Poster H-5

Accurate Global Alignment Statistics: An Application to Searching for Protein Domains

Authors:
Maricel G. Kann (*NCBI*)
Sergey L. Sheetlin (*NCBI*)
Yonil Park (*NCBI*)
Stephen H. Bryant (*NCBI*)
John L. Spouge (*NCBI*)

Short Abstract: The most powerful approach for inferring function of protein sequences is the transfer of annotation using sequence comparison methods. This paper develops general, accurate statistical approximations for semi-global alignment and other biological applications, and in searching a protein domain database, its methods retrieved domains better than current local alignment methods.

Long Abstract:
With complete genome sequencing now routine, biology faces a fundamental problem, namely, large-scale automatic annotation of gene function. Since proteins are composed of structural and functional units called domains, a gene can be annotated from domain databases by aligning domains to the gene's protein sequence. Ideally, protein subsequences should be aligned to complete domains, in a "semi-global alignment". Local alignment tools, which align subsequences to only pieces of domains, dominate the field of automatic annotation, because they have heuristics and accurate statistics to screen large databases rapidly and to evaluate the search results. In many motif searches, however, including searching for complete protein domains within a protein sequence, local alignment has two shortcomings. First, it is distracted by strong but incomplete motif matches. Second, it does not align motifs completely or define their boundaries. In this work, we compare the protein domain retrieval of a new semi-global alignment tool GLOBAL (GLObal Blocks Aligned Locally) and introduce an accurate P-value approximation key to semi-global alignment and many other biological applications. In searching domain databases, the P-value retrieved relevant domains better than current local alignment methods. Because GLOBAL derives from a general technique retaining both heuristics and accurate statistics, our results suggest a broad re-evaluation of the role of semi-global alignment in annotation. The implementation of GLOBAL as the default tool at NCBI for searching the CDD is currently under examination.