Structural Studies on the Basic-Helix-Loop-Helix Region from MyoD

by

Philip Chun-Ming Ma

A.B., Biochemical Sciences
Harvard University, 1986

Submitted to the Department of
Biology in Partial Fulfillment of
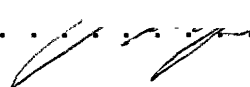the Requirements for the
Degree of

DOCTOR OF PHILOSOPHY

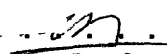at the
Massachusetts Institute of Technology

May 1994

© 1994 Philip C.M. Ma
All rights reserved

The author hereby grants to MIT permission to reproduce and to
distribute publicly paper and electronic copies of this thesis
document in whole or in part.

Signature of Author . . . .
Department of Biology
April 22, 1994

Certified by . . . .
Carl O. Pabo, Professor of Biophysics and Structural Biology
Thesis Supervisor

Accepted by . . . .
Frank Solomon, Professor of Biology
Chairman, Biology Graduate Committee

Science

STRUCTURAL STUDIES ON THE BASIC-HELIX-LOOP-HELIX REGION OF MYOD

by
Philip Chun-Ming Ma

submitted to the Department of Biology in partial fulfillment of the
requirements for the degree of Doctor of Philosophy

ABSTRACT

This thesis describes the structure determination of the MyoD basic-helix-loop-helix (bHLH) region bound specifically to its DNA site, and discusses the implications of the structure for understanding recognition and regulation.

Chapter 1 provides an introduction to the biology of the myogenic proteins, of which MyoD is a member. Many of the important biological activities of the myogenic proteins, including DNA binding, dimerization, and control of transcription, are mediated through the bHLH regions of these proteins.

Chapter 2 describes the cocrystallization experiments with MyoD peptides containing the bHLH region and a series of synthetic DNA oligonucleotides. The purity of the MyoD peptide and the sequence and length of the synthetic DNA oligonucleotide used for the cocrystallization were crucial for obtaining crystals suitable for structure determination.

Chapter 3 describes the determination of the structure of the MyoD/DNA complex. We used the method of multiple isomorphous replacement using DNA which had 5-iodouracil substituted for thymine. Purification of the iodine-substituted DNA oligomers was critical for obtaining isomorphous crystals suitable for phasing.

Chapter 4 presents the structure of the MyoD/DNA complex. We focus on structural aspects of DNA recognition, and transcriptional activation. We find that two of the three residues, Ala 114 and Thr 115, which are important for positive control, are buried within the major groove of the DNA, and not easily accessible for protein-protein interactions. Analysis of the structure suggests that these two residues may affect transcriptional activation by determining the position of a neighboring arginine. The experiments from this chapter will be published as "Crystal Structure of MyoD bHLH-DNA Complex: Perspectives on DNA Recognition and Implications for Transcriptional Activation" (Ma, P.C.M., Rould, M.A., Weintraub, H., and Pabo, C.O. (1994) *Cell* (in press)).

Chapter 5 expands upon aspects of the MyoD/DNA structure which were not discussed in detail in Chapter 4. In particular we focus on the structural determinants for dimer formation among the bHLH proteins.

Chapter 6 summarizes the structural analysis from Chapters 4 and 5 and suggests possible directions for future research.

Thesis supervisor: Carl O. Pabo

This thesis is dedicated to my parents,

Dr. Hung-Fook Ma and Theresa Ma

ACKNOWLEDGMENTS

with the coordinates of the Max bHLH-ZIP-DNA complex, and have also helped us interpret our structure of MyoD.

Within the Pabo lab, I have benefited from the help and expertise of many members, past and present. I am especially grateful to Dr. Mark A. Rould for teaching me crystallography, for having unbelievable patience in answering my many questions, for being a constant source of help in solving and interpreting the structure, and for being a good friend within the lab. Mark, you will always be the guru of crystallography for me. To my baymate, Harvey Greisman, I must offer thanks for putting up with me for three years. Harvey deserves a note of congratulations for listening to my reactionary rantings on politics for three years, and still not wavering in his core set of beliefs. Edward Rebar, Juli Klemm, Wenqing Xu, Ernest Fraenkel, Monicia Elrod-Erikson, and Kristen Chambers also deserve kudos for putting up with me, and for many stimulating discussions on science and other matters. Nikola Pavletich was a source of inspiration for me, especially in his scientific intensity. Lena Nekludova was always available for help whenever I had problems with the computers, which was quite often. I would not have been able to make it through the administrative kudzu of ordering supplies, making applications to meetings, and being a graduate student without the help of Kristine Kelley and Amy Dunn. Thank you both for your patience.

Outside of the lab, I have several people to thank for making life more enjoyable. I shared many laughs with Patrick Waller and Carl Gordon over the years, many of them having to do with my mix-ups in lab. Thanks for helping me laugh at my mistakes, so I wouldn't be so tempted to cry over them. Vera Maljkovic has been a good friend since college, even though I still mispronouce her name. Thanks for humoring in this, and many other things. Lastly to Margaret, thank you for your love and support over the years. Intense benchwork has a way of making you lose perspective on life; thanks for keeping my field of vision clear.

BIOGRAPHICAL NOTE

Philip Chun-Ming Ma

Education

Massachusetts Institute of Technology 1988 to present
Ph.D. candidate (Biology). Research in Alexander Varshavsky's lab
on the genetics of selective protein degradation in yeast. Research
in Carl Pabo's lab on the biochemistry and structure of lambda O
protein. Thesis research in Carl Pabo's lab on the crystallization
and structure determination of MyoD bound specifically to DNA.
Teaching assistant for Physical Biochemistry, 1989 and 1991.

University of Oxford 1986-1988
MPhil. Degree in Economics, June 1988.

Harvard University 1982-1986
A.B. Degree in Biochemical Sciences, *magna cum laude,* June 1986.

Achievements
1982 Governor General of Canada's Medal for the province of
Saskatchewan. 1986 Canadian Rhodes Scholar.

Publications

Wiren, K.M., Ivashkiv, L., Ma, P.C.M., Freeman, M., Potts, J.T.,Jr., and
Kronenberg, H.M. (1989). Mutations in signal sequence cleavage
domain of preproparathyroid hormone alter protein translocations,
signal sequence cleavage, and membrane-binding properties.
Molecular Endocrinology *3,* 240-250.

Ma, P.C.M, Rould, M.A., Weintraub, H., and Pabo, C.O. (1994). Crystal
Structure of MyoD bHLH-DNA Complex: Perspectives on DNA
Recognition and Implications for Transcriptional Activation. Cell.
In Press.

TABLE OF CONTENTS

LIST OF TABLES

Note: Table legends are at the end of each chapter

FIGURE LEGENDS

Note: Figure legends are at the end of each chapter.

Chapter 2

Chapter 3

Chapter 4

CHAPTER 1

The Regulatory Network of bHLH Proteins
Involved in Myogenesis

Proteins involved in gene regulation often have a modular architecture, being composed of several domains. By studying the function of these separate domains in addition to studying how these domains interact with each other, we can better understand the mechanisms of gene regulation. The transcriptional activity originally identified as TFIID, for example, now has been separated into a TATA-binding protein (TBP), which binds DNA, and several TATA-associated factors (TAFs), which appear to interact with and modulate the activity of RNA polymerase II (Dynlacht et al., 1991; Hoey et al., 1993; Weinzierl et al., 1993). Modularity of protein structure and function becomes even more pronounced in the control of gene expression beyond the basal level. The MyoD (Davis et al., 1987) family of proteins, which also includes Myf-5 (Braun et al., 1989b), MRF4 (Rhodes and Konieczny, 1989) and myogenin (Edmondson and Olsen, 1989), regulates myogenesis by controlling tissue- and temporally-selective gene expression in myoblasts and myotubes; these proteins are not present in fibroblasts or other non-muscle cell types (Weintraub et al., 1991). The myogenic proteins interact with each other, either directly or indirectly, and appear to take part in an autoregulatory loop of gene expression. MyoD can activate its own expression as well as the expression of myogenin, while myogenin and Myf-5, in turn, can activate their own expression and that of MyoD (Braun et al., 1989a; Thayer et al., 1989; Edmondson et al., 1992). MyoD appears to play a central role in the early stages of myogenesis. Expression of the MyoD gene alone is necessary and sufficient to induce myogenesis in 10T 1/2 fibroblasts (Weintraub et al., 1991). Moreover, for this induction of

myogenesis, only 68 amino acids of MyoD are required. This 68
amino acid region, known as the 'basic-helix-loop-helix' (bHLH)
motif (first identified by sequence analysis (Murre et al., 1989a)),
shows homology to the other myogenic proteins, as well as to the
oncogenic *myc*-family of proteins, to proteins involved in
Drosophila sex and neuronal development, and to the E2A proteins
E47 and E12, which bind the immunoglobulin enhancer sequence
(Murre et al., 1989b).

In addition to MyoD itself, expression of other members of the
MyoD family (Myf-5, myogenin, and MRF4) in fibroblasts and a
variety of other cell types can convert these different cell types
into muscle (Buckingham, 1992; Wright, 1992). It was thus thought
that the MyoD gene family played master regulatory roles in
skeletal muscle determination (Weintraub et al., 1991). In contrast
to the results of the cell culture experiments, which suggested a
fairly straightforward role for the MyoD family genes in muscle
development, the results of recent gene knockout experiments in
mice reveal a more complex mechanism for myogenic gene regula-
tion. Mice with homozygous knockouts of either *Myf-5* or *MyoD*
produced close to normal amounts of muscle (Braun et al., 1992;
Rudnicki et al., 1992). The two mutant mice strains had distinct
phenotypes. For the *Myf-5*-negative mice, although there was
seemingly no effect on muscle development, the mice failed to
develop a normal rib cage. In fact, the rib cage development was so
impaired in the *Myf-5*-negative mice that the mice had severe
respiratory problems and died immediately after birth. (Braun et al.,

1992). For the *MyoD*-negative mice, there was survival after birth, but a lower viability than for the wild-type (Rudnicki et al., 1992). The results of the single gene knockout experiments for *MyoD* and *Myf-5* thus seemed to conflict with the results of the cell culture experiments with respect to the importance of the MyoD gene family for muscle development. This apparent paradox was resolved recently in a double knockout experiment involving both the *MyoD* and *Myf-5* genes. Double homozygous mutants for *MyoD* and *Myf-5*, produced from a genetic cross of the two single homozygous mutants, had an interesting phenotype (Rudnicki et al., 1993). They were born alive, but were immobile and died soon after birth. Closer examination of the mice revealed that they did not express skeletal muscle-specific genes, did not have any skeletal muscle formation, and did not express any markers for skeletal muscle, revealing that myoblasts did not form in these mice. It thus appears that *MyoD* and *Myf-5* are at least partially redundant for the determination of skeletal myoblasts, but that expression of at least one of the genes is absolutely necessary for myoblast formation.

In addition to the single and double gene knockout experiments for *MyoD* and *Myf-5*, there has also been a recent study of a gene knockout for *myogenin* (Hasty et al., 1993). Like the *MyoD/Myf-5* double knockout mice, the *myogenin*-negative mice died immediately after birth with a severe reduction of skeletal muscle. The mutant mice expressed normal levels of MyoD mRNA, indicating that myoblasts had formed, but histological analysis revealed that these myoblasts failed to differentiate fully into myotubes. Thus, it

appears that MyoD or Myf-5 are necessary for myoblast determi-
nation, but that myogenin is needed for proper differentiation of the
myoblasts into muscle (Weintraub, 1993).

Given the partially overlapping functions of the myogenic
proteins, as revealed by cell culture and gene knockout experiments,
it is not surprising that the MyoD gene family would share some
structural similarities as well. Sequence analysis of the MyoD gene
family reveals a particularly strong homology in the bHLH motif
among these proteins. There is an 81% strict homology and 73%
sequence similarity among 60 of the 68 amino acids that comprise
the bHLH motif in MyoD, Myf-5 and myogenin. All of the myogenic
proteins oligomerize, bind DNA, and activate transcription. The
bHLH motif is necessary for all of these functions and sufficient for
oligomerization and DNA-binding (Weintraub et al., 1991). The
oligomerization function is of primary importance because bHLH
proteins, like the basic-leucines zipper (bZIP) proteins such as the
yeast transcription factor GCN4, only bind DNA as homo- or hetero-
dimers. The myogenic proteins appear to function best as hetero-
dimers with E47 or E12. MyoD forms homodimers less efficiently
than it forms heterodimers with E12 or E47 (Sun and Baltimore,
1991; Fairman et al., 1993). Once MyoD homodimers are formed,
however, they appear to bind DNA with a similar, but lower
efficiency than MyoD-E47 heterodimers (Sun and Baltimore, 1991;
Blackwell and Weintraub, unpublished).

It appears, then, that the activity of the MyoD family of pro-
teins may be positively regulated through interaction with the E-
proteins. Other experiments reveal that the MyoD family may be
subject to significant negative regulation as well. The Id (inhibitor
of DNA binding) protein contains a helix-loop-helix (HLH) motif, and
has been show to prevent DNA binding of other HLH-containing
proteins (Benezra et al., 1990). Id lacks a functional basic region
and can inhibit the DNA-binding function of proteins from the MyoD
or E2 family through the formation of inactive heterodimers
(Benezra et al., 1990; Sun et al., 1991). In addition to Id, other
proteins have been shown to negatively regulate the activity of the
myogenic proteins. Many of these negative regulators are involved
with the control of the cell cycle. In the development of muscle
cells, terminal differentiation of cells is mutually exclusive to cell
proliferation (Olson et al., 1991). Before myoblasts fuse to form
myotubes, they withdraw from the cell cycle (Miller,1992). During
this withdrawal from the cell cycle, myogenesis and the activity or
expression of the myogenic proteins is inhibited by a number of
negative regulators. The Ras and Fos oncogenes inhibit both MyoD
transcription and MyoD activity (Lassar et al., 1989). Jun can
physically interact through the helix-loop-helix region of MyoD and
the leucine zipper region of Jun; this physical interaction is thought
to inhibit the activity of both genes (Bengal et al., 1992). Direct
and/or indirect interactions between Transforming growth factor-
beta (TGF-β) and myogenin (Martin et al., 1992) and between the E1A
gene product and myogenin (Taylor et al., 1993) have also been
demonstrated. In both cases, the interaction results in an inhibition

of myogenin activity and is mediated through the basic-helix-loop-
helix region of the protein. It is clear then, that the activities of
MyoD and the other members of the family are extensively regulated
by interactions with other proteins. A large number of positive and
negative signals probably act directly on the MyoD family proteins
during the entire muscle development process, from determination
to final differentiation.

