The Primary Structure of Rubrerythrin, a Protein with Inorganic Pyrophosphatase Activity from *Desulfovibrio vulgaris*

**COMPARISON WITH HEMERYTHRIN AND RUBREDOXIN**

(Received for publication, June 5, 1991)

Jozef J. Van Beeumen†‡, Gonzalez Van Driessche§, Ming-Yih Liu¶, and Jean LeGall†

From the †Laboratory of Microbiology, State University of Ghent, Ghent 9000, Belgium and the §Department of Biochemistry, School of Chemical Sciences, University of Georgia, Athens, Georgia 30602

The complete polypeptide chain of rubrerythrin from the sulfate reducing bacterium *Desulfovibrio vulgaris*, strain Hildenborough NCIB 8303, was found by protein chemical techniques to consist of 191 residues and to have the amino acid sequence MKSLKSRTEKNILTAFAGESQRNRNYFQGQAKKDGFVQISDIF-AETADQEREHAKRLFKPLEGGLEIVAAFPAGA-GIADTHANIIAAGAHETYMFPSAFRARY-EEGYBIELRVFASIAVACEFHEKRPFLDFARNK-KEGVRFLREQATKWRNCNGYVHEGTAPEL-CPACAHPKAHFELLGINW. The C-terminal part of the protein (position 153 → 191) shows the typical sequence features of rubredoxin, a protein with a non-native iron center also present in the same and other *Desulfovibrio* species. Based on the known three-dimensional structure of *D. desulfuricans* rubredoxin, we propose that the C-terminal part of rubrerythrin is folded in a similar way and suggest that the deletion of the extra 10 residues is compatible with the same basic rubredoxin-fold.

After characterization of the C-terminal region, and in contrast to what could be expected from previously published spectroscopic analyses, the N-terminal region 1–152 of rubrerythrin appears to have no sequence similarity with the eukaryotic protein hemerythrin which is known to contain a binuclear iron center bound by 5 histidine ligands. However, the N-terminal region of rubrerythrin does contain 5 histidine residues but they are differently spaced along the polypeptide chain. We suggest that at least one of the 5 histidine residues located in the rubrerythrin-like center of rubrerythrin may be liganded to one iron atom of the hemerythrin-like center. This paper is the first sequence report of a protein with pyrophosphatase activity although the physiological substrate for the rubrerythrin may be no inorganic pyrophosphate.

Inorganic pyrophosphatase is an essential enzyme for the activation of sulfate by sulfate reducing bacteria (1). Pyrophosphatase itself is formed along with adenosyl-phosphosulfate from sulfate and one molecule of ATP through action of ATP-sulfurylase:

\[
\text{SO}_4^{2-} + \text{ATP} \rightarrow \text{APS} + \text{PP}_i
\]

In order for the reaction to proceed, pyrophosphate must be hydrolyzed by pyrophosphatase to two molecules of inorganic phosphate. A recent survey of inorganic pyrophosphatase in *Desulfovibrio vulgaris* has shown that two enzymes are responsible for this activity (2). One enzyme has a very high turnover and contains a dialyzable atom of zinc; the other, which is 35 times less active, is “rubrerythrin.”

The name of rubrerythrin was given to this protein because an extensive spectroscopic study has shown that it contains two rubredoxin-like centers and one hemerythrin-like binuclear iron center in a dimer of 43,800 molecular weight (3). Rubredoxin is a very thoroughly characterized nonheme iron center containing protein in biology (4). It is the simplest member of the iron-sulfur class of metalloproteins and serves as the prototype of the proteins that bind high-spin iron in the tetrahedral coordination by sulfur atoms. Its small size \((M_r \leq 6000)\) has made rubredoxin a particularly attractive candidate for a variety of physical and chemical studies (4–6). As far as we know, rubredoxins only occur in anaerobic bacteria (4). Hemerythrin is another relatively small size iron center containing protein that has been thoroughly studied (7). Its polypeptide chain has a molecular mass of 13,500 daltons and contains a binuclear iron center. Hemerythrin constitutes the prototype of a whole class of binuclear iron sites in proteins with different functions such as ribonucleotide reductase (8), purple acid phosphatase (9), protein A of methane-monooxygenase (10) and the formation of the ferriin core (11). This type of protein has not yet been found in bacterial species. Mössbauer and EPR data of rubrerythrin have indicated that the rubredoxin centers are indistinguishable from those in rubredoxin (3). Redox titrations monitored by EPR, however, have shown that they have a midpeak redox potential of 230 mV (±10 mV) which is 250 mV higher than that in normal rubredoxins.

