PREDICTION OF STRUCTURALLY CONSERVED REGIONS OF D-SPECIFIC HYDROXY ACID DEHYDROGENASES BY MULTIPLE ALIGNMENT WITH FORMATE DEHYDROGENASE

Carlota Vinals, Eric Depiereux and Ernest Feytmans

Facultés Universitaires Notre Dame de la Paix, Department of Biology, rue de Bruxelles, 61
5000 Namur - Belgium

Received January 13, 1993

We propose a multiple alignment of the sequence of formate dehydrogenase with the D-specific 2-hydroxy acid dehydrogenases family. Structurally conserved regions are predicted for those sequences corresponding to important regions of the catalytic and the coenzyme binding domains defined from the known three-dimensional structure of the formate dehydrogenase, namely the nicotinamide binding site (ßD to ßF) and the ßA-loop-αB region containing the typical glycine pattern of the adenosine binding site, the catalytic histidine/aspartic acid pair and an arginine probably involved in the interaction with the carboxyl group of the substrate.

The reduction of pyruvate to the L(-)- or D(+)-lactate isomer is catalyzed by different NAD-dependent enzymes, the L-lactate and D-lactate dehydrogenases (1). Several sequences and structures of L-lactate dehydrogenases have been determined and characterized (2-6) before obtaining the first D-lactate dehydrogenase sequences in 1991 (7,8). Since all the L-lactate dehydrogenases seem to belong to the same evolutionary family (9-11), first results obtained on D-lactate dehydrogenases (7, 8, 11,12) indicate that the latter form a distinct family, probably with a distinct evolutionary origin. Practically, this means also that structural information can not be obtained from L-lactate dehydrogenases to predict any structural particularity of the D-lactate dehydrogenases. An alignment of sequences related to the D-lactate dehydrogenase from *Lactobacillus delbrueckii* has been proposed (13), namely with D-2-hydroxyisocaproate dehydrogenase from *L. casei* (14), glycerate dehydrogenase (or hydroxypyruvate reductase) from cucumber (15), D-3-phosphoglycerate dehydrogenase from *E. coli* (16) and erythronate 4-phosphate dehydrogenase from *E. coli* (17). This alignment strongly suggests that these sequences form what can be called the 'D-specific 2-hydroxy acid dehydrogenases family'. It seems also that these enzymes do not share any homology with L-specific 2-hydroxyacid dehydrogenases such as L-lactate, L-malate or L-hydroxyisocaproate dehydrogenases. Recently, the sequence and three-dimensional structure of the formate dehydrogenase from *Pseudomonas* sp. 101 have been solved (18, 19). We show that formate dehydrogenase has relevant sequence similarities with D-lactate dehydrogenase, although it is not part of the hydroxy acid dehydrogenases. This structure provides the first opportunity of getting some prediction about the possible structure of the D-lactate dehydrogenase, as well as other homologous proteins. As similarities between the targets and the considered template remain relatively sparse (~20%), a

0006-291X/93 $4.00

multiple alignment method is used. This method produces an accurate prediction of structurally conserved regions in homologous sequences (20).

EXPERIMENTAL PROCEDURES

The amino acid sequence of the formate dehydrogenase has been compared with the GenEMBL and Swissprot protein and nucleotide sequence databanks with the FastA and TFastA algorithms (21). Similarities (optimized score > 100) are found with D-lactate dehydrogenases (E.C.1.1.1.28) from *L. bulgaricus* (8,11) and *L. plantarum* (12), D-2-hydroxyisocaproate from *L. casei* (14), vancomycin resistance protein from *Enterococcus faecium* (22), glycerate dehydrogenase (E.C.1.1.1.29) from cucumber (15), D-3-phosphoglycerate dehydrogenase (E.C.1.1.1.95) from *E. coli* (16) and erythronate 4-phosphate dehydrogenase from *E. coli* (17).

The whole set of sequences is analyzed with the software Match-Box for simultaneous alignment of several protein sequences (20, 23). In this approach, the comparison of the physicochemical profiles of all the 7-residue segments of sequence allows (i) to perform an analysis of the global similarity between the sequences by factor analysis, (ii) to compare the observed similarities with the ones expected by chance between unrelated sequences and (iii) to perform between- and within group alignments. The algorithm delineates boxes including only segments that are all similar one with each other according to a relatively severe statistical threshold. A separate analysis is performed with L-lactate dehydrogenase from *Bacillus stearothermophilus* and D-lactate dehydrogenase from *L. delbrueckii* to test their expected relatedness by the same approach.