Protein-protein interactions have an important role in con-
trolling the activities of the myogenic proteins. Moreover, protein-
protein interactions are intimately involved in the activity of the
myogenic proteins themselves. The members of the MyoD family of
proteins bind DNA as homodimers or as heterodimers with the E-
proteins E12 and E47 (Murre et al., 1989b; Davis et al., 1990; Lassar
et al., 1991). In addition to stabilizing the structures of the indi-
vidual proteins [the MyoD monomer is to a large extent, unstruc-
tured (Anthony-Cahill et al., 1992)], the formation of homo- and
heterodimers also helps determine the specificity of DNA binding.
All bHLH proteins recognize a consensus DNA binding site known as
the E-box (CANNTG) that is present in the regulatory regions of
many tissue-specific genes; these include the muscle specific
genes for the myosin light chain and the muscle creatine kinase
(Weintraub et al., 1991). MyoD and E47 show preferential binding to
different DNA sites. (Blackwell and Weintraub, 1990). Moreover,
the MyoD-E47 heterodimer favors a DNA-binding site comprising the
separate half-sites of the MyoD and E47 homodimers rather than a
new whole site. Thus, the pairing of MyoD with different bHLH

proteins may be a means to discriminate between different E-box sites.

In summary, the activity of myogenic proteins are subject to the control of a number of bHLH and non-bHLH proteins through direct interactions, the DNA binding activity of the myogenic proteins themselves is dependent on protein-protein interactions, and through binding to different DNA sites, the myogenic proteins directly control the expression of a number of muscle specific genes, and other genes as well [MyoD has recently been shown to bind to the enhancer element of the c-fos promoter and by so doing, block the transcription of the c-fos gene (Trouche et al., 1993)]. For the family of myogenic proteins, of which MyoD is one of the best studied members, many of the protein-protein and all of the protein-DNA interactions are mediated by the bHLH domain. Extensive biochemical and genetic studies have shown that the basic region of the bHLH motif contacts the DNA, while the HLH region is responsible for oligomerization (Davis et al., 1990). For MyoD and the other myogenic proteins, there is a particularly rich supply of mutagenesis data which documents the role of many of the amino acids contained within the bHLH domain (Davis et al., 1990; Brennan et al.,1991; Winter et al., 1992). Residues important for DNA binding, and oligomerization, as well as for transcriptional activation have been determined by these studies. Sequence homology among the bHLH domains across a wide spectrum of eukaryotic proteins suggests that the tertiary fold of the bHLH domain would be fairly well conserved.

When we began our studies, there was little known about the tertiary structure of bHLH (or bHLH-ZIP) domains. Since it was known that bHLH and bHLH-ZIP domains bound DNA as dimers, many of the initial models of these proteins were based on four helix bundles. Some of these models had parallel bundles (Anthony-Cahill et al., 1992; Vinson and Garcia, 1992; Halazonetis and Kandil, 1992; Davis and Halazonetis, 1993), while other models featured anti-parallel bundles (Starovasnik et al., 1992; Gibson et al., 1993). Since little was known about the structure of the bHLH domains, and since there was a wealth of biochemical and genetic data on MyoD and the other myogenic proteins, we had strong motivations to undertake structural studies of the MyoD bHLH domain. We hoped that a detailed structure of the MyoD bHLH domain bound to DNA would provide a framework for answering three important structural questions concerning the bHLH family of proteins: 1) How does the bHLH basic region distinguish between different E-box sites; 2) How does the bHLH dimerization region distinguish between different oligomeric partners and what stabilizes these dimeric interactions; 3) How might the structure of the bHLH domain contribute to transcriptional activation and how does the structure of the MyoD complex explain the biochemical and genetic data on transcriptional activation in myogenic proteins?

References

Anthony-Cahill, S.J., Benfield, P.A., Fairman, R., Wasserman, Z.R., Brenner, S.L., Stafford, W.F., III, Altenbach, C., Hubbell, W.L., and DeGrado, W.F. (1992). Molecular characterization of helix-loop-helix peptides. Science *255*, 979-983.

Benezra, R., Davis, R.L., Lockshon, D., Turner, D.L., and Weintraub, H. (1990). The protein Id: a negative regulator of helix-loop-helix DNA binding proteins. Cell *61*, 49-59.

Bengal, E., Ransone, L., Scharfmann, R., Dwarki, V.J., Tapscott, S.J., Weintraub, H., and Verma, I.M. (1992). Functional antagonism between c-Jun and MyoD proteins: a direct physical association. Cell *68*, 507-519.

Blackwell, T.J., and Weintraub, H. (1990). Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. Science *250*, 1104-1110.

Braun, T., Bober, E., Buschhausen, D.G., Kohtz, D.G., Grzeschik, K.H., and Arnold, H.H. (1989a). Differential expression of myogenic determination genes in muscle cells: possible autoactivation by the Myf gene products. EMBO J. *8*, 3617-3625.

Braun, T., Buschhausen, D.G., Bober, E., Tannich, E., and Arnold, H.H. (1989b). A novel human muscle factor related to but distinct from MyoD1 induces myogenic conversion in 10T1/2 fibroblasts. EMBO J. *8*, 701-709.

Braun, T., Rudnicki, M.A., Arnold, H.-H., and Jaenisch, R. (1992). Targeted inactivation of the muscle regulatory gene *Myf-5* results in abnormal rib development and perinatal death. Cell *71*, 369-382.

Brennan, T.J., Chakraborty, T., and Olson, E.N. (1991). Mutagenesis of the myogenin basic region identifies an ancient protein motif critical for activation of myogenesis. Proc. Natl. Acad. Sci. USA *88*, 6675-5679.

Buckingham, M. (1992). Making muscle in mammals. Trends Genet. *8*, 144-148.

Davis, L.J., and Halazonetis, T.D. (1993). Both the helix-loop-helix and the leucine zipper motifs of c-myc contribute to its dimerization specificity with Max. Oncogene *8*, 125-132.

Davis, R.L., Cheng, P.F., Lassar, A.B., and Weintraub, H. (1990). The MyoD DNA binding domain contains a recognition code for muscle specific gene activation. Cell 60, 733-746.

Davis, R.L., Weintraub, H., and Lassar, A.B. (1987). Expression of a single transfected cDNA converts fibroblasts to myoblasts. Cell *51*, 987-1000.

Dynlacht, B.D., Hoey, T., and Tjian, R. (1991). Isolation of coactivators associated with the TATA-binding protein that mediate transcriptional activation. Cell *66*, 563-576.

Edmondson, D.G., Cheng, T.C., Cserjeski, P., Chakraborty, T., and Olson, E. (1992). Analysis of the myogenin promoter reveals an indirect pathway for positive autoregulation mediated by the muscle-specific enhancer factor MEF-2 . Mol. Cell. Biol. *12*, 3665-3677.

Edmundson, D.G., and Olson, E.N. (1989). A gene with homology to the *myc* similarity region of *MyoD1* is expressed during myogenesis and is sufficient to activate the muscle differentiation program. Genes Dev. *3*, 628-640.

Fairman, R., Beran-Steed, R.K., Anthony-Cahill, S.J., Lear, J.D., Stafford, W.F., DeGrado, W.F., Benfield, P.A., and Brenner, S.L. (1993). Multiple oligomeric states regulate the DNA binding of helix-loop-helix peptides. Proc. Natl. Acad. Sci. USA *90*, 10429-10433.

Gibson, T.J., Thompson, J.D., and Abagyan, R.A. (1993). Proposed structure for the DNA-binding domain of the helix-loop-helix family

of eukaryotic gene regulatory proteins. Protein Engineering *6*, 41-
50.

Halazonetis, T.D., and Kandil, A.N. (1992). Predicted structural similarities of the DNA binding domains of c-myc and endonuclease Eco RI. Science 255, 464-466.

Hasty, P., Bradley, A., Morris, J.H., Edmondson, D.G., Venuti, J.M., Olson, E.N., and Klein, W.H. (1993). Muscle deficiency and neonatal death in mice with a targeted mutation in the *myogenin* gene. Nature *364*, 501-506.

Hoey, T., Weinzierl, R.O.J., Gill, G., Chen, J.-L., Dynlacht, B.D., and Tjian, R. (1993). Molecular cloning and functional analysis of Drosophila TAF110 reveal properties expected of coactivators. Cell *72*, 247-260.

Lassar, A.B., Davis, R.L., Wright, W.E., Kadesch, T., Murre, C., Voronova, A., and Baltimore, D. (1991). Functional activity of myogenic HLH proteins requires hetero-oligomerization with E12/E47-like proteins in vivo. Cell *66*, 305-315.

Lassar, A.B., Thayer, M.J., Overell, R.W., and Weintraub, H. (1989). Transformation of activated *ras* or *fos* prevents myogenesis by inhibiting expression of MyoD1. Cell *58*, 659-667.

Martin, J.F., Li, L., and Olson, E.N. (1992). Repression of myogenin function by TGF-β 1 is targeted at the basic helix-loop-helix motif and is independent of E2A products. J. Biol. Chem. *267*, 10956-10960.

Miller, J.B. (1992). Myoblast diversity in skeletal myogenesis: how much and to what end? Cell *69*, 1-3.

Murre, C., McCaw, P.S., and Baltimore, D. (1989a). A new DNA binding and dimerization motif in immunoglobulin enhancer binding, *daughterless, MyoD,* and *myc* proteins. Cell 56, 777-783.

Murre, C., McCaw, P.S., Vaessin, H., Caudy, M., Jan, L.Y., Jan, Y.N., Cabrera, C.V., Buskin, J.N., Hauschka, S.D., Lassar, A.B., Weintraub, H., and Baltimore, D. (1989). Interactions between heterologous helix-loop-helix proteins generate complexes that bind specifically to a common DNA sequence. Cell *58*, 537-544.

Olson, E.N., Brennan, T.J., Chakraborty, T., Cheng, T.C., Cserjesi, P., Edmondson, D., James, G., and Li. L. (1991). Molecular control of myogenesis: antagonism between growth and differentiation. Mol. Cell. Biochem. *104*, 7-13.

Rhodes, S.J., and Konieczny, S.F. (1989). Identification of *MRF4*: a new member of the muscle regulatory factor gene family. Genes Dev. *3*, 2050-2061.

Rudnicki, M.A., Braun, T., Hinuma, S., and Jaenisch, R. (1992). Inactivation of *MyoD* in mice leads to up-regulation of the myogenic HLH gene *Myf-5* and results in apparently normal muscle development. Cell *71*, 383-390.

Rudnicki, M.A., Schnegelsberg, P.N.J., Stead, R.H., Braun, T., Arnold, H.-H., and Jaenisch, R. (1993). MyoD or Myf-5 is required for the formation of skeletal muscle. Cell *75*, 1351-1359.

Starovasnik, M.A., Blackwell, T.K., Laue, T.M., Weintraub, H., and Klevit, R.E. (1992). Folding topology of the disulfide-bonded dimeric DNA-binding domain of the myogenic determination factor MyoD. Biochemistry *31*, 9891-9903.

Sun, X.-H., Copeland, N.G., Jenkins, N.A., and Baltimore, D. (1991). Id proteins Id1 and Id2 selectively inhibit DNA binding by one class of helix-loop-helix proteins. Mol. Cell. Biol. *11*, 5603-5611.

Sun, X.H., and Baltimore, D. (1991). An inhibitory domain of E12 transcription factor prevents DNA binding in E12 homodimers but not in E12 heterodimers. Cell 64, 459-470.

Taylor, D.A., Kraus, V.B., Schwarz, J.J., Olson, E.N., and Kraus, W.E. (1993). E1A-mediated inhibition of myogenesis correlates with a direct physical interaction of E1A$_{12S}$ and basic helix-loop-helix proteins. Mol. Cell. Biol. *13*, 4714-4727.

Thayer, M.J., Tapscott, S.J., Davis, R.L., Wright, W.E., Lassar, A.B., and Weintraub, H. (1989). Positive autoregulation of the myogenic determination gene MyoD1. Cell *58*, 241-248.

Trouche, D., Grigoriev, M., Lenormand, J.-L., Robin, P., Leibovitch, S.A., Sassone-Corsi, P., and Harel-Bellan, A. (1993). Repression of c-*fos* promoter by MyoD on muscle cell differentiation. Nature 363, 79-82.

Vinson, C.R., and Garcia, K.C. (1992). Molecular model for DNA recognition by the family of basic-helix-loop-helix-zipper proteins. The New Biologist *4*, 396-403.

Weintraub, H. (1993). The MyoD family and myogenesis: redundancy, networks, and thresholds. Cell *75*, 1241-1244.

Weintraub, H., Davis, R., Tapscott, S., Thayer, M., Krause, M., Benezra, R., Blackwell, T.K., Turner, D., Rupp, R., Hollenberg, S., Zhuang, Y., and Lassar, A. (1991). The *myoD* gene family: nodal point during specification of the muscle cell lineage. Science *251*, 761-766.

Weinzierl, R.O.J., Dynlacht, B.D., and Tjian, R. (1993). Largest subunit of *Drosophila* transcription factor IID directs assembly of a complex containing TBP and a coactivator. Nature *362*, 511-517.

Winter, B., Braun, T., and Arnold, H.H. (1992). Co-operativity of functional domains in the muscle-specific transcription factor Myf-5. EMBO J. *11*, 1843-1855.

Wright, W.E. (1992). Muscle basic helix-loop-helix proteins and the regulation of myogenesis. Curr. Opin. Gen. Dev. *2*, 243-248.

CHAPTER 2

Crystallization of a MyoD/DNA Complex

Crystallization of the MyoD bHLH-DNA complex presented a number of challenges. The first part of this chapter describes how microheterogeneity in the protein samples (Lorber and Giegé, 1992) was tested for and when discovered, how steps were taken to reduce or eliminate these problems. The second part of this chapter describes the survey of different conditions and DNA sequences in the successful cocrystallization of MyoD bound specifically to DNA. We found that varying both the length and the sequence of the DNA (Jordan et al., 1985), particularly at the 5' and 3' ends, was essential for obtaining the diffraction quality crystals used in the structure determination. Finally, the third part of the chapter describes the packing arrangements in the asymmetric unit of the MyoD-DNA cocrystal. The packing arrangement seen in our MyoD-DNA crystals is different from that seen in most other protein-DNA cocrystals which have the DNA duplexes stacking end-on-end to generate a pseudo-continuous helix (Jordan et al., 1985). The distinct crystal packing seen in the MyoD-DNA complex provides a possible explanation of why the initial selection of DNA sequences for the crystallization trials was unsuccessful and underscores the importance of testing a variety of different DNA sequences in crystallization trials of protein-DNA complexes.

A. Analysis, Preparation, and Purification of MyoD and DNA

1. Bacterial strains and plasmids

The *E. coli* expression host BL21(DE3) contains the plasmid
pLysS in order to express low levels of T7 lysozyme and thus,
stabilize plasmids under the control of an inducible promoter for T7
polymerase (Studier, 1991). The expression plasmid pRK171a
(McLeod et al., 1987; Figure 1) contains a restriction site polylinker
downstream of a phage T7 promoter. Deletion and truncation
mutants of MyoD which contained the bHLH region (Tapscott et al.
1988, Harold Weintraub, personal communication) were subcloned
into the polylinker of pRK171a. All plasmids, subclones and
expression hosts used in our studies of MyoD were provided by
Harold Weintraub.