In the present paper, we report the complete primary structure of rubrerythrin which allows to make some interesting comparisons with both rubredoxins and hemerythrin.

**EXPERIMENTAL PROCEDURES AND RESULTS**

Fig. 1 shows the complete amino acid sequence of *D. vulgaris* rubrerythrin. The upbuilding of this proposal was as follows.

1 Portions of this paper (including “Experimental Procedures” and portions of “Results,” Figs. S.1 and S.2, and Tables S.1–VII) are presented in miniprint at the end of this paper. Miniprint is easily read with the aid of a standard magnifying glass. Full size photocopies are included in the microfilm edition of the Journal that is available from Waverly Press.
Initially, N-terminal sequence analysis of the native protein allowed to determine 46 out of the first 51 residues. Chemical cleavage of the apoprotein with cyanogen bromide and dilute formic acid (Fig. 2) were then chosen as the first method to obtain peptides because compositional analysis of the protein (Table I) revealed the presence of only 2 or 3 methionine and 12 or 13 aspartic acid residues, feeding the hope to obtain large peptides. The choice appeared to be justified since a major hydrolysis peptide (H8) not only could be sequenced for 44 steps but also appeared to give an overlap with the last 12 identified residues of one major cyanogen bromide peptide (CN2) which itself could be sequenced for 48 cycles. The first major cyanogen bromide peptide started off with the second residue of the protein so that it could be concluded that the second methionine in rubrerythrin occurred further down the polypeptide chain than residue 63, also because peptide H6 had prolonged the knowledge of the N-terminal region up to that position. Sequence analysis of peptide H10 indicated an overlap from Edman cycle 20 onwards with peptide CN2, and

1 The abbreviations used are: HPLC, high performance liquid chromatography; PTH, phenylthiohydantoin amino acid; PDMS, plasma desorption mass spectrometry; SDS, sodium dodecyl sulphate; TFA, trifluoroacetic acid.

FIG. 1. Amino acid sequence of rubrerythrin from *D. vulgaris*. Peptides obtained by chemical cleavage with cyanogen bromide and dilute formic acid and by enzymatic cleavage with Glu-C endoproteinase and Lys-C endoproteinase are indicated as CN, H, S, and K, respectively. The indication K’ refers to a second digest with the latter protease using more enzyme for a longer incubation time. The numbering of the peptides is that of the order of elution in the HPLC chromatograms (Figs. 2-5 and Fig. S.1). Horizontal arrows indicate the assignment made from automated sequence analysis. The small vertical arrows mean that the sequence runs were deliberately stopped. The cysteines at the positions 158, 161, 174, and 177 were identified as the pyridylethyl derivatives.

FIG. 2. Separation of the peptides obtained from partial acid hydrolysis of 45 nmol of apoprotein. For detailed information, see the Miniprint Section.
TABLE I

Amino acid analysis of rubrerythrin

The results were obtained after hydrolysis during 24 h in the absence (A) and the presence (B) of dodecanethiol. C is the composition deduced from the amino acid sequence. The analyses were carried out on, respectively, 115 and 150 pmol of material.

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>A (%)</th>
<th>B (%)</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asp</td>
<td>12.05</td>
<td>13.09</td>
<td>13</td>
</tr>
<tr>
<td>Glu</td>
<td>27.39</td>
<td>27.51</td>
<td>27</td>
</tr>
<tr>
<td>Ser</td>
<td>6.03</td>
<td>6.15</td>
<td>7</td>
</tr>
<tr>
<td>Gly</td>
<td>14.99</td>
<td>15.43</td>
<td>15</td>
</tr>
<tr>
<td>His</td>
<td>6.52</td>
<td>6.71</td>
<td>8</td>
</tr>
<tr>
<td>Arg</td>
<td>11.80</td>
<td>13.63</td>
<td>14</td>
</tr>
<tr>
<td>Thr</td>
<td>5.86</td>
<td>6.11</td>
<td>7</td>
</tr>
<tr>
<td>Ala</td>
<td>26.29</td>
<td>25.87</td>
<td>27</td>
</tr>
<tr>
<td>Pro</td>
<td>4.97</td>
<td>5.42</td>
<td>6</td>
</tr>
<tr>
<td>Val</td>
<td>5.90</td>
<td>5.52</td>
<td>6</td>
</tr>
<tr>
<td>Met</td>
<td>2.03</td>
<td>2.37</td>
<td>2</td>
</tr>
<tr>
<td>Ile</td>
<td>10.89</td>
<td>10.80</td>
<td>12</td>
</tr>
<tr>
<td>Leu</td>
<td>11.22</td>
<td>11.15</td>
<td>11</td>
</tr>
<tr>
<td>Phe</td>
<td>14.19</td>
<td>14.99</td>
<td>14</td>
</tr>
<tr>
<td>Lys</td>
<td>12.68</td>
<td>14.38</td>
<td>11</td>
</tr>
<tr>
<td>Trp</td>
<td>0.50</td>
<td>0.50</td>
<td>2</td>
</tr>
<tr>
<td>Cys</td>
<td>ND*</td>
<td>ND*</td>
<td>4</td>
</tr>
</tbody>
</table>