RESULTS AND DISCUSSION

The first step of the sequence analysis provides the distribution of the cumulated frequencies of the physicochemical distance between the segments of sequences (Fig.1). The distribution obtained for the original sequences is compared with the one obtained for the same sequences after randomizing them. A higher frequency of small classes of distances in the original sequences points out a greater similarity than expected by chance. When the similarity departs from randomness, this graph allows also to fix an adequate cutoff to discriminate reliable matches from random noise in the alignment procedures. Figure 1A presents the distribution cumulated for all the sequences included in the analysis. The difference between the distributions obtained for original and randomized sequences clearly indicates that at least some of them are significantly similar. The less related sequences are then fetched and compared. Figure 1B shows that a significant difference is still observed. L- and D-lactate dehydrogenases are compared in a separate analysis and in this case no difference is found (Fig. 1C). The same result is obtained when comparing L- and D-hydroxyisocaproate dehydrogenases (data not shown). This is in accordance with the results obtained previously by other authors (7, 12, 13) and validates the similarities observed between the aligned sequences.

In a second step a factor analysis is performed on the set of sequences, after a first rough alignment. This algorithm computes, for each pair of sequences, the probability of matching any segment of one sequence with at least one segment in the other sequence, with a maximal admitted shift between the segments compared (23). The similarity matrix obtained is analyzed by principal coordinates analysis. The space of factors 2, 3 and 4 (the first factor being trivial) allows a qualitative analysis of the similarities between the sequences, the distance between two sequences

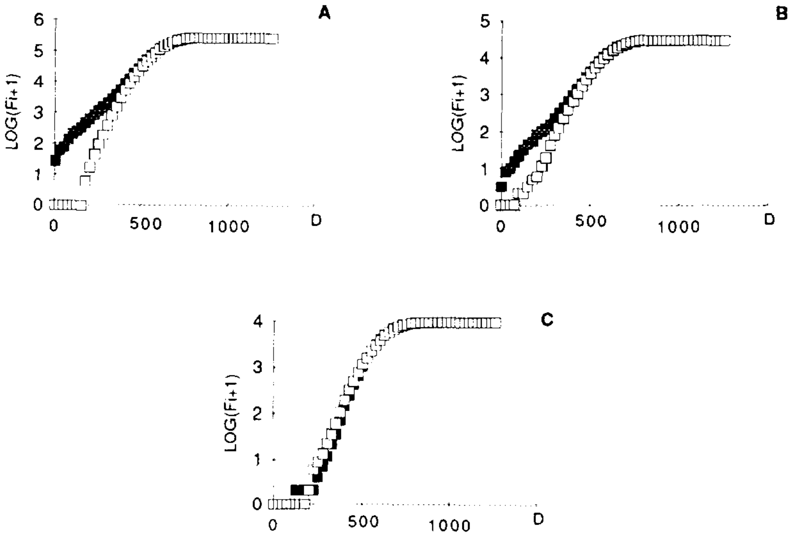

Figure1. Comparison between matches observed for the original sequences (closed squares) and the same sequences after randomization of the residues (open squares). Fi are the cumulated frequencies of the physicochemical distance (D) between segments of (A) all the dehydrogenases of the alignment (only one of the two sequences of D-lactate dehydrogenase of *L. delbrueckii*), (B) D-lactate dehydrogenase from *L. delbrueckii*, D-3-phosphoglycerate dehydrogenase from *E. coli* and vancomycin resistance protein from *E. faecium*, and (C) L-lactate dehydrogenase from *Bacillus stearothermophilus* and D-lactate dehydrogenase from *L. delbrueckii*.

in the space being negatively correlated with their overall similarity. Figure 2 suggests that the D-lactate dehydrogenases and the D-2-hydroxyisocaproate dehydrogenase form a homogeneous group. Among them, we find two versions of the D-lactate dehydrogenase of *Lactobacillus delbrueckii* (8,11) which differ by only 10 substitutions. The remaining five sequences are more dispersed. As tested above, even the less related sequences show significant similarities. The groups detected by factor analysis are then used to perform within- and between group alignments.

