Identification by computer sequence analysis of transcriptional regulator proteins in *Dictyostelium discoideum* and *Serratia marcescens*

Rita Grandori and Chris Sander[1]*
Dipartimento di Fisiologia e Biochimica Generali, Sezione di Biochimica Comparata, Universita' degli Studi di Milano, Italy and [1]European Molecular Biology Laboratory, D-6900 Heidelberg, FRG

Received January 28, 1991; Revised and Accepted April 10, 1991

ABSTRACT

We have performed computer searches in the database of known protein sequences for proteins similar in sequence to bacteriophage regulatory proteins of known 3-D structure. The searches are more selective than other methods due to the use of a length-dependent threshold in sequence similarity, above which structural homology is implied with high certainty. Two probable DNA binding proteins were identified which are predicted to have a three-dimensional structure very similar to bacteriophage cro and repressor proteins. Approximate three-dimensional model coordinates are available from the authors. Both proteins contain the helix-turn-helix sequence motif typical of a wide class of DNA binding proteins and their function is deduced by analogy to sequence-similar proteins of known function. We predict that the *Y.SmaI* protein in the restriction-modification enzyme gene locus of the enterobacterium *serratia marcescens* is a regulator of endonuclease expression; and, that the vegetative specific gene *VSH7* of the slime mold *dictyostelium discoideum* codes for a regulator of gene expression specific for the slime mold growth phase before the onset of the developmental program. Point mutations that would have a strong effect on growth regulation phenotype are suggested. The *VSH7* protein would be the first eukaryotic representative of the cro / phage repressor class.

INTRODUCTION

A common structural DNA binding motif present in the four known and deposited crystal structures of bacteriophage gene regulating proteins (repressors) is a small domain of four alpha-helices, of which two helices and several loops make physical contact with the DNA (1−4). A corresponding sequence pattern, termed the helix-turn-helix or HTH motif, has been derived from the multiple sequence alignment of more than 91 sequences of proteins presumed to bind DNA in a similar fashion (5), including eukaryotic homeobox proteins. These proteins are involved in a wide spectrum of functions: transcriptional control (e.g. LacI, MetR), recombination/ transposition (e.g. Tn21 TnpR, IS1 InsA), bacteriophage transcriptional control (e.g. λ Cro, 434 Cro, 434 CI, P22 C2), RNA polymerase activity (sigma factors, e.g. RpoD, KpNtrA), DNA replication/chromosome partition (e.g. F SopB), or eukaryotic developmental regulation (homeobox proteins, e.g. Dm Entp).

In the known 3D structures of the bacteriophage gene regulating proteins, certain strongly conserved residues appear indispensable for maintaining the structure of the domain. For the amino-terminal domain of phage 434 repressor (3), R1-69 (Protein Data Bank code 1R69), these include R10, Q17 and E35, involved in a tight hydrogen-bonding network between the ends of helices 1, 2 and 3. In the wider class of DNA binding proteins with the HTH motif, which includes homeobox proteins, the most conserved residues are, in R1-69 numbering, A21, G25 and I31. Structurally, A21 and G25 appear to be essential for maintaining the tight turn and the precise relative orientation of helices 1 and 2 required for DNA binding, while I25 makes strong contacts in the hydrophobic core formed by helices 1 to 4 (3).

Database searches for similar sequences can be powerful tools for the elucidation of previously unknown protein structure and function by analogy, provided that the alignment procedure is accurate and significance of the sequence similarity can be established. Recently, some progress has been made in the accuracy of the dynamic programming alignment procedure used here and in the definition of a threshold in terms of sequence similarity above which sequence alignments can be trusted to imply similarity of three-dimensional structure (see Methods). An related and partially complementary method is the definition of selective sequence patterns descriptive of a functional class of proteins. A pattern for the DNA-binding helix-turn-helix motif (6) was recently updated based on 91 protein sequences (5).

Using these facts, databases and improved software tools, we are able to advance plausible predictions for the function of two proteins previously sequenced but not yet characterized in their function (7−9).