2. Overexpression and Purification of MyoD bHLH Domain

The MyoD fragments which we used in our crystallization
studies all contained the bHLH domain and were derived from the
MyoD deletion mutant DM:4-101;TM:167 (Tapscott et al., 1988). The
initiator ATG of this mutant was engineered to form part of an *NdeI*
site, while the termination codon formed part of a *HindIII* site so
that the MyoD fragments could be subcloned into the appropriate
sites of the polylinker of pRK171a. We used five different MyoD
peptide fragments (provided by Harold Weintraub) in our studies
(Figure 2). The first of these fragments comprised 68 amino acids

and contained the entire bHLH region from mouse MyoD (residues 102 to 166), along with the three amino acids Met-Glu-Leu at the N-terminus. The second and third fragments of MyoD were based on the first and were made in order to avoid oxidation problems (see below). These fragments contained a C135Y and a C135S substitution respectively. The fourth and fifth fragments were identical to the second and third except that they were shorter at the N-terminus by 8 amino acids; hence they contained residues 108 to 166, along with an N-terminal Met for a total length of 60 amino acids. All five fragments of MyoD were tested for DNA binding activity by gel shift experiments. The fragment which we used in our final crystals was the 68 residue peptide with the C135S substitution (MD-bHLH_C135S).

Cells containing the pRK171a plasmid with the MyoD inserts were grown at 37 °C in LB broth containing ampicillin (200 μg/ml) and chloramphenicol (50 μg/ml). When the A_{600} reached 0.5, expression of the protein was induced by adding IPTG (isopropyl β-D-thiogalactopyranoside) to 0.5 mM in order to induce expression of the T7 RNA polymerase which in turn started the expression of the bHLH fragment. Three hours after the IPTG induction, the cells were harvested, resuspended in lysis buffer (25 mM HEPES, pH 7.6, 1 mM EDTA, 5 mM DTT, 200 mM NaCl, 1 mM PMSF (phenylmethyl-sulfonyl fluoride), 1 μg/ml pepstatin, 1 μg/ml benzamide, 1 μg/ml leupeptin, 10% (v/v) glycerol), and lysed by repeated freeze-thawing and sonication. After collecting the supernatant of the centrifuged cells, polyethyleneimine (PEI) was added to a final

concentration of 0.7% in order to partially purify the protein from
the cellular DNA. After spinning down the PEI precipitate, the
supernatant was brought to 0.6 M ammonium sulfate over the course
of 1 hour while stirring at 4 °C. The pellet from this precipitation
was resolubilized in lysis buffer w/o glycerol and dialyzed against
the same solution in preparation for column chromatography.

Two reversed-phase HPLC chromatography steps were used to
purify the MyoD peptides. For the first step, the resuspended and
dialyzed ammonium sulfate pellet was loaded onto a Vydac C_4
preparative column (2.2 x 25.0 cm), 10-μm particle size. The
protein was eluted with a gradient of 5 to 95% acetonitrile/0.1%
trifluoroacetic acid (TFA) at a column temperature of 40 °C. The
protein usually eluted at around 40% acetonitrile. The eluate was
then dialyzed successively against lysis buffer and distilled water
and concentrated by partial lyophilization (being careful not to
lyophilize to dryness). Several steps were taken at these stages to
guard against oxidation of the protein (for the C135Y and C135S
mutants, these steps were not necessary). First, immediately after
collecting the protein fractions from the reversed phase columns,
DTT was added to a final concentration of 5 mM before freezing the
eluates in a dry ice/ethanol bath. DTT was also present in all of the
dialysis solutions, and whenever possible, lyophilization was done
in an anaerobic chamber with < 1 ppm of O_2 (Coy Laboratory
Products, Grass Lake, MI). For the second reserved-phase
chromatography step, we used a Vydac C_{18} semi-preparative column
(1.0 x 25.0 cm), 5-μm particle size. The elution gradient was

similar to, but more shallow than that used with the first reversed-phase column. After dialysis and concentration of the eluate from this second chromatography step, the protein was judged to be greater than 98% pure as judged by SDS gel electrophoresis (single band on an overloaded silver stained gel).

3. Synthesis and Purification of Oligodeoxyribonucleotides

We used the solid-phase phosphoramidite method on an Applied Biosystems (Foster City, CA) model 392 DNA/RNA synthesizer for producing all of the oligodeoxyribonucleotides (oligonucleotides) in our studies. We generally used 1 μm and 10 μm columns from Applied Biosystems, as well as the manufacturer's chemicals and protocols. We did not remove the 5'-dimethoxytrityl (DMT) protecting group after the last cycle in order to aid in purification of the oligonucleotide. The purification of the DNA was a modification of a protocol previously described (Jordan et al., 1985) involving two chromatographic steps of reversed-phase HPLC. In the first step, each strand of DNA is purified with the DMT group attached. Since many of the oligonucleotides used in our studies were symmetrical or close to symmetrical, we took care to perform all chromatographic steps at 50 °C in order to destabilize unwanted secondary structure. After the first chromatography step, the DMT group was cleaved with a brief acid treatment (10 minutes) at room temperature using trichloroacetic acid (1.1%, pH 1.5). To protect against acid depurination of the DNA, we stopped the acid treatment with the addition of triethylamine,

and then immediately dialyzed the solution against 10 mM

triethylammonium bicarbonate (pH 7.0) followed by distilled water,

before lyophilizing the sample. The second reversed phase HPLC

step was done after the DMT group had been removed. After the two

chromatography steps, we annealed the separate DNA strands into

duplexes. As mentioned previously, many of the oligonucleotides

used in our studies were self-complementary. In order to prevent

the formation of hairpins, we performed the annealing steps at a

high oligonucleotide concentration (about 3 mM) so as to drive the

formation of duplexes versus hairpins (Joachimiak and Sigler,

1991).

4. Oxidation of Cysteines in MyoD

Previous attempts to crystallize the MyoD-DNA complex

resulted in a number of crystals, all of which failed to diffract

(Beishan Liu and Carl Pabo, unpublished results). Before starting

the next round of crystallization, we wanted to determine if there

were any systematic microheterogeneities in the protein and/or

DNA samples which may have prevented the formation of diffraction

quality crystals. Formation of disulfide bonds may disrupt normal

folding, oligomer contacts, or preferred crystal packing

arrangements (Van der Laan et al., 1989; Pavletich and Pabo, 1991).

The bHLH domain of mouse MyoD contains a cysteine residue at

position 135, which is at the C-terminus of the helix 1. In order to

determine if oxidation may have been contributing to heterogeneity

in our crystals, we dissolved some previously grown crystals of the

MyoD-DNA complex (Beishan Liu, personal communication) and by running the protein products on an SDS gel determined that some of the MyoD peptide had formed a covalent dimer.

Problems with oxidation of the MyoD bHLH peptide have been corroborated in a recent NMR study of MyoD (Starovasnik et al., 1992). The NMR study used a fragment of MyoD which was identical to our MD-bHLH fragment (Figure 2). In the study, the authors were unable to determine an NMR structure of reduced MyoD, but were able to determine the structure of an oxidized homodimer of MyoD covalently linked through a disulfide bond between Cys135 residues from each monomer. The oxidized MyoD peptide was tested for DNA binding activity and this was found to be reduced by 100-1000 fold. Unlike our crystallization studies with MyoD, the NMR study only involved the MyoD bHLH peptide alone and not the complex of MyoD bound to DNA. In the NMR study, the authors proposed an antiparallel four helix bundle structure for the oxidized MyoD homodimer which was markedly different from the parallel four helix bundle structure seen in the cocrystal structures of the bHLH-ZIP domain of Max bound to DNA (Ferré-D'Amaré et al., 1993), the bHLH domain of USF bound to DNA (the leucine zipper, which is normally present in this bHLH-ZIP protein, was deleted for crystallization purposes; Ferré-D'Amaré et al., 1994), the bHLH domain of E47 bound to DNA (Ellenberger et al., 1994), and the bHLH domain of MyoD bound to DNA. The antiparallel four helix bundle structure of the oxidized MyoD homodimer is incompatible with the mode of DNA binding seen in the three cocrystal structures. We

presume that the antiparallel structure is an artifact of the
disulfide bonds, and we do not believe that it is biologically
relevant.

After concluding that oxidation of the MyoD peptides would be
a serious impediment to obtaining diffraction quality crystals, we
took several approaches to avoid oxidation of the MyoD peptide. At
first, we tried to grow our cocrystals of MyoD in an anaerobic
chamber with < 1 ppm of O_2 (Coy Laboratory Products, Grass Lake,
MI). Growing the MyoD-DNA crystals under these conditions
produced crystals which diffracted, albeit weakly (Table 1). Since
it was experimentally awkward to set up a very large number of
crystallization trials in the anaerobic chamber, we soon switched
to using the MyoD fragments MD-bHLH_C135Y and MD-bHLH_C135S
(provided by Harold Weintraub; Figure 2) in our crystallization
trials. As mentioned above, the replacement of the cysteine by
tyrosine and serine did not alter the DNA binding activity of the
MyoD bHLH peptide, as measured by the gel shift assay. We note
also that among the myogenic proteins, there is a fair amount of
diversity in the identity of residue 135, suggesting that there may
not be strong functional constraints on the identity of this residue.
In fact, one of the two forms of MyoD from Xenopus has a tyrosine
at position 135, while myogenin (77% identity with MyoD in the
bHLH region) has a serine at this position.

5. Probing for Proteolysis and Post-Synthetic Modifications

Aside from the oxidation of cysteines, we searched for other common sources of microheterogeneity such as partial proteolysis. We used several independent analytical techniques to probe for possible contaminants and microheterogeneities in our MyoD peptides. All of the MyoD peptides which we tested in our crystallization studies were judged to be pure by SDS gel electrophoresis: the samples gave a single band on an overloaded silver stained gel. With the help of Jonathan Weissman (M.I.T.), we also tested the C135S-substituted peptides by capillary zone electrophoresis on an Applied Biosystems machine (Foster City, CA) which indicated that the peptides were homogeneous (Figure 3). Electrospray mass-spectroscopy, amino acid composition analysis and N-terminal sequencing were also performed on the MyoD peptides (William Lane, Harvard University). These tests suggested that there were no serious problems of proteolysis or postsynthetic modifications in our C135Y and C135S substituted MyoD peptides.

6. Flexible Regions in the MyoD bHLH Domain

Even when proteins are judged to be free of oxidation, proteolysis or post-synthetic modifications, problems in crystallization may still occur due to the presence of flexible

regions in the protein. These flexible regions may significantly increase the degrees of conformational freedom for the protein and prevent the formation of well-ordered crystals. In the NMR studies of MyoD, the flexibility that was present in the reduced peptide of MyoD (but not in the oxidized homodimer) prevented the determination of the tertiary structure for the MyoD peptide (Starovasnik et al., 1992). For the bZIP DNA binding domains found in proteins such as GCN4, fos, jun, and C/EBP there is considerable evidence to suggest that much of the domain is unstructured in the absence of DNA, but upon DNA binding, there is an induction of alpha-helical structure (O'Neil et al., 1990, 1991; Weiss et al., 1990; Patel et al., 1990). Given that both the bZIP and the bHLH domains consist of a basic region attached to a dimerization motif, we wanted to test if there were any flexible regions in the bHLH domain and if so, would the flexible regions in the bHLH domain become better ordered when the bHLH domain was bound to DNA.

We used circular dichroism (CD) spectroscopy to look at the interaction of the MyoD bHLH domain with an oligonucleotide containing its specific DNA-binding site. With the help of Susan Marquese (who had been in Robert Sauer's laboratory at M.I.T.), we took CD spectra of the MyoD bHLH peptide in the absence and presence of the DNA probe. CD spectra were taken on an Aviv model 60DS spectrapolarimeter. For the protein sample, we used purified MD-bHLH at 5.5 μM concentration (quantitated by measurement of A_{205}) in 25 mM potassium phosphate (pH 7.0), 50 mM potassium chloride, 0.1 mM EDTA, and 0.2 mM DTT (to prevent oxidation of

MyoD). The DNA sample was a 21 base pair oligonucleotide at 5.5
μM concentration, purified according to the methods listed above
and containing the sequence 5' ATAGTAGACAGCTGTCCCAAC 3' on one
strand, and the sequence 5' TGTTGGGACAGCTGTCTACTA 3' on the
other strand (we had used this oligonucleotide in our crystallization
trials; Table 1). In the MyoD/DNA samples, we mixed 5.5 μM of MyoD
with 5.5 μM of DNA (thus, the DNA was added at 2-fold molar excess
to dimeric protein). Both the DNA samples and the MyoD/DNA
samples were in the same buffer as that of the MyoD peptide alone.
CD Spectra were taken at 25 °C using a 1 mm path-length cell, 1.5
nm bandwidth, 1.0 average time per point, and 0.5 nm scan steps.
Measurements were taken over a range of 200 to 350 nm. We did
not perform smoothing on the measurements. For all measurements,
we subtracted a baseline spectra of the CD sample buffer.

Our CD measurements showed that the 68 amino acid MyoD
peptide in the absence of DNA showed only a modest alpha-helical
content (Figures 4 and 5). We measured the molar ellipticity at 222
nm, and estimated the fractional helix contents from θ_{222}
measurements, using the formula -40,000(1-2.5/n) to represent
100% helix (n is the number of amino-acid residues in the peptide),
and θ_{222} = 0 to represent 0% helix (Chakrabartty et al., 1991). For
the MyoD peptide alone, θ_{222} = -13,251 deg cm^2/dmol which is
consistent with an estimated helical content of 34%. When we
added the DNA probe in 2-fold molar excess to the MyoD, and then
subtracted the DNA spectrum from the MyoD-DNA spectrum (Figures
4 and 5), we got a θ_{222} value of -33,977 deg cm^2/dmol, which is

consistent with an estimate of 88% helical content. Interestingly,

an independent CD study of a MyoD bHLH peptide gave a

θ_{222} of -14,000 deg cm^2/dmol for the peptide alone at 5 μM, and a

θ_{222} of -30,000 deg cm^2/dmol when a 1.2 fold excess of a DNA

probe was added to the peptide (Anthony-Cahill et al., 1992).

Significant inductions of alpha-helical structure in bHLH-ZIP

proteins upon specific binding of DNA have also been reported

(Fisher et al., 1993; Ferré-D'Amaré et al, 1994). Our results, plus

those of others, demonstrate that there is significant flexibility in

the basic region of MyoD. Our CD studies have shown that the fold of

the bHLH region is significantly different in the presence and

absence of DNA. In addition to the interest of this result by itself,

our solution CD studies of MyoD also suggested to us that it would

be difficult to grow well-ordered crystals of the MyoD bHLH domain

in the absence of DNA (and may perhaps explain why it was so

difficult to obtain good NMR spectra of MyoD peptide when it was

not in an oxidized homodimer form (Starovasnik et al, 1992)). In

contrast, our CD studies suggest that when MyoD is bound to DNA,

much of the bHLH peptide has an ordered alpha-helical structure.

