Reprinted from

FEBS Federation of European
Biochemical Societies
12th Meeting Dresden 1978

Volume 51

GENE FUNCTIONS

Edited by S. ROSENTHAL et al

PERGAMON PRESS OXFORD and NEW YORK 1979

COMPLEMENTARITY AND RECOGNITION CODE BETWEEN REGULATORY PROTEINS AND DNA

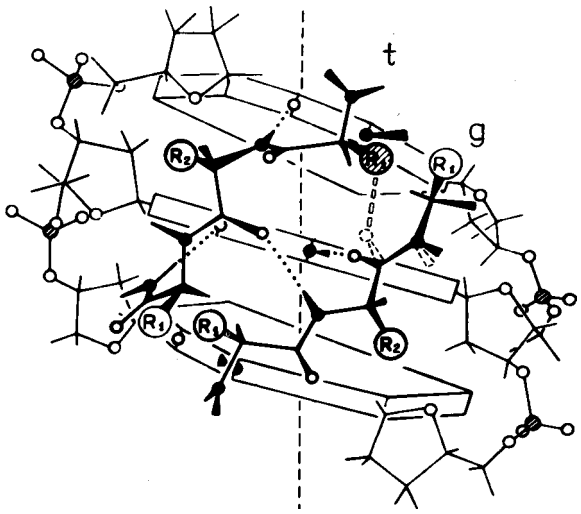
G.V. Gursky, A.S. Zasedatelev,
V.G. Tumanyan, A.L. Zhuze, S.L. Grokhovsky
and B.P. Gottikh
Institute of Molecular Biology Academy of
Sciences of the USSR, Moscow, USSR

ABSTRACT

Arguments are summarized which support the existence of a universal recognition code.

Is there a correspondence (code) between the protein and nucleic acid sequences implicated in specific binding interactions? This communication deals with this particular aspect of the recognition problem. We shall demonstrate that for a number of specific protein-nucleic acid complexes there is a remarkable correspondence between the protein and nucleic acid sequences which suggests the existence of a universal recognition code. At present it seems plausible that protein sites involved in specific interactions with double-helical DNA and RNA consist of two polypeptide chain segments forming a right hand-twisted antiparallel β-sheet (1-6). Earlier (4-6) we have suggested that the β-sheet, upon specific protein binding to DNA, undergoes a transition to a structure illustrated in Fig.1.

This structure may be considered as a deformed antiparallel β-sheet in which half of hydrogen bonds are broken and substituted by hydrogen bonds connecting the polypeptide chain backbone NH and CO groups with the DNA base pairs. The adenine N3 and pyrimidine 02 atoms belonging to one and the same polynucleotide strand serve as acceptor sites for hydrogen bonding to the backbone NH groups of the t-chain segment whereas the guanine 2-amino groups in the opposite polynucleotide strand act as hydrogen bond donors for the interaction with the backbone CO groups of the g-chain segment (see Pig. 1).



AT—coding residues:
Ser. Thr, Asn.Gln.Cys, His

GC—coding residues;
Gly. Ala.Leu.Val, I le. Met,
Phe, Tyr, Trp,Asp,Glu,Lys, Arg

Pig. 1. General structural motif within the recognition protein site involved in specific binding interactions with DNA

Symbols: O , oxygen; • , nitrogen; @, phosphorus. The dotted lines represent hydrogen bonds. The hydrogen bonding between the guanine 2-amino group and the backbone CO group of the g-chain segment is controlled by amino acid residues located in the t-chain segment in position R_1 (hatched circle). Indicated are the amino acid residues coding for AT and GC base pairs. In the presence of an AT-coding residue (such as serine) the hydrogen bond with guanine 2-amino group is weakened or broken, and a new hydrogen bond is formed between the serine side chain hydroxyl and the backbone CO group. In the complex, the $N-C^{\alpha}-C'$ sequence in the polypeptide chain segments coincides with the C3'-C5' direction in the corresponding polynucleotide chains.

