A new family of bacterial regulatory proteins

David J. Haydon and John R. Guest

The Krebs Institute, Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield, U.K.

Received 11 December 1990
Accepted 19 December 1990

Key words: Helix-turn-helix motif; Transcriptional regulator; DNA-binding protein; Sequence comparison; FadR; GenA; GntR; HutC; KorA; P30; PhnF

1. SUMMARY

A new family of bacterial regulatory proteins has been identified by sequence similarity. The family contains the repressor of the *Bacillus subtilis* gluconate operon (GntR), the regulators for histidine utilization in *Pseudomonas putida* (HutC<sub>Pp</sub>) and *Klebsiella aerogenes* (HutC<sub>Ka</sub>), the repressor (FadR) of fatty acid degradation in *Escherichia coli*, a regulator involved in the conjugal transfer of the broad host range plasmid pIJ101 (KorA), and three proteins of unidentified function in *E. coli* (GenA, P30 and PhnF). The proteins share amino acid sequence similarities in a 69-residue N-terminal region. A helix-turn-helix motif is predicted in the most highly-conserved segment of each protein suggesting that they are members of a new family of helix-turn-helix DNA-binding proteins.

2. INTRODUCTION

Many transcriptional regulatory proteins contain a helix-turn-helix motif originally identified in the λCrot protein and associated with site-specific DNA-binding [1,2]. In some of these proteins the presence of extensive sequence similarity has allowed specific families with common ancestries and similar modes of action to be recognised, e.g. the λCrot, LysR and CRP families [2–5].

The GntR protein (243 residues, *M*<sub>r</sub> 28,277) is the transcriptional repressor of the gluconate operon (*gntRKPZ*) of *Bacillus subtilis* [6,7], and in previous studies it was shown to resemble two unidentified proteins of *Escherichia coli*, P30 and GenA [8]. These are the products of genes which are located immediately adjacent to the operons encoding the respective 2-oxoglutarate and pyruvate dehydrogenase complexes, *sucABCD-g30* and *genA-aceEF-lpd* [8,9], and it was suggested that GenA and P30 may be analogous regulatory proteins controlling the synthesis of the corresponding complexes [8].

Database searches using matrix-based and consensus-based methods have now revealed a further...
five proteins which possess sequences similar to GenA, P30 and GntR, and have therefore been assigned to the GntR family of regulatory proteins.

3. METHODS

The UWGCG package of computer programs was used on the SEQNET facility of the SERC laboratory at Daresbury. A multiple alignment of the conserved 69-residue N-terminal segments of GENA, P30 and GNTR was prepared with LINE-UP and assembled into a profile matrix with PROFILEMAKE. This matrix was used to search the NBRF-Protein Database V26.0 with PROFILESEARCH. In addition, a 69-residue consensus sequence was generated from six sequences by PROFILE and used with the TFASTA program to screen the EMBL-DAILY DNA database. Both methods were employed in order to ensure the widest search of available sequences. Secondary structure predictions were derived from the joint output of a combination of methods [10].

4. RESULTS

A matrix-based search of 23196 sequences in the NBRF-Protein database was used to detect proteins with sequences similar to GenA, P30, and GntR. This approach has the advantage that the matrix of residue scores includes all of the residues in an alignment, whereas a consensus-based search simply relies upon the most frequent residue at each position. A matrix-based search can therefore emphasise similarity within conserved regions while tolerating diversity in the variable regions. This search revealed three proteins (FadR, KorA and PhnF) which together with those used to construct the matrix, form a distinct group possessing high similarity scores.

FadR is a regulator of fatty acid degradation in E. coli: it functions as a transcriptional repressor for several genes e.g. fadL, fadD, fadE, fadAB, and aceBAK [11]. KorA is a regulatory protein encoded by the broad host range plasmid pJ101 in Streptomyces [12,13]. It controls plasmid transfer and prevents the lethal effects of the corresponding tra (kilA) gene. PhnF is a protein of unknown function which is encoded by the phnF gene in the phn (psiD) locus of E. coli. The phn genes determine alkylphosphonate uptake and carbon-phosphorus lyase activity [14].

Using a consensus derived from the above sequences the EMBL-DAILY DNA database was searched with the TFASTA program and this revealed the HutC protein of Klebsiella aerogenes [15] as an additional member of the GenA/P30/GntR group. The HutC<sub>Ka</sub> protein is known to be homologous with the corresponding Pseudomonas putida protein, HutC<sub>PP</sub> [15,16]. Both are regulators of the corresponding histidine utilization genes, but neither was in the protein database.