This suggested to us that we should be able to obtain well-ordered

crystals of the protein-DNA complex given the appropriate DNA site

and the absence of significant microheterogeneities.

B. Crystallization of the MyoD-DNA complex: The Effects of Systematic Variation in DNA Sequence and Length

The sequence and length of the DNA used in cocrystallization trials have been shown to have significant effects on the quality of the cocrystals produced (Anderson et al, 1984; Jordan et al., 1985, Schultz et al, 1990; Liu et al., 1990; Wolberger et al., 1991a). In most protein-DNA cocrystals (the lambda cro repressor-operator complex is one of the few exceptions (Brennan et al., 1990)), the DNA segments stack end on end to form a continuous or pseudo-continuous helix (Jordan et al., 1985). In order to maximize the chances that the DNA sequences have a repeating unit in the crystal which is consistent with the 10.5 base pairs per helical turn that DNA has in solution, many oligonucleotides in cocrystallization studies have lengths which are consistent with having an integral or half-integral number of helical turns in the DNA (Anderson et al., 1984; Jordan et al., 1985). One study showed that differences in the DNA length of as little as one base pair can dramatically effect the quality of the crystals (Jordan et al., 1985). The same study suggested that the sequence identity at the 5' and 3' ends of the DNA can have a large effect on crystal quality. In particular, the study recommended that DNA fragments with one or two complementary overhanging nucleosides at the 5' and 3' ends would stabilize end on end stacking of the DNA segments in a pseudocontinuous helix, and thus, help provide stable crystal contacts. Our cocrystallization trials with MyoD strongly confirm the importance of testing a variety of different DNA sequences and lengths. One surprise from

our studies, however, was that the DNA packing contacts in our final crystal did not follow the usual end on end, pseudocontinuous DNA helix pattern seen in other crystals. Thus, our cocrystallization studies of MyoD may be of some general relevance to the problem of cocrystallizing protein-DNA complexes. This idea will be discussed in more detail in the section describing the crystal packing contacts in our structure.

1. Selection of DNA Sites and Results of Cocrystallization Trials

In all, we tested 34 different DNA sequences in our cocrystallization trials with MyoD (Table 1). We were able to obtain crystals with most of the DNA sequences tested, and we tested many of these crystals for X-ray diffraction. A number of these crystals diffracted, but only a few of these crystals gave relatively ordered diffraction patterns. Our structure was determined from the crystal which diffracted to the highest resolution even though the unit cell size of this crystal suggested that there were several complexes in the asymmetric unit (a = 222.8 Å, b = 70.8 Å, c = 30.0 Å, $P2_12_12$ space group; see Chapter 3 for further discussion of structure determination). In our first crystallization experiments, we tried to reproduce the results of previous crystallization trials with MyoD (Beishan Liu and Carl Pabo, unpublished results). In these earlier experiments, crystallization attempts with MyoD (which was probably partially oxidized) and the oligonucleotides D21-1 and D21-2 gave crystals which did not diffract. After we had determined that oxidation of

the cysteines was a potential problem with crystallization, we tried to grow crystals of reduced MyoD with the D21-1 and D21-2 oligonucleotides in an anaerobic chamber. We were unable to obtain crystals of MyoD with these oligonucleotides under conditions similar to those which gave crystals in an air oxidizing environment, thus raising the possibility that the partial oxidation of MyoD substantially changes the crystal packing contacts and/or folding arrangement of the peptide. At this point, we decided that in addition to performing crystallization trials under anaerobic conditions and/or using the C135Y and C135S substitutions, we would test a greater variety of DNA sites. We based the design of our DNA sites on recent binding site selection experiments on MyoD which suggested that the optimal binding sites for the MyoD homodimer would be a symmetrical site based on the sequences 5' AACAGCTGTT 3' or 5' GACAGCTGTC 3' (Figure 1 of Chapter 4; Blackwell and Weintraub (1990); Blackwell and Weintraub, unpublished results).

The binding site selection experiments suggested that the MyoD homodimer contacted a region spanning at least 10 nucleotides, and thus, we planned to test oligonucleotides which had DNA binding sites of at least that length. Our first DNA sites were designed to have an integral number of helical turns, and had overhanging ends with one or two complementary nucleosides in order to promote formation of a pseudocontinuous helix (M21-2, M20-2, M21-3, M20-3 in Table 1). We were able to grow cocrystals in the anaerobic chamber with all four of these DNA sites, but these

crystals diffracted quite poorly. We also realized that although the central bases of CAGCTG were symmetric within these four DNA sites, the flanking base pairs were not perfectly symmetric. Thus, overall, the four DNA sites were only pseudo-symmetric and not perfectly symmetric. While this lack of perfect symmetry may not have significantly affected binding of the MyoD peptide to the DNA sites, it was a potential source of crystallographic disorder. Previous crystallization studies had indicated that nearly symmetric DNA sites may crystallize in two different orientations in the unit cell, where one orientation of the DNA differs from that of the other by a 180° rotation through the pseudo-dyad axis of the nearly symmetric DNA (DiGabriele et al., 1989; Wolberger et al., 1991b). A mixture of two different orientations for the DNA in the unit cell may be a source of crystallographic disorder, and thus, we were concerned that the four DNA sites M21-2, M20-2, M21-3, and M20-3 gave poorly diffracting cocrystals because of disorder arising from pseudo-symmetric DNA sites. In the rest of our crystallization trials, then, we chose DNA sites that were perfectly symmetric, or would result in repeating units that were perfectly symmetric, or which had sufficient asymmetries in the flanking bases such that we expected the DNA helices to pack in a unique direction in the crystal. We tested three oligonucleotides that had significant asymmetries in the flanking bases (M22-1, M14-5, and M14-9 in Table 1). Only one of these oligonucleotides, M22-1, gave cocrystals and these were of poor quality. We therefore decided to concentrate on symmetric oligonucleotides or asymmetric oligonucleotides that would generate symmetric repeating units.

One of the problems with using perfectly symmetric oligonucleotides for crystallization studies is that one can not easily vary the spacing between adjacent DNA binding sites without changing the length of the oligonucleotide. With an asymmetric oligonucleotide of any given length, it is possible to form a number of repeating units that have different spacings between adjacent DNA binding sites. With a symmetric oligonucleotide of the same length, however, one cannot change the spacing between adjacent DNA binding sites unless 1) the symmetric oligonucleotide is long enough to accommodate 2n number of sites with spacing between the sites, or if 2) adjacent sites within the symmetric oligonucleotide are allowed to overlap. For MyoD, with its binding site of at least ten bases, the first condition above would be difficult to satisfy for practical reasons (synthetic DNA sites greater than 20 base pairs are difficult to purify), while satisfying the second condition could bring back the problems of alternate packing in the crystal (MyoD could bind to either one of two overlapping sites, and a mixture could result in the crystal).

In order to circumvent some of the limitations of using symmetric oligonucleotides in crystallization trials, we designed a series of asymmetric oligonucleotides which would nevertheless generate a repeating unit that was symmetric because the asymmetric oligonucleotides contained 2 base pair overhangs which were two-fold symmetric (Figure 6). All of the oligonucleotides tested in this series (M21-5, M20-5, M22-2, M21-6, M21-7 in Table 1) contained a 12 base pair symmetric DNA binding site for MyoD: 5'

CAACAGCTGTTG 3'. Our design of these oligonucleotides allowed for variation of the spacings between adjacent 12 base pair binding sites within an oligonucleotide of a given length (compare M21-5 with M21-6 and M21-7 in Table 1; Figure 6). These spacings between adjacent binding sites had a big effect on crystallization since we were only able to obtain cocrystals with the oligonucleotide M21-5, and not with the four other oligonucleotides in the series. Unfortunately, the MyoD/M21-5 crystals did not diffract well, indicating that there was still substantial disorder within these crystals.

While we were testing the strategy of using asymmetric oligonucleotides with two-fold symmetric overhangs, we also tested a series of perfectly symmetric oligonucleotides for cocrystallization trials (M10-2, M12-1, M14-1, M14-2 in Table 1). We obtained cocrystals with the 14 base pair oligonucleotides; these were of the best quality of all the cocrystals tested to that point. We collected low resolution data sets from both the M14-1 and M14-2 cocrystals, but were unable to grow cocrystals with these oligonucleotides which diffracted beyond 3.5 to 4 Å resolution. The M14-1 oligonucleotide had a two base pair complementary overhang, whereas the M14-2 oligonucleotide was blunt ended. Since we were able to obtain modestly ordered cocrystals with both of these oligonucleotides, we decided to test 14-mer and 15-mer oligonucleotides with different sequences at the flanking bases. In particular, we wanted to optimize the crystal packing of the DNA by

systematically testing different types of DNA termini: 1) blunt ends, 2) complementary overhangs, 3) blunt-ends with mismatches at the termini, 4) non-complementary overhangs which are identical on the top and bottom strands; these overhangs could not form Watson-Crick base pairs, but could potentially interact with a base pair at the end of a neighboring helix to form a base triplet arrangement at the junction of the DNA helices (Schultz et al., 1990; Luisi et al., 1991). In addition to M14-2, oligonucleotides in the first category included M14-6, M14-7, M14-8, M14-11, and M14-12. The second category contained the previously mentioned M14-1, along with M15-4. The oligomers M14-13 and M14-14 were in the third category while the fourth category contained M14-3, M14-4, M15-1, M15-2, M15-3, M15-5, and M15-6.

Our best cocrystals were grown with oligonucleotides from the first category: symmetric 14-mers with blunt ends. The quality of crystals was very dependent on the identity of the base pairs at the DNA termini. Thus, cocrystals grown with oligonucleotides containing C-G base pairs at the ends, such as M14-2, M14-7, and M14-11, differed considerably in morphology and diffraction quality from cocrystals grown with oligonucleotides containing A-T base pairs at the ends, such as M14-6, M14-8, and M14-12. Moreover, the effect on crystallization of the sequence identity at the DNA termini was not limited to a simple C-G base pair versus A-T base pair difference. The identity of the individual nucleosides at the 5' and 3' termini had a significant effect on crystal quality. Thus, M14-2, which had a C at its 5' end and a G at its 3' end gave better

cocrystals than M14-7 and M14-11, both of which had G's at their 5' ends and C's at their 3' ends. M14-12 gave the best crystals of all, with a T at the 5' end and A at the 3' end; M14-6 and M14-8 which had the identities of the 5' and 3' ends exactly reversed from those of M14-12 gave significantly poorer crystals. It is interesting that a number of studies which have used quantum mechanical methods to calculate stacking energies of different base pairs predict that stacking of adjacent C-G base pairs should be more energetically favorable than stacking of adjacent A-T base pairs (Saenger, 1984; Kudritskaya and Danilov, 1976; Ornstein et al., 1978). It may be significant that our best crystals came from a blunt-ended oligonucleotide, M14-12, that would be predicted to have rather weak stacking energy. We shall see later on in this chapter how weak stacking energies between adjacent DNA's may have contributed to the particular crystal packing arrangements seen in our structure.

2. Crystallization Using Different MyoD Peptide Fragments

During the course of our crystallization experiments, we tried several different peptides, and also tried peptides with substitutions for the Cys 135. The C135Y and C135S substitutions which we used in our crystallization studies substantially widened the range of crystallization conditions we could test since the substitutions allowed us to work outside of the anaerobic chamber. We were able to grow cocrystals with the substitution mutants, as well as with the original MyoD peptides under anaerobic conditions.

The cocrystals were of similar quality, although cocrystals grown
with the C135S substituted MyoD appeared to give slightly better
quality crystals (judged by visual appearance) than the C135Y
substituted crystals. Thus, we performed most of our
crystallization trials with the C135S substituted MyoD peptides. In
addition to the substitutions at the Cys 135 residue, we also tested
the effects of shortening the MyoD bHLH peptide at the N-terminus.
Sequence comparisons with other bHLH and bHLH-ZIP proteins
(Murre et al., 1989), along with mutational analysis on the MyoD
bHLH region (Davis et al., 1990) suggested that the residues N-
terminal to Ala 108 were not necessary for the function of the bHLH
region. We thus tested the two MyoD peptides sMD-bHLH_C135Y and
sMD-bHLH_C135S (Figure 2), which were shorter than our original
bHLH peptide by 8 residues, in crystallization trials. We were
unable to grow good crystals with the shorter peptides, and thus,
we concentrated on using MD-bHLH_C135S as the main MyoD peptide
in our experiments. (Later analysis showed that the N-terminal
residues formed critical crystal packing contacts.)

3. Conditions for Crystallizations

We were able to grow cocrystals of MyoD from a variety of
crystallization conditions. The best conditions appeared to vary
with the oligonucleotide used for the cocrystallization. Thus, the
best crystals for cocrystals grown with M14-1 were grown at 4 °C
in 5-10% polyethylene glycol (PEG) 3350, 100 mM Malonic Acid, pH
5.5, 20-200 mM $MgCl_2$, and 50-100 mM NaCl or Na_3 Citrate. The

cocrystals grown with M14-2 grew at room temperature over a narrow pH range (5.0 - 6.0). There was no need to add precipitating agents in order to grow crystals. Instead, crystals grew directly out of an amorphous precipitate formed when the MyoD peptide was added to DNA. The crystals grew with a variety of buffers: citrate, acetate, Bis-Tris, malonate, succinate, MES, and cacodylate, but the best crystals grew with citrate or MES. The rate of crystal growth was controlled by using volatile buffers in hanging drop experiments or by growing crystals in microdialysis buttons. Crystals grown with M14-12 were of two forms: 1) a monoclinic form that grew out of crystallization conditions similar to those used in cocrystallizing MyoD with M14-2, that is, vapor diffusion of volatile buffers to generate a pH gradient and 2) an orthorhobmic form that grew out of 10-15% PEG 4000, 100 mM Tris, pH 8.5, 100 mM NaCitrate, 20 mM $BaCl_2$ at room temperature; our structure was determined from the orthorhombic crystal form. It may be significant that out of the four crystal forms from which data were collected, the three monoclinic forms (for M14-1, M14-2 and M14-12) were all grown at relatively low pH values (5 to 6), whereas the orthorhombic crystal form, which diffracted the best, grew at relatively high pH values (8 to 8.5). Thus, out of all of the possible variables in crystallization conditions (aside from the identity of the DNA oligomer), it would appear that the pH probably played the most significant role in obtaining well-ordered cocrystals of MyoD, consistent with the notion that the pH is one of the most important factors in the crystallization of macromolecules (Zeppenzauer, 1971; McPherson, 1982).

