Sequence similarity between *Rhodospirillum rubrum* ribulose bisphosphate carboxylase/oxygenase and the large subunit of the plant enzyme

Gerald R. Reeck and David C. Teller*

Department of Biochemistry, Kansas State University, Manhattan, KS 66506 and **Department of Biochemistry, University of Washington, Seattle, WA 98195, USA*

Received 8 February 1983

Computerized methods were used to analyze published sequence information from *Rhodospirillum rubrum* ribulose-1,5-bisphosphate carboxylase/oxygenase (RUBPCase) and the sequence of the large subunit of spinach RUBPCase. The sequences of 7 peptides from the bacterial enzyme were compared to the entire sequence of the spinach large subunit to find optimal alignments and test statistical significance. Several of the *R. rubrum* RUBPCase peptides align unambiguously with the spinach sequence, and the alignment of the largest peptide is clearly significant in a statistical sense. The total of 91 positions aligned with *R. rubrum* peptides includes 30 identities. Our analysis strongly suggests that the entire sequence of the *R. rubrum* enzyme, when completed, will be found to exhibit statistically significant similarity to the large subunit of plant RUBPCase.

Ribulose bisphosphate carboxylase/oxygenase *Spinach* *Rhodospirillum rubrum* *Computerized sequence comparison* *Sequence similarity* *Homology*

1. INTRODUCTION

Ribulose-1,5-bisphosphate carboxylase/oxygenase, henceforth referred to as RUBPCase, occurs in both eukaryotic and prokaryotic organisms. It is ubiquitously distributed in plants, where it occurs in the chloroplasts. The plant enzyme is composed of 16 subunits [1]: 8 large (M_r 56 000) and chloroplast DNA-encoded; and 8 small (M_r 15 000) and nuclear DNA-encoded. The large subunit contains the enzyme's active site [2], and there is no known function for the small subunit. RUBPCase from photosynthetic bacteria differ in subunit organization. That from *Rhodospirillum rubrum* (the most thoroughly studied bacterial RUBPCase) occurs as a dimer of two similar (possibly identical) subunits, each of M_r 56 000 [3].

R. rubrum RUBPCase is an attractive object for study because of its relatively simple architecture and because of the potential for genetic manipulation of the organism. How relevant studies on the *R. rubrum* enzyme will be for understanding plant RUBPCase, however, depends upon the extent of similarity between the enzymes from the two sources. There are good reasons to expect a homologous relationship between the *R. rubrum* RUBPCase and the plant large subunit: they catalyze the same reactions and presumably serve the same physiological roles. The *R. rubrum* enzyme's subunit is nearly the same size as the catalytic large subunit of the plant RUBPCase. The C-3 fixation of CO_2 carried out by RUBPCase is such a fundamentally important reaction that one would expect it to have been functional in early life forms, including the common ancestor to higher plants and photosynthetic bacteria.

Nevertheless, with the 7 peptides that have been isolated and sequenced from *R. rubrum* RUBPcase [4–6], which constitute >20% of the total sequence, little similarity to the plant large subunit sequences has been apparent from visual inspection. We have re-investigated the possibility of sequence similarity between the *R. rubrum* RUBPCase and the large subunit of plant RUBPCase us-

Published by Elsevier Science Publishers

00145793/83/0000–0000/$3.00 © Federation of European Biochemical Societies

ing computerized methods developed in [7]. These methods are capable of detecting weak similarities that are not obvious by visual inspection and, just as importantly, of providing a rigorous evaluation of the statistical significance of a proposed alignment. Our analysis strongly suggests that the *R. rubrum* enzyme is in fact similar in a statistically significant way to the large subunit of plant RUBPCase.

2. METHODS

The plant RUBPCase sequence that we used was that inferred [8] from the nucleotide sequence of the spinach gene for the large subunit. To that protein's sequence we compared the 5 Cys-containing peptides of [4] and the pyridoxal phosphate-modified peptides of [5,6]. We will refer to the Cys-containing peptides as peptides 1–5, starting with the peptide listed first in the summary of [4]. We extended peptide 2 by 6 residues on its NH_2-terminus using sequence information obtained [9] from automated Edman degradation of the intact *R. rubrum* enzyme.

In quantifying alignments, we used 3 schemes that assign scores to all pairs of amino acid residues. The schemes of McLachlan [10] and Dayhoff [11] are each based on the frequency with which amino acids have been interchanged in the evolution of homologous proteins. A third scoring scheme we have used is based on minimum base differences (MBD's) between amino acid residues [12]. As in [10], we have found this scoring system considerably less sensitive in general than are the McLachlan and Dayhoff systems.

