Vol. 128, No. 1, 1985 BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS
April 16, 1985 Pages 470-476

Conserved Amino Acid Sequence Domains in Alpha-Amylases from Plants, Mammals, and Bacteria

John C. Rogers

Departments of Internal Medicine and Biology, Division of Hematology-Oncology, Washington University School of Medicine, St. Louis, MO 63110

Received February 13, 1985

Although alpha-amylases from mammals, plants, and bacteria have common functions, the amino acid sequences of enzymes from these three, evolutionarily distant groups of organisms are not known to share common homologies, and active sites have not been identified. Here I demonstrate that there are three sequence domains common to all alpha-amylases that are aligned and spaced at similar intervals along the length of each protein. The first domain in the barley enzymes appears to contain a calcium binding site. These common domains may represent important functional regions, perhaps the active sites.

Alpha amylases catalyze hydrolysis of α,1-4-linked glucose polymers at most internal bonds and are ubiquitously found in the plant and animal kingdoms. These enzymes were among the earliest to be purified and their mechanisms of action and substrate specificities have been widely studied (1), but, in contrast, very little is known about potential active sites in alpha-amylases. All are metalloenzymes and tightly bind one mole of calcium per mole of enzyme (1). The recent availability of alpha-amylase amino acid sequences derived from cDNA clones offered an approach to identifying functionally important regions of these proteins, since any commonly conserved sequences among evolutionarily very distant organisms might represent critical domains. I therefore compared (2) the amino acid sequences of alpha-amylases from barley (3,4), Bacillus species (5,6), and mammals (7,8). Here I report that all contain three different, well conserved sequence domains spaced at similar intervals; comparison of these domains using an alignment program suggests that the similarities among the different sequences are not due to

0006-291X/85 $1.50

chance. A portion of the first domain in the barley sequences is closely related to established calcium binding sites in other proteins (9,10). The regions between these domains show no apparent homology among the different organisms.

Materials and Methods

Amino acid sequences for the types A and B barley alpha-amylase isozymes were derived from cDNA sequences in this laboratory as described (3,4). All other published alpha-amylase sequences were obtained from the National Biomedical Research Foundation Protein Sequence Databank carried in the Washington University Medical School DEC VAX-11 computer. Computer analysis of sequence homologies utilized the programs of Dayhoff, et al (2); the application of individual programs for specific comparisons is described in Results.

Results and Discussion

The initial screen for homologous regions utilized the RELATE program (2) in which all possible segments of 20 residues in the type A barley amylase sequence were compared to all segments of the same length in other amylase sequences. The rationale was to identify regions that were shared by enzymes from all three groups of organisms: plants, mammals, and bacteria. While the presence of such shared sequences on an initial screen would not in itself prove that they were important, it would allow further anaylsis to determine whether the apparent homologies were likely to be due only to chance. Their spatial organization and similarity to sequences with known functions might then allow an appraisal of whether they were likely to have any functional importance. In this analysis, for any comparison a score is calculated where the numerical value contributed by each match of amino acids is derived from a previously defined matrix of values (2). My analyses used the mutation data matrix, derived from mutational changes in families of closely related sequences (2). Scores obtained from the table of values in RELATE analyses were used to prepare DOTMATRIX (2) output for visual analysis; all 20 residue sequence segments with a match score of >20 (representing the top 0.1% of scores) were plotted, and the results are presented in Fig. 1.

In this figure the results of comparisons between mouse salivary and hepatic preamylase (8) and B. subtilis amylase (5) (Fig. 1,A), between type A

