# IDENTIFYING PROTEINS FROM MULTIPLE PEPTIDE SEQUENCES

Aaron J. Mackey and William R. Pearson

*Dept. of Microbiology and Dept. of Biochemistry, University of Virginia, Charlottesville, VA, USA*
*amackey@virginia.edu, wrp@virginia.edu*

As complete proteome sequences become available, functional genomics and expression studies seek to identify unknown proteins from limited peptide sequence information. To address this problem, we have developed two programs, FASTS and FASTF, which are designed to identify orthologous proteins from multiple short peptide queries. Using modified FASTA heuristics, FASTS searches protein and translated DNA databases with a query of multiple, noncontiguous peptide sequences to identify related proteins. FASTF searches sequence databases with queries that are the aggregate sequence of several noncontiguous peptides, as is obtained by Edman sequencing of a mixture of peptide species, performing the additional task of deconvoluting the residues at each position. Using a test set of peptide queries, we evaluated the relative performance of these two algorithms and determined the minimal amount of sequence information necessary to robustly infer homology of related proteins. When a single peptide fragment is used for a query, FASTS is equally effective as FASTA at finding homologies; however, a FASTS search with two peptides of length 5 is not as effective as a FASTA search with one peptide of length 10. Generally, as the number of fragments increase, FASTS requires fewer residues per fragment to perform equivalently, e.g. FASTS requires two peptides of length 7, or three peptides of length 5, to perform as well as FASTA with a single 10 residue sequence. In contrast, effective FASTF performance requires as many, if not more, residues per fragment as the number of fragments increases, reflecting the greater cost of deconvolution with more fragments. For related sequences at an evolutionary distance of PAM20 (appropriate for most orthologous mammalian sequences), FASTS can find more than 80% of homologies with fragments containing only four or five residues, while FASTF requires at least seven or more residues per fragment. Scoring matrices with high information content (e.g. PAM20 rather than BLOSUM50) are more effective at finding highly similar relatives than matrices with lower information content. Using these as the default scoring matrices, FASTF and FASTS can quickly and reliably identify proteins that have diverged in the last 100 million years using 10 to 20 residues of peptide sequence data in two or three peptides. These programs provide an efficient strategy for the rapid identification of unknown proteins from limited sequence information.