Since ⅢA sequences recognized by various repressors and activators exhibit an asymmetric distribution of guanine bases between the two polynucleotide strands, the recognition protein sites are complementary to regulatory base pair sequences. Earlier (4-6) we have suggested that a more detailed complementarity allowing for regulatory

proteins to recognize their specific binding sites on DNA is governed by the side chain-backbone interactions in the recognition protein sites. These interactions may control the hydrogen bonding between the peptide groups of recognition protein site and DNA base pairs. From consideration of hydrogen bonding possibilities for various amino acid side chains a conclusion can be drawn that only four outward-pointing R_1 residues surrounding a given CO group of the g-chain segment (see Fig. 1) can in principle control its hydrogen bonding state. The real situation is probably even simpler since the most favourable conditions for these side chain-backbone interactions are realized for the side chains of Ser, Thr, Asn, Gln, Cys and His present in the t-chain segment in position R_1 indicated in Pig. 1. This points toward a linear recognition code, in which these six residues code for AT base pairs while other residues, such as Gly, Ala, Val, Leu, lle,Met, Phe, Trp, Asp, Glu, Tyr and presumably also Lys and Arg, being located in this position, are unable to interact by this mechanism and code for GC pairs. However, it should be noted that lysine, arginine and glutamine residues lying in other positions R_1 around a given CO group of the g-chain segment may form hydrogen bonds with this group and serve as AT-coding residues.

We have applied the linear code summarized in Fig. 1 in order to determine which regions of the lac repressor polypeptide chain are implicated in specific interactions with the lac operator (7). We have found that the repressor polypeptide chain segment ranging from Thr 19 to Val 30 exhibited a correspondence with the base pair sequence on the left side of the lac operator at six separate positions (Fig. 2A). It is well-known that these regions of lac repressor (8) and operator (9) are implicated in specific binding. A prominent feature of the lac operator sequence is the presence of a twofold rotation symmetry which probably dictates symmetrical attachment of the lac repressor subunits to DNA. Figure 2 shows that the base pair sequence on the right operator side does not fit the repressor protein sequence in two positions. This suggests that the repressor subunit interacting on the left-hand side of lac operator forms a greater number of favourable specific contacts with base pairs than does the repressor subunit interacting on the right operator side.

Figure 2B shows the result of similar analysis for S8 ribosomal protein (10) and its specific interaction site on ribosomal 16S RNA (11). Again, GC- and AU-coding residues occur at the appropriate positions in the proposed complex. There is a correspondence between the protein and nucleic acid sequences at five separate positions. This correspondence enables us to predict that

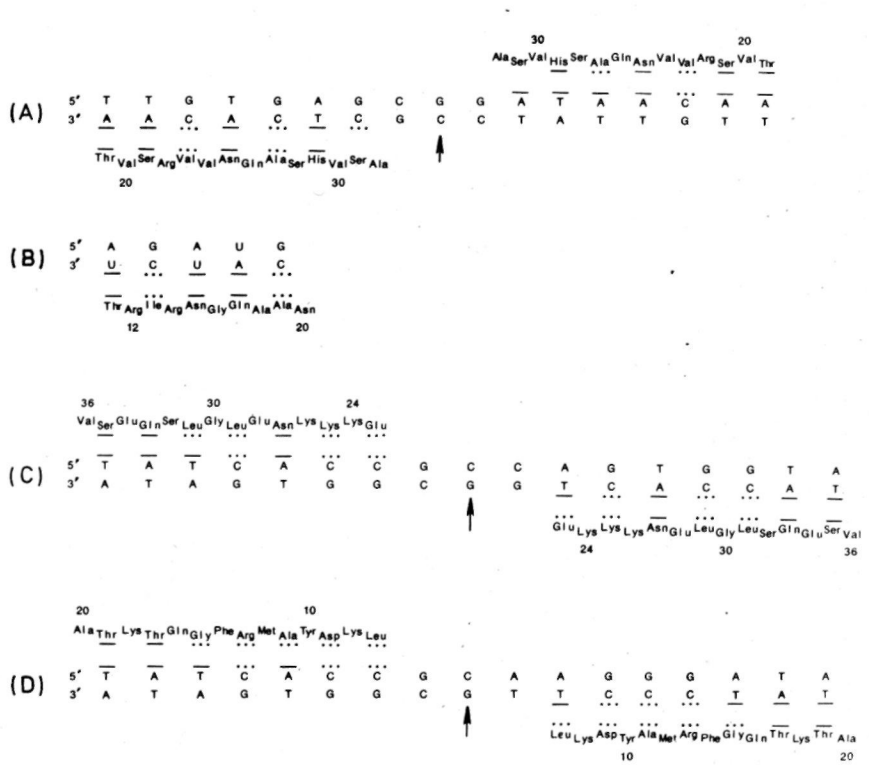


Fig. 2. Proposed correspondences between the protein and nucleic acid sequences for several specific protein-nucleic acid complexes

