# P4-33 大規模化合物のスケッチ表現によるクラスタリング法

田部井 靖生(JST), 津田 宏治(産総研)

- **化合物データーベースの大規模化**
  - **NCBI PubChemに約2千5百万の化合物**

  ⇒ **クラスタリングは薬の発見に有用**

- **索引による近傍検索からJarvis-Patrickクラスタリングを行うのが一般的**

  ⇒ **$O(n^2)$の時間 (n:データー数)**

- **数千万規模のデーターに適応可能な近傍検法とクラスタリング法の必要性**
  - 両手法に関する研究は少ない (Cao et al., 2010のみ)

# P4-33 大規模化合物のスケッチ表現によるクラスタリング法

**最小値独立置換法**
**:文字列データーへ射影**

**+**

**複合ソート法(Uno,08)**
**:文字列上の高速全点間近傍検索**

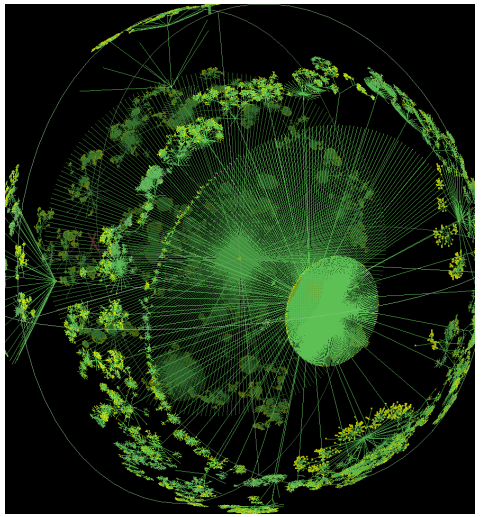
**大規模クラスタリング**

**SketchSortM**
**:高速な全点間傍検索法**

**+**

**オンラインMST**

- 他の近傍検索法よりも高速
- 全PubChemデーターをクラスタリング



(SketchSortはACML 2010に採録)