4. Effects of Salts and Additives on Crystallization

All of the MyoD cocrystals grew out of relatively low salt concentrations (usually < 100 mM). The identity of the salt used was fairly important: in most cases, using Na Citrate helped to improve the quality of the cocrystals. Aside from Na Citrate, other additives did not have much of an effect on the quality of the crystals. Divalent cations, such as $MgCl_2$, $CaCl_2$, and $BaCl_2$ were usually tested when screening crystallization conditions since previous studies of protein-DNA complexes had shown the importance of multivalent cations in influencing the quality of the crystals formed (Joachimiak et al., 1987; Wolberger and Harrison, 1987; Schultz et al., 1990). In our crystallization trials with MyoD, we found that although the presence of divalent cations appeared to improve the quality of the crystals in some cases, the marginal improvement which they added was small. Other common additives in crystallization studies such as glycerol, MPD, spermine, and n-Octyl-beta-D-glucopyranoside had no significant effect on the quality of our cocrystals.

C. Packing Contacts in the MyoD Cocrystals

Solution of the crystal structure (see Chapters 3 and 4) revealed that the crystal packing contacts in the unit cell of the MyoD-DNA structure do not involve end on end stacking of the DNA's to generate a pseudocontinuous helix (a packing arrangement which

is commonly seen in protein-DNA complexes (Joachimiak and Sigler, 1991)). The asymmetric unit of the crystal contains two copies of a MyoD bHLH dimer bound to a 14 bp DNA site (Figure 7). Within this asymmetric unit, the two blunt ended DNA oligomers stack against each other. In the context of the unit cell, however, the ends of the DNA oligomers pack against the MyoD protein, as well as against the ends of other DNA oligomers (Figure 8). Specifically, the end of one DNA helix packs against helix 2 and the loop from a MyoD protein in a neighboring asymmetric unit (Figure 9). The DNA-DNA and DNA-protein packing contacts, as well as a number of protein-protein contacts, together comprise a complicated network of packing contacts in our crystal (Figure 8). It is interesting that the Max bHLH-ZIP-DNA cocrystal also has a rather intricate packing arrangement for the protein and the DNA; as with the MyoD cocrystal, the DNA in the crystal of the Max complex does not form a pseudocontinuous helix (Ferré-D'Amaré et al., 1993).

The elaborate packing arrangements seen in our crystal and that of the Max complex may have been due, in part, to the extensive solvent accessible surfaces found in the bHLH and bHLH-ZIP structures. These structures have long, extended alpha-helices oriented essentially perpendicular to the DNA helical axis, and thus, are significantly different from the more compact globular structures seen in the helix-turn-helix, homeodomain, and zinc finger family of DNA-binding domains (Jordan and Pabo, 1988; Kissinger et al., 1990; Pavletich and Pabo, 1991). Another factor which may have contributed to the packing schemes seen in the

MyoD and Max crystals, is the nature of the blunt-ended DNA oligomers found in the two crystals. As discussed earlier, the crystallization of the MyoD complex was very dependent upon both the length and sequence of the DNA sites used in the crystallization trials. We needed to test a large number of different DNA sites before we obtained well-diffracting crystals. The intricate packing scheme seen in our final crystal is consistent with a strict requirement for both DNA length and sequence. In particular, we note that the packing of one end of a DNA helix against helix 2 and the loop of a MyoD peptide from a neighboring asymmetric unit (Figure 9) would not be predicted to be as favorable if the ends of the DNA helices had overhanging ends; complementary Watson-Crick and/or triple helix interactions formed by these overhanging ends would promote end on end stacking of DNA duplexes from neighboring asymmetric units (Jordan et al., 1985, Otwinowski et al., 1988).

Overall, our experience with the cocrystallization of MyoD, together with an analysis of the packing seen in the final crystal, reinforces the importance of testing a large variety of different DNA sites in cocrystallization trials for protein-DNA complexes. We note, with some humility, that in spite of our attempts to rationally design the DNA sites so as to optimize crystal packing, the final packing scheme for our crystals was quite unexpected. We had originally designed our I4 bp DNA site to have a half-integral number of turns in the DNA (Anderson et al., 1984; Jordan et al., 1985), so that the DNA oligomers would stack end on end in a

pseudo-continuous helix. The Max complex had a 22 bp DNA site in the cocrystal (Ferré-D'Amaré et al., 1993) which, like the MyoD cocrystal, did not have the DNA pack in a pseudo-continuous helix. The unexpected crystal packing schemes seen in the MyoD and Max cocrystals illustrate the diversity of packing arrangements in crystals of protein-DNA complexes. It may be the case that the long extended alpha-helices found in the bHLH and bHLH-ZIP structures make it especially difficult to rationally design oligonucleotides for crystallization. In this regard, we note that the crystal packing schemes seen in the E47-DNA complex (a bHLH protein; Ellenberger et al., 1994) and in the USF-DNA complex (USF is a bHLH-ZIP protein, but the crystal contains a truncated bHLH peptide with the naturally occurring leucine zipper deleted for crystallization purposes; Ferré-D'Amaré et al., 1994), have DNA duplexes which stack end on end to form pseudocontinuous helices. Even within the context of a few bHLH and bHLH-ZIP structures then, we see a variety of different packing schemes. Thus, with a limited ability to predict packing *a priori*, the crystal grower must rely on testing DNA sites of different lengths and sequences (especially at the ends) in order to maximize the chances for obtaining diffraction quality crystals of bHLH(/Z) - DNA complexes, and by extension, other protein-DNA complexes as well (Jordan et al., 1985).

References

Anderson, J., Ptashne, J., and Harrison, S.C. (1984). Cocrystals of the DNA-binding domain of phage 434 repressor and a synthetic phage 434 operator. Proc. Natl. Acad. Sci. USA 81, 1307-1311.

Anthony-Cahill, S.J., Benfield, P.A., Fairman, R., Wasserman, Z.R., Brenner, S.L., Stafford, W.F., III, Altenbach, C., Hubbell, W.L., and DeGrado, W.F. (1992). Molecular characterization of helix-loop-helix peptides. Science 255, 979-983.

Blackwell, T.K., and Weintraub, H. (1990). Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. Science 250, 1104-1110.

Brennan, R.G., Roderick, S.L., Takeda, Y., and Matthews, B.W. (1990). Protein-DNA conformational changes in the crystal structure of a lambda cro-operator complex. Proc. Natl. Acad. Sci. USA 87, 8165-8169.

Chakrabartty, A., Schellman, J.A., and Baldwin, R.L. (1991). Large differences in the helix propensities of alanine and glycine. Nature 351, 586-588.

Davis, R.L., Cheng, P.-F., Lassar, A.B., and Weintraub, H. (1990). The MyoD DNA binding domain contains a recognition code for muscle-specific gene activation. Cell *60*, 733-746.

DiGabrielle, A.D., Sanderson, M.R., and Steitz, T.A. (1989). Crystal lattice packing is important in determining the bend of a DNA dodecamer containing an adenine tract. Proc. Natl. Acad. Sci. USA *86*, 1816-1820.

Ellenberger, T., Fass, D., Arnaud, M., and Harrison, S.C. (1994). Crystal structure of transcription factor E47: E-box recognition by a basic region helix-loop-helix dimer. Genes Dev. In press.

Ferré-D'Amaré, A.R., Pognonec, P., Roeder, R.G., and Burley, S.K. (1994). Structure and function of the b/HLH/Z domain of USF. EMBO J. *13*, 180-189.

Ferré-D'Amaré, A.R., Prendergast, G.C., Ziff, E.B., and Burley, S.K. (1993). Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain. Nature *363*, 38-45.

Fisher, D.E., Parent, L.A., and Sharp, P.A. (1993). High affinity DNA-binding myc analogs: recognition by an alpha-helix. Cell *72*, 467-476.

Joachimiak, A., and Sigler, P.B. (1991). Crystallization of protein-DNA complexes. Methods in Enzymology *208*, 82-99.

Joachimiak, A., Marmorstein, R.Q., Schevitz, R.W., Mandecki, W., Fox, J.L., and Sigler, P.B. (1987). Crystals of the *trp* repressor-operator complex suitable for X-ray diffraction analysis. J. Biol. Chem. *262*, 4917-4921.

Jordan, S.R., and Pabo, C.O. (1988). Structure of the lambda complex at 2.5 Å resolution: Details of the repressor-operator interactions. Science *242*, 893-899.

Jordan, S.R., Whitcombe, T.V., Berg, J.M., and Pabo, C.O. (1985). Systematic variation in DNA length yields highly ordered repressor-operator cocrystals. Science *230*, 1383-1385.

Kissinger, C.R., Liu, B., Martin-Blanco, E., Kornberg, T.B., and Pabo, C.O. (1990). Crystal structure of an engrailed homeodomain-DNA complex at 2.8 Å resolution: A framework for understanding homeodomain-DNA interactions. Cell *63*, 579-590.

Kudritskaya, Z.G., and Danilov, V.I. (1976). Quantum mechanical study of bases interactions in various associates in atomic dipole approximation. J. Theor. Biol. *59*, 303-318.

Liu, B., Kissinger, C.R., and Pabo, C.O. (1990). Crystallization and preliminary X-ray diffraction studies of the engrailed homeodomain and of an engrailed homeodomain/DNA complex. Biochemical and Biophysical Research Communications *171*, 257-259.

Lorber, B., and Giegé, R. (1992). Preparation and handling of biological macromolecules for crystallization. In Crystallization of Nucleic Acids and Proteins, A. Ducruix and R. Giegé, eds. (New York, New York: Oxford University Press), pp. 19-42.

Luisi, B.F., Xu, W.X., Otwinowski, Z., Freedman, L.P., Yamamoto, K.R., and Sigler, P.B. (1991). Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. Nature *352*, 497-505.

McLeod, M., Stein, M., and Beach, D. (1987). The product of the *mei3+* gene, expressed under control of the mating-type locus, induces meiosis and sporulation in fission yeast. EMBO J. 6, 729-736.

McPherson, A. (1982). Preparation and Analysis of Protein Crystals (Malabar, Florida: Krieger Publishing Company).

Murre, C., McCaw, P.S., and Baltimore, D. (1989). A new DNA binding and dimerization motif in immunoglobulin enhancer binding, *daughterless*, *MyoD*, and *myc* proteins. Cell *56*, 777-783.

O'Neil, K.T., Hoess, R.H., and DeGrado, W.F. (1990). Design of DNA binding peptides based on the leucine zipper motif. Science *249*, 774-778.

O'Neil, K.T., Shuman, J.D., Ampe, C., and DeGrado, W.F. (1991). DNA-induced increase in the alpha-helical content of C/EBP and GCN4. Biochemistry *30*, 9030-9034.

Ornstein, R.L., Rein, R., Breen, D.L., and MacElroy, R.D. (1978). An optimized potential function for the calculation of nucleic acid interaction energies. I. Base stacking. Biopolymers *17*, 2341-2360.

Otwinowski, Z., Schevitz, R.W., Zhang, R.G., Lawson, C.L., Joachimiak, A., Marmorstein, R.Q., Luisi, B.F., and Sigler, P.B. (1988). Crystal structure of *trp* repressor/operator complex at atomic resolution. Nature *335*, 321-329.

Patel, L., Abate, C., and Curran, T. (1990). Altered protein conformation on DNA binding by fos and jun. Nature *347*, 572-574.

Pavletich, N.P., and Pabo, C.O. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. Science *252*, 809-817.

Saenger, W. (1984). Principles of Nucleic Acid Structure (New York, New York: Springer-Verlag).

Schultz, S.C., Shields, G.C., and Steitz, T.A. (1990). Crystallization of *Escherichia coli* catabolite gene activator protein with its DNA binding site: The use of modular DNA. J. Mol. Biol. *213*, 159-166.

Starovasnik, M.A., Blackwell, T.K., Laue, T.M., Weintraub, H., and Klevit, R.E. (1992). Folding topology of the disulfide-bonded dimeric DNA-binding domain of the myogenic determination factor MyoD. Biochemistry 31, 9891-9903.

Studier, F.W. (1991). Use of bacteriophage T7 lysozyme to improve an inducible T7 expression system. J. Mol. Biol. 219, 37-44.

Tapscott, S.J., Davis, R.L., Thayer, M.J., Cheng, P.-F., Weintraub, H., and Lassar, A.B. (1988). MyoD1: a nuclear phosphoprotein requiring a myc homology region to convert fibroblasts to myoblasts. Science 242, 405-411.

Van der Laan, J.M., Swarte, M.B.A., Groendijk, H., Hol, W.G.J., and Drenth, J. (1989). The influence of purification and protein heterogeneity on the crystallization of p-hydroxybenzoate hydroxylase. Eur. J. Biochem. 179, 715-724.

Weiss, M.A., Ellenberger, T.E., Wobbe, C.R., Lee, J.P. Harrison, S.C., and Struhl, K. (1990). Folding transition in the DNA-binding domain of GCN4 on specific DNA binding. Nature 347, 575-578.

Wolberger, C., and Harrison, S.C. (1987). Crystallization and X-ray diffraction studies of a 434 cro-DNA complex. J. Mol. Biol. 196, 951-954.

Wolberger, C., Pabo, C.O., Vershon, A.K., and Johnson, A.D. (1991a). Crystallization and preliminary X-ray diffraction studies of a MATα2-DNA complex. J. Mol. Biol. *217*, 11-13.

Wolberger, C., Vershon, A.K., Liu, B., Johnson, A.D., and Pabo, C.O. (1991b). Crystal structure of a MATα 2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. Cell *67*, 517-528.

Zeppenzauer, M. (1971). Formation of large crystals. Methods in Enzymology *22*, 253-266.

Table Legends

Table 1: DNA sequences used in cocrystallization experiments with MyoD.

We summarize the results of cocrystallization experiments between various MyoD peptides and the DNA oligomers shown in this table. For each of the oligomers shown, we have highlighted in bold the E-box bases: **CAGCTG** , **CAGGTG**, or **CACCTG**. The column labeled "Diffraction" refers to the diffraction limit of the crystal tested.

Figure Legends

Figure 1: Expression plasmid pRK171a.

Schematic of the expression plasmid (provided by Harold Weintraub) used for the overexpression of the MyoD bHLH peptides.

Figure 2: Primary sequence of MyoD peptides tested in crystallization studies.

Sequence alignment of the peptides (produced from cDNA clones provided by Harold Weintraub) used in our cocrystallization trials. Amino acid residues are given in one-letter code.

Figure 3: Capillary zone electrophoresis on MyoD.

Analysis of the MD-bHLH_C135S peptide shows that the peptide runs as a single peak. (Analysis performed with the help of Jonathan Weissman, M.I.T).

Figure 4: CD difference spectroscopy in the presence of DNA.

CD absorption spectrum for 5.5 μM of MyoD (open squares), 5.5 μM of DNA (closed squares), and the difference spectrum obtained from subtracting the DNA component of 5.5 μM MyoD plus 5.5 μM DNA (closed diamonds).

Figure 5: CD absorption spectroscopy of MyoD in the absence and presence of DNA.

The figure is similar to that of Figure 4. For clarity, we don't show the absorption spectrum for DNA. The ordinate of the graph is measured in units of molar ellipticity (in units of $deg*cm^2/dmol$).

Figure 6: Asymmetric DNA oligomers can form part of a symmetric repeating unit.