We used two methods to find and evaluate alignments between the *R. rubrum* peptides and the spinach large subunit. The one used most in this work we call a 'peptide search'. This is in essence a specialized form of the 'diagonal search' that we used to compare portions of two large sequences [7]. In the peptide search, a peptide's sequence is compared to all sequences of its length contained in a much longer sequence, in this case that of the spinach RUBPCase large subunit. We test the statistical significance of an alignment obtained in a peptide search by a Monte Carlo (randomization) procedure similar to that which we have used for other tests of statistical significance (details in [7]). Comparison of alignments between the real sequences with alignments between computer-generated sequences of the same lengths and amino acid compositions allows us to calculate an estimated level of significance as in [7]. The level of significance is the probability that a score equal or greater than the score of a proposed alignment could have resulted simply from chance. When peptide 3 was gapped (section 3), alignment of an amino acid residue from the large subunit sequence with a dash (the symbol for a 1-residue gap) was assigned a score of 0. The dashes were retained in the randomization of the gapped peptide and scored in the same way when aligned to the randomized large subunit sequences.

The Sankoff algorithm [13], modified as in [7] for comparison of protein sequences, is a second method used in this work. Given any two sequences, the Sankoff algorithm finds the optimal alignment between them under a constraint, imposed by the investigator, on the number of allowed gaps. It is also used in a Monte Carlo approach to evaluate the statistical significance of an alignment and to determine the number of gaps justified on statistical grounds [7].

3. RESULTS

The best evidence for sequence similarity between the *R. rubrum* RUBPCase and the large subunit of spinach RUBPCase is provided by peptide 3, the longest (32 residues) of the Cys-containing peptides isolated and sequenced in [4]. Peptide 3 in its entirely aligns well with residues 46–77 of the spinach large subunit. Of the 444 possible alignments of the 32-residue peptide with the spinach protein, that alignment is the best in both the McLachlan and MBD scoring schemes. The level of significance of the alignment, estimated by comparison with scores from randomized sequences of the lengths and compositions of peptide 3 and the spinach large subunit, is 0.003 or 0.008 under the McLachlan or MBD scoring schemes, respectively. Thus, as judged by use of both of these scoring systems, the alignment of peptide 3 with residues 46–77 of the spinach large subunit is clearly statistically significant.

With the Dayhoff scores, the alignment of peptide 3 with positions 46–77 ranks 8th among all possible alignments. That it does not rank higher is due in part to the relatively low scores in the

Dayhoff scheme for certain identities (particularly Ala and Ser identities) and in part from the poor quality of the alignment of the COOH-terminal 5 residues of peptide 3 with positions 73–77 of the spinach large subunit. If a 2-residue gap is inserted into peptide 3 before its last 5 residues, the peptide search program carried out on the 34-residue, gapped peptide identifies the alignment with positions 46–79 of the spinach protein as the best under each of the 3 scoring schemes. We have no statistical justification for introducing this 2-residue gap, but we include it in the alignment shown in fig. 1 because of the satisfying way in which it brings the results of the 3 scoring schemes into agreement on the placement of peptide 3. A final alignment of the last 5 residues of that peptide must await completion of the *R. rubrum* RUBPCase sequence. Their placement is not central, however, to the conclusion we want to draw: peptide 3 can be aligned in a statistically significant way with a portion of the sequence of the spinach RUBPCase large subunit.

Peptide 2 (extended as in section 2) is known to be the NH_2-terminal sequence of *R. rubrum* RUPBCase [4, 9]. All of peptide 2 must be aligned with residues 1–45 of the spinach protein, since residue 46 is the start of the alignment with peptide 3. The Sankoff algorithm is ideally suited to find the optimal alignment between peptide 2 and residues 1–45 of the spinach large subunit. Using the McLachlan scoring scheme with randomizations as a guide to gapping [7], we found that a 2-gap alignment is best (fig. 1). It is not good enough to be statistically significant, however: level of significance = 0.38. The strongest portion of the alignment involves the last 8 residues of peptide 2. The peptide search program, when carried out with that 8-residue segment, demonstrated that its alignment with residues 33–40 of the spinach large subunit (as shown in fig. 1) has the top scores among all possible alignments under the Dayhoff and McLachlan schemes. (In the latter, this alignment was tied in score with one other alignment, however.) If we exclude all alignments that involve portions of the large subunit sequence after position 45, the alignment of the last 8 residues of peptide 2 with residues 33–40 is the best alignment in all 3 scoring schemes. We therefore believe that this alignment is correct and will not be affected by further sequence information. We note that completing the *R. rubrum* sequence will not provide any additional information on the placement of residues 1–20 of peptide 2. Further insight into their alignment will come only with the elucidation of other bacterial RUBPCases.

