SCOP: Sequence, Structure, and Function

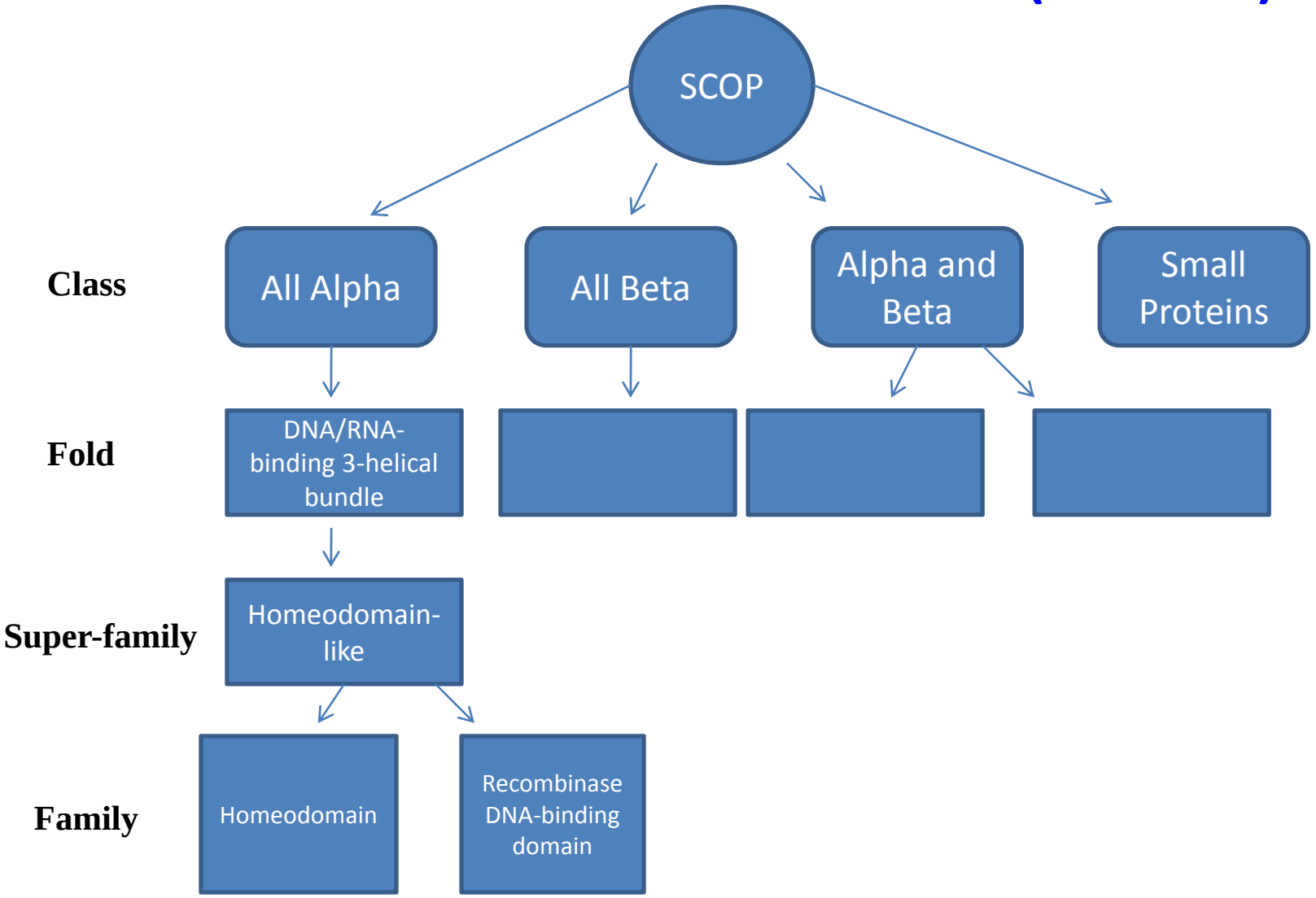
Renqiang (Martin) Min

Mar 22, 2011

SCOP Hierarchy

- Family: Proteins are clustered together based on one of two criteria implying common evolutionary origin: (1) 30% or more sequence identity (2) lower sequence identities but functions and structures are very similar (manual inspection)
- Superfamily: families whose proteins have low sequence identities but whose structures and functional features suggest tat a common evolutionary origin is probable are placed together in superfamilies.
- Fold: superfamilies and families whose proteins have same major secondary structures in same arrangement with the same topological connections.
- Class: just for covenience, different folds are grouped into classes based on the secondary structures they are composed of: alpha, beta, alpha/beta, alpha-beta, multi-domain proteins, small proteins,