Schematic showing how the symmetric two base-pair overhangs of an asymmetric DNA oligomer can lead, through the base pairing of the overhangs, to symmetric repeating units.

Figure 7: Asymmetric unit of the MyoD/DNA cocrystals.

The asymmetric unit of the MyoD/DNA cocrystals contain two copies of a dimer of MyoD bound to a 14bp DNA site (see Chapters 3 and 4 for details on structure determination). The DNA sites are shown in gold and red. The alpha carbon traces from each one of the four MyoD peptides (from the two dimers) are shown in a different color: white, green, pink, and blue.

Figure 8: Crystal packing contacts in the unit cell of the MyoD/DNA cocrystals.

Figure 8a shows the unit cell, in which the view is of the plane formed by the **a** and **b** unit cell edges. We show the arrangement of the DNA within the unit cell, leaving out the protein molecules for now. Notice that the DNA oligomers are grouped by two, since each asymmetric unit contains two copies of a 14 bp DNA oligomer. Also, the DNA oligomers do not stack end on end to form a pseudo-continuous helix through the unit cell. The view in Figure 8b is rotated 90° along the axis following the **a** unit cell edge. Thus, the view is of the plane formed by the **a** and **c** unit cell edges. As in Figure 8a, we show only the DNA oligomers. Figure 8c shows an identical view to that in Figure 8a, except now, both the protein and the DNA molecules within the unit cell are shown. Notice that there are numerous protein-protein, and protein-DNA contacts between members of different asymmetric units within this unit cell.

Figure 9: Crystal packing contacts between molecules belonging to two adjacent asymmetric units.

We show the contents of two adjoining asymmetric units. The two MyoD dimer/DNA complexes in the center of the unit cell are from different asymmetric units. Notice the protein-DNA contacts between helix 2 and the loop regions of the MyoD dimers and the terminal base pairs of the DNA.

Oligomer	Peptide	Crystals	Diffraction
1) D21-1 5' AGCAGGTGTTCAGCAGGTGTT 3' 3' CGTCCACAAGTCGTCCACAAT 5'	MD-bHLH	crystals (aerobic conditions only)	poor diffraction (few spots visible)
2) D21-2 5' AGCAGGTGTTGCAGCAGGTGT 3' 3' CGTCCACAACGTCGTCCACAT 5'	MD-bHLH	crystals (aerobic conditions only)	poor diffraction
3) M21-2 5' ATAGTAGACAGCTGTCCCAAC 3' 3' ATCATCTGTCGACAGGGTTGT 5'	MD-bHLH MD-bHLH_C135Y	crystals (anaerobic conditions for MD-bHLH)	weak diffraction (~ 7Å)
4) M20-2 5' ATAGTAGACAGCTGTCCCAA 3' 3' ATCATCTGTCGACAGGGTTT 5'	MD-bHLH MD-bHLH_C135Y	crystals (anaerobic conditions for MD-bHLH)	poor diffraction
5) M21-3 5' ATCACAACAGCTGTTGCTCTA 3' 3' AGTGTTGTCGACAACGAGATT 5'	MD-bHLH	crystals (anaerobic conditions for MD-bHLH)	poor diffraction
6) M20-3 5' ATCACAACAGCTGTTGCCTA 3' 3' AGTGTTGTCGACAACGGATT 5'	MD-bHLH	crystals (anaerobic conditions for MD-bHLH)	poor diffraction
7) M21-5 5' ATCGCAACAGCTGTTGGACTA 3' 3' GCGTTGTCGACAACCTGATGC 5'	MD-bHLH_C135Y MD-bHLH_C135S	crystals	disordered diffraction pattern
8) M20-5 5' ATCGCAACAGCTGTTGGACT 3' 3' GCGTTGTCGACAACCTGAGC 5'	MD-bHLH_C135Y	no crystals	n/a

Table 1

Oligomer	Peptide	Crystals	Diffraction
9) M22-1 5' ATAGTAGA**CAGCTG**TCCCAACT 3' 3' ATCATCTGTCGACAGGGTTGAT 5'	MD-bHLH_C135Y	small crystals	not tested
10) M22-2 5' ATCGGCAA**CAGCTG**TTGGTACT 3' 3' GCCGTTGTCGACAACCATGAGC 5'	MD-bHLH_C135Y	no crystals	n/a
11) M14-1 5' TACAA**CAGCTG**TTG 3' 3' GTTGTCGACAACAT 5'	MD-bHLH_C135Y MD-bHLH_C135S sMD-bHLH_C135Y sMD-bHLH_C135S	Crystal Form I a = 62.4 Å α = 90° b = 143.9 Å β = 103.9° c = 132.1 Å γ = 90° (no crystals for sMD peptides)	diffraction (~3.5 Å)
12) M14-2 5' CCAA**CAGCTG**TTGG 3' 3' GGTTGTCGACAACC 5'	MD-bHLH_C135Y MD-bHLH_C135S sMD-bHLH_C135Y sMD-bHLH_C135S	Crystal Form II a = 55.9 Å α = 90° b = 64.2 Å β = 102.6° c = 140.4 Å γ = 90° (no crystals for sMD peptides)	diffraction (~3.5 Å)
13) M14-3 5' GGCAA**CAGCTG**TTG 3' 3' GTTGTCGACAACGG 5'	MD-bHLH_C135Y MD-bHLH_C135S	no crystals	n/a
14) M14-4 5' AACAA**CAGCTG**TTG 3' 3' GTTGTCGACAACAA 5'	MD-bHLH_C135Y MD-bHLH_C135S	crystals	poor diffraction

Table 1 (cont.)

Oligomer	Peptide	Crystals	Diffraction
15) M14-5 5' AGAGCACCTGTCTA 3' 3' CTCGTGGACAGATT 5'	MD-bHLH_C135Y MD-bHLH_C135S	no crystals	n/a
16) M21-6 5' ATGCAACAGCTGTTGGTACTA 3' 3' CGTTGTCGACAACCATGATGC 5'	MD-bHLH_C135S	no crystals	n/a
17) M21-7 5' ATCGACAACAGCTGTTGGCTA 3' 3' GCTGTTGTCGACAACCGATGC 5'	MD-bHLH_C135S	no crystals	n/a
18) M10-2 5' TAACAGCTGT 3' 3' TGTCGACAAT 5'	MD-bHLH_C135Y MD-bHLH_C135S	no crystals	n/a
19) M16-1 5' TAGCAACAGCTGTTGC 3' 3' CGTTGTCGACAACGAT 5'	MD-bHLH_C135Y	no crystals	n/a
20) M15-1 5' ACCAACAGCTGTTGG 3' 3' GGTTGTCGACAACCA 5'	MD-bHLH_C135S	crystals	poor diffraction
21) M15-2 5' GCCAACAGCTGTTGG 3' 3' GGTTGTCGACAACCG 5'	MD-bHLH_C135S	crystals	disordered diffraction pattern
22) M15-3 5' CCCAACAGCTGTTGG 3' 3' GGTTGTCGACAACCC 5'	MD-bHLH_C135S	crystals	disordered diffraction pattern

Table 1 (cont.)

Oligomer	Peptide	Crystals	Diffraction
23) M15-4 5' TCCAACAGCTGTTGG 3' 3' GGTTGTCGACAACCA 5'	MD-bHLH_C135S	no crystals	n/a
24) M14-6 5' ACAACAGCTGTTGT 3' 3' TGTTGTCGACAACA 5'	MD-bHLH_C135S	crystals	disordered diffraction pattern
25) M14-7 5' GAGACAGCTGTCTC 3' 3' CTCTGTCGACAGAG 5'	MD-bHLH_C135S	crystals	poor diffraction
26) M14-8 5' AAAACAGCTGTTTT 3' 3' TTTTGTCGACAAAA 5'	MD-bHLH_C135S	crystals	disordered diffraction pattern
27) M14-9 5' CCCCCAGCTGTTTT 3' 3' GGGGGTCGACAAAA 5'	MD-bHLH_C135S	no crystals	n/a
28) M14-11 5' GCAACAGCTGTTGC 3' 3' CGTTGTCGACAACG 5'	MD-bHLH_C135S	crystals	weak diffraction (~5Å)
29) M12-1 5' CAACAGCTGTTG 3' 3' GTTGTCGACAAC 5'	MD-bHLH_C135S	no crystals	n/a
30) M15-5 5' AACAACAGCTGTTGT 3' 3' TGTTGTCGACAACAA 5'	MD-bHLH_C135S	crystals	poor diffraction

Table 1 (cont.)

Oligomer	Peptide	Crystals	Diffraction
31) M14-12 5' TCAACAGCTGTTGA 3' 3' AGTTGTCGACAACT 5'	MD-bHLH_C135S	Crystal Form III a = 68.5 Å α = 90° b = 61.0 Å β = 101.9° c = 58.9 Å γ = 90°	diffraction (~3.5 Å)
		Crystal Form IV a = 222.8 Å α = 90° b = 70.8 Å β = 90° c = 30.0 Å γ = 90° $P2_12_12$ space group	diffraction (~2.6 Å)
32) M15-6 5' TACAACAGCTGTTGT 3' 3' TGTTGTCGACAACAT 5'	MD-bHLH_C135S	crystals	poor diffraction
33) M14-13 5' AAAACAGCTGTTTA 3' 3' ATTTGTCGACAAAA 5'	MD-bHLH_C135S	no crystals	n/a
34) M14-14 5' TAAACAGCTGTTTT 3' 3' TTTTGTCGACAAAT 5'	MD-bHLH_C135S	crystals	weak diffraction (~6 Å)