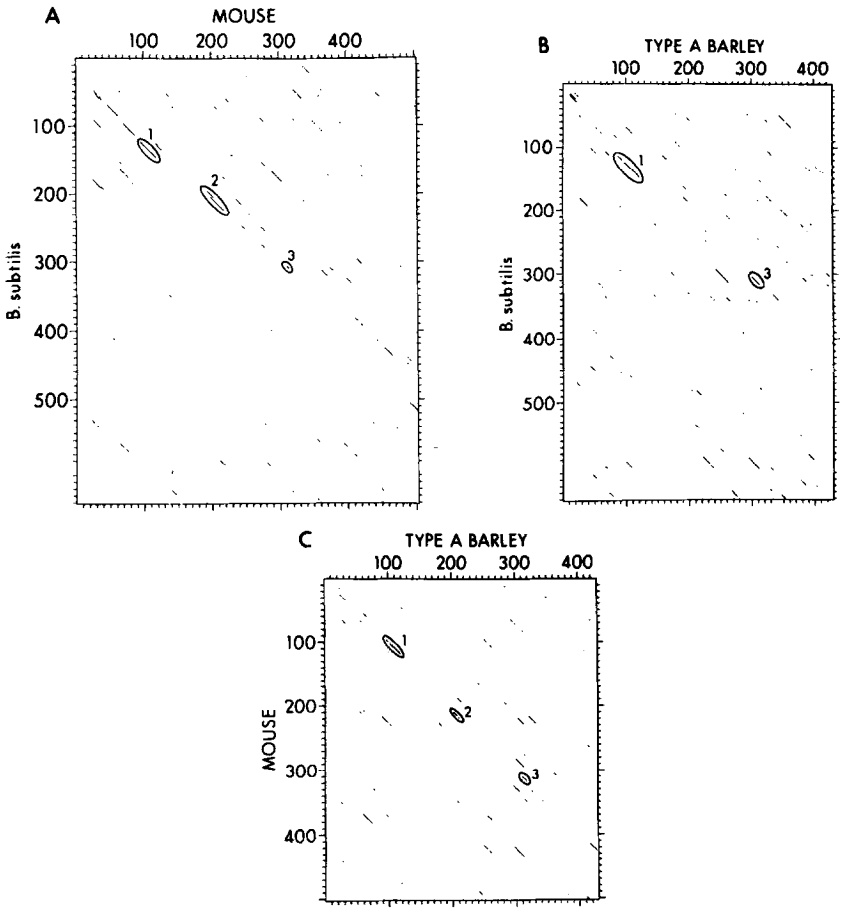

Figure 1: Comparison of the barley type A alpha amylase sequence (3), mouse salivary and hepatic preamylase(8) and B. subtilis amylase (5) to each other utilizing the DOTMATRIX program as detailed in the text. Circled are regions (identified by positions as 1, 2, and 3) that were shared on at least two comparisons.

barley and B. subtilis (B), and between type A barley and mouse preamylases (C) are presented. The purpose of the comparisons was to identify regions that were shared in common by the three types of amylase sequences. It can be seen that there are three regions of similarity shared by at least 2 of the 3 sequences (numbered 1-3 in panels A-C); these regions align on a diagonal indicating that they are spaced at similar intervals in each of the sequences. There is some background in the plots, but only regions 1-3 were reproducibly present on more than one comparison and demonstrated significant homology with ALIGN (see below). Similar results were obtained in comparisons with pig pancreatic amylase (7) (data not presented).

In order to determine whether the presence of these conserved regions might be due to chance alone, they were compared with the ALIGN program (2). This program aligns two sequences optimally and calculates a score for that match utilizing matrix data as described for RELATE. That match score is compared with the distribution of maximum scores for 100 random permutations of the two sequences, and a final alignment score is derived; the alignment score represents the number of standard deviations by which the real sequence match score exceeds the average random match score. Dayhoff, et al (2) provide data on the probabilities of scores for a normal distribution, from which a statement regarding the significance of a given match can be derived. To put these comparisons in perspective, when the entire sequences of the B. subtilis, mouse, or porcine enzymes are aligned with the type A barley sequence using this program, the alignment obtained with the B. subtilis sequence is significantly (3.4 standard deviation units) above the mean random score, but the alignment scores (1.6 and -0.9 respectively) for the other two are not.

The data from these comparisons are presented in Fig. 2. To provide a means of comparison that differed from the matrix utilized in the DOTMATRIX plots, these alignments were scored with the genetic code matrix, which reflects the similarities of the codons utilized for each amino acid (2); similar results were obtained with the mutation data matrix (data not presented). To provide a common basis for comparison, in each set the alignments were made with the type A barley amylase sequence, and the alignment score (in standard deviation units) is presented at the end of the sequence utilized for comparison. Since the type B barley amylase sequence is very similar, only the amino acids that differ are listed above the type A sequence. It should be noted that Chandler, et al (9), obtained a sequence for their type B cDNA clone that differs at positions 195-197: Ile-Gly-Phe; these residues are indicated in parentheses in Fig. 2B. The numbers on either end indicate the positions of the first and last residues that define the particular "domain". Where sequences from two or more different organisms

