

Reprint Series  
5 June 1992, Volume 256, pp. 1443-1445

**SCIENCE**

## **Exhaustive Matching of the Entire Protein Sequence Database**

Gaston H. Gonnet, Mark A. Cohen, and Steven A. Benner\*

# Exhaustive Matching of the Entire Protein Sequence Database

Gaston H. Gonnet, Mark A. Cohen, Steven A. Benner\*

The entire protein sequence database has been exhaustively matched. Definitive mutation matrices and models for scoring gaps were obtained from the matching and used to organize the sequence database as sets of evolutionarily connected components. The methods developed are general and can be used to manage sequence data generated by major genome sequencing projects. The alignments made possible by the exhaustive matching are the starting point for successful *de novo* prediction of the folded structures of proteins, for reconstructing sequences of ancient proteins and metabolisms in ancient organisms, and for obtaining new perspectives in structural biochemistry.

A decade has passed since questions were raised (1) about the general validity of conclusions drawn from alignments of protein sequences (2). Today, virtually every biochemical analysis routinely begins with, contains, or concludes with an alignment of sequences of proteins that are presumed to be homologous (3). Alignments are also the starting point for methods of predicting *de novo* the secondary structure of proteins (4-6), for all knowledge-based structure predictions (7), for estimating the number of different types of protein folds (8), for interpreting data from the human genome project (9), and for resolving phylogenetic issues (3, 10).

Despite the varied applications of sequence alignments, it has proved difficult to construct sequence alignments correctly. This is not because of inadequate theory; an algorithm that achieves the optimal alignment of two homologous protein sequences was provided over 20 years ago by Needleman and Wunsch (11). Rather, the problem arises because there are simply too many sequence data to analyze and because the parameters needed to correctly score mutations, deletions, and insertions are unavailable.

Today, mutations (mismatches) in an alignment are usually scored with a mutation matrix developed by Dayhoff and her co-workers in the 1970s (12). However, this matrix was derived from alignments of an extremely small set of proteins that are very similar in sequence, and is therefore unsuitable for alignments between two proteins whose sequences are sufficiently similar to suggest that they might be homologous, but not similar enough to make homology obvious.

The difficulties in constructing alignment routines are further complicated by the requirement that they handle deletions and insertions. Even random sequences can be aligned if gaps are introduced at no

penalty. Most alignment programs therefore assign penalties to gaps of the form ( $ak + b$ ), where  $k$  is the length of the gap and  $a$  and  $b$  are arbitrarily chosen constants. There is no justification, either theoretical or empirical, for this treatment. Indeed, many of the questionable conclusions drawn from alignments arise because of inappropriately placed gaps. Conversely, correctly placed gaps provide information that is critical to the *de novo* prediction of the folded structure of proteins from sequence data alone (4-6). Such information led to the remarkably accurate predictions of the folded structures of tryptophan synthase and protein kinase before crystallographic information was available (4-6).

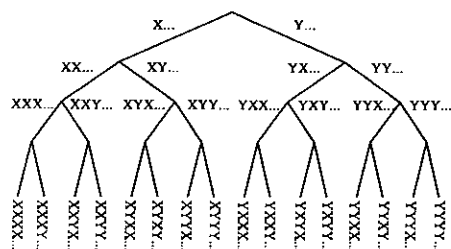
The amount of sequence data presently available should make it possible to do an exhaustive matching of the entire sequence database (defined here as the result of an attempted Needleman-Wunsch alignment of every subsequence in the database with every other subsequence), obtain empirical probabilities of mutations between amino acids, determine empirical gap scoring penalties, and use these to obtain high-quality alignments. This has not been done previously because the Needleman-Wunsch algorithm is slow (about one pairwise comparison per second). Because a typical contemporary database (such as MIPS Version 64) contains 8,344,353 ( $n$ ) amino acids, exhaustive matching of all subsequences could involve some  $35 \times 10^9$  pairwise

comparisons (on the order of  $n^2$ ) and more than  $10^6$  years of computer time. Not surprisingly, an exhaustive matching of a modern sequence database with the Needleman-Wunsch algorithm has been thought to be essentially impossible (13).

