Methods for searching protein sequence databases have become important tools for the molecular biologist. Because distantly related proteins may share only isolated regions of similarity, e.g. in the vicinity of an active site, such methods usually seek "local alignments" of segments from the query and database sequences. These alignments are generally assessed by means of an amino acid "substitution matrix" that assigns scores to aligning every pair of amino acids. Over the years much effort has been devoted to defining, analyzing and refining such matrices, with the hope of finding the one best suited to distinguishing biological relationships from similarities due merely to chance.

While a wide variety of rationales have been advanced for various scoring systems, recent statistical results show that all matrices may be seen in a common light. Specifically, any substitution matrix is implicitly a log-odds matrix, optimized for a certain set of amino acid pair "target frequencies". With proper scaling, the scores in such a matrix may be viewed as bits of information, or evidence, for the hypothesis of relatedness over that of chance similarity. In order to rise above background noise, the score of an alignment needs to exceed the number of bits required to specify the starting positions of the alignment's two segments.

Since the choice of a substitution matrix reduces to the choice of appropriate amino acid pair target frequencies, how can these frequencies be specified? Given a model of molecular evolution, such as that proposed by Dayhoff and coworkers, one may calculate the expected frequencies at any given evolutionary distance. The popular "PAM-250" substitution matrix, for example, is derived in exactly this manner. Database sequences related to a query, however, may be removed by any evolutionary distance, and short but strong similarities may be just as interesting as long but weak ones. It is argued that in protein database searches, the PAM-120 matrix is the most appropriate to employ for recognizing distant relationships of the most typical lengths. To recognize short and long homologies, this matrix should be supplemented by the PAM-30 and traditional PAM-250 matrices, but further matrices should be of only marginal utility. These arguments are illustrated with biological examples.