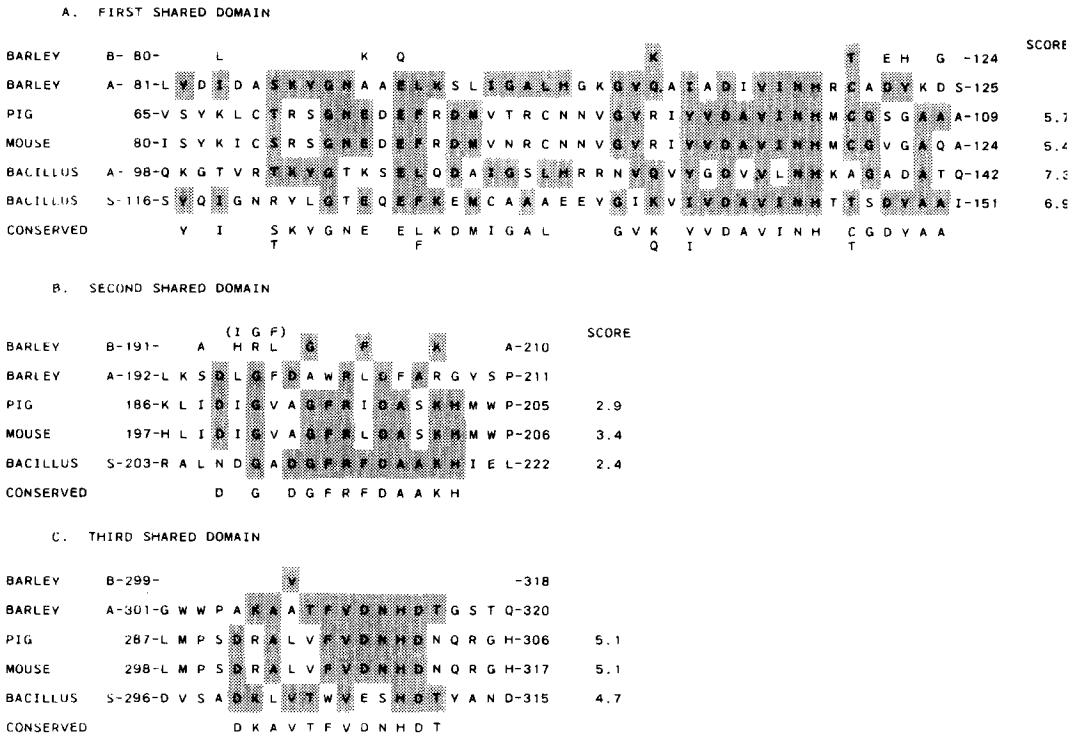

Figure 2: Alignment of the homologous regions identified as in Fig. 1. The protein sequences utilized are Barley type B (4) and type A (3), pig pancreatic (7), mouse hepatic and salivary (8), Bacillus amyloliquifaciens (Bacillus A) (6), and Bacillus subtilis (Bacillus S) (5) alpha-amylases. The numbers at either end of the sequences represent the positions of the first and last residues, respectively. The genetic code matrix (2) was utilized for scoring with a gap penalty of 8, but similar results were obtained using the mutation data matrix (data not presented). The alignment score is explained in the text and represents the results of a comparison between that sequence and the corresponding type A barley sequence. The "conserved" sequence represents residues that are present at that position in two or more of the different types of organisms.

(e.g. plant and mammal, or plant and bacteria) contained the same residue at a given position, that residue is presented at the bottom of each section ("conserved").

The alignment scores obtained from these comparisons indicate that the probabilities that they were due to chance alone range from $<10^{-2}$ (score 2.4) to $<10^{-12}$ (score 7.3). This fact, coupled with the observation that these "domains", in each instance, are spaced at similar positions and similar intervals along the length of each of the proteins, strongly suggests that they may have some important functional role. In this regard it is

PROTEIN	X	Y	Z	-Y	-X	-Z	
TYPE A AMYLASE	83-D	I	D	A	S	K	Y G N A A E- 94
CALMODULIN	93-D	K	D	G	N	G	Y I S A A E-104
TROPONIN C	105-D	K	N	A	D	G	Y I D L E E-116
MYOSIN L CHAIN	137-D	K	E	G	D	T	V G M G A E-148
B. SUBTILIS AMYLASE	116-Y	Q	I	G	N	R	Y L G T E Q-127
PIG PANCREATIC AMYLASE	67-Y	K	L	C	T	R	S G N E D E- 78