* To whom correspondence should be addressed

REFERENCES

1. Jordan, S.R. and Pabo,C.O. (1988) Science **242**, 893–899.
2. Aggarwal, A.K., Rodgers, D.W., Drottar, M., Ptashne, M. and Harrison, S.T. (1988) Science **242**, 899–907.
3. Mondragon, A., Subbiah, S., Almo, S.C., Drottar,M. and Harrison, S.C. (1989) J.Mol.Biol. **205**, 189–200.
4. Wolberger,C., Dong, Y., Ptashne, M. and Harrison,S. (1988) Nature **335**, 789–795.
5. Dodd,I.B. and Egan,J.B. (1990) Nucleic Acids Res. **18**, 5019–5026.
6. Ohlendorf,D.H., Anderson,W.F. and Matthews,B.W. (1983) J.Mol.Evol. **19**, 109–114.
7. Heidmann, S. Seifert, W., Kessler, C and Domdey, H. (1989) Nucleic Acids Res. **17**, 9783–9796.
8. Singleton, C.K., Manning, S.S. and Ken, R. (1989) Nucleic Acids Res. **17**, 9679–9692.
9. Singleton, C.K., Manning, S.S. and Feng, Y. (1988) Mol.Cell.Biol. **8**, 10–16.
10. Sander, C. and Schneider, R. (1991) Proteins **9**, 56–68.
11. Pearson W.R. and Lipman, D.J. (1988) PNAS **85,** 2444–2448
12. Kabsch, W. and Sander, C. (1983) Biopolymers **22**, 2577–2637.
13. Falk, G. and Walker,J.E. (1988) Biochem. J. **254**, 109–122.
14. Gaur,N.K., Dubnau,E. and Smith, I. (1986) J. Bacteriol. **168**, 860–869.
15. Gaur, N.K., Cabane K., Smith I. (1988) J. Bacteriol. **170**, 1046–1053
16. Van Kaer, L., Gansemans, Y., Van Montagu, M. and Dhaese, P. (1988) EMBO J. **7**, 859–866.
17. Sauer, R.T., Pan, J., Hopper, P., Hehir, K., Brown, J. and Poteete,A.R. (1981) Biochem. **20**, 3591–3598.
18. Walder, R.Y. and Walder, J.A . (1987) in Gene Amplification and Analysis, ed. Chirikjian JG, Elsevier, New York, Vol. 5, pp. 209–226.
19. Sullivan KM and Saunders JR (1989) MolGenGenet **216**, 380–387.

similarity is 33%/54aa/1gap, 2 percentage points above the threshold for this length; this hit ranks number 5 in our search, up from a FASTA rank of 113. Other, slightly lower ranking hits from our search were discarded because the structurally essential residues were not conserved (data not shown).

Y.SmaI

The Y.SmaI or sORF gene from enterobacterium *serratia marcescens* is predicted to code for a protein of 84 aa (7). It is the third open reading frame in a class-II restriction-modification system, located next to the genes for the putative M.SmaI DNA modification enzyme (292 aa.) and the R.SmaI restriction endonuclease (247 aa.). The open reading frame is transcribed: a polycistronic message has been found (7) in which the last ten nucleotides of Y.SmaI overlap the first ten nucleotides of the endonuclease, suggesting joint regulation; the methylase is transcribed separately. However, the Y.SmaI protein product has not been detected, nor is there biochemical or genetic evidence about its function.

In the sequence alignment of Y.SmaI protein, relative to 1R69, identically conserved residues are the structurally 'essential' R10, Q17 and E35 involved in a hydrogen bonding network, as well as A21 and G25 involved in the helix1-helix2 turn. In addition, there is a conserved cluster (9/16) of residues in the region connecting helices 4 and 5. In our search, the closest relatives of Y.SmaI among the HTH proteins are bacteriophage phi-105 repressors RPC$BPPH1 (or IMRE$BPPH1), the bacillus subtilis inhibitor of sporulation SIN$BACSU, bacteriophage P22/P21 C2 repressor RPC2$BPP22, all of which are repressors, and the purple bacterial URF4 (or DNU4$RHORU), probably a DNA binding protein.

Based on these observations, the most plausible function of Y.SmaI gene is that of regulation of expression of the endonuclease gene, with which it is cotranscribed. It is plausible that the endonuclease should be subject to negative control because of its toxicity for unmodified DNA. As far as we know this would be the first example of a regulatory protein in a restriction-modification system, representing a new kind of regulatory mechanism for endonuclease expression. In two known examples of negative control (or delay) of class-II endonuclease expression simultaneous transcription of endonuclease and methylase is prevented by overlapping promoters (PstI gene locus of *E.coli*) (18) or by overlapping coding regions (NgoPII locus of *neisseria gonorrhoeae*) (19). One possible role for the coordinated transcription and, probably, translation of Y.SmaI and endonuclease is negative feedback that controls the enzyme concentration inside the cell.

VSH7

This slime mold protein was cloned from a set of vegetative specific mRNAs. It was classified as belonging to the H subclass of vegetative specific genes. The mRNA of most genes in this subclass is superinduced if protein synthesis is inhibited by cycloheximide (9). However, regulatory proteins implicated in this phenomenon have not yet been identified.