We report the exhaustive matching of an entire protein sequence database. Neither the Needleman-Wunsch algorithm nor any of the rigor that it implies was sacrificed. The key to matching in a reasonable time lies in the step preceding the application of the Needleman-Wunsch algorithm: a reorganization of the sequence data by indexing on a patricia tree (14) (Fig. 1). In an indexed database, pairs of identical sequences are found instantaneously because they lie together on the tree. Similar sequences lie near each other in the tree. Thus, all pairs of sequences that might be significantly similar can be found in an indexed database by far fewer than  $n^2$  matching operations and aligned with the Needleman-Wunsch algorithm. Thus, our exhaustive matching required only 405 days of CPU time and was obtained in the background (otherwise idle CPU capability) from up to six workstations running in parallel for only 19 weeks.

Classical mutation matrices and gap penalties were used in the first phase of the exhaustive matching. A liberal target score ensured that every match with potentially significant sequence similarity was examined. The  $6.5 \times 10^6$  matched pairs of subsequences that were found in the first phase were then refined by running the Needleman-Wunsch algorithm from the point where each match began in one direction along the sequence alignment to the point where the alignment was optimized (or the sequences exhausted), running the algorithm in the reverse direction to achieve the same goal, and repeating the process until the alignment score was no longer improved. After refining,  $1.7 \times 10^6$  matches remained, each optimally aligned, which were then used to calculate new mutation matrices and a model for scoring gaps. These new scoring parameters were then used to further refine the matches to self-consistency. The parameters provide

**Fig. 1.** Reorganization of sequences to form semi-infinite strings placed in "alphabetical order" (left to right in this diagram) on a patricia tree (13), idealized here for sequences built from just two letters. Reorganization time is almost linear with database size and requires negligible computation. Exhaustive matching is achieved by comparing patricia subtrees from the top. Time is saved because the matching of patricia subtrees is aborted when the score falls below a liberally chosen similarity limit. Because all subsequences in the database are indexed, the fact that two similar protein sequences do not begin identically does not diminish the generality of the search.



Institute for Scientific Computation and Institute for Organic Chemistry, Swiss Federal Institute of Technology, Universitatstr. 16, 8092 Zurich, Switzerland.

\*To whom correspondence should be addressed.



observed  $-3/2$  power dependence of probability on gap length.

The final outcome of the exhaustive matching is a reorganized database that can be rapidly searched using the DARWIN (Data Analysis and Retrieval With Indexed Nucleotide/Peptide Sequences) system (19). Each of the  $1.7 \times 10^6$  aligned pairs of subsequences that result from the exhaustive matching is characterized by an evolutionary distance measured in PAM units. DARWIN, taking a PAM distance from the user, rapidly reconstructs the entire database in the form of sets of "connected components," entries joined by a match with every other entry in the component at or below the user-designated PAM. Because the PAM distances are accompanied by a statistical variance, evolutionary trees (20, 21) constructed from these distances by DARWIN are rigorous; they are accompanied by a probability score for the most probable connectivity, probabilistic sequences for the ancestral proteins at the nodes of the tree, and a multiple alignment.

At very low PAM distances, the connected components include very similar sequences, multiple entries in the database, and entries that differ only because of sequencing or entry error. At increasing PAM distances, however, connected components grow to include families and superfamilies of proteins. Repetitive sequences are the only feature that significantly joins apparently nonhomologous entries into connected components. From the total number of connected components plotted as a function of PAM distance (Fig. 4), the number of different protein types in the database can be estimated. Even conservative estimates indicate the existence of several thousand separate families of proteins (8). Finally, from these connected components, proteins and metabolisms can be reconstructed for various ancestors of modern organisms (10). Several of these reconstructed ancient proteins have now been prepared and studied in these laboratories (22).