* ND, not determined.

Fig. 3. HPLC chromatographic analysis of the peptides obtained after treatment of 30 nmol of aporubrerythrin with Glu-C endoproteinase. See Miniprint Section for details.

Fig. 4. Rechromatography of fraction S4, after pyridylethylation, on a narrow-bore reversed-phase PTC column. For detailed information, see the Miniprint Section.

Fig. 5. Separation of peptides obtained after incubation of 15 nmol of aporubrerythrin with Lys-C endoproteinase (E/S = 1/40). Changes in the sensitivity settings are indicated as a (0.08 absorbance units at full scale), b (0.32 absorbance units at full scale), and c (0.16 absorbance units at full scale). The separation of the digest mixture obtained using more enzyme and a longer incubation time is given in the Miniprint Section (Fig. S.1).

one of the two components in the fraction H12, peptide H12A, appeared to have the same sequence from residue 15 onwards as the N-terminal sequence of peptide H10. At that moment of analysis, and apart from the N-terminal region of the rubrerythrin (1-63), a continuous polypeptide chain of 113 residues was established, covering the region Leu6'-Lys181, although the identity of the residues 75, 77, 78, and 79 remained to be confirmed as well as the residues 124, 125, 129, 130, and 179 to be identified from another digest.

Two cleavages on the apoprotein were then carried out more or less simultaneously, one with the Glu-C endoproteinase of Staphylococcus aureus V8 (Fig. 3), one with Lys-C endoproteinase from Achromobacter lyticus, both in the presence of sodium dodecyl sulfate. Under the conditions used (see Miniprint for details) several Glu-X and/or Lys-X peptide bonds appeared to have remained partially uncleaved. This unexpected phenomenon, however, was very beneficial for the further sequence determination of the rubrerythrin. Peptide S8 showed that the partial hydrolysis peptides H6 and H12A succeeded each other in the sequence with only a distance of 7 residues between the last residue determined for H6 (Phe63) and the N-terminal residue for H12A (Leu6'). Because of the purity of S8, the sequence of the first 11 residues of H12A could now unambiguously be established. Peptide S8 contained no less than four uncleaved Glu-X bonds with X = Arg, His, Gly, and Ile. Anyway, the sequence result for peptide S8 now allowed to propose the sequence of rubrerythrin from its N terminus up to lysine 181, with the residues 28, 34, 67, and the above mentioned residues 124, 125, 129, 130, and 179 left to be identified.
At this point, a comparison with the literature data revealed that the region from residue 155 onwards showed similarity with bacterial rubredoxin sequences (see "Discussion"). For that reason, the sequence run of peptide H8 was repeated after pyridylethylation, and cysteine residues could be demonstrated at the positions 21, 24, 37, and 40 of H8. Another interesting S. aureus fraction of the rubrerythrin was S4. It appeared to contain 5 peptides (Table S.IV) of which four, by sequence analysis of the mixture, could be recognized to start off, respectively, at Met101, Gly107, Ile117, and Leu128. Because the latter peptide overlapped with 7 of the last 9 determined residues of H8, we separated the pyridylethylated mixture S4 by reversed-phase high performance liquid chromatography on a small-bore column (Fig. 4) yielding the five peptides in a pure form. Peptide S4(PE)B proved to be 13 residues long and to contain cysteine at the presumed positions 2 and 5. It also proved the identity of Hid7'. Another peptide, S4(PE)A, could be sequenced up to a C-terminal tryptophan residue although the amino acid composition obviously did not indicate the presence of this residue. However, the analysis showed that the peptide could not be very much longer than 6 residues, or a multiple thereof, and that it did not contain a glutamic acid residue. Peptide S4A was therefore postulated to be a good candidate for being the C-terminal Glu-C peptide. The digest also allowed to identify the residues 28 and 34 (in peptide S11A and K1A) and 124, 125, and 130 (in S10).

