(72) **Alignment of Low-Complexity Glycoprotein Sequences: Composition-Modified Scoring Matrices Allow Alignment of Yeast Cell Wall Proteins**

Juan Coronado 1, Oliver Attie1, Susan L. Epstein 2, Wei-Gang Qiu1 and Peter N. Lipke1

[1] Department of Biology, Hunter College of CUNY, New York, NY 10021,

[2] Department of Computer Science, Hunter College of CUNY, New York, NY 10021.

Although low-complexity sequences are extremely common in glycoproteins and comprise >10% of the proteome, reliable and accurate homology detection and alignment have not been achieved. The alignment of such sequences is essential for structural and evolutionary studies. BLAST and FASTA use the default scoring matrix BLOSUM62, which is optimized for sequences with diverse amino acid composition, that is, high Shannon entropy. With these tools, low-complexity sequence alignments place identical residues in nonhomologous positions and thereby generate anomalously high scores with small e-values. These low e-values are a consequence of deviations from the extreme value distribution of alignment scores, a phenomenon called low-complexity corruption. This corruption prevents BLOSUM62-based BLAST and FASTA from identifying correct homologs for a test query set of low entropy: wall proteins in Saccharomyces cerevisiae. We have devised strategies to restore the extreme value distribution of alignment scores by altering matrix score elements to compensate for amino acid frequencies in any query sequence. That is, matrix score elements were reduced for alignment pairs with high probability of occurrence, and in some cases scores were increased for low-probability events. Empirical tests of these modified matrices used cell-wall proteins as queries in a data set consisting of the yeast proteome plus randomized pseudo-protein

sequences of conserved entropy. BLAST or FASTA using the best-performing matrix modification, called gtQ, reliably identified 95 homologs of a set of cell wall query proteins, where BLOSUM62-based searches identified 0–15 even if low-complexity regions were masked out. The gtQ modifications decreased sensitivity by a modest 15% for searches with high-complexity sequences. These composition-modified matrices generated alignment scores that complied with the expected extreme value distribution and generated e-values more accurate that did BLOSUM62 for queries with cell wall proteins. In both BLAST and FASTA searches, five types of modified matrix identified a consistent set of yeast proteins as homologous to the cell-wall test set and consistently rejected nonwall proteins and randomized pseudo-protein sequences. The modified matrices were computationally efficient and generated alignment scores that conformed to the extreme value distribution. Therefore, selective reduction of scores for high-frequency residue pairs yielded searches with high sensitivity and discrimination for low-complexity sequences. The PERL and shell scripts, customized databases, and supplemental sensitivity curves presented in this article can be obtained at http://wallace.hunter.cuny.edu/~jc/docs/