Proc. Natl. Acad. Sci. USA
Vol. 84, pp. 4355–4358, July 1987
Biochemistry

Profile analysis: Detection of distantly related proteins

(amino acid/sequence comparison/protein structure/globin structure/immunoglobulin structure)

MICHAEL GRIBSKOV*, ANDREW D. MCLACHLAN†, AND DAVID EISENBERG*

*Molecular Biology Institute and Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90024; and †Medical Research Council, Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England, United Kingdom

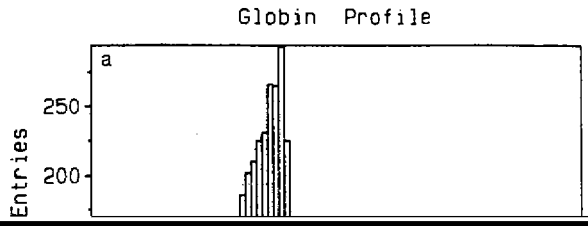
Communicated by Paul Boyer, February 17, 1987 (received for review November 19, 1986)

ABSTRACT **Profile analysis is a method for detecting distantly related proteins by sequence comparison. The basis for comparison is not only the customary Dayhoff mutational-distance matrix but also the results of structural studies and information implicit in the alignments of the sequences of families of similar proteins. This information is expressed in a position-specific scoring table (profile), which is created from a group of sequences previously aligned by structural or sequence similarity. The similarity of any other sequence (target) to the group of aligned sequences (probe) can be tested by comparing the target to the profile using dynamic programming algorithms. The profile method differs in two major respects from methods of sequence comparison in common use: (*i*) Any number of known sequences can be used to construct the profile, allowing more information to be used in the testing of the target than is possible with pairwise alignment methods. (*ii*) The profile includes the penalties for insertion or deletion at each position, which allow one to include the probe secondary structure in the testing scheme. Tests with globin and immunoglobulin sequences show that profile analysis can distinguish all members of these families from all other sequences in a**

Common methods for detection of similarity depend on pairwise alignment of sequences—for example, the dot matrix method (9, 10) or dynamic programming methods (11–14). Another class of methods are the rapid database searching methods (15, 16). All of these normally test every sequence in the database independently against a single probe sequence without using information implicit in the alignments of families of related sequences or including information available from structural studies. [An exception is the family comparison dot matrix method (9), which, however, does not allow for insertion or deletion.] Profile analysis brings in both structural and family information at the expense of a modest increase in computation time.

METHODS

Construction of the Profile (PROFMAKE). Profile analysis has two steps (Fig. 1*a*): (*i*) construction of the profile with the program PROFMAKE, and (*ii*) comparison of the profile with a database of sequences or a single sequence (program PROFANAL). The starting point for the creation of a profile



alignment given replacement scores and penalties for insertions and deletions (11–14). The major modification to these algorithms for use with profiles lies in the scoring system. In the unmodified algorithm, the score at a given position in the alignment score matrix is based on the comparison of the amino acid residues at the corresponding positions in the two sequences. In profile analysis, the score is read from the column of the profile corresponding to the amino acid residue

DISCUSSION

Selectivity of Profile Analysis. An ideal method for detecting homologous proteins would separate a database of sequences into two groups with no overlap in scores between them: the homologous proteins and all other proteins. Fig. 2 suggests that profile analysis is powerful by this criterion. This selectivity comes from (*i*) the information implicit in aligned sequences, encoded in the flexible scoring system of the profile, and (*ii*) the ability of dynamic programming methods to position gaps, as guided by the penalties in the profile. The essence of the profile is that both the gap penalty and amino acid preference are position dependent. The position-dependent gap penalty introduces structural information, such as the known locations of secondary structure elements. The position-dependent amino acid preference introduces information about the character of the allowed side chains in each position.

Comparison with Other Methods. The profile method is useful for learning whether a protein sequence belongs to a known family of sequences. The method differs from both rapid database methods and standard dynamic programming methods in that these methods are designed for pairwise, rather than family, comparisons. Dynamic programming methods have been applied to align three sequences (24) but may be hard to apply for large numbers of sequences. With dynamic programming methods, information from a family of proteins can be included by comparing the members of the family by twos or threes and then synthesizing an overall alignment from the individual alignments. This tedious process is replaced in profile analysis by the position-specific scoring table.

The profile method shares characteristics of template methods. Template (20, 25) or fingerprint (27) methods fit a sequence to a rigid pattern of amino acid residues with no gaps allowed. This rigidity can be softened by breaking the template into segments separated by variable-length regions where any residue is allowed (functionally equivalent to gaps). The size of these regions is determined either by fitting each segment independently and checking that the order and spacing of the segments is reasonable (20), or by making a different template for every possible allowed spacing (27).

