

## **New developments to SCANPS: High performance parallel iterated protein sequence searching with full dynamic programming and on-the-fly statistics.**

Geoffrey J. Barton, Caleb Webber and Stephen M. J. Searle.

EMBL-European Bioinformatics Institute (EBI)  
Wellcome Trust Genome Campus  
Hinxton, Cambridge, CB10 1SD, UK

email: geoff@ebi.ac.uk  
Tel: +44 1223 494414  
Fax: +44 1223 494496

SCANPS (SCAN Protein Sequence) is a program written to perform algorithms related to the Smith-Waterman [1] local similarity search. It runs on variety of conventional hardware [2,3], and can be used to search protein or DNA sequences against large protein sequence databases. In this poster, we present significant enhancements to the program that: (1) Add new on-the-fly statistics based on the score distribution found with each search: the new probability estimates correct for length effects through the use of non-linear fitting methods. (2) Iterative protein sequence searching with full dynamic programming: the results of a search with a single sequence are built automatically into a multiple sequence alignment; a profile calculated from this alignment is then used to probe the database again; searching stops either when the results of a search converge, or after a pre-determined number of iterations. (3) Parallel processing on Silicon Graphics and Intel Linux hardware: three different types of parallelism have been applied to improve performance; OpenMP [4] on Silicon Graphics Origin hardware gives linear speedup to 32 processors [5]; similar performance is obtained through an MPI [6] implementation on shared-memory machines; on-chip parallelism by exploiting Intel MMX and SSE instructions on a 650MHz Intel Pentium III Coppermine CPU gives 164 million cell updates/second (MCU/s) for a non-affine gap algorithm and 71 MCU/s for an affine gap algorithm. For comparison a speed of 47 MCU/s was achieved on a 300MHz MIPS R12000 for the non-affine gap search. When coupled with MPI parallelism SCANPS provides an efficient solution for high performance sequence searching on a low-cost Intel-PC Linux farm.

[1] Smith, T.F. and Waterman, M.S. (1981), *J. Mol. Biol.*, **197**, 723.

[2] Barton, G. J. (1992), *Science*, **257**, 1609.

[3] Barton, G. J. (1993), *Comp. Appl. Bio. Sci.*, **9**, 729-734.

[4] Dagum, L. and Menon, R. (1998), *IEEE Comp. Sci. & Eng.*, **5**, 42-50.

[5] <http://barton.ebi.ac.uk/servers/scanps.html>

[6] Message Passing Interface Forum (1994), *Int. J. Supercomputer App.*, **8**, 3-4.

[7] Peleg, A. and Weiser, U. (1996), *IEEE Micro.*, **16**, 42-50.