Diversity Filtering

Christos Kannas
University of Cyprus

Outline

- Introduction
- Methodology
- Implementation

Introduction

- The need to select all the diverse molecules from a dataset (based on a threshold).

- Divide the dataset into diverse molecules and similar molecules .

SEVENTH FRAMEWORK PROGRAMME
University of Cyprus Department of Computer Science
LiSIs Life Sciences Informatics System
RDKit Open-Source Cheminformatics and Machine Learning

Methodology

- 2D Fingerprints

- Similarity Metric: Tanimoto, Dice
  - Similarity Matrix
  - Diagonal has 1…
    - Make diagonal 0, or
    - Skip it… ☺

- Max/Mean/Min Similarity (row/column based)

- Divide molecules in to 2 datasets
  - One with diverse molecules (below similarity threshold)
  - One with similar molecules (above similarity threshold)

Implementation 1/4

- Diversity Score Function [O(n^2)]
  - Inputs:
    - Query Molecules == Reference Molecules
    - Similarity Metric [Tanimoto, Dice]
    - Scoring Method [Max, Mean, Min]
  - Output:
    - Diversity Score

Implementation 2/4

- Show source code for fingerprint similarity/diversity…

Implementation 3/4

- Filtering Engine [O(n)]
  - Inputs:
    - Molecules + Diversity Score
    - Threshold
  - Outputs:
    - Diverse Molecules
    - Similar Molecules

SEVENTH FRAMEWORK PROGRAMME

University of Cyprus
Department of Computer Science

LiSIs Life Sciences Informatics System

RDKit Open-Source Cheminformatics and Machine Learning

Implementation 4/4

- Show source code for diversity filtering…