Amino acid residues coding for AT and GC base pairs are underlined by continuous and dotted lines, respectively. Vertical arrows indicate the positions of twofold symmetry axes. (A) lac repressor and lac operator; (B) S8 ribosomal protein and its interaction site on ribosomal 16S RNA; (C) λCI repressor and its strongest affinity binding site OL_1; (D) λ cro protein and its strongest affinity binding site OR_3.

the segment of the S8 polypeptide chain ranging from Thr 11 to Asn 20 is involved in specific binding to ribosomal 16S RNA (6). This prediction agrees with the recent observations showing that the fragment containing the N-terminal 47 residues of S8 protein binds to the same specific site on 16S RNA as does the intact S8 protein (12)

Protein sequences have been determined very recently for λ CI and λ cro repressors (13, 14). These proteins can recognize the same IIIA sequences (15) although their amino acid sequences have no obvious sequence-homologies. This fact strongly suggests that the recognition code is highly degenerated with respect to protein sequences. It is well-known that the λ CI repressor binds to two operators, O_L and O_R, each of which consists of three non-identical sites recognized by the repressor (16). The best correspondence is found between the repressor polypeptide chain segment ranging from Glu 23 to Val 36 and the strongest affinity binding sites OL_1 and OR_1 for the repressor (Pig. 2C). The nucleotide sequences for these binding sites are identical except the AT —» TA replacement in one position. The extent of correspondence between the protein and DNA sequences for various repressor binding sites is found to correlate with the affinity of the repressor for these sites. In our tentative analysis of this system (6), we took a great value to the presence of negatively charged residues in this part of the repressor polypeptide chain and suggested that the t-chain segment of the repressor recognition site is likely to be formed from residues present in the middle part of the repressor polypeptide chain. We now think that this proposal was wrong. Prom inspection of molecular models we have found that many of negatively charged residues in the sequence 23-36 can be neutralized by positively charged residues lying in the same sequence. Genetic evidence indicates that the operator-DNA binding site for the repressor involves the N-terminal 80 residues (17).

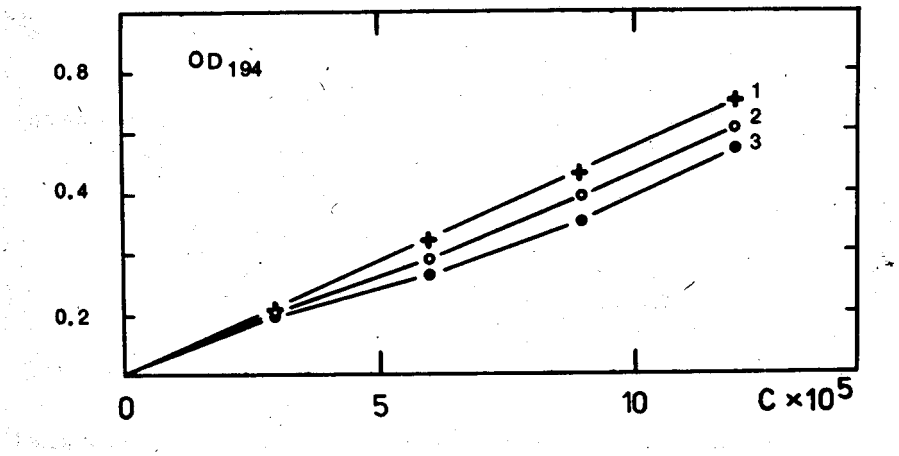
Similar analysis for the binding of the cro protein shows that the amino-terminal part of the protein is the only region which exhibits a correspondence with the base pair sequences in O_L and O_R operators. The best correspondence is found between the protein sequence and the nucleotide sequence in the binding site O_{R3}, (Fig. 2D). This agrees with the recent observations that OR_3 is the highest affinity binding site for the cro protein (15). Since the cro protein consists of 66 residues, it is unlikely that this correspondence between the protein and DNA sequences is a random coincidence.

