Gene duplications in *H. influenzae*

SIR — The sequences and structures of proteins characteristic of late evolution indicate that most, and possibly all, have arisen through a process of gene duplication, mutation and, in some cases, recombination[1]. At some point, this process became more important for the evolution of modern organisms than the *ab initio* creation of new proteins. The recent sequencing of the complete genome of the bacterium *Haemophilus influenzae*[2], which identified some 1,680 protein sequences (two-fifths of the number expected for *Escherichia coli* and about one-fortieth that in humans), has allowed the estimation that at least one-third of the proteins in this small bacterial genome arose from gene duplication.

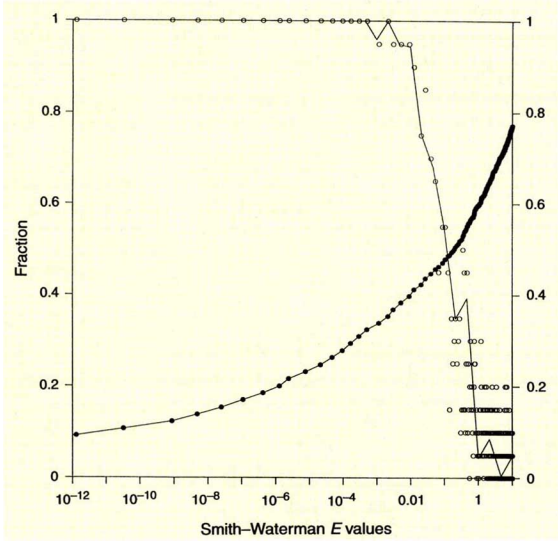
We carried out an all-against-all comparison of the *H. influenzae* protein sequences using the Smith–Waterman algorithm[3]. Although this is the most accurate method for comparing pairs of sequences[4], calculations of this kind have intrinsic limitations. Proteins can diverge to a point where their common evolutionary origin and function cannot be seen from simple sequence comparisons, even though it may be obvious from the close similarities of their three-dimensional structures. Before examining the *H. influenzae* sequences, therefore, we assessed the algorithm's capacity to determine homology from sequence by testing it against a set of proteins of known three-dimensional structure.

The evolutionary relationships of protein domains, determined by comparisons of their three-dimensional structures, are described in the Scop database[5]. We took from this database all sequences that have less than 40% identity and carried out an all-against-all comparison using the Smith–Waterman algorithm. The matched sequences were sorted by the ascending order of their E values (the expectation of finding such a match by chance), and placed in bins of 20. For each bin, we found the fraction of the 20 for which an evolutionary relationship is also clearly indicated by their three-dimensional structures (see figure). For $E<10^{-3}$, there are no false positives, whereas at $E>0.1$ most sequence matches are not supported by structural comparisons. For $E \leq 0.01$, the evolutionary relationships derived from sequence comparisons differ from those derived from structure comparisons by less than 1 case in 100. E values between 0 and 0.01, however, detect only 40% of the evolutionary relationships found by structural comparisons (see figure).

Although these values may not be identical for *H. influenzae* sequences, they demonstrate clearly that E scores of less than 0.01 reliably detect true relationships, but underestimate the number of evolutionary relationships that actually occur.

We next carried out an all-against-all comparison of the *H. influenzae* sequences, using the Smith–Waterman algorithm, and clustered into families all sequences whose E values are less than 0.01. At this level, no relatives are found for 980 *H. influenzae* sequences; the remaining 700 form 208 families (see table). Functions for two-thirds of the *H. influenzae* proteins have been predicted on the basis of their sequence similarities to proteins whose functions are known[1,6]. Among the remaining third, we find 80 that are related to *H. influenzae* proteins with putative functional assignments and with whom they probably share related functions.

Labedan and Riley[7] compared 1,264 *E. coli* protein sequences comprising just under a third of the total expected for this bacterium. They estimate that 38–45% of these sequences have arisen from gene duplications. Our examination of the complete set of sequences from the much smaller *H. influenzae* genome found that at least 30% of the proteins have arisen from processes that involve gene duplications. Simple sequence comparisons underestimate the extent to which this occurs and so the true proportions will be significantly greater. It is clear that gene duplications played a major role in the development of even the simplest forms of life.



The capacity of the Smith–Waterman algorithm accurately to find protein relationships determined on the basis of three-dimensional structures. We compared the sequences of domains from the Scop database[5] that have less than 40% sequence identity using the Smith–Waterman algorithm. For groups of 20 sequence comparisons, the fraction for which evolutionary relationships are indicated by similarities in the three-dimensional structure of the domains is shown (○), along with the cumulative fraction of the structurally determined evolutionary relationships detected by the Smith–Waterman algorithm (●). Both fractions are given as a function of increasing values of E, the expectation of finding such a match by chance.

NUMBER AND SIZE OF PROTEIN FAMILIES IN *H. INFLUENZAE*

n, No. of sequences in a family	No. of families of size *n*
1	980
2	130
3	46
4	10
5	5
6	5
7	2
8	2
9	1
10	2
13	1
15	1
35	1
42	1
43	1

Total no. of families: 1,188
Total no. of sequences: 1,680

Sequence comparisons were carried out using the SSEARCH implementation of the Smith–Waterman algorithm with ln-scaled scores and default parameters[4]. All sequences with E values ≤ 0.01 were clustered into families assuming symmetry and transitivity of relationships.

Steven E. Brenner*
Tim Hubbard†
Alexey Murzin†
Cyrus Chothia*
**MRC Laboratory of Molecular Biology, and*
†*Cambridge Centre for Protein Engineering,*
Hills Road, Cambridge CB2 2QH, UK

1. Patthy, L. *Curr. Opin. struct. Biol.* **4,** 383–392 (1994).
2. Fleischmann R. D. *et al. Science* **269,** 496–512 (1995).
3. Smith, T. F. & Waterman, M. S. *J. molec. Biol.* **147,** 195–197 (1981).
4. Pearson, W. R. *Protein Sci.* **4,** 1145–1160 (1995).
5. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C *J. molec. Biol.* **247,** 536–540 (1995).
6. Casari, G. *et al. Nature* **376,** 647–648 (1995).
7. Labedan, B. & Riley, M. *J. Bact.* **177,** 1585–1588 (1995).

Scientific Correspondence

Scientific Correspondence is intended to provide a forum in which readers may raise points of a scientific character. Priority will be given to letters of fewer than 500 words and five references. Manuscripts can be submitted to London or Washington.