Some ideas for library design

March 22, 2016

Matching pursuit

See `https://en.wikipedia.org/wiki/Matching_pursuit` for background. The idea here is to consider library design as a (sparse) signal approximation problem. For our purpose, a signal is simply a weighted (PageRank iterated) scaffold network. We can formulate library design as follows: Let f denote the desired signal (e.g., scaffolds derived from a set of actives in one or more assays). For a given library $D = \{g_1, g_2, \ldots, \}$ of scaffolds, we seek to approximate f as

$$\tilde{f} = \sum_n a_n g_n,$$

where a_n is the associated weight of scaffold g_n. Our goal is to select a set of molecules that minimize $\|f - \tilde{f}\|$. To calculate $\tilde{f}$, we use the greedy matching pursuit algorithm as given in the wikipedia reference.

Bayesian library

Given a set of assay annotations $A = \{a_1, a_2, \ldots, a_n\}$ defined over a collections of assays, we would like to identify which of the annotations are likely to apply to a set of unscreened molecules. We formulate this in the context of Bayesian as follows. Let $F = \{f_1, f_2, \ldots, f_k\}$ be the set of features (e.g., structural keys) derived from a set of molecules M. The posteriori probability of a_i for a new set of molecules L is simply

$$p(a_i|L) = \frac{p(L|a_i)P(a_i)}{P(L)}$$

or its naïve Bayes formulation

$$p(a_i|L) = \frac{\prod_k p(g_k|a_i)P(a_i)}{P(L)}$$

where g_k is the distribution of the feature f_k over L. If f_k is binary, then one possible way to encode g_k is the Bernoulli distribution with mean m estimated from M.