Cleavage with Lys-C endoprotease (Fig. 5) was sufficient to complete the sequence determination of rubrerythrin. Peptide K7 confirmed that residue 44 of the protein is indeed aspartic acid, and that residue 62 is lysine as already found from analysis of peptide S8. The Lys-Phe, as well as a Lys-Arg bond (position 58, protein numbering) had apparently not been cleaved by the protease in peptide K7, although the sequence analysis of the fractions K5 and K6 proved that a partial cleavage had taken place. The most interesting peptide of the Lys-C digest was contained in fraction K3. Its major component showed to have a N-terminal sequence identical to the last 4 residues of S4(PE)B. Moreover, it could not be sequenced further than a tryptophan residue at Edman cycle 10, a feature which could be explained by the complete washout effect during sequence analysis. Alternatively, it could be explained by the fact that this tryptophan was the C-terminal residue of the peptide and, as already anticipated after sequencing S4A, of the protein as a whole. The identity of Gly67 could only clearly be established from peptide K'2, obtained after a second but longer cleavage of rubrerythrin with the Lys-C protease.

Definite evidence for the C-terminal sequence of the rubrerythrin was obtained by combining the results of carboxypeptidase P action on the native protein, and precise molecular mass measurements of the peptides S4A and K3A by 206Cf plasma desorption mass spectrometry. The spectra for the latter analyses are given in Fig. 6. The molecular mass of 1202.7 confirmed the proposed sequence of K3A, including one tryptophan residue (calc. 1199.61), whereas the mass of 715.4 is in complete agreement with the sequence of the hexapeptide S4A (calc. 715.39). Carboxypeptidase P released only 2 residues from the native protein, asparagine and tryptophan (Fig. S.II), although at a very slow rate. The polypeptide chain of rubrerythrin was therefore concluded to be 191 residues long and to have a tryptophan residue as C-terminus. The protein has a slight excess of 3 acidic over the total number of 25 basic residues.

---

**Fig. 6.** PDMS spectrum of the C-terminal peptides S4A (A) and K3A (B). Values represent the species [MH]+.

**Fig. 7.** A, alignment of the C-terminal region of ruberythrin (1) with rubredoxin from D. desulfuricans (2) and D. vulgaris (3). B, schematic presentation of the α-carbon positions of D. vulgaris rubredoxin (modified from Ref. 4). The side-chain carbon atoms of the cysteine residues at the positions 6, 9, 39, and 42 are shown as small filled circles, the sulfur atoms of the iron cluster as small open circles, and the iron atom as a large dotted circle. D. desulfuricans rubredoxin has the same overall structure with the hairpin loop 20–26 missing. We propose that also the C-terminal part of ruberythrin has the same general folding of the carbon backbone but with an additional deletion of the region shown as heavy lines.
Finally, it may be noticed that many of the peptides obtained in this study were sufficiently pure to allow amino acid analysis. As can be found in the Miniprint Tables S.III, S.V, and S.VII, the data indeed corroborated the proposed sequence even though several analyses were carried out on not more than 10 pmol of material.

**DISCUSSION**

Rubrerythrin is the first example of a protein in which, by means of extensive spectroscopic analyses, the presence of both a rubredoxin-type and a hemerythrin-type of polypeptide chain have been discovered (3). The primary structure determined here shows that the combined name of rubrerythrin is justified, at least as far as the rubredoxin part of the protein is concerned. Rubrerythrin is 191 residues long and contains 4 cysteine residues located in the C-terminal part of the protein at the positions 158, 161, 174, and 177, respectively. The 4 residues are part of two typical rubredoxin clusters, namely -Cys-Xaa-Yaa-Cys-Gly-Tyr- and -Cys-Pro-Xaa-Cys-. There are 12 residues between the last cysteine of the first cluster and the first cysteine of the second. In Fig. 7A we compare the C-terminal region of rubrerythrin with the amino acid sequence of two rubredoxins for which the three-dimensional structure is known in great detail: rubredoxin from *D. vulgaris* (strains 27774) (4) and rubredoxin from the same strain from which the rubrerythrin has been isolated (D. vulgaris) (12). The carbon backbone of the spatial structures themselves are given in Fig. 7B. The main difference between the two rubredoxins is that the region between the two cysteine clusters is shorter by 7 residues in the *D. vulgaris* protein. This deletion corresponds to a hairpin loop present in the protein of *D. vulgaris* which has 29 residues between the second and the third cysteine. If we now assume that the three-dimensional structure of the rubredoxin part of rubrerythrin has the same general fold as that of the authentic rubredoxin and that therefore the four cysteines can effectively be aligned as given in Fig. 7A, it is apparent that in the rubrerythrin the extra 10 residues gap (very likely after a GAP sequence) relative to *D. vulgaris* rubredoxin is due to an extra deletion as indicated with a heavy lining in Fig. 7B. Remember that *D. vulgaris* rubredoxin is unique among this class of proteins in having an important deletion in the sequence.