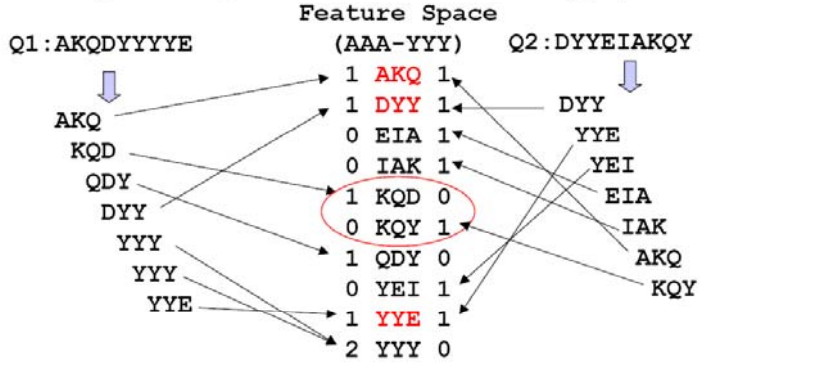
Protein Classification (SCOP)



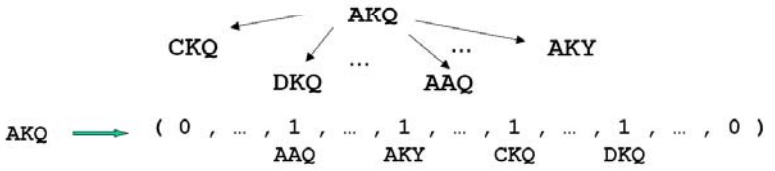
Protein Representations

- Position Specific Weight Matrix (PSWM)
- Pairwise sequence similarity
- Bag-of-Word analogous representation

- Feature map indexed by all possible k-length subsequences ("*k-mers*") from alphabet Σ of amino acids, $|\Sigma| = 20$



- For k-mer s, the *mismatch neighborhood* $N_{(k,m)}(s)$ is the set of all k-mers t within m mismatches from s
- Size of mismatch neighborhood is $O(|\Sigma|^m k^m)$

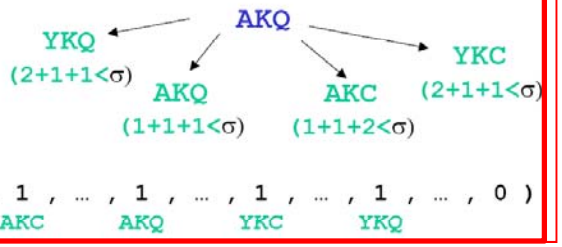


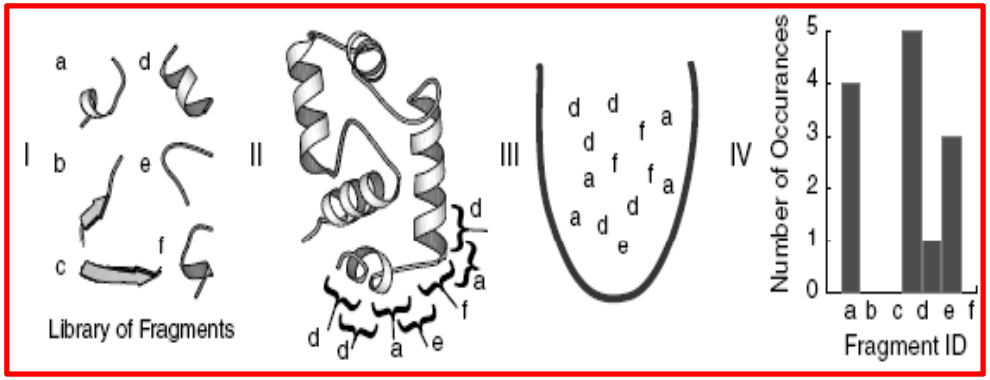
$$\Phi^{emp}(x) = \left(e^{-\lambda S(x,X_1)}, e^{-\lambda S(x,X_2)}, \ldots, e^{-\lambda S(x,X_P)}\right)^T$$

- Use profile $P(x) = \{p_j(b), b \in \Sigma, j = 1 \ldots |x|\}$ to define *position-dependent* mutation neighborhoods:
- E.g. k=3, σ=5 and a profile of negative log probabilities

	A	K	Q	…
A	1	3	4	…
C	5	4	1	…
D	4	4	4	…
…	…	…	…	…
K	4	1	4	…
…	…	…	…	…
Q	3	4	1	…
…	…	…	…	…
Y	2	4	3	…

$$M_{(k,\sigma)}(P(x[j+1:j+k])) = \{b_1 b_2 \cdots b_k : -\sum_i \log(p_{j+i}(b_i)) < \sigma\}$$





Structure and Function Prediction

- Embed the vector representations of proteins into low-dimensional space for efficient query, visualization, and functional analysis
- Identify sequence regions highly important for structures or functions