Table 1 (cont.)
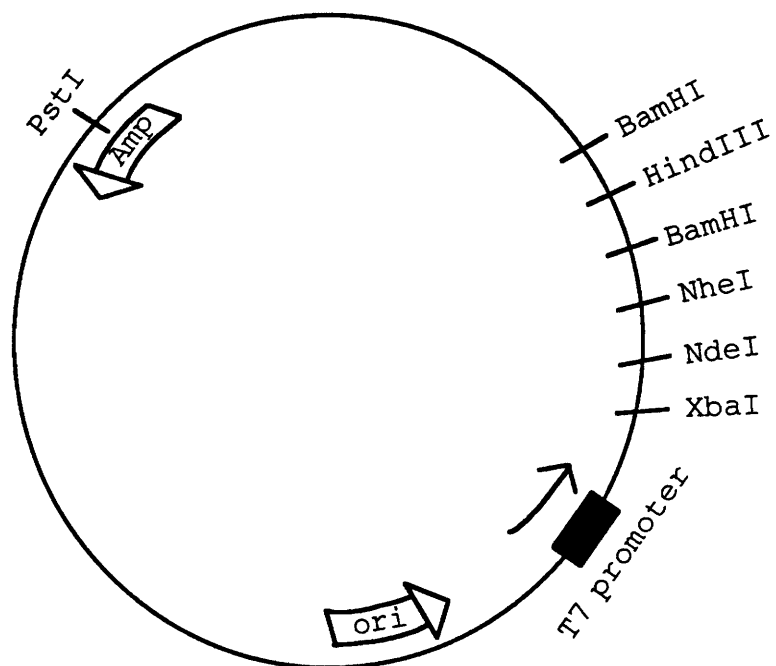
pRK171a

size ~2.4 kilbases

Figure 1

MD-bHLH

MELKRKTTNADRRKAATMRERRRLSKVNEAFETLKRCTSSNPNQRLPKVEILRNAIRYIEGLQALLRD

MD-bHLH_C135Y

MELKRKTTNADRRKAATMRERRRLSKVNEAFETLKRYTSSNPNQRLPKVEILRNAIRYIEGLQALLRD

MD-bHLH_C135Y

MELKRKTTNADRRKAATMRERRRLSKVNEAFETLKRSTSSNPNQRLPKVEILRNAIRYIEGLQALLRD

sMD-BHLH_C135Y

MADRRKAATMRERRRLSKVNEAFETLKRYTSSNPNQRLPKVEILRNAIRYIEGLQALLRD

sMD-BHLH_C135S

MADRRKAATMRERRRLSKVNEAFETLKRSTSSNPNQRLPKVEILRNAIRYIEGLQALLRD

Figure 2

MyoD 0.8 mg/ml
CZE pH 2.0 30°C 8/21/92
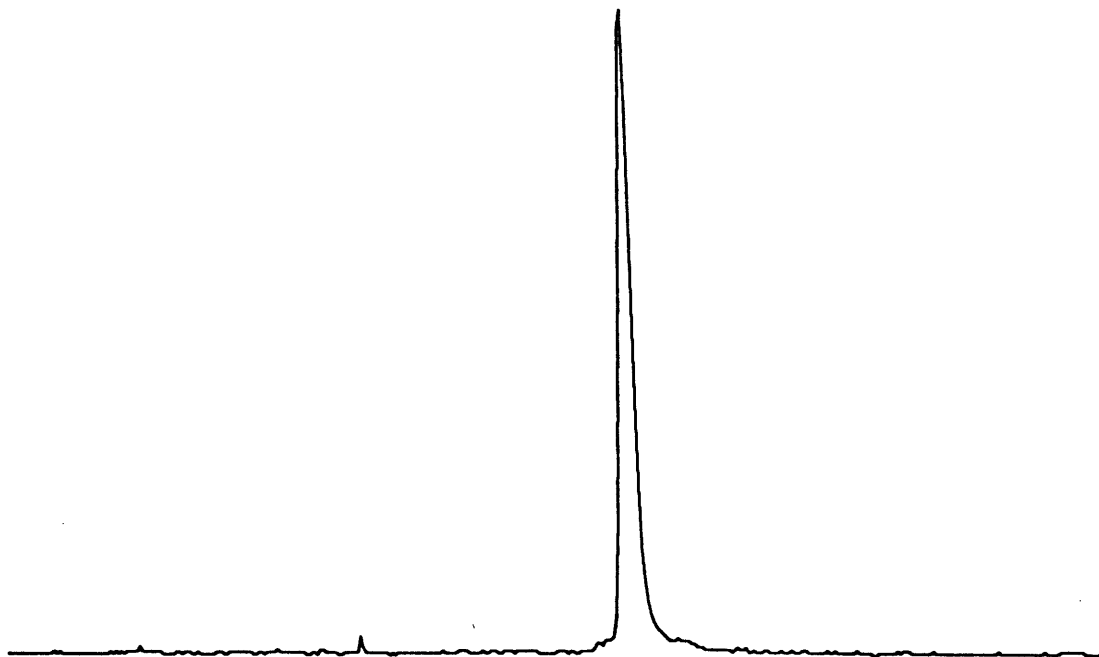
Figure 3
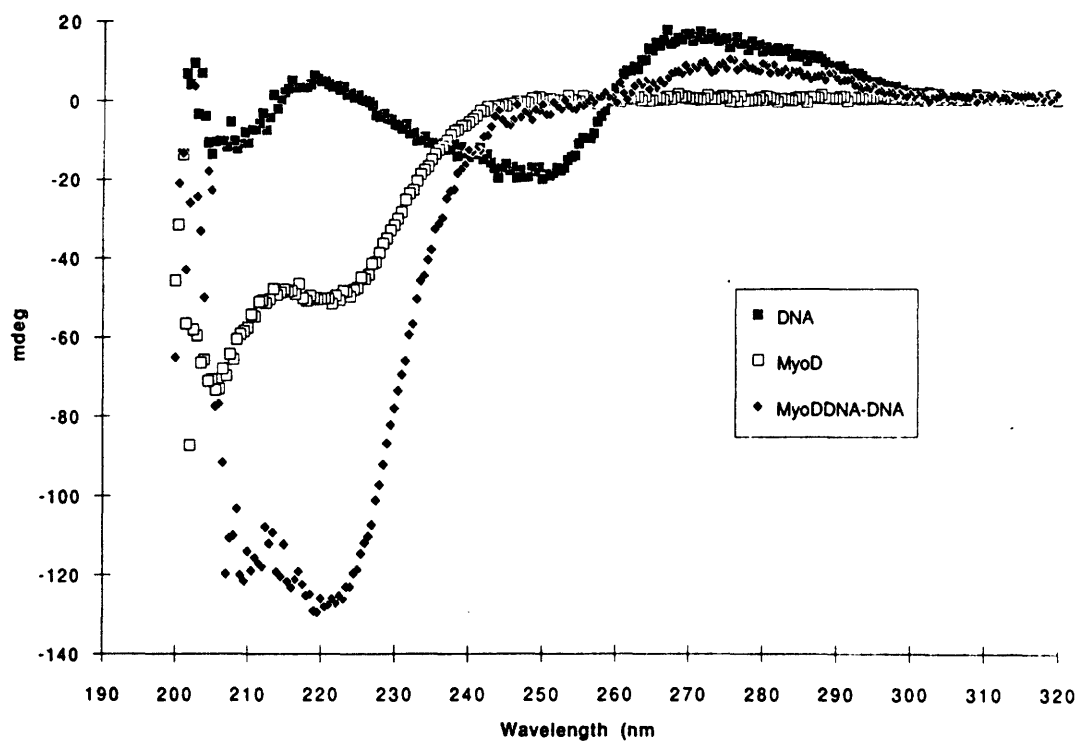
Figure 4
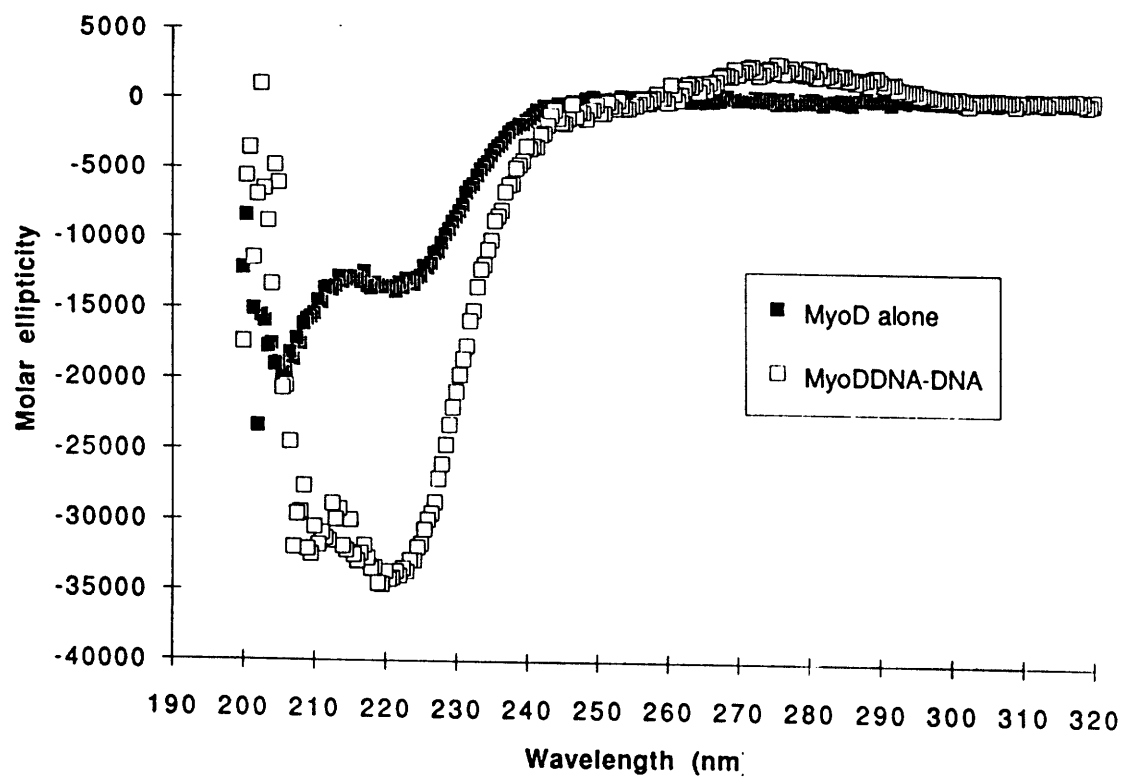
Figure 5

The following oligomers are asymmetric:

```
M21-5   5'   ATCGCAACAGCTGTTGGACTA      5'
        3'       GCGTTGTCGACAACCTGATGC  3'

M21-6   5'   ATGCAACAGCTGTTGGTACTA      5'
        3'       CGTTGTCGACAACCATGATGC  3'

M21-7   5'   ATCGACAACAGCTGTTGGCTA      5'
        3'       GCTGTTGTCGACAACCTGTGC  3'

M20-5   5'   ATCGCAACAGCTGTTGGACT       5'
        3'       GCGTTGTCGACAACCTGAGC   3'

M22-2   5'   ATCGGCAACAGCTGTTGGTACT      5'
        3'       GCGGTTGTCGACAACCATGAGC  3'
```

BUT they can form a part of a
symmetric repeating unit

For example:

```
      M21-5
ATCGCAACAGCTGTTGGACTACGTAGTCCAACAGCTGTTGCG
  GCGTTGTCGACAACCTGATGCATCAGGTTGTCGACAACGCTA
                                M21-5

      M21-6
ATGCAACAGCTGTTGGTACTACGTAGTACCAACAGCTGTTCG
  CGTTGTCGACAACCATGATGCATCATGGTTGTCGACAAGCTA
                                M21-6

      M21-5
ATCGACAACAGCTGTTGGCTACGTAGCCAACAGCTGTTGTCG
  GCTGTTGTCGACAACCGATGCATCGGTTGTCGACAACAGCTA
                                M21-5
```

Figure 6

Figure 7
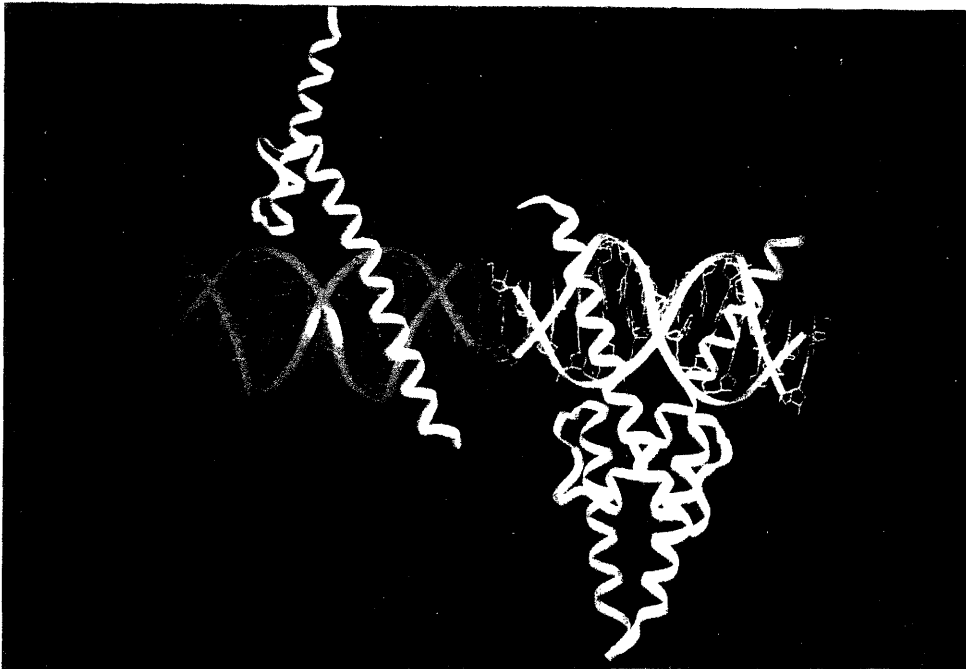
a)
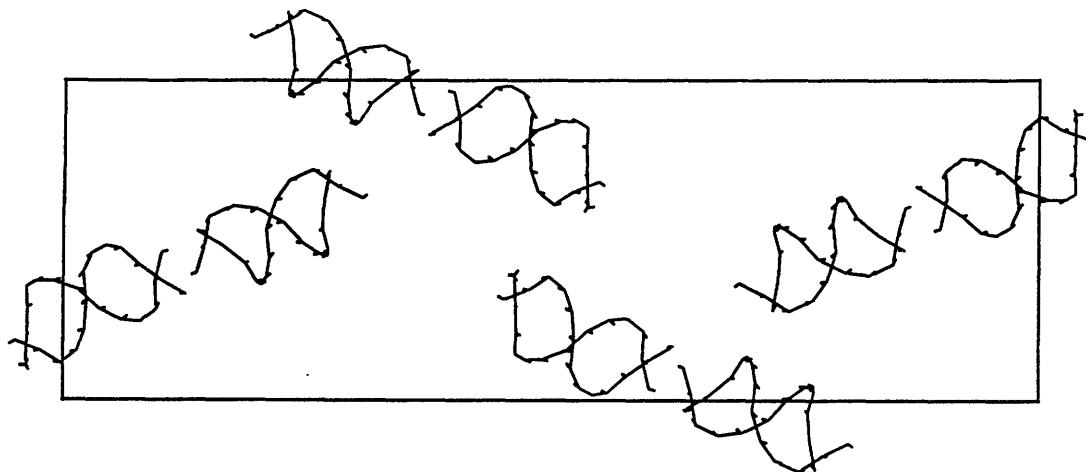
b)
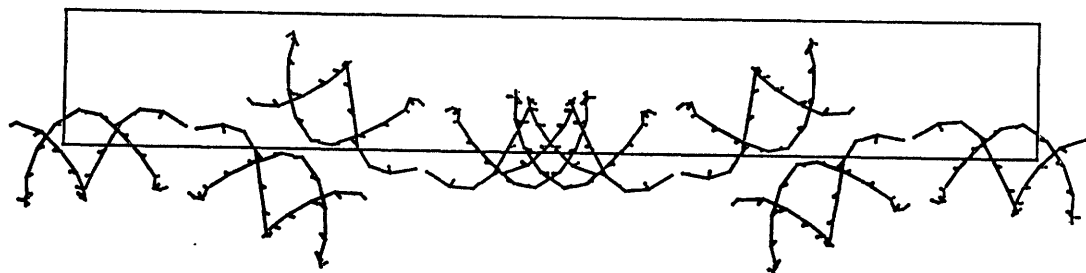
Figure 8

c)
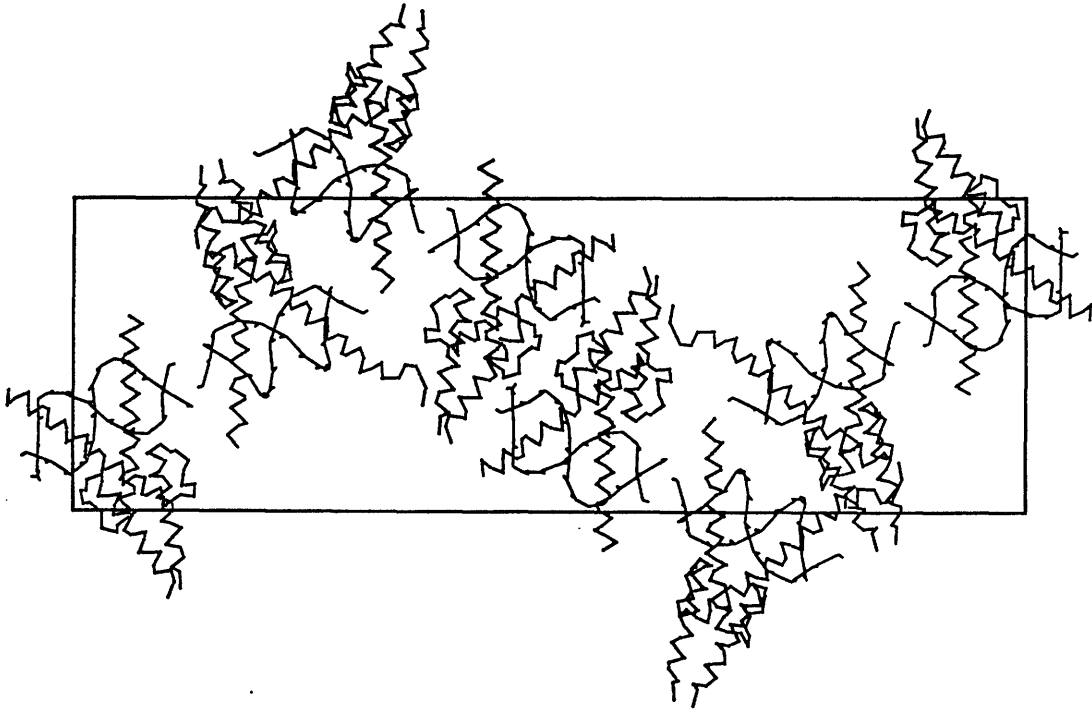
Figure 8
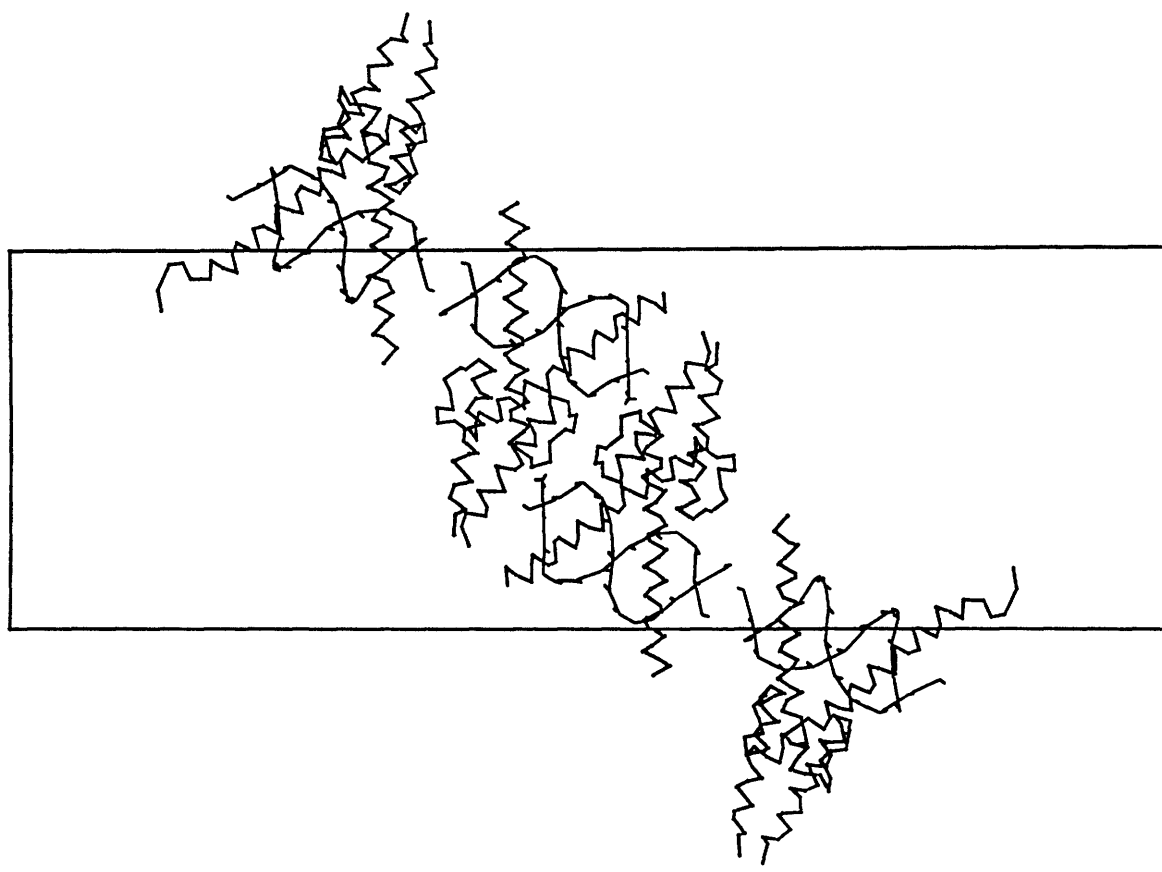
Figure 9