The rubredoxin part of rubrerythrin is also different from the rubredoxins in that there is no tryptophan residue between the two cysteine clusters. This aromatic residue is thought to participate in the stabilization of the rubredoxin center (13). Interestingly, there is a tryptophan residue at position 156 of rubrerythrin corresponding to position 4 of the rubredoxins where an aromatic residue is always found. We suggest that Trp$^{156}$ may take over this stabilizing role in rubrerythrin and that the gap and the different location of the tryptophan contribute to the higher site symmetry of the rubredoxin center that has been observed by resonance Raman spectrometric analysis (13).

In contrast to the C-terminal part which shows moderate but distinct sequence similarities with rubredoxins, it is quite difficult to demonstrate sequence similarities for the N-terminal part of rubrerythrin with any other known protein. Even hemerythriins, as could have been expected from the physicochemical data, do not seem to contain sequence patterns that can be recognized in the nonrubredoxin part of rubrerythrin (14–16). Hemerythrin is known to have a four anti-parallel up-down helix structure (17) with a binuclear iron center liganded by 5 histidine residues. It is generally understood that these residues are evolutionarily conserved. In Fig. 8 we have indicated the precise positions of these 5 histidines in a linear presentation of the hemerythrin polypeptide chain. Rubrerythrin contains a total of 8 histidines and also five of them are located in the N-terminal nonrubredoxin part of the molecule. The spacing between these histidines, however, is not identical to that in hemerythriins. On itself, this does not, of course, exclude the possibility that part of the N-terminal region 1–152 of rubrerythrin has a similar three-dimensional folding as hemerythrin (108 residues for the protein of *Phascolosia gouldii* (13)), since the different spacing may be due to insertions or deletions. When we align the first two histidines of hemerythrin (spaced by 27 residues) with the same residues in the rubredoxin-part (spaced by 26 residues) there is no sequence similarity between the two proteins, even after introducing gaps. The only similarities of any significance can be found if one aligns the third histidine of hemerythrin (His$^{73}$) with the first histidine of rubrerythrin. This alignment is given in Fig. 8 as a possibility (c). The amino acids flanking these histidines give a sequence similarity of 75% but the comparable region comprises only 4 residues. Fig. 8 also presents the two other possibilities of alignment that give some sequence similarity in the polypeptide...
chain. Both possibilities include, however, that part of the N-terminal region of hemerythrin overlaps with the C-terminal rubredoxin part of rubrerythrin. Solution (b) is centered around histidine 84 of rubrerythrin and histidine 25 of hemerythrin with a stretch of 9 residues then giving 55% sequence similarity. The identical residues are noted in the figure. Alignment solution (a) gives a similar figure of similarity (50%) for the sequence region Arg119-Glu206 of rubrerythrin, but does not include even one of the five histidines of hemerythrin or the hemerythrin-like center of rubrerythrin.

The fact that three out of the eight histidines in rubrerythrin are located in the C-terminal rubredoxin-like center of rubrerythrin raises the possibility that one or more of these residues could be involved as ligands of the hemerythrin-like center. Such a possibility would include that the two centers of rubrerythrin would be in close proximity of each other, and would well explain the high redox potential (+230 mV) of the rubredoxin-like center in rubrerythrin compared to that in authentic rubredoxin (−20 mV). Moreover, both the iron stoichiometry (3) and preliminary x-ray analysis data (18) suggest that the binuclear hemerythrin-center is liganded through two monomers. In such case, the center should be more symmetrical than in authentic hemerythrin. This extra factor of symmetry could be realized if the two iron atoms from the center would be liganded by six instead of five histidines, with three ligands being donated per monomer. Anyone of the three histidines in the rubredoxin-like center can then be a ligand candidate for one of the iron atoms in the binuclear center. If such a hemerythrin-like structure would finally turn out to be real following high resolution x-ray analysis, it could explain why there is no sequence homology between the N-terminal part of rubrerythrin and hemerythrin.