The final alignment is presented in figure 3. The boxes outline the matching regions. Those boxes represent alignments within the group discussed above, between the group and the non-grouped sequences, and between the non-grouped sequences. When boxes are superimposed, the greatest intersection is placed on the front. Residues outside the boxes or separated by horizontal lines are not aligned. In the text, the numbering of the amino acids is the one of formate dehydrogenase, and the position in the alignment is indicated within brackets. The alignment clearly shows that regions common to all the sequences are limited to a few but very reliable boxes. Some extra informations about the structure of the formate dehydrogenase, which are not part of the results of the sequence alignment, are also reported. This information, compared with the boxes, points out a high degree of conservation of the coenzyme binding domain, mainly the nicotinamide binding site (ßD to ßF) and the ßA-loop-αB containing the typical GXGXXG(17X)D pattern of the adenosine binding site

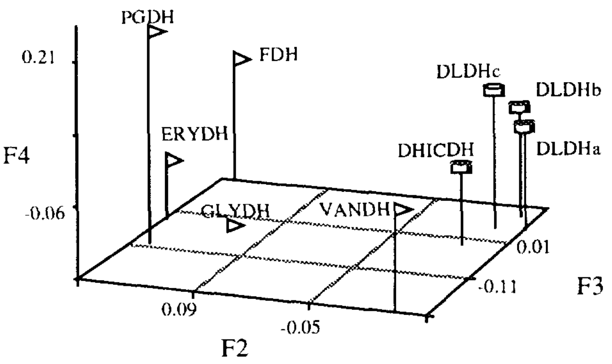

Figure 2. Three-dimensional representation of the factors F2 to F4 of the factor analysis of formate dehydrogenase from *Pseudomonas* (FDH), D-lactate dehydrogenase of *L. delbrueckii* (DLDHa, DLDHb) and from *L. plantarum* (DLDHc), D-2-hydroxyisocaproate dehydrogenase from *L. casei* (DHICDH), vancomycin resistance protein from *E. faecium* (VANDH), glycerate dehydrogenase from cucumber (GLYDH), D-3-phosphoglycerate dehydrogenase from *E. coli* (PGDH), and erythronate 4-phosphate dehydrogenase from *E. coli* (ERYDH). Cylinders represent the grouped sequences, flags the ungrouped sequences.

(24), X representing any amino acid. The connecting parts between the two domains, where the active site is located (19), namely β8-loop-αA and βG-loop-α8 are also highly conserved, as well as the β7-α4 region of the active site. Note that no conservation is pointed out in the α5-α6 region, which is an intersubunit contact region, penetrating deep into the adjacent subunit. Assuming that at least D-lactate and D-hydroxyisocaproate dehydrogenases are dimeric (7, 12, 25), this suggests that the assembly of subunits could be different for each protein, or that this interaction is highly non-specific. We remark the good conservation of the suggested catalytic histidine of formate dehydrogenase, namely H336(371), as well as of D249(271), K274(300), N281(307), R284(310), D289(315) and L297(323). Other residues, like the P97(99)-F98(100) pair of the active site of formate dehydrogenase, do not seem to be conserved, at least in our alignment. The coenzyme binding domain is structurally conserved among many NAD(P)-dependent dehydrogenases (26), including the new D-hydroxyacid dehydrogenase-like family. The typical pattern of coenzyme binding is generally present. Note that the glycine pattern is slightly modified for the glycerate dehydrogenase (GXGXXG(18X)D) and for the vancomycin resistance protein (GXGXXG(17X)SR), the last indicating a possible NADP dependence (22). This region has some structural similarity with other NAD-dependent dehydrogenases, as for L-lactate or malate dehydrogenases (19).

The residues of the formate dehydrogenase active site are aligned with the corresponding residues of the other enzymes. The catalytic site of such dehydrogenases can be schematized as mainly formed by conserved loops and turns, mostly located at the interdomain interface: the residues involved in the catalytic activity are distributed along the whole sequence, and the degrees of conservation are variable. Although the substrates of these enzymes are quite diverse, they all share