A template can be considered a special case of a profile in which any amino acid occurring in the probe sequences is given a score of 1.0, and in which the insertion/deletion penalty is set high in regions corresponding to segments (to prevent gaps), and low in the regions between segments. In contrast, profile analysis assigns positive scores even to target amino acid residues that are not observed in the probe and permits gaps within segments if a much better alignment can be obtained. Profile analysis thus includes template and fingerprint methods as special cases.

Extensions of the Method. Any set of properties that can be represented as similarity or difference scores for pairs of amino acids can be used to construct profiles. The scoring system used in the examples shown here is based on observed frequencies of replacement in homologous proteins. Other properties such as hydrophobicity, α or β structural preference (28), or side-chain volume can be used as scoring tables.

A possible eventual use for the profile method is to infer information on three-dimensional structure from sequence. Creation of a set of profiles for a variety of protein families will offer a library of structural motifs. Comparison of any newly discovered sequence with the library may yield information on structural motifs within the protein.

Copies of this program may be obtained from the authors at the University of California at Los Angeles. Programs are available in a format compatible with the University of Wisconsin Genetics Computer Group (UWGCG) software package or in an independent implementation. Program development was aided by the UWGCG procedure library (26).

We thank Drs. A. M. Lesk and C. Chothia for discussions of their template method for comparison. This work was supported by grants from the National Science Foundation (PCM 82-07520), National Institutes of Health (GM 31299), the University of California Biotechnology Research and Education Program, the Simon Guggenheim Foundation to D.E., and the National Cancer Society (PF-2649) and Lita Annenberg Hazen to M.G.

1. Fitch, W. M. (1966) *J. Mol. Biol.* **16,** 9–16.
2. McLachlan, A. D. (1972) *J. Mol. Biol.* **62,** 409–424.
3. Doolittle, R. F. (1981) *Science* **214,** 149–159.
4. Blundell, T. L., Sibanda, L. & Pearl, L. (1983) *Nature (London)* **304,** 273–275.
5. Sweet, R. M. (1986) *Biopolymers* **25,** 1565–1577.
6. Kabsch, W. & Sander, C. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 1075–1078.
7. Sweet, R. M. & Eisenberg, D. (1983) *J. Mol. Biol.* **171,** 479–488.
8. Dickerson, R. E. & Geis, I. (1983) *Hemoglobin* (Benjamin/Cummings, Menlo Park, CA).
9. McLachlan, A. D. (1983) *J. Mol. Biol.* **169,** 15–30.
10. Maizel, J. V., Jr., & Lenk, R. P. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 7665–7669.
11. Needleman, S. B. & Wunsch, C. D. (1970) *J. Mol. Biol.* **48,** 443–453.
12. Smith, T. F. & Waterman, M. S. (1981) *Adv. Appl. Math.* **2,** 482–489.
13. Sellers, P. H. (1974) *SIAM J. Appl. Math.* **26,** 787–793.
14. Boswell, D. R. & McLachlan, A. D. (1984) *Nucleic Acids Res.* **12,** 457–464.
15. Wilbur, W. J. & Lipman, D. J. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 726–730.
16. Lipman, D. J. & Pearson, W. R. (1985) *Science* **227,** 1435–1442.
17. Dayhoff, M. O. (1979) in *Atlas of Protein Sequence and Structure*, eds. Schwartz, R. M. & Dayhoff, M. O. (Natl. Biomed. Res. Found., Washington, DC), Vol. 5, Suppl. 3, pp. 353–358.
18. Gribskov, M. & Burgess, R. R. (1986) *Nucleic Acids Res.* **14,** 6745–6763.
19. Staden, R. (1984) *Nucleic Acids Res.* **12,** 505–519.
20. Taylor, W. R. (1986) *J. Mol. Biol.* **188,** 233–258.
21. Lesk, A. M. & Chothia, C. (1980) *J. Mol. Biol.* **136,** 225–270.
22. Wakabayashi, S., Matsubara, H. & Webster, D. A. (1986) *Nature (London)* **322,** 481–483.
23. Yanagi, Y., Yoshikai, Y., Leggett, K., Clark, S. P., Aleksander, I. & Mak, T. W. (1984) *Nature (London)* **308,** 145–149.
24. Murata, M., Richardson, J. S. & Sussman, J. L. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 3073–3077.
25. Taylor, W. R. & Thornton, J. M. (1984) *J. Mol. Biol.* **173,** 487–514.
26. Devereux, J., Haeberli, P. & Smithies, O. (1984) *Nucleic Acids Res.* **12,** 387–395.
27. Wierenga, R. K., Terpstra, P. & Hol, W. G. J. (1986) *J. Mol. Biol.* **187,** 101–107.
28. Chou, P. Y. & Fasman, G. D. (1978) *Annu. Rev. Biochem.* **47,** 251–276.