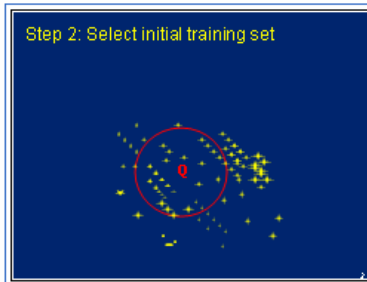
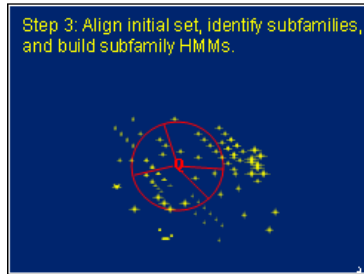


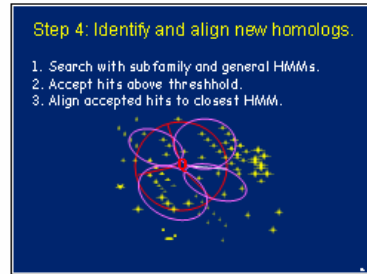
1



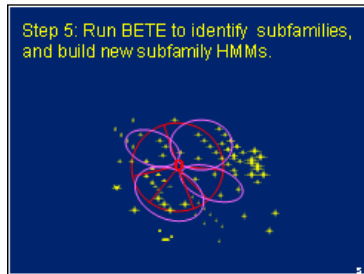
2



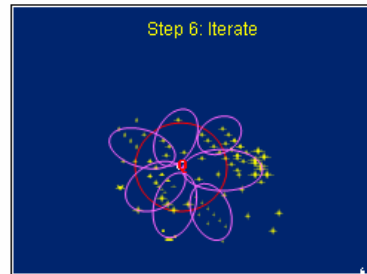
3



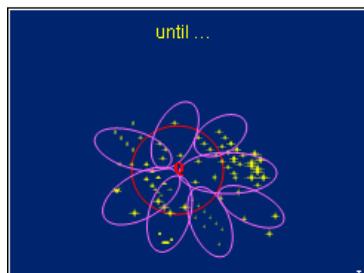
4



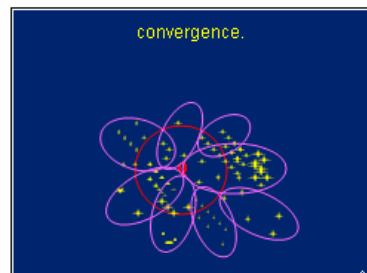
5



6



7



8

FlowerPower clustering and alignment.

Step 1: Given a query sequence, gather a set of putative homologs by searching the UniProt sequence database using PSI-BLAST. This set (the *SearchDB*) will be used for selection of sequences for the cluster and multiple sequence alignment. In some cases (as shown in this toy example), all the *SearchDB* sequences will be accepted; in other cases, only a fraction of the *SearchDB* sequences will be accepted. If a structural (or seed) alignment is provided, skip to Step 3b.

Step 2: Select sequences from the *SearchDB* satisfying two criteria: (1) aligned region is $\geq 70\%$ length of both the query sequence and the hit, and (2) pairwise identity over region aligned is $\geq 30\%$.

Step 3: (a) Align these sequences using a progressive/iterative alignment method (we currently employ ClustalW). (b) Run BETE to find subfamilies and construct subfamily HMMs.

Step 4: (a) Score the *SearchDB* with the subfamily HMMs; tentatively accept all sequences with significant E-values. (b) Align each sequence to the subfamily HMM giving it the strongest score. Concatenate all alignments to create a multiple sequence alignment of all sequences. (c)

Produce statistics for the alignment, such as average pairwise identity across subfamilies, percentage of columns conserved above a particular threshold, number of gapped positions and inserted residues (between HMM nodes) on average within subfamilies. Remove sequences whose alignments do not pass quality control parameter settings (coverage, percent ID, etc.).

Step 5: Run BETE to identify subfamilies, and construct subfamily HMMs.

Iterate Steps 4 and 5 until no new sequences are detected within the specified score cutoff, or the alignment statistics (produced in Step 4c) indicate the alignment quality has reached the desired level.

Further expansion of clusters. If no new sequences are identified in the last iteration, and alignment statistics indicate additional homologs are needed to improve HMM sensitivity, the last general HMM can be used to identify additional homologs by scoring either the *SearchDB* (assuming the PSI-BLAST search produced good results) or a large sequence database such as NR.