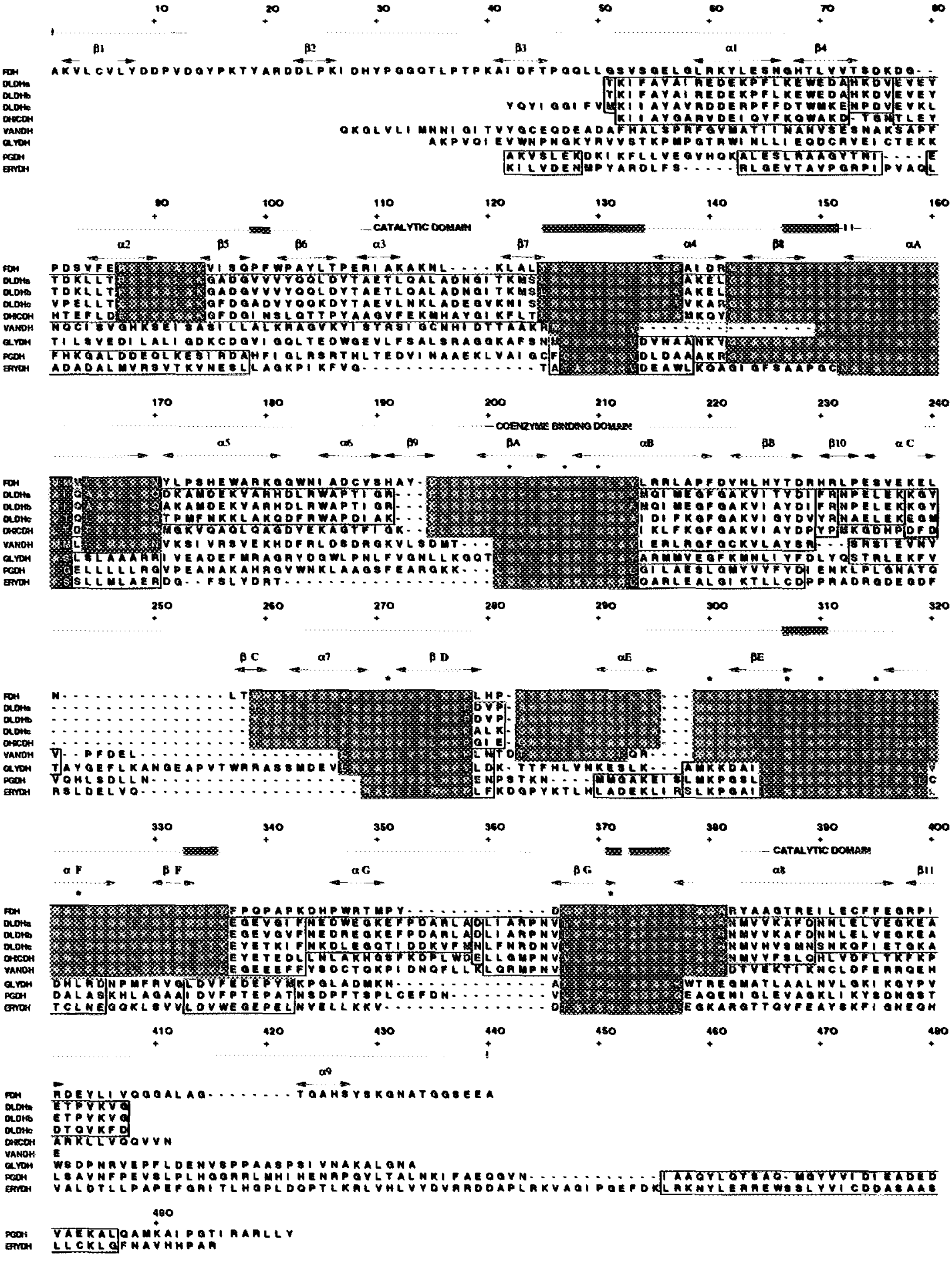

Figure 3. Multiple alignment of FDH with DLDHa, DLDHb, DLDHc, DHICDH, VANDH, GLYDH, PGDH and ERYDH. Abbreviations are the same as in figure 2. Boxes delineate the matching regions. The complete boxes are coloured in red. DLDHa, DLDHb, DLDHc and DHICDH are included in a group (see factor analysis).The boxes between FDH and at least the four grouped sequences are coloured in green. Arrows represent the secondary structure of FDH. The domains and the active site of FDH are represented respectively by a thin and a thick line above the sequence. Asterisks point out the perfectly conserved residues.

a common carboxyl group, except for the substrate of the vancomycin resistance protein which is unknown. This suggests that the conserved R284(310) is a key residue in the binding of the substrate, which can be compared to the binding of pyruvate to the R171 of the L-lactate dehydrogenase (1). This residue is included in a well conserved pattern.

The catalytic histidine/aspartic acid pair, although not being the typical DXXR(23X)H motif found in many L-hydroxyacid dehydrogenases as L-lactate dehydrogenase, L-malate dehydrogenase, L-2-hydroxyisocaproate dehydrogenase, is present in formate dehydrogenase and generally conserved among the aligned sequences. There are several conserved aspartic acid residues which could be candidates for the interaction with the catalytic histidine (D249(271), D289(315)), but this function seems to be assumed in the formate dehydrogenase by D308(335), a residue which is not perfectly conserved and not included in a complete match.