Figure 3: Comparison of known calcium binding domains to a portion of the conserved domain 1 of the barley amylases. The calmodulin, troponin C, and myosin light chain sequences were pulled from the Protein Sequence Databank in register with the barley sequence on the top line (text). The corresponding sequences from the porcine and B. subtilis enzymes are presented at the bottom.

interesting to note that carboxyl and imidazole side chains are thought to be important functional groups in porcine pancreatic amylase (1). In both domain 1, corresponding to type A residues 122 and 127, and domain 3, corresponding to residues 315 and 314, all of the sequences have aspartic acid and histidine, respectively. The availability of the cloned cDNAs may permit site specific mutagenesis/expression vector experiments in the future to test whether these are in fact part of the enzymatic active sites.

In a further attempt to define functions for these domains, the sequences were used to search the National Biomedical Research Foundation Protein Data Bank using the SEARCH (2) program. In each instance the different domain sequences selected corresponding amylase sequences, as expected, but there was no other pattern for functional classes of proteins selected by any set of domain sequences, with one exception. Amino acids 81-105 from the type A barley amylase (the N-terminal portion of domain 1) selected calcium-binding sequences from calmodulin (10), troponin C (11), and myosin light chains (12); these sequences were in register with the barley residues presented in Fig. 3. In this figure, the notation at the top indicates the ligands coordinating with calcium as octahedral vertices. These were identified from x-ray crystallographic studies of carp muscle calcium-binding parvalbumin (reviewed in 13) and applied to three-dimensional models of troponin C (13) and calmodulin (14) from sequence homologies. Outstanding common features are

aspartic acid residues at X and Y coordinates and glutamic acid at the -Z coordinate (13,14). The residue between the Y and Z coordinates is usually glycine, but alanine may substitute (14); serine at Z and tyrosine at -Y are not uncommon (13,14). It is therefore reasonable to hypothesize that the barley amylase sequence aligned at the top may be the site responsible for binding the essential Ca^{2+} ion. The corresponding sequences from B. subtilis and pig pancreatic amylases (Fig. 3, bottom) do not conform to the models described above for established calcium binding sites, although some of the features are present; the absence of all features would not preclude a calcium binding function, however (13). The prospect of high resolution x-ray crystallographic data from the porcine enzyme (15) should answer this question in the future.

Acknowledgements

This work was supported in part by grant No. 83-CRCR-1-1332 from the U.S.D.A. I thank Drs. Jens Birktoft and Mark Boguski for helpful comments regarding analysis of sequence homologies.

References

1. Thoma, J. A., Spradlin, J. E., and Dygert, S. (1971) Meth. Enzymol. 5,115-189.
2. Dayhoff, M. O., Barker, W. C., and Hunt, L. T. (1983) Meth. Enzymol. 91 524-545.
3. Rogers, J. C., and Milliman, C. (1983) J. Biol. Chem. 258, 8169-8174.
4. Rogers, J. C. J. Biol. Chem., in the press.
5. Yang, M., Galizzi, A., and Henner, D. (1983) Nucl. Acids Res. 11, 237-249.
6. Chung, H. S., and Friedberg, F. (1980) Biochem. J. 185, 387-395.
7. Kluh, I. (1981) FEBS Lett. 136, 231-234.
8. Hagenbuchle, O., Bovey, R., and Young, R. A. (1980) Cell 21, 179-187.
9. Chandler, P. M., Zwar, J. A., Jacobsen, J. V., Higgins, T. J. V., and Inglis, A. S. (1984) Plant Molec. Biol. 3, 407-418.
10. Putkey, J. A., Ts'ui, K. F., Tanaka, T., Lagace, L., Stein, J. P., Lai, E. C., and Means, A. R. (1983) J. Biol. Chem. 258, 11864-11870.
11. van Eerd, J.-P., and Takahashi, K. (1976) Biochemistry 15, 1171-1180.
12. Frank, G., and Weeds, A. G. (1974) Eur. J. Biochem. 44, 317-334.
13. Kretsinger, R. H. (1976) Ann. Rev. Biochem. 45, 239-266.
14. Dedman, J. R., Jackson, R. L., Schreiber, W. E., and Means, A. R. (1978) J. Biol. Chem. 253, 343-346.
15. Payan, F., Haser, R., Pierrot, M., Frey, M., Astier, J. P., Abadie, B., Duee, E., and Buisson, G. (1980) Acta Cryst. B36, 416-421.