From the sequence alignment (fig. 1), the closest (33%/54aa/1 gap) relative of VSH7 of known 3D structure is the bacteriophage 434 cro protein. The sequence similarity is sufficient to imply homology of 3D structure (10). Most structurally essential residues (see above) are conserved identically, with the exception of G25N (R1-69 numbering). There is particularly strong local similarity to the phage 434 cro proteins in the region at end of helix 1 up to beginning of helix 2: 11 out of 14 residues are identical. In the sequence database search the closest (27%/70 aa) HTH relative of VSH7 is DNU4$RHORU from the purple bacterium *rhodospirillum rubrum* (13), coded for by an open reading frame (URF4) near ATP synthase subunits. The precise function of DNU4 is unfortunately not known; perhaps it regulates the expression of nearby genes.

From these observations, the most plausible function of VSH7 is to regulate the expression of genes which affect the transition from the vegetative to the developmental phase. It either helps assure the proper level of expression of genes required for maintenance of the growth phase or helps repress the expression of genes required for initiation of subsequent developmental phases. An interesting analogy is with the inhibitor of sporulation from *bacillus subtilis* (SIN$BACSU), also one of the closest known relatives (33%/ 60 a.a) of phage 434 repressor (PRC1$BP343): overexpression of that protein inhibits the transition from vegetative growth to sporulation in *bacillus subtilis* (14,15). A somewhat more complicated role is suggested by the excessive levels of transcription of genes in the H subclass of VS specific genes when protein synthesis is inhibited (9), probably as the result of elimination of a negative regulator of transcription: VSH7 may be such a negative regulator of expression of subclass H genes. As far as we know, VSH7 is the first cro-like eukaryotic protein identified.

CONCLUSIONS

Although the predicted function of the two proteins appears very plausible, experimental verification or falsification will have the last word. Our prediction may facilitate the design of experiments designed to test the role of these proteins in gene regulation. To name only one example: if altered genes can be reintroduced into cells, point mutations in residues analogous to those that make contact with DNA in the repressors of known structure (fig. 1), for either protein, should disrupt DNA binding and thus have a strong effect on growth phenotype (without significantly altering the native fold of the protein domain). We suggest to test mutations in the positions of Q28,Q29, E32 and Q33 of R1-69 repressor, that is R41,T42, G45, S46 in Y.SmaI and P65, G66, N69, E70 in VSH7 or in solvent-exposed residues on the protein surface immediately adjacent to these positions.

NOTE ADDED IN PROOF

T.Tao et al. (J.Bact. 173, 1367−1375 (1991)) have independently obtained evidence for the regulatory and DNA binding role of small open reading frames of the SmaI as well as of the EcoRV, PvuII, BamHI restriction-modification systems, consistent with our first result. Note that our second result, the sequence similarity involving the VSH7 *dictyostelium* gene, lies at the same level of significance.

ACKNOWLEDGEMENTS

We are grateful to Reinhard Schneider, Peter Rice, Gerrit Vriend, Peer Bork, and Brigitte Altenberg for interesting discussions and for help with database searches; to Toby Gibson, Peter Sibbald, Des Higgins and Ian Mattaj for comments on the manuscript; and to Charles Singleton and Hans Domdey for sharing their most recent experimental results.

METHODS

The database alignment search for similar protein sequences was performed using the dynamic programming alignment search MaxHom (Sander and Schneider, unpublished) with selection of significant hits according to the threshold for structural homology (10). This threshold is based on an extensive survey of sequence alignments between proteins of known structure and quantifies the very strong but often neglected dependence of the significance of alignment scores on the length of the aligned subsequences. For example, a sequence identity of 26% is probably significant for an alignment length of 100 residues, but not for one of 60 residues. As a control, the fast database search of Pearson and Lipman was used (11).

As an additional check on the correctness of the hypothetical relationship resulting from the alignment search, the two proteins were scored for the presence of an HTH (helix-turn-helix) consensus motif by the method of Dodd and Egan (5) as implemented by Peter Rice (unpublished). The HTH pattern scores were good, i.e. they were in the range of scores for proteins in the consensus dataset of Dood and Eagan. Finally, full sequence alignment (MaxHom) with higher weights on residues known to be conserved in the family of repressor proteins gave full confidence in the overall similarity of three-dimensional structures.

Secondary structure assignments were calculated from the 3D coordinates using DSSP (12). The protein sequence database was Swissprot release 16 (Amos Bairoch, University of Geneva and EMBL Data Library), with 18364 sequences.

