Computer Speed and Protein Databank Searching

Geoffrey J. Barton
Laboratory of Molecular Biophysics
South Parks Rd.
Oxford OX1 3QU
UK.
Tel: 865-275368
Fax: 865-510454
e-mail: GEOFF@UK.AC.OX.BIOP

This letter appeared in *Science*, 257, 1609, 1992.

Dear Sir,

Despite the well known advances in computer performance that have occurred over recent years, it is still a commonly and erroneously held belief that rigorous sequence comparison methods are too expensive to use for protein databank scanning. In their recent article in *Science*, Gaston et al [1] suggested that the full self comparison of the protein databank would require 10^6 *years* of computer time using a rigorous method [2]. Whilst this figure may be a misprint, the implication is that such a task is beyond the capabilities of today's workstations.

The Smith-Waterman [3] local similarity algorithm, in common with other rigorous methods [2, 4] requires MN steps to calculate the optimum score for aligning two sequences of length M and N including a consideration of insertions and deletions. In order to provide a fair test of modern computers, I have implemented this algorithm in the C-language, taking care to optimize the most frequently executed parts of the program.

When run on a Hewlett-Packard 730 workstation, and using the 141 residue human α – hemoglobin as a query, the program takes 368 seconds to scan the 25044 sequences (8375696 residues) in the SwissProt release 22 databank. Six minutes is a reasonable scan time that compares favorably with IntelliGenetics Inc, BLAZE, one of the speediest rigorous scanning programs. For the same query, BLAZE runs a factor of 17.5 faster, but only on a dedicated 4096 processor MasPar computer. The HP730 time corresponds to 3.2 million array operations per second. The complete rigorous self comparison of SwissProt 22 that Gonnet *et al.* consider impossible, would require 3.5 x 10^{13} such operations, or 4.5 months CPU time to complete on an HP730. Simple distributed processing techniques could reduce this figure to a few weeks [5], and single processor computers from DEC that will be shipping in the Fall and promise speeds up to four times the HP730 would provide similar times on a single workstation.

Since the mid 1980's the speed of typical laboratory and institutional computers has increased by a factor of 70, whilst their cost has reduced by a factor of 10. In contrast, over the same period, the databank of known protein sequences has only increased by a factor of 8. If we ignore the cost/performance gains, today's conventional computers are at least 9 times faster at scanning today's databank, than the machines in 1985 were on the contemporary databank. It seems likely that single processor computer technology will continue to keep ahead of the protein databank until the large scale automation of DNA sequencing becomes a reality.

References

[1] Gonnet, G. H., Cohen, M. A. and Benner, S. A. 1992, *Science* **256**, 1443–1444.

[2] Needleman, S. B. and Wunsch, C. D. 1970, *J. Mol. Biol.* **48**, 443–453.

[3] Smith, T. F. and Waterman, M. S. 1981, *J. Mol. Biol.* **147**, 195–197.

[4] Sellers, P. H. 1974, *SIAM J. Appl. Math.* **26**, 787–793.

[5] Barton, G. J. 1991, *Comput. Appl. Biosci.* **7**, 85–88.