CONCLUSIONS

Several tools of sequence analysis proposed by the software Match-Box allowed us to evaluate the similarity, to group and to align a set of D-lactate dehydrogenases with several related sequences. Our results indicate relevant sequence similarities between D-specific hydroxyacid dehydrogenases and formate dehydrogenase. The structurally conserved regions predicted in the set of sequences analyzed give a good insight into the structural shape of this new family of proteins and suggest that all the regions critical for the formate dehydrogenase activity are highly conserved at least in the D-lactate dehydrogenases. This alignment represents the first necessary step for the knowledge-based modeling of the D-lactate dehydrogenase three-dimensional structure. The comparisons carried out with the L-specific hydroxyacid dehydrogenases confirm the hypothesis that these two sets of proteins are unrelated and probably evolutionary distinct.

ACKNOWLEDGMENT

We thank Dr. V. S. Lamzin (EMBL, Hamburg) for kindly providing the cristallographic coordinates of formate dehydrogenase.

REFERENCES

1. Holbrook, J. J., Liljas, A., Steindel, S. J. and Rossman, M. G. (1975) in The Enzymes (P. D. Boyer, Ed.), Vol. XI, pp. 191-292. Academic Press, New York.
2. Abad-Zapatero, C., Griffith, J. P., Sussman, J. L. and Rossmann, M. G. (1987) J. Mol. Biol. **198**, 445-467.
3. Piontek, K., Chakrabarti, P., Schaer, H.-P., Rossmann, M. G. and Zuber, H. (1990) Proteins: Struct. Funct. Genet. **7**, 74-92.
4. Buehner, M. and Hecht, H. J. (1987) Z. Kristallogr., **178**, 44.
5. Hogrefe, H. H., Griffith, J. P., Rossmann, M. G., and Goldberg, E. (1987) J. Biol. Chem. **262**, 13155-13162.
6. Grau, U. M., Trommer, W. E. and Rossmann, M. G. (1981) J. Mol. Biol. **151**, 289-307.
7. Taguchi, H. and Ohta, T. (1991) J. Biol. Chem. **266**, 12588-12594.
8. Bernard, N., Ferain, T., Garmyn, D., Hols, P. and Delcour, J. (1991) FEBS Lett. **290**, 61-64.

9. Li, S. S. L., Fitch, W. M., Pan, Y. C. E. and Sharief, F. S. (1983) J. Biol. Chem. **258**, 7029-7032.
10. Hediger, M. A., Frank, G. and Zuber, H. (1986) Biol. Chem. Hoppe-Sleyer 367, 891-903.
11. Kochhar, S., Chuard, N. and Hottinger, H. (1992) Biochem. Biophys. Res. Com. **185**, 705-712.
12. Le Bras, G. and Garel, J.-R. (1991) FEMS Microbiol. Lett. **79**, 89-94.
13. Kochhar, S., Hunziker, P. E., Leong-Morgenthaler, P. and Hottinger, H. (1992) Biochem. Biophys. Res. Com. **184**, 60-66.
14. Lerch, H.-P., Blöcker, H., Kallwass, H., Hoppe, J., Tsai, H; and Collins, J. (1989) Gene **78**, 47-57.
15. Greenler, J. McC., Sloan, J. S., Schwartz, B. W. and Becker, W. M. (1989) Plant Mol. Biol. **13**, 139-150.
16. Tobey, K. L. and Grant, G. A. (1986) J. Biol. Chem. 261, 12179-12183.
17. Schoenlein, P. V., Roa, B. B. and Winkler, M. E. (1989) J. Bacteriol. **171**, 6084-6092.
18. Popov, V. O., Shumilin, I. A., Ustinnikova, T. B., Lamzin, V. S. and Egorov, T. A. (1990) Bioorg. Khim. **16**, 324-335.
19. Lamzin, V. S., Aleshin, A. E., Strokopytov, B. V., Yukhnevich, M. G., Popov, V. O., Harutyunyan, E. H. and Wilson, K. S. (1992) J. Biochem. **206**, 441-452.
20. Pearson, W. R. and Lipman, D. J. (1988) Proc. Natl. Acad. Sci. USA **85**, 2444-2448.
21. Arthur, M., Molinas, C., Dutka-Malen, S. and Courvalin, P. (1991) Gene **103**, 133-134.
22. Depiereux, E. and Feytmans, E. (1992) CABIOS **8**, 501-509.
23. Depiereux, E. and Feytmans, E. (1991) Protein Engng. **4**, 603-613.
24. Wierenga, R. K., Terpstra, P. and Hol, W. G. J. (1986) J. Mol. Biol. **187**, 101-107.
25. Hummel, W., Schütte, H. and Kula, M.-R. (1985) Appl. Microbiol. Biotechnol. **21**, 7-15.
26. Eklund, H. and Branden, C.-I. (1987) in Pyridine Nucleotide Coenzymes (Dolphin, D. N. Y., Ed.) pp 51-98. Wiley, New York.