RESULTS

Multiple alignment

The results of the database searches are summarized as a multiple sequence alignment (fig. 1). The eight proteins shown are a subset of the approximately 100 known proteins containing the HTH motif. For this subset the HSSP threshold indicates safe sequence similarity either to Y.SmaI (sequences #1,5,6,7 and 8 in fig. 1) or to YSH7 (sequences #2 and 5). For example, the sequence similarity between Y.SmaI and R1-69 is 32% identical residues over a length of 59 residues with an insertion/deletion gap of one residue—hereinafter written as 32%/60aa/1gap. For this length of alignment, 32% is 3 percentage points above the homology threshold of 29% (10). The Y.SmaI / R1-69 alignment ranks number 9 out of more than 18000 in the database search when scores are ranked by their distance from the homology threshold. The same alignment ranks number 174 in the control search using FASTA (11) (ktup=1), too low to be noticed in routine searches. For VSH7 and phage 434 cro the sequence

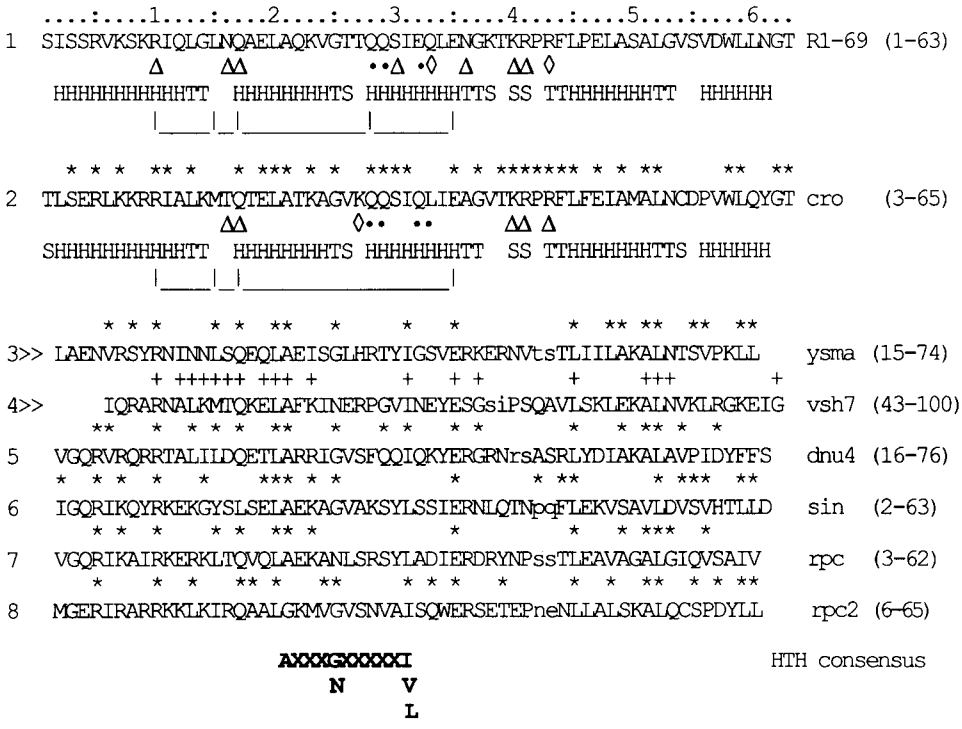

Figure 1. Multiple sequence alignment of proven and putative DNA binding regulatory proteins. 1 and 2 are repressors of known 3D structure; 3 and 4 (marked >>) are the subject of this paper; 5 to 8 are there closest relatives among all proteins of known sequence. Protein identifiers are 1: repressor 1R69 (RPC1$BP434) (2,3), 2: cro protein 2CRO (RCRO$BP434) (4), 3: YSMA$SERMA (7), 4: VSH7$DICDI (8), 5:DNU4$RHORU (11), 6: inhibitor of sporulation SIN$BACSU (14,15), 7: repressor RPC$BPPH1 (16), 8: C2 repressor RPC2$BPP22 (17) (Protein Data Bank and/or Swissprot identifiers; species codes are BP434 = *bacteriophage 434*, SERMA = *serratia marcescens;* DICDI = *dictyostelium discoideum;* RHORU = *rhodospirillum rubrum;* BACSU = *bacillus subtilis;* BPHH1 = *bacteriophage phi-105*). For the two proteins of known structure (1R69 and 2CRO), secondary structure is given in terms of H = helix, T = H-bonded turn, S = bent loop, blank = extended loop (12). Residues known to contact DNA backbone (△), base-pairs (●) and both (◇) are indicated for 1R69 and 2CRO. The H-bond network connecting helices 1,2 and 3 is indicated by thin lines below the sequences. Above each sequence residues are marked as identical to 1R69(*) or 2CRO(+). In each case sequence identity relative to the more similar protein of the two is indicated (* or +). The consensus pattern for H-T-H motif (simplified from ref. 5) is at the bottom. Sequences are in the one letter amino acid code in upper case, except for residues adjacent to insertion/deletion gaps which are in lower case, e.g NVtsTL means that there is an insertion between T and S (of unspecified length).