CHAPTER 3

Details of the Structure Determination
of the MyoD bHLH-DNA Complex

As described in the previous chapter, by using the DNA site M14-12, we were able to grow cocrystals with MyoD that diffracted to 2.6 Å resolution. Our general approach to structure determination relied on using modified DNA in our cocrystals to produce substitutions for solving the phase problem. Specifically, we synthesized DNA oligomers in which 5-iodouracil and/or 5-iodocytosine were introduced as substitutions for thymine and cytosine respectively (Anderson et al., 1987; Wolberger et al., 1988; Jordan and Pabo, 1988; Aggarwal et al., 1988). We tested a number of these modified DNA oligomers in cocrystallization trials, and found that two of these oligomers produced crystals which were useful for phasing. Model building of the protein-DNA complex was done with the computer graphics program FRODO (Jones, 1978) and refinement of the model used the programs XPLOR (Brunger et al., 1987; Brunger, 1992) and TNT (Tronrud et al., 1987). This chapter summarizes our results and methods from the structure determination, focusing on areas that were not discussed in detail in the 'Experimental Procedures' section of Chapter 4 (which concentrates on describing and analyzing the structure of the complex). After discussing the preparation of heavy atom derivatives using DNA oligomers with halogenated substitutions, we discuss our data collection and data reduction strategies. The unusual shape of the unit cell (a = 222.8 Å, b = 70.8 Å, c = 30.0 Å) required extra care in the data collection and data reduction. A discussion of the phasing procedures, in which we used heavy atom parameter refinement and solvent flattening techniques, is next,

followed by some comments on our model building. A brief discussion of the refinement procedures ends this chapter.

A. Preparation of Heavy Atom Derivatives

We tested a number of different DNA sites in which thymine and/or cytosine bases were substituted by pyrimidines with iodine at the 5 position (Table 1). With the two-fold symmetry in our DNA, we expected that the DNA would be symmetrically substituted. The two DNA duplexes which gave crystals suitable for phasing, 11+11' and 12+12', each contained two iodines. We tested a total of nine substituted oligomers in cocrystallization trials under conditions identical to those which gave native cocrystals. Five of the oligomers tested produced crystals. Of the four oligomers which did not produce crystals, three of these were substituted at bases inside the core 6 bp CAGCTG region and thus, might have interfered with DNA binding by the MyoD peptide; the fourth oligomer was substituted at the 5' base position, which could have affected how the DNA stacked against other DNAs and/or MyoD peptides. For the crystallization trials, as well as for the synthesis and purification steps, we were careful to limit the exposure of the iodinated DNA to UV and visible light, working in the dark whenever possible and/or using aluminum foil to cover photochemically sensitive reagents. In addition to guarding against photochemical reactions of the iodine, we were careful not to overreact the iodine-substituted oligonucleotides with ammonia during the base

deprotection steps, or with tricholoracetic acid during the detritylation steps.

We found that the iodine-substitutions were quite sensitive to exposure to ammonia during the base deprotection steps. Our initial derivative crystal (12+12', first crystal) did not give an easily interpretable difference Patterson map (viewed by Harker sections). We suspected that the iodine-substituted DNA used in this derivative crystal may have been damaged during synthesis and/or purification. Earlier conversations with Stephen Burley (Rockefeller University) had suggested that iodine-substituted DNA may be especially sensitive to ammonia treatment during the base deprotection steps; he also suggested to us that we try ion-exchange chromatography for purification of the DNA since this was a standard practice in his laboratory (personal communication). We decided to test if ion-exchange chromatography could be used to detect impurities in our iodine-substituted DNA. Figure 1 shows a chromatograph from the trityl-off reverse-phase HPLC step in our purification of the DNA with iodine substitutions at positions 12+12'. The trityl-off chromatography step is normally the final step in our purification procedure for synthetic oligonucleotides (see Chapter 2). The chromatograph in Figure 1 shows a single dominating peak with a slight shoulder at the front end of the peak. Before we had discussed the problems of purifying iodinated DNA oligonucleotides with Stephen Burley, we had hoped that a conservative collection of the peak shown in Figure 1 would result in a relatively pure product; we used this DNA to grow our initial

derivative crystal. After we grew these initial derivative crystals and had discussions with Stephen Burley, we retested the purity of our iodine-substituted oligomers by running them through an ion-exchange column. We used a Mono Q anion exchange column (Pharmacia LKB, Piscataway, New Jersey) with running buffer containing 10 mM NaOH, pH 12. Column elution was done by means of a gradient of 200 mM to 1000 mM NaCl. Figure 2b shows a chromatograph of what we had thought was purified DNA oligomer with iodines at 12+12'. The chromatograph shows that there was a significant impurity in our substituted DNA. An ion-exchange run of purified DNA without iodine substitutions (M14-12) suggested that the impurity was probably due to some modification of the iodines during synthesis and/or purification (compare Figure 2a with Figure 2b). When we mixed the unsubstituted M14-12 DNA together with the substituted 12+12' DNA and ran the mixture on the ion-exchange column, we saw that the impurities in Figure 2b and 2c seemed to elute at around the same point in the chromatograph as the unsubstituted DNA did (Figures 2a, 2b, and 2c), which also suggested that the iodines in the substituted DNA had been modified during synthesis and/or purification.

At this point, we reviewed our DNA synthesis and purification protocols to look for steps where damage to the iodinated substitutions might occur. After DNA synthesis, our normal procedure was to remove the alkaline-labile base-protecting groups by treating them with a saturated ammonia solution for 8 to 15 hours at 55°; this procedure followed that recommended by the

manufacturers of the phosphoramidites (Applied Biosystems, Foster City, CA). Given the problem of impurities with the iodine-substituted oligos, we decided to test the effects of using fast oligonucleotide deprotecting (FOD) phosphoramidites (Applied Biosystems, Foster City, CA) for the synthesis of the iodine-substituted DNA. FOD phosphoramidites use a different protecting group than the standard phosphoramidites: dimethylformaminidine and isobutyryl groups for FODs vs. benzoyl and isobutyryl groups for standard phosphoramidites. The result of using different base-protecting groups is that FODs can be deprotected by treatment with saturated ammonia at 55° for only 1 hour versus 8 - 15 hours for the standard chemistry. We suspected that the reduced exposure to saturated ammonia with DNA oligomers synthesized with FOD phosphoramidites might limit damage to the iodine-substitutions. To test this, we synthesized iodine-substituted oligomers using the FOD phosphoramidites. We purified these oligomers according to our usual protocol and then ran the purified products on the ion-exchange column. The chromatograph showed that using the FOD phosphoramidites and limiting the exposure of the oligomer to saturated ammonia during base deprotection significantly decreased the amount of impurity in the iodine-substituted oligomers (Figure 2d). Thus, from that point on, we always used FOD phosphoramidites for the synthesis of iodine-substituted oligomers and were careful to test samples of our oligomers for impurities using ion-exchange chromatography. With these additional procedures, we felt that it was not necessary to include an additional purification step of ion-exchange chromatography after

the two reverse-phase HPLC purification steps. We also decided at this point that given the subtle differences in impurity that could arise in iodine-substituted oligomers during synthesis and purification, we would treat all derivative crystals independently during data collection and reduction.

B. Data Collection and Reduction

1. Mounting of Crystals and Initial Analysis

The best MyoD cocrystals grew out of 10-15% PEG 4000, 100 mM Tris, pH 8.5, 100 mM NaCitrate, 20 mM $BaCl_2$ at room temperature. The largest crystals measured 0.3 mm x 0.3 mm x 0.4 mm, and in spite of having well-defined faces and edges, were not homogenous in morphology. The crystals were very sensitive to changes in solvent content. During mounting, we found that adding any type of buffer or solution, including well buffer, to the hanging drop produced cracks in the crystals. Thus, we mounted the crystals directly out of the hanging drops using thin-walled glass capillary tubes of 1.0 mm or 0.7 mm diameter (Charles Supper Company, Natick, MA). After mounting, crystals were examined for their diffraction characteristics by taking still 'photographs' on a Rigaku RAXIS-IIC imaging plate area detector.

The crystals have unit cell dimensions of a = 222.8 Å, b = 70.8 Å, c = 30.0 Å, and are of the orthorhombic space group $P2_12_12$.

Initial determination of the lattice parameters was done by first
collecting three 10° wedges of data (in the phi direction; Δphi = 1°)
with each 10° wedge separated by at least 30° in phi. After
collecting the wedges of data, the positions of the individual spots
on the oscillation photographs were converted to positions in
reciprocal space using conversion programs developed by Mark Rould
(extract_peaks.for, peaks_to_reciprocal_space_coordinates.for,
M. Rould). We then measured the distances and angular separation of
the individual spots in reciprocal space with the aid of the
computer graphics program FRODO (Jones, 1978) and this allowed us
to directly determine the reciprocal lattice parameters. We
inferred from the spot pattern and the fact that all angles were
very close to 90° that we had a primitive orthorhombic space group;
we confirmed this by testing for the presence of the two-fold
symmetry operators. The identification of the two two-fold screw
axes was made only after a native data set was collected and
examined for systematic absences; solution of the difference
Patterson function allowed us to definitively identify the space
group as $P2_12_12$.

We used the lattice parameters to help estimate the
stoichiometric contents of the asymmetric unit of the crystal along
with the solvent content. If we assume that there are two copies of
a MyoD dimer bound to a 14 bp DNA site in each asymmetric unit, we
get a V_M of 2.39 $Å^3$/dalton and a corresponding solvent content of
55%. The value for V_M is within the range of 1.6 $Å^3$ Da^{-1} to
3.5 $Å^3$ Da^{-1} found for most protein crystals (Matthews, 1968).

2. Data Collection Strategy

With the oscillation method of data collection, there are a few important considerations in choosing the orientation of the crystal axes and angular range of phi and Δphi: 1) completeness of the data in the minimum number of frames; 2) adequate redundancy for internal scaling; 3) absorption and 4) convenience in crystal mounting (Wonacott, 1977; Harrison et al., 1985). The unusual shape of the MyoD cocrystal's unit cell, with its very long **a** axis (222.8 Å), and short **c** axis (30.0 Å), suggested that the crystals should be mounted with the **a** axis aligned parallel to the spindle axis so as to minimize spot overlap. Unfortunately, we found that the features of the crystals (i.e. faces and edges) did not have a reproducible correspondence with the axes of the unit cell. Thus, it was difficult for us to determine the orientation of the crystal axes from the crystal morphology alone. If after taking still photographs, we determined that a crystal was mounted such that the **a** axis was aligned perpendicular to the spindle axis, we could not realign the crystal in a position more favorable for efficient data collection without remounting the crystal (our area detector system only has a single axis of rotation, phi). As the MyoD cocrystals were quite fragile, they usually did not survive remounting without significant damage. Thus, for crystals aligned such that the **a** axis was perpendicular to the spindle axis, we minimized spot overlap during data collection by using Δphi's of only 0.5°. Wonacott gives the following formula as a guide to choosing the maximum Δphi which will avoid overlapping reflexions

for an orthorhombic cell with the a* axis along the X-ray beam
when ϕ = 0 (Wonacott, 1977):

$$\Delta\phi = [(\lambda/a) \, (r/\lambda) \, (180/\pi)] - \Delta$$

where a = 222.8 Å, r = maximum resolution = 2.6 Å, Δ = minimum
rocking range of the crystals.

The MyoD cocrystals have values of Δ somewhere between 0.3
and 0.4. Given this value for Δ, Wonacott's formula suggests that
our maximum Δphi is between 0.27° and 0.37°. Thus, using a Δ phi of
0.5° would likely result in a number of overlapping reflexions in the
higher resolution shells which would have to be discarded during
measurement. We chose a lower limit for Δphi of 0.5° because the
DENZO software (Z. Otwinowski) we used to reduce the data required
a minimum number of fully recorded reflexions in order to process
the data effectively. A smaller value than 0.5° for Δphi would
result in too many partially recorded reflections.

3. Data Reduction

We used the program DENZO (Z. Otwinoski) for our data
reduction. Crystal and camera parameters were refined, and
intensity measurements were made by using profile fitting of the
recorded spots. We handled partially recorded reflexions by
merging successive oscillation frames and integrating the spot
intensities over the merged oscillation frames (merge-denzo.for,

M. Rould). Data were scaled using the program SCALEPACK (Z. Otwinowski) after dividing the merged oscillation frames (divide-merged-denzo.for, M. Rould). We applied a single scale factor for each 5° wedge of data. No explicit corrections were made for absorption or crystal decay. For each of our two native data sets, we combined the data from two separate crystals. Given our concerns about variations in impurity in the iodine-substituted DNA, we treated each derivative crystal individually (attempts to merge data from two different crystals for the 12+12' derivative resulted in high values for R_{sym}, reinforcing our decision to treat each derivative crystal individually). Tables 2 through 7 show final outputs from the SCALEPACK program which summarize the data redundancy, R_{sym} values, and intensities for each resolution shell of the data sets. In order to reduce systematic