## REFERENCES AND NOTES

1. R. F. Doolittle, *Science* 214, 149 (1981).
2. A. W. F. Edwards and L. L. Cavalli-Sforza, *Ann. Hum. Genet.* 27, 104 (1963); E. Zuckerkandl, *Protides Biol. Fluids* 12, 102 (1964); \_\_\_\_\_ and L. Pauling, in *Evolving Genes and Proteins*, V. Bryson and H. J. Vogel, Eds. (Academic Press, New York, 1965), p. 97; J. S. Farris, *Syst. Zool.* 19, 83 (1970); W. Fitch, *ibid.* 20, 406 (1971).
3. R. F. Doolittle, Ed., *Methods Enzymol.* 183, 1 (1990).
4. S. A. Benner, *Adv. Enzyme Regul.* 28, 219 (1989); \_\_\_\_\_ and D. Gerloff, *ibid.* 31, 121 (1991).
5. D. R. Knighton *et al.*, *Science* 253, 407 (1991); J. M. Thornton, T. P. Flores, D. T. Jones, M. B. Swindells, *Nature* 354, 105 (1991).
6. T. Niermann and K. Kirschner, *Protein Eng.* 4, 137 (1990).
7. T. L. Blundell, B. L. Sibanda, M. J. E. Sternberg, J. M. Thornton, *Nature* 326, 347 (1987).
8. W. R. Taylor, *Methods Enzymol.* 183, 456 (1990); R. L. Dorit *et al.*, *Science* 250, 1377 (1990).
9. T. D. Yager, D. A. Nickerson, L. E. Hood, *Trends Biochem. Sci.* 16, 454 (1991); M. C. Rechsteiner, *ibid.*, p. 455.
10. S. A. Benner, A. D. Ellington, A. Tauer, *Proc. Natl. Acad. Sci. U.S.A.* 86, 7054 (1989).
11. S. B. Needleman and C. D. Wunsch, *J. Mol. Biol.* 48, 443 (1970).
12. M. O. Dayhoff, R. M. Schwartz, B. C. Orcutt, in *Atlas of Protein Sequence and Structure*, M. O. Dayhoff, Ed. (National Biomedical Research Foundation, Washington, DC, 1978), vol. 5, suppl. 3, p. 345.
13. W. Taylor, *Nature* 353, 388 (1991).
14. G. H. Gonnet, *Handbook of Algorithms and Data Structures* (Addison-Wesley, London, 1984).
15. A PAM 1 mutation matrix (the matrix describing the probability of mutations in a pair of proteins that have diverged by 1 accepted point mutation per 100 residues) can be formally extrapolated to a PAM 250 matrix by raising the PAM 1 matrix to the 250th power by matrix multiplication.
16. S. E. Altschul, *J. Mol. Biol.* 219, 555 (1991).
17. Because dynamic programming with nonlinear gap penalties is problematic, a linear fit of the equation given in the text is useful:  
$$10 \log(P) = -39.21 + 7.75 \log(\text{PAM distance}) - 1.65(k - 1)$$
  
For users of alignment programs that do not permit variation with different PAM distances, the following scoring is recommended:  
$$10 \log(P) = -20.63 - 1.65(k - 1)$$
  
Although this scoring is less accurate, its parameters are sufficiently different from those found as defaults in most alignment programs to make its use advisable.
18. P. Flory, *Principles of Polymer Chemistry* (Cornell Univ. Press, Ithaca, NY, 1953); D. A. Brant and P. J. Flory, *J. Am. Chem. Soc.* 87, 2788 (1965).
19. G. H. Gonnet and S. A. Benner, *Tech. Rep.* 154, *Departement Informatik* (Eidgenossische Technische Hochschule, Zurich, 1991). DARWIN is available in a version that operates on a Sun workstation under Unix.
20. W. R. Taylor, *Comput. Appl. Biosci.* 3, 81 (1987).
21. J. Hein, *Mol. Biol. Evol.* 6, 649 (1989).
22. J. Stackhouse, S. R. Presnell, G. M. McGeehan, K. P. Nambiar, S. A. Benner, *FEBS Lett.* 262, 104 (1990).
23. Dedicated to F. W. Westheimer on the occasion of his 80th birthday. M.A.C. was supported by a fellowship from the Wellcome Trust. Part of this work was presented at the July 1990 meeting of the Institute for Advanced Biological Studies. We thank Digital Equipment Corporation for donation of computer equipment and Sandoz AG for partial support of this work.

24 January 1992; accepted 1 April 1992