1. Introduction

Conserved amino acid sequence motifs in enzymes often indicate involvement in the binding of metal ion(s) and/or in the binding and/or reactions of substrate(s). The four very early proteinaceous amino acids are glycine (G), alanine (A), aspartic acid (D) and valine (V) as was demonstrated with the clarification of the stepwise evolution of the genetic code (Eigen and Schuster, 1979), and in agreement with the quantities obtained in the earlier, classical work (Miller, 1953 and on) showing proteinaceous amino acid production under possible prebiotic conditions. Aspartic acid stands out as the unique very early amino acid containing an additional, highly reactive free charge, suitable for cation binding. Active site motifs with a very high content of any or all of these four amino acids may well be of early evolutionary significance.

2. Very Early Sequence Motifs

Some “early” proteins appear to harbor or reflect very early sequence motifs. For example, certain enzymes involved in inorganic pyrophosphate (PPi) metabolism, as well as in that of ATP, seem to be of particular significance in this connection. Special attention has recently been given to the integrally membrane-bound, proton-pumping PPi synthase, which in bacterial photophosphorylation is the first and still only known alternative to the ubiquitous ATP synthase in biological electron transport coupled phosphorylation (Baltscheffsky et al., 1999). The putative active site of the PPi synthase from the purple photosynthetic bacterium
**Rhodospirillum rubrum**, in loop 5–6 (between transmembrane segments 5 and 6) has two nonapeptidyl sequences (DVGADLVGK and DNVGDNVGD), which are strongly conserved in the homologous enzyme family and which contain unusually many very early amino acids. Importantly, these sequences have charged amino acids regularly arranged in positions 1, 5 and 9, five of these six being aspartic acid. We have discussed the possible roles of these charged amino acids in the putative active site for the binding and reactions of the inorganic phosphates (Baltscheffsky et al., 1999, 2001) and have described this with a preliminary sketch (Baltscheffsky et al., 2001). Notably, other loops may also be involved in the active site for these reactions (Schultz and Baltscheffsky, 2003).

Recently, we counted the total number of occurrences of the sequence pattern D-A/G/V-A/G/V-A/G/V-D-A/G/V-A/G/V-D (as a possible predecessor of the nonapeptidyl sequence DVGADLVGK) in the Swissprot and TrEMBL databases (Boeckmann et al., 2003) of October 2003. In total, we found 35 occurrences, which is four-fold more than could be expected by chance, indicating that this pattern is overrepresented, probably due to its functional or structural importance. We also checked the patterns

\[
\text{D-A/G/V-A/G/V-D and D-A/G/V-A/G/V-G-D-A/G/V-A/G/V-G-D.}
\]

These patterns showed to be even more overrepresented to the extent of 15-fold and 27-fold, respectively. The extremely strong overrepresentation of the G-D pattern appears to be in agreement with our preliminary observation that this pattern occurs particularly frequently in certain enzymes involved in energy transfer. We also compared the somewhat more general patterns D/E-(A/G/V)\(_n\)-D/E-(A/G/V)\(_n\)-D/E, where \(n\) was set to 2, 3 or 4. For \(n = 2\), we arrived at the expected number of sequences, while for \(n = 3\) and \(n = 4\), the occurrences were between four- and seven-fold, also indicating some evolutionary selection caused by functional or structural requirements.

A search for nonapeptidyl sequences in the Swissprot and TrEMBL databases, with D in positions 1, 5 and 9 and the other, very early A/G/V in the other positions, revealed that since 1996, several such sequences have been shown. DGGGDGGGD was an early found (Wen and Tseng, 1996) and dominating sequence, but also, from the year 2000 and on, sequences such as DAGGDAGGD, DAAGDAAGD, DVGGDAGGD and DVAGDVGGGD have been reported. Several duplications at the tetrapeptide level and subsequent mutations would appear to have occurred. We interpret these findings as supporting our belief that the nonapeptide DVGADLVGK in PP\(_1\) synthase may have evolved from DVGADLVD or, earlier, DVGADVGAD. The introduction of K should have been useful in connection with the binding of anionic phosphate oxygen.

Among the “very early” sequence motifs at active sites, that are found in some other PP\(_1\) metabolizing enzymes, and also, more or less similar, in some ATP metabolizing enzymes, we shall discuss those found in phosphofructokinases, which show the motif GGDD in those operating with PP\(_1\) and GGDG in those operating with ATP (Moore et al., 2002). In the 174GGDD177 sequence of the PP\(_1\)-dependent phosphofructokinase from *Borrelia burgdorferi*, 174GGD are three conserved active site residues and D177 is one of four residues suggested to be directly involved in PP\(_1\) discrimination, by preventing ATP from binding.