An interesting possibility to test the validity of the proposed code is to study the binding of β-polypeptides

to natural and synthetic nucleic acids with defined base sequences. We have synthesized several oligopeptides (such as oligo(L-valine) and oligo(L-threonine) of various degree of polymerization and measured their binding to nucleic acids. All these oligopeptides have a limited solubility in aqueous solution and in water-methanol mixtures. They tend to form aggregates of various size which are in concentration-dependent equilibria with each other and with monomers. Oligo(L-valine)$_5$ exhibit a greater tendency to form a β-structure than do oligo(L-threonine)s of the same size. If these oligopeptides exist in a β-conformation as revealed by CD spectra, they exhibit a general affinity for nucleic acids. The binding is accompanied by changes in the UV and CD spectra in the region of 190-240 nm where the amide groups of the oligopeptide molecules absorb the light. However, only minor spectral changes occur in the DNA absorbance band with a peak at 260 nm. The binding has been also detected by gel filtration technique. The strongest affinity binding sites for oligo(L-Thr)$_{30}$-OCH$_3$ (30 is the average degree of polymerization) on poly(dA)•poly(dT) are saturated when about four threonine residues are bound per base pair.

Oligopeptides in a single-stranded form exhibit a low affinity for the nucleic acid. A shift in the raonomer-dimer equilibria is observed on adding DNA at relatively low oligopeptide concentrations where the dimers are predominantly in equilibria with the monomers. If concentrated ($2 \cdot 10^{-2}$ M) aqueous solution of oligo(L-threonine) is diluted 50 to 1000 times, a drop in the molar extinction coefficient of the oligopeptide is observed at 194 nm with an accompanying decrease in the molar dichroism at 218 nm. These changes reflect a slow disaggregation process with a characteristic time of about 20 hours. In the presence of DNA, these spectral changes are smaller, thus, indicating a shift in the aggregation equilibria toward a β-form of oligo(L-threonine). Figure 3 shows that poly(dA)•poly(dT) and poly(dI)•poly(dC) have a greater effect on the aggregation equilibria than poly(dG)•poly(dC) under the same conditions. These experiments suggest that the oligo(L-threonine) exhibits a binding preference for poly(dA)•poly(dT) and poly(dl)•poly(dC).

Other experiments show that the oligo(L-valine) containing 5 to 7 valine residues exhibits an opposite order of preferences binding more strongly to poly(dG)•poly(dC) than to poly(dA)•poly(dT). Although the origin of these preferences remains unclear, these observations are in accordance with the predictions of the proposed code.



Pig. 3. Effect of various synthetic *ШАв* on the aggregation equilibria in solutions of oligo(L-Thr)$_{30}$-OCH$_3$

Aliquots of concentrated oligo(L-threonine) solution ($2 \cdot 10^{-2}$ M) were added to a series of tubes containing synthetic DNAэ in 0.01 SSC (1 SSC contains 0.15 M NaCl, 0.015 M sodium citrate). After storage for 24 hr at 20°, the difference spectra of the complexes were monitored against solutions of the corresponding nucleic acids. OD_{194} is the measured difference in the optical density at 194 nm in 1 cm-path cell. C is the molar concentration of oligo(L-threonine) (on threonine basis). Curve 1, oligo(L-threonine) in the presence of poly(dA)•poly(dT); curve 2. oligo(L-threonine) in the presence of poly(dG)• poly(dC); curve 3, oligo(L-threonine) in the absence of DNA. Data obtained for oligo(L-threonine) in the presence of poly(dI)•poly(dC) and poly(dA)•poly(dT) were coinciding. Concentrations of nucleic acids were $1.2 \cdot 10^{-4}$ M (base pairs).

We have also synthesized a number of distamycin A (DM) analogs with two methylpyrrole carboxamide units (DM2) linked through a glycine residue to oligo(L-valine) chains of various size. The antibiotic distamycin A binds in the minor DNA groove by hydrogen bonding to the adenine N3 and pyrimidine O2 atoms lying in one and the same polynucleotide Chain (18,19).The antibiotic molecule carries four AT-specific reaction centres each being associated with the antibiotic amide group. Since the antibiotic

molecule somewhat imitates the t-chain segment in the model shown in Fig. 1 we have synthesized a number of DM analogs containing various number of methylpyrrole carboxamide groups and investigated their binding to natural and synthetic DNAs (20).

