# BioE/MCB/PMB C146/246, Spring 2005
# Problem Set 5

**1.** Low complexity regions must be masked because they violate the assumptions for computing signi£cance using the EVD distribution. Although low complexity regions with similar composition may occur in unrelated sequences, they may greatly increase the score, leading to a large number of false positives. Coiled coil regions represent one type of low-complexity protein motif, with a periodicity of 7 and biases for hydrophobic residues.

**2.** Unknown protein 1 has the identical phylogenetic pro£le to CheZ and CheY. It is probably involved in chemotaxis. Since unknown protein 2 is one bit different to both the Type I pilins and the Type IV pilins, it is probably a pilin of some sort. The phylogenetic pro£le method provides no quantitative measure of statistical signi£cance, so we can only treat these functional assignments as conjectures.

**3.** Uknown protein 1 contains domains from two tryptophan biosynthesis enzymes, so it is probably a multifunctional tryptophan biosynthesis enzyme itself. Unknown protein 2 combines domains involved in pyrimidine biosynthesis.

**4.** A. Proteins with similar sequences may nevertheless change their functions in different species. A simple annotation transfer to a new sequence may fail to take this change into account.

One scenario: If a gene duplicates in one species, both copies are orthologs of the single copy in some other, well-anotated species. One paralog may have changed its function, but both would be annotated based on similarity to the single gene in the second species.

Errors in annotation may be propagated if functional annotation is based on sequence similarity alone.

Improvements include GO annotations that include the evidence for the annotation, and increased use of phylogenomics methods to annotated genes. New problems include the large number of newly sequenced genomes that have been annotated by similarity, causing more incorrect annotations to be propogated.

B. Little information can be con£dently deduced about the hypothetical protein's function. The only common characteristic among the molecular functions of the top BLAST hits is ATP binding (an SH2/SH3 adaptor protein is a special case of a transmembrane protein tyrosine kinase). The majority of th etop hits contain the molecular signatures of a protein kinase. Since the biological process and cellular component characteristics of GO are less conserved than the molecular functions, we cannot conclude much about these characteristics for the hypothetical protein.

**5.** E should be assigned the "purple" function. A and C are in the same species, so they must be paralogs. A duplication probably occurred in the lineage between where *Plasmodium* split off and where *C. elegans* and *S. cerevisiae* diverged from each other. Then both copies were retained in *C. elegans* but one, the ortholog of protein C, was lost in *S. cerevisiae*. Thus, B and C are also paralogs.