Of the remaining Cys-containing peptides from

```
                                            10                  20                  30                        40
Spinach large subunit:     M S P Q T E T K A S V E F K A G V K D Y K L T Y Y T P E Y E T L - - - D T D I L A A F
                                 | : :       : : : : | :                             |   |         : : : |   | :
R. rubrum Peptide 2:         M D Q S S R Y V N L A L K E - - - - - - - - - - - - - - E D L I A G G E H V L C A Y

                                            50                  60                  70                  80
Spinach large subunit:     R V S P Q P G V P P E E A G A A V A A E S S T G T W T T V W T D G L T N L D R Y
                                     : |         :     |     | | | | | | | |     : |   |         : :   |
R. rubrum Peptide 3:                 A G Y G Y V A T A A H F A A E S S T G T D V E V C T T - - B B F T R

                                           380                 390
Spinach large subunit:     T P G V L P V A S G G I H V W H M P A L
                                     | :   | | | : :     :
R. rubrum Peptide 4:           A C T P I I S G G M N A L R

                                           160                 170                 180
Spinach large subunit:     P P H G I Q V E R D K L N K Y G R P L L G C T I K P K L G L
                                       :   |   : : :     |     : : |     | | | |
R. rubrum Lys-peptide:               V L G R P E V D - - G G L V V G T I I K P K
```

Fig. 1. Alignment of 4 *Rhodospirillum rubrum* RUBPCase peptides with sequences in the large subunit of spinach RUBPCase. Peptides 2, 3 and 4 are Cys-containing peptides isolated and sequenced in [4]. The 'Lys-peptide' is the pyridoxal phosphate-modified peptide isolated and sequenced in [6]. In the alignments, pairs of identical residues are indicated by solid lines. Dashed lines connect pairs of residues that have higher-than-average scores in the McLachlan scoring scheme [10].

R. rubrum RUBPCase, only peptide 4 can be aligned with some confidence with the spinach large subunit. The alignment shown in fig. 1 was the best for this peptide under the Dayhoff and McLachlan scoring schemes and was tied in score with one other alignment in the MBD scoring scheme. Randomizations gave estimated levels of significance of 0.16, 0.08 and 0.45 in the Dayhoff, McLachlan and MBD scoring schemes, respectively. The alignment is thus marginally significant at best, but the fact that the 3 scoring schemes agree on the best placement of peptide 4 suggests that its alignment as shown in fig. 1 will not be changed when the entire *R. rubrum* RUBPCase sequence is known.

The active site peptide of [5] was placed in different top-scoring alignments in the McLachlan and Dayhoff scoring schemes. Each 'best' alignment was second best in the other scoring scheme, however. The two alignments match the NH_2-terminal residue of the peptide against residues 320 and 432 of the spinach large subunit. In neither of these alignments is the active-site Lys residue of the peptide matched with a basic residue, but in both alignments it is one residue removed from a basic residue in the spinach enzyme (His-325 or Arg-439).

In [6] an active-site peptide was isolated from *R. rubrum* RUBPCase that, like the peptide in [5], contains a pyridoxal phosphate-modified Lys. In the peptide of [6], this Lys is part of a 4-residue stretch (KPLK) that matches exactly a 4-residue sequence of the spinach enzyme, the first residue of which is also modified by pyridoxal phosphate. On the basis of the chemical modification results, one should align the *R. rubrum* peptide of of [6] with the region of the spinach large subunit sequence that contains the KPLK tetrapeptide. An unambiguous alignment would not be possible without this chemical modification information, however, since the alignment with this region of the spinach large subunit competes (unfavorably at some peptide lengths) with an alignment of the active-site peptide starting at residue 320 of the spinach enzyme. The 4-residue identity (KPLK) is not in itself a definitive criterion for positioning the active-site peptide of [6] with the spinach large subunit. The level of significance of the 4-residue identity is 0.11 or 0.05 under the McLachlan or Dayhoff scheme, respectively. (These estimates were obtained for an alignment of length 4 residues from a peptide of the length and composition of the active-site peptide of [6] with a sequence of the length and composition of residues 80–475 of the spinach large subunit.) The 1-gap alignment shown in fig. 1 was found by the Sankoff algorithm, operating on the active-site peptide and on residues 151–180 of the spinach large subunit. The strongest portion of the alignment (involving the COOH-terminal 12 residues of the peptide) aligns with a level of significance of 0.14 or 0.02 under the McLachlan or Dayhoff schemes, respectively. The alignment of the NH_2-terminal 8 residues of the peptide is not statistically significant.