Prom the analysis of experimental binding isotherms obtained for these analogs the following conclusions were drawn: i) the binding free energy to poly(dA)•poly(dT) depends linearly on the number of amide groups in the DM analogs; ii) the free energy change associated with the bonding of a single amide group is equal to 2 kcal/mole in the case of binding to poly(dA)•poly(dT) and is about 1 kcal/mole in the case of binding to poly(dG)•poly(dC); iii) DM analogs carrying oligo(L-valine) chains tend to form dimers through the formation of an antiparallel β structure between the oligo(L-valine) chains and exhibit a greater extent of binding specificity than DM itself.

We have found that for relatively short oligo(L-valine) chains containing about 5 to 10 valine residues, the monomeric and dimeric forms of the peptide analogs are in concentration-dependent equilibria. The CD spectra for the peptide analogs in the region of 190-240 nm (Fig. 4) can be represented as a superposition of two reference spectra: one characteristic of β-polypeptides and one usually observed for polypeptides in a random-coiled conformation.

The CD measurements show that these analogs in the monomeric and dimeric forms bind to DNA showing positive band with a peak at about 320 nm which is characteristic of the bound form of DM2. Fig. 4 shows a typical titration curve obtained for the binding to poly(dA)•poly(dT) of the DM2 peptide analog containing about 10 valine residues. The initial slope of the curve coincides with that expected if all the added antibiotic molecules were bound to poly(dA)•poly(dT). A well-defined saturation level of binding allows one to determine the size of the binding site which is equal to about 6 base pairs. Since DM2 covers four base pairs upon binding, the observed increase in the site size can be ascribed to the presence of the oligo(L-valine) portion of the peptide analog. Since the two linked antibiotic molecules exhibit a greater binding specificity than distamycin itself, this approach may have a general applicability to chemical designing of sequence-specific ligahds. If further experiments confirm that peptide fragments in these complexes can interact specifically with DNA base pairs, this approach will also be useful for direct testing of the proposed code.

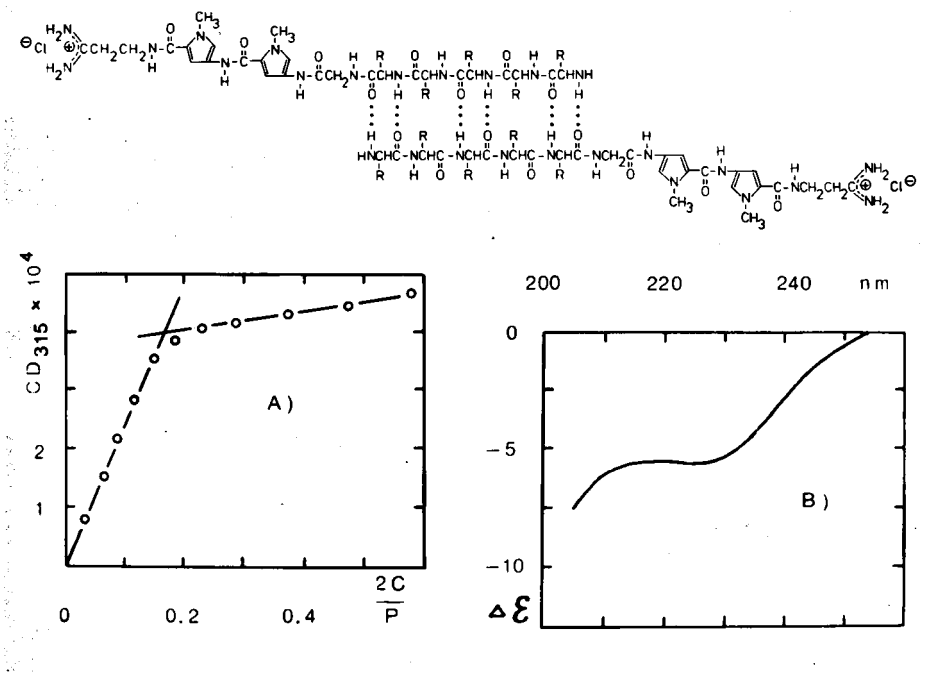


Fig. 4» Binding of distamycin peptide analogs to DNA

(A), CD amplitude of the complexes between the DM2 peptide analog containing 10 valine residues and poly(dA)• poly(dT) as a function of 2C/P, the ratio of moles of DM2 to moles of DNA base pairs. CD_{315} is the measured dichroism in 1 cm-path cell. The concentration of DNA was 1.1 • 10^{-4} M (base pairs). (Б), CD spectrum of the peptide analog containing 10 valine residues in the absence of DNA. ^ε is the measured dichroism expressed per mole of DM2 and per 1 cm pathlength. The concentration of the peptide analog of DM2 was 2.05 • 10^{-4} M. Indicated is the proposed structure for the dimeric form of a peptide analog of DM2 containing five valine residues.