In the given status of data it is difficult to speculate on the evolutionary origin of rubrerythrin. The presence of a binuclear center indicates that the N-terminal part of the protein and hemerythrin could have a common evolutionary origin. However, hemerythrin is a eukaryotic protein that occurs in oxygen-respiring worms, whereas rubredoxin is found in sulfate-reducing bacteria and rubrerythrin, so far, only in one ecological marine sediments of the sulfate-reducing bacteria and such conditions are a prerequisite for the possibility of gene transfer.

Finally, we would like to stress the point that, although rubrerythrin has pyrophosphatase activity in vitro, the physiological substrate for the protein may not be inorganic pyrophosphate. Some phosphatases such as the binuclear iron-center containing acid phosphatase from bovine spleen also hydrolyze pyrophosphate and are thus rather substrate-unspecific (19). On the other hand, the presence of several inorganic pyrophosphatases in the same organism, as in *D. vulgaris*, has also been reported in the phototrophic bacterium *Rhodospirillum rubrum* which contains both a soluble and a membrane-bound enzyme (6). The latter is thought to be able to translocate protons in a manner corresponding to that of ATPase.

Acknowledgments—We thank M. Howard for his skillful technical help and B. C. Prickril for preliminary sequence determination.

REFERENCES

Preparation of the ruberythrin
Growth of the bacterial cells, preparation of extracts and protein purification were as described in [3]. The iron clusters of ruberythrin were removed following the procedure described in [30]. After lyophilization, the apoprotein was mixed with water giving a white suspension.

Preparation and purification of peptides
Partial acid hydrolysis of the aporuberythrin at 105°C during 3 hours was carried out on 45 nmol of material using 2.5% formic acid as hydrolysing agent. The final volume was 300 µl.
An enzymatic digest of 30 nmol of the aporuberythrin with Staphylococcus aureus V8 protease (Miles, IL) was performed at an enzyme/protein ratio of 1/10 for 6 hours in 0.1 M ammonium bicarbonate, pH 8, supplemented with 0.5% SDS, in a final volume of 200 µl. A first digest with Lys-C endoproteinase from Aspergillus niger (Wako, Japan) was performed on 15 nmol of aporuberythrin at an enzyme/protein ratio of 1/40 for 4 hours at 37°C in 0.1 M Tris Cl buffer, pH 8, also in the presence of 0.5% SDS. A second digest with the same enzyme, on 15 nmol of the cluster-free ruberythrin as starting material was carried out at an enzyme/protein ratio of 1/20 under the same conditions but for a total period of 17 hours. Digest volumes were 100 µl each time.
Peptides resulting from the different cleavage procedures were separated by high performance liquid chromatography on a microcapillary C4 column (Vydac, 4.6 x 250 mm, 214TP4). The chromatographic equipment consisted of a RIKEN condition Controller, a 870 Chromatographic Pump and a UV spectrophotometric detector set at 220 nm (all parts from DuPont Instruments, USA). A Rhodapy equipment was equipped with a 200 µl loop. The gradient was always made starting with 0.1% TFA in MeOH-water (M:0, USA) with increasing fractions of 0.1% TFA in 70% ammonium. The course of the gradient is drawn on the Figures 2:5 and S.1. The generated peptides were collected manually, as a rule of 1 ml, in polypropylene tubes (Kartell, USA). The collected fractions were dried in the Speed Vac Concentrator (Savant, USA) and stored at -18°C. After use, the peptides were dissolved in 40 to 100 µl 0.1% TFA in MeOH-water. Some peptide mixtures were rechromatographed on a so-called PCT C18 column of 2.5 x 250 mm using either a 100A Separation System or a 140A Solvent Delivery system equipped with a 1000S Diode Array Detector (all equipment from Applied Biosystems, USA). Column eluents were as described for the Vydac-column but the elution rate was only 0.2 ml/min.

Carboxypeptidase treatment
Treatments of native ruberythrin with carboxypeptidase P (Roche, BBE) was carried out on 875 nmoles of material. Samples were taken at several times, alkylated and analyzed on a 420A Derivator, coupled to a 100A Separation System (see Amino acid analysis, without prior precipitation of the proteins and the peptides). Each digestion was performed in 300 µl of 0.1 M sodium acetate buffer, pH 4.6, as an enzyme/substrate ratio of 1/75.