4. DISCUSSION

Our analyses suggest strongly that *R. rubrum* RUBPCase is more similar in amino acid sequence to the large subunit of spinach RUBPCase than would be expected from chance. This supports the hypothesis that the two proteins evolved from a common ancestor. The level of similarity is far lower than that among the large subunit sequences of plant RUBPCase [8] and falls in a range that is difficult to detect by visual inspection. Similarity at this level when extended over large stretches of sequence can, however, achieve a very high level of statistical significance. It is not merely coincidental that the most convincing evidence for statistically significant similarity between the bacterial and plant enzymes is obtained with the longest peptide of known sequence (peptide 3) from the bacterial protein. Ambiguities in aligning some of the other peptides does not undermine the notion that the complete sequences of the *R. rubrum* and plant enzyme will exhibit similarity that is very significant from a statistical standpoint. Indeed, for sequences with the level of similarity that appears to exist between the *R. rubrum* and plant RUBPCases, poor statistical significance is expected in attempting to align a relatively small peptide from one protein against a part of the entire sequence of the other protein. Such ambiguities will be largely, if not entirely, eliminated under the constraints established when the *R. rubrum* sequence is completed: segments of the sequence must maintain the same order in the alignment as they do in the intact protein sequence. At this time, only in the case of peptide 2 (the NH_2-terminal peptide) does such a constraint hold. The chemical modification information establishes a similar constraint on the alignment of the reactive site peptide [6].

Our analysis leads to a prediction that ought to be readily tested. Automated Edman degradation of the intact *R. rubrum* RUBPCase, if carring ≥ 35 residues, should extend into the sequence of peptide 3. Such a finding would confirm our alignment of peptide 3 and further substantiate the hypothesis of sequence similarity between the *R. rubrum* RUBPCase and the large subunit of plant RUBPCases.

ACKNOWLEDGEMENTS

This work was supported by NIH grant GM-13401 to D.C.T., by Kansas Agricultural Experiment Station, and by NIH grants GM-29203 and CA-17782 to G.R.R. who is the recipient of NIH Research Career Development Award CA-00425. Publication number 83-33-J, Kansas Agricultural Experiment Station.

REFERENCES

[1] Rutner, A.C. (1970) Biochem. Biophys. Res. Commun. 39, 923–929.

[2] Nishimura, M. and Akazawa, T. (1973) Biochem. Biophys. Res. Commun. 54, 842–848.

[3] Tabita, F.R. and McFadden, B.A. (1974) J. Biol. Chem. 249, 3459–3464.

[4] Stringer, C.D., Norton, I.L. and Hartman, F.C. (1981) Arch. Biochem. Biophys. 208, 495–501.

[5] Robison, P.D., Whitman, W.B., Waddill, F., Riggs, A.F. and Tabita, F.R. (1980) Biochemistry 19, 4848–4853.

[6] Herndon, C.S., Norton, I.L. and Hartman, F.C. (1982) Biochemistry 21, 1380–1385.

[7] De Haën, C., Swanson, E. and Teller, D.C. (1976) J. Mol. Biol. 106, 639–661.

[8] Zurawski, G., Perrot, B., Bottomley, W. and Whitfeld, P.R. (1981) J. Nucleic Acids Res. 9, 3251–3270.

[9] Schloss, J.V., Phares, E.F., Long, M.V., Norton, I.L., Stringer, C.D. and Hartman, F.C. (1979) J. Bacteriol. 137, 490–501.

[10] McLachlan, A.D. (1971) J. Mol. Biol. 61, 409–424.

[11] Dayhoff, M.O. (1972) Atlas of Protein Sequence and Structure, vol. 5, National Biomedical Research Foundation, Silver Springs MD.

[12] Cantor, C.R. and Jukes, T.H. (1966) Proc. Natl. Acad. Sci. USA 56, 177–184.

[13] Sankoff, D. (1972) Proc. Natl. Acad. Sci. USA 69, 4–6.