REFERENCES

(1) Ch.W. Carter & J. Kraut, Proc. Nat. Acad. Sci. USA 71, 283 (1974).

(2) S.-H. Kim, J.L. Sussman & G.M. Church, in Structure and Conformation of Nucleic Acids and Protein-Nucleic Acid Interactions, eds. M. Sundaralingam

& S.T. Rao, p. 571, University Park Press, Baltimore (1975).

(3) G.M. Church, J.L. Sussman & S.-H. Kim, Proc. Nat. Acad. Sci. USA 74, H58 (1977). ~

(4) G.V. Gursky, V.G. Tumanyan, A.S. Zasedatelev, A.L. Zhuze, S.L. Grokhovsky & B.P. Gottikh, Mol. Biol. USSR 9, 635 (1975).

(5) G.V. Gursky, V.G. Tumanyan, A.S. Zasedatelev, A.L. Zhuze, S.L. Grokhovsky & B.P. Gottikh, Mol. Biol. Rep. 2, 413 (1976).

(6) G.V. Gursky, V.G. Tumanyan, A.S. Zasedatelev, A.L. Zhuze, S.L. Grokhovsky & B.P. Gottikh, in Nucleic Acid-Protein Recognition, ed. H.J. Vogel, p. 189, Academic Press, New York (1976).

(7) G.V. Gursky, V.G. Tumanyan, A.S. Zasedatelev, A.L. Zhuze, S.L.Grokhovsky & B.P. Gottikh, Mol. Biol. Rep. 2, 427 (1976).

(8) B. Muller-Hill, B. Gronenborn, J. Kania, M. Schlotmann & K. Beyreuther, in Nucleic Acid-Protein Recognition, ed. H.J. Vogel, p. 219, Academic Press, New York (1976).

(9) W. Gilbert, J. Gralla, J. Majors & A Maxam, in Protein-Ligand Interactions, eds. H. Sund and G. Blauner, p. 193, de Gruyter, Berlin (1975).

(10) H. Stadler, FEBS Letters 48, 114 (1974).

(11) E. Ungewickell, R. Garrett, C. Ehresmann, P. Stiegler & P. Fellner, Eur. J. Biochem. 51, 165 (1975).

(12) J. Bruce, E.J. Pirpo & H.W. Schaup. Nucleic Acid Res. 4, 3327 (1977).

(13) R.T.Sauer & R. Anderegg, Biochemistry 17, 1092 (1978).

(14) M.W. Hsiang, R.D. Cole, Y. Takeda & H. Echols, Nature 270, 275 (1977).

(15) A. Polkmanis, Y. Takeda, J. Simuth, G. Gussin & H. Ecohls, Proc.Nat. Acad. Sci. USA. 73, 2249 (1976).

(16) T. Maniatis, M. Ptashne, K. Backman, D. Kleid, S. Plashman, A. Jeffrey & R. Maurer, Cell 5, 109 (1975).

(17) A.B. Oppenheim & D. Noff, Virology 64, 553 (1975).

(18) Ch. Zimmer, Prog. Nucleic Acid Res. Mol. Biol. 15, 285 (1975).

(19) A.S. Zasedatelev, A.L. Zhuze, Ch. Zimmer, S.L. Grokhovsky, V.G. Tumanyan, G.V. Gursky & B.P. Gottikh, Dokl. Akad. Nauk SSSR 231, 1006 (1976).

(20) A.S. Krylov, S.L. Grokhovsky, A.S.Zasedatelev, A.L. Zhuze, G.V. Gursky & B.P. Gottikh, Dokl. Akad. Nauk SSSR 239, 732 (1978).

Reprinted from
FEBS Federation of European
Biochemical Societies
12th Meeting Dresden 1978

Volume 51

GENE FUNCTIONS

Edited by S. ROSENTHAL et al

PERGAMON PRESS OXFORD and NEW YORK 1979