Chemical modification of peptides
Some peptides for which the presence of cysteine residues was anticipated (see under "RESULTS") were pyridylethylated in the gas phase following the procedure described in [21] or in the liquid phase following the method described in [22].

Sequence analysis and amino acid analysis
Automated Edman degradation were performed in a 477A pulsed liquid sequencer followed by on-line analysis of the phenylthiohydantoin amino acids on a 120A PTH-analyzer (Applied Biosystems, USA). Amino acid analysis of proteins and peptide samples was carried out on a 420A Derivator with on-line analysis of the phenylthiohydantoin amino acids on a so-called PCT C18 column of 2.5 x 250 mm using a 130A Separation system (Applied Biosystems, USA). Hydrolysis of the samples was performed in glass tubes of 5 x 55 mm placed in a hydrolysis vial (Millipore, USA) and using 66% HCl as hydrolysing agent in a covered vial. The duration of the hydrolysis was 24 hours at 100°C. In the case of aporuberythrin, a hydrolysis was also carried out after addition of 10% dodecane (Sigma, USA), final concentration [23].

Mass analysis
Time of flight mass spectrometry was carried out using a BioIon 20K plasma desorption spectrometer with Calilfornia-252 source. The principle of the method is described in [24].

RESULTS
In the following tables with sequence results, the net increase of the given residues at each Edman cycle compared to the previous step are given in picomoles. Values smaller than 10 picomoles are reported with a decimal. Because of the collision of PTH-Tyr with diphenylurea, the values for tryptophan are not absolute and are mentioned between brackets. When two peptides in a mixture have the same residue at a given position this is indicated by mentioning that residue halfway the columns with the data for each peptide. When the measured content of more than two peptides, the same amino acid in two peptides is sometimes indicated with "x" in one of the peptides (as e.g. in Table SIV). When a sequence was deliberately stopped, this is indicated as ".." The sign "--" at the bottom of the columns means that no more amino acids could be identified at that cycle nor at two subsequent cycles. The position of the peptide in the total sequence of the proteins is indicated by the column number of each column. When there is no certainty about the identity of the C-terminals in the peptide, this fact is shown by leaving a blank space.

Sequence analysis of the native protein

The N-terminal sequence of the native decuterated protein was determined up to residue 53 with lack of assignment at the positions 28, 34, 42, 47 and 50. The fact that the level of the background of all the PTH-amino acids was constant from one Edman cycle to the next one, combined with the very small individual increases of the PTH amounts at these 5 positions did not permit further unambiguous identifications.

Peptides from cyanogen bromide cleavage
In the cleavage with cyanogen bromide of the native protein tryptic in two peptides which were separated by SDS-polyacrylamide gel electrophoresis. One electrophoretic band (ECNE) could be sequenced for 48 residues and histidine as N terminus. Except for the positions 27, 23, 24, 28, 29, and 50, the sequence was given in Fig. 1 covering the region I01-I49. The second band of only slightly different mobility from the first one contained the N-terminal sequence of the protein. The quantitative data for the sequence runs found to be low at the moment this paper has been written up.

Peptides from the partial acid hydrolysates (Fig. 2, Table SIII and SIV)
The major peak, H13, in the separation pattern of the partial hydrolysate of the aporuberythrin contains 53 residues and is the C-terminal H13 peptide. We sequenced in turn, one time before and a second time after pyridylmethylation in the gas phase with Edman work, in the case of pyridylmethyated pepti de, identification of the residues 21, 24, 37 and 40 as the pyridyl derivatives of cysteine. The longest H1 peptide, H12A, is composed of 69 residues and covers the region 69-117. It contains an unlabeled Ser. The peptide that did originate from the cleavage of this bond is H10. Its sequence determination revealed the unidentified residues 23, 27, 28 and 29 of I9A in Ala, His, His and Glu respectively, and also extended the sequence of the latter peptide with 22 more residues. The flanking peaks of H10 both contained peptides with the same initial sequence as H10 and the amino acid composition of these two samples is in agreement with the total number of some 50 residues. H10 and H11, therefore, very likely have the same total length but should have used up one of the two peptides of the mixture H3 (H3A). The two smaller peaks H1 and H2 seen earlier in the chromatogram are both 7 residues long, but differ in polarity. Peptide H2 covers the region Gly38-Arg44, whereas peptide H1 covers the region from His45 to Asp49, and contains one more acidic residue. H1 is also the N-terminal part of peptide H3 which gives a clear overlap with the (I3) residues described during the sequence run of the native protein, thereby extending the knowledge of the N-terminal sequence of the ruberythrin up to residue 61.

Peptides from the digest with S. eureca protease (Fig. 3 and Table SIV and SV)
As could be expected from its width, peak SIV contained several peptides. Sequence analyses revealed a major peptide (A) which appeared to be the N-terminal 5-peptide of the ruberythrin. Two other peptides, (B) and (C), could be recognized as starting respectively at Met19 and Tyr36 of peptide H10, giving information up to position 139 (protein numbering). An unambiguous sequence identification of the parts of the protein, however, was obtained by analysing the pure peptide S10, which on the basis of its amino acid composition contained some 50 residues including 9 glucosidic acids, only one of them
Peptide eluted as protein. and three minor peaks were obtained as amino acid sequence (Table S.VII). The amino acid composition of S4A (Table S.V) is consistent with the expected stoichiometry of the sequence shown in Fig. 4. The C-terminal peptide (S4A-D) of the protein eluted with the highest concentration, but also the 2 smaller peaks eluted somewhat earlier (A and B) have the same amino acid sequence.

Peptide S4PEBD was conserved in its sequence. With the exception of two minor peaks, one at position 2 and one at position 3. These results were confirmed by sequence analysis of peptide S4PEBD which contained the sequence of S4PEBD from residue 7 onwards. The results of an unverified Glu-tryptan bond; the tryptan bond was then hydrolyzed and the peptides S4PEBD were consequently digested at the Edman cycles 8 and 11 in S4PEBD.

The third major peptide (S7) yielded no sequence information. As the amino acid analysis revealed the absence of peptide material, the peak must be due to a UV-absorbing impurity. The smaller peak S8 contained useful information as it covered the region around Thr198, and, in particular, allowed the ending of GLH. In this peptide, a GLH-Aa, a GLH-Hic, a GLU-tryptan and a GLU-tryptan bond had not been cleaved.

Peptide S5 can be considered to be a subfragment of S8. It resulted from a cleavage of the GLHAA GVH bond and confirmed the sequence residues 139-Aa 2. This result was unexpected, but the sequence of the Edman cycle resulted in the sequence residues 139-Aa 2. In peptide S5, the sequence residues 139-Aa 2 were not cleaved. The third major peak (S7) yielded no sequence information. As the amino acid analysis revealed the absence of peptide material, the peak must be due to a UV-absorbing impurity.

The smaller peak S8 contained useful information as it covered the region around Thr198, and, in particular, allowed the ending of GLH. In this peptide, a GLH-Aa, a GLH-Hic, a GLU-tryptan and a GLU-tryptan bond had not been cleaved.

Peptide S5 can be considered to be a subfragment of S8. It resulted from a cleavage of the GLHAA GVH bond and confirmed the sequence residues 139-Aa 2. This result was unexpected, but the sequence of the Edman cycle resulted in the sequence residues 139-Aa 2. In peptide S5, the sequence residues 139-Aa 2 were not cleaved. The third major peak (S7) yielded no sequence information. As the amino acid analysis revealed the absence of peptide material, the peak must be due to a UV-absorbing impurity.

The smaller peak S8 contained useful information as it covered the region around Thr198, and, in particular, allowed the ending of GLH. In this peptide, a GLH-Aa, a GLH-Hic, a GLU-tryptan and a GLU-tryptan bond had not been cleaved.
**D. vulgaris, Rubrerythrin, Primary Structure**

**Table S.V**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

Sequence data on the peptides generated by cleavage of the apoprotein with Lys-C endoproteinase

**Table S.VI**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

Sequence data on the peptides generated by cleavage of the apoprotein with Lys-C endoproteinase

**Table S.VII**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

Sequence data on the peptides generated by cleavage of the apoprotein with Lys-C endoproteinase

**Fig. S.1** Separation of a second digest on 15 nanomoles of apopruberythrin with Lys-C endoproteinase at prolonged period (1) instead of 4 hours as in Fig. 3.

**Fig. S.11** Release of amino acids from native ruberythrin after incubation for different lengths of time with carboxypeptidase F.

**Fig. S.2** Release of amino acids from native ruberythrin after incubation for different lengths of time with carboxypeptidase F.