

SimCluster: reconhecimento de padrões em dados composicionais

Ricardo Vêncio¹

Resumo

Na área de análise de expressão gênica existe uma grande confusão entre métodos estatísticos desenvolvidos para análise de dados obtidos por sequenciamento e obtidos por hibridização. Com o barateamento dos custos dos sequenciadores automáticos de DNA (next-gen sequencing) esse problema tende a ficar ainda mais importante dada a popularização das abordagens de sequenciamento quantitativa. Os dados obtidos por sequenciamento são eminentemente dados composicionais, e portanto, não é natural trata-los como vivendo num espaço Euclidiano usual. Métodos estatísticos específicos e densidades de probabilidade adequadas ao espaço Simplex se fazem necessárias. Exemplos são Dirichlet e Logistic-Normal. Neste trabalho, desenvolvemos um método, implementado numa ferramenta computacional eficiente, o SimCluster, para reconhecimento não-supervisionado de padrões (clustering) de dados modelados naturalmente no espaço Simplex.

¹Universidade de S. Paulo, Faculdade de Medicina de Ribeirão Preto, rvencio@usp.br.