum P_i (1 - P_i) = 1 - \sum P_i^2
\]

色々ある決定木の方法

- 分割基準 エントロピー, Gini index
- 枝切り 前切り, 後切り
- Boosting, Bagging, Forest
- ルールと木
- 整形 同一枝パターンのマージ

決定木は何に適するか？

- 最大の問題点はXOR
- 最適の決定木は計算困難
- 葉を見るのか, 木を見るのか

CASE, MultiCASE

- 構造式中の線形連結部分 + 置換基
- 左 2種 (実線部) + 活性Bay reigion
- 右端 実線部 + 発ガン性抑制

網羅的に節点を見る

- これも計算困難
- しかしContingency tableを一歩進めたレベルまでの網羅的列挙は実用でもできる
- 深いレベルまで実現するのが相関ルール前段のライヌス展開
  - 相関ルール後段のルール選択はもっともも深いルール表現を得る
  - ための一つの試み

Cascade ModelとDatascape

- 何でも比較して対照的ならそこが面白いところ
- BSS値
- Ridge
- ルールの組織化

A Link, Distribution of Veiled Items & a Rule

<table>
<thead>
<tr>
<th>y</th>
<th>n</th>
<th>BSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>60</td>
<td>40</td>
</tr>
<tr>
<td>C</td>
<td>00</td>
<td>00</td>
</tr>
<tr>
<td>D</td>
<td>64</td>
<td>36</td>
</tr>
<tr>
<td>E</td>
<td>24</td>
<td>76</td>
</tr>
</tbody>
</table>

IF [B:y] added on [A:y] THEN [D:y; E:n]
Cases: 100 → 60%
BSS = 6.67

No need to generate [A:y, B:y, D:y, E:n]

Datascape 4 Ridge: Topographic Expression of a Rule using BSS Values

<table>
<thead>
<tr>
<th>y</th>
<th>n</th>
<th>BSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>60</td>
<td>40</td>
</tr>
<tr>
<td>C</td>
<td>00</td>
<td>00</td>
</tr>
<tr>
<td>D</td>
<td>64</td>
<td>36</td>
</tr>
<tr>
<td>E</td>
<td>24</td>
<td>76</td>
</tr>
</tbody>
</table>

Cascade Model

Rules

Interpretation

Scheme of Analysis

Structural formulae (SDF file)

MM-AM1-Geo, ClogP

Linear substructures

Applied to Dopamine D1, D2, D3, D4 Antagonists 1227 compounds

Symposium on Structure Activity Relationships
Linear Fragments by CHEMITEM v.2

左上をanchorとして右端のN2まで順にfragmentを左の4つに生成

記述子名生成（左右反転して辞書解析）数 配列数，H, 水素原子
O2H C3-H C3-H-C3H
O2H-C3-C3H-C3H-C3H
O2H-C3-C3H-C3H-C4H
O2H-C3-C3H-C4H-C4H

ヘテロ原子とolefin 炭素をanchor とし，長さ10以下の記述子をすべて生成

MergeしてDopamine分子の記述子

すべての記述子をMergeして各分子中での有無によりnを付与

Computation by DISCAS

- Parameters
  - minsup = 0.01, thres = 0.1, thr-BSS = 0.015, min-rlv = 0.7.
- Computation in the analysis of the D1 activity
  - PC with 1.7GHz Pentium IV, 1GB
  - Lattice generation: 9 min., 76,441 nodes
  - Rule organization: 4.5 min.
  - 51 candidate links, 12 optimized and 2 principal rules,
    1st rule 4 UL relative, 3 Lrelative rules.
    2nd rule 1UL relative, no Lrelative rules.
- Number of nodes for each level
  - 級別
    - 0: 108, 1: 5095, 2: 52718, 3: 100520
- Rule organization: 4.5 min.
  - 51 candidate links, 12 optimized and 2 principal rules,
    1st rule 4 UL relative, 3 Lrelative rules.
    2nd rule 1UL relative, no Lrelative rules.

The 1st Principal Rule for Dopamine D1

Rule 1: Cases: 422 -> 84; BSS= 33.5


THEN D1: 0.81 0.19 ==> 0.18 0.82 (off on) BSS:33.5

Ridge information

- BBSSrate NewBSS Up(off on) --> Low (off on)
- Pre:outside [C3H:C3-C4H-C4H-N3: n] -3.20 12.9 1.00 0.00 / 15
- Pre:outside [C3H-C4H-C4H-N3: n] -1.40 21.0 0.92 0.08 / 382
- New:inside [C3-C4H-N3-C4H-C4H-N3: n] -.797 5.97 0.67 0.33 / 193
- New:inside [C3: n] -.645 1.51 0.71 0.29 / 245
- New:inside [C3-N3: n] -.621 1.84 0.70 0.30 / 246

The 1st Principal Rule for Dopamine D1

Rule 1: Cases: 422 -> 84; BSS= 33.5

IF [C3H:C3-C4H-C4H-N3: y] added on [C4H-C4H-C4H-N3: n] [C3:C3H:C3H-C3-F1: n] [C4H-C4-N3-C4H-C4H-N3: n]

THEN D1: 0.81 0.19 ==> 0.18 0.82 (off on) BSS:33.5

D1:Rule1の視覚化

Linear Fragments by CHEMITEM v.2

- 記述子生成のパラメータ
  - Anchorとしての原子：すべてのヘテロ原子とsp3以外の炭素
  - 最長結合数：10
  - 再端原子で配位数記述の有無：水素原子記述の有無、数
  - 中間原子で：枝分かれ直前の原子記述
  - 配位数記述の有無：水素原子記述の有無、数
  - 生成された記述子数とthres=0.12での採用記述子数

- 解析では、配位数Hの記述あり、thres=0.10の74種を採用

D1:Rule1

2002/10/3
D1活性のまとめ
1. 主条件は、ドーパミン分子構造中の1級アミンが3級アミンに変化した部分構造とD1受容体への結合部位
2. Nとphenyl基間のアルキル鎖長が3では、D1受容体に結合しない。
3. ルール2のbridge情報が示唆）
4. ルールの適用法は、含窒素5価芳香環が存在しない場合には示されない。
5. LUMOが高い場合、Ar-Clが存在すれば活性だが、無い場合は不活性となる。
6. ルール2のbridge情報）
7. Ph-FやAlkyl-piperadino構造の活性への関与も示唆されたが、未段階で解釈不十分。

D4:Rule1-L1
1. N3Hが存在する条件で下記のルール
2. 前提条件を主条件がよく似ている
3. D2でも大きな活性変化がD4内D2
4. 水素結合の有無が影響？

教訓
1. 記述子の重要性
2. 良い記述子ならCASEでも解る
3. 解釈者が高くても解釈困難では困る
4. 記述子数を増やしすぎると計算困難
5. 化学者のセンスが必要
6. ルールだけを見ても解釈困難
7. 周辺分布の特徴を理解する必要
8. 化合物の構造を解読する必要
9. 今後の課題
10. 化学者の仮説との相互作用が大事
11. 利用者のインテリジェンス向上

結論
1. Workbenchの構成
2. MDDR on ISIS, ChemFinder
3. Desalt, Cache, ChemDraw on Excel
4. MM, AM1, CLogP
5. ChemItem, DISCAS
6. Spotfire
7. 知識ベースの公開
8. GPCRを手始めに多数の活性を理解
9. WWWで知識ベース公開
10. Workbenchの公開
11. オープンソースへの出場者受け入れ
12. ソフトウェアの公開