![Fig. 1. Multiple sequence alignment for the N-terminal segments of the GntR family of regulatory proteins. The putative LAcro-like helix-turn-helix motifs are bracketed. A consensus for the alignment was derived with the PROFILESEGMENT option of PROFILE, and the same method was used to obtain the consensus (HTH) for the helix-turn-helix sequences compiled by Dodd and Eagan [3]. Asterisks (*) denote perfect matches between the HTH consensus and the most highly conserved part of the alignment consensus. The number of amino acid residues in each protein subunit is indicated in parentheses.](https://academic.oup.com/femsle/article-abstract/79/2-3/291/445851)
A multiple alignment of the N-terminal segments of the eight related proteins and the corresponding consensus sequence are shown in Fig. 1. Insertion of only two single-residue gaps into the GntR sequence was necessary to give the best alignment. A matrix showing the pairwise sequence similarities in the N-terminal segments is shown in Fig. 2. The proteins all contain between 236 and 248 amino acid residues per subunit ($M_r$, 26,972–29,610) but the sequence similarity does not extend beyond the segments shown.

The secondary structures predicted for the aligned regions of each protein were similar, but apart from being uniformly rich in $\alpha$-helices, the predicted structures for the residual segments were not similar. Representative predictions for the aligned segments of GntR and GenA are shown in Fig. 3, and it is clear that helix-turn-helix structures are strongly predicted in the most highly conserved regions.

A search was made for $\lambda$Cro-like helix-turn-helix motifs using the weighted matrix of Dodd and Egan [3]. This identified one potential $\lambda$Cro-like DNA-binding site in each of the aligned sequences with the following probabilities: GenA, 60%; HutC$_{K_a}$, 56%; HutC$_{P_p}$, P30 and PhnF, 7%; FadR, 6%; GntR and KorA, < 1%, where scores of > 1% are significant [3]. The probability score obtained with the corresponding region of the alignment consensus was 100%. Significantly, these sites coincide with the regions in which helix-turn-helix structures are strongly predicted (Figs. 1 and 3). The low probabilities obtained with the known regulators, GntR and KorA, could reflect the fact that the $\lambda$Cro family may be disproportionately represented in the matrix for helix-turn-helix prediction [3]. This could bias the distribution of preferred residues so as to exclude some members of the GntR family.

5. DISCUSSION

The similarities in sequence, molecular size, and predicted secondary structure, combined with the known regulatory functions of the GntR, FadR, HutC and KorA proteins suggest that the eight proteins represent a family of transcriptional regulators, the GntR family. This is supported by

<table>
<thead>
<tr>
<th></th>
<th>FadR</th>
<th>GenA 48% (35%)</th>
<th>KorA 30% (20%) 46% (29%)</th>
<th>P30 41% (29%) 55% (35%) 48% (33%)</th>
<th>GntR 36% (22%) 46% (27%) 43% (24%) 48% (30%)</th>
<th>HutC$_{K_a}$ 36% (28%) 52% (39%) 48% (38%) 59% (43%) 49% (31%)</th>
<th>HutC$_{P_p}$ 42% (32%) 49% (38%) 39% (28%) 58% (46%) 45% (30%) 81% (74%)</th>
<th>PhnF 36% (26%) 43% (36%) 41% (30%) 43% (36%) 34% (22%) 45% (35%) 41% (29%)</th>
</tr>
</thead>
</table>

Fig. 2. Pairwise similarities and identities for the sequences aligned in Fig. 1. Similarities were calculated using the Schwartz and Dayhoff relationship [17] with a threshold of > 0.6, and identities are shown in parentheses.
the presence of a putative λCro-like DNA-binding motif within the conserved regions of most of the sequences. Furthermore, the structural similarities flank the highly conserved 20-residue motifs, suggesting that they all have related 69-residue N-terminal structural domains with analogous DNA-binding functions. These putative domains are approximately the same size as the λCro protein (66 residues) and the DNA-binding domains of many transcriptional regulators, e.g. CRP (70 residues). The DNA-binding domains of such regulators are generally attached to effector-binding domains, which are approximately the same size as the dissimilar 170-residue C-terminal segments in the GntR family of proteins. These segments presumably represent the effector-binding domains and their lack of sequence similarity is consistent with the wide distribution and diversity
of function within the GntR family. The DNA binding-sites (operator sequences) are known for GntR and HutC_pp and it is significant that there are identical bases at 11 out of 13 positions.

GntR ATACTTGATACAAAGTAT
HutC_pp CTTGTATATACATA.

and that 7 of the 11 residues in the DNA-recognition helices are also identical (Fig. 1). In view of the strong sequence conservation at several positions in the putative recognition helices (Fig. 1), it is anticipated that the operators for all members of the GntR family will be related and exhibit an informative distribution of conserved and discriminatory nucleotide to amino acid interactions.

ACKNOWLEDGEMENTS

This work was supported by a studentship from the Science and Engineering Research Council (D.J.H.). We are indebted to Dr. G.C. Russell for helpful discussions.

REFERENCES