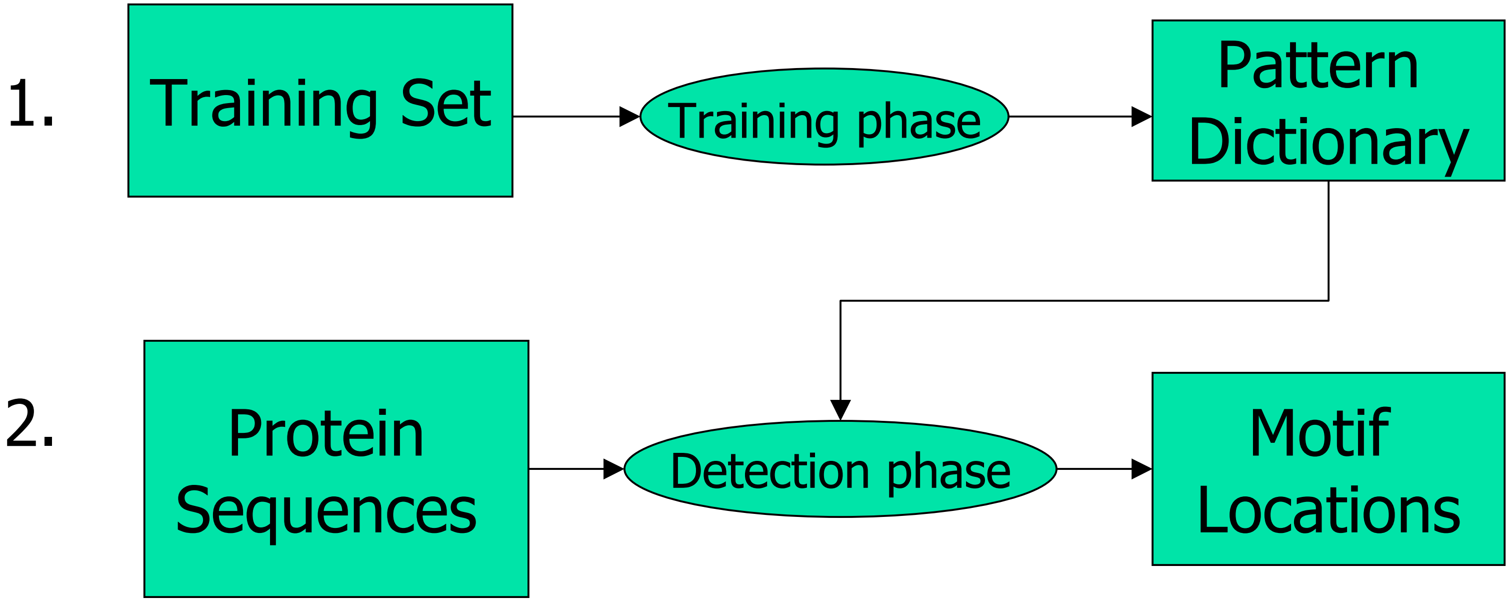
HTH Detection Training Set Selection Based on Phylogenetic Trees

HTH Motif detecion

- HTH Motifs
- HTH Motif Dection
  - HMM
  - DE (Dodd & Egan)
  - GYM

GYM



Pattern: **{C4, P4, S5, L7}**

Support: **the number of protein sequences in the training set in which the pattern appears.**

Maximal Patterns: **not contained in any other significant patterns.**

Difficulties in Training Set Selection

- Some errors or inaccuracy may exist

- Some sequences might be redundant due works at independent labs or by mutation

- There might be an excessive # of motifs of a specific structure in the set.

Unbiased Training Set

- Goal - Evenly distributed in structure
- Reduction/Avoidance of
  - Pure Spurious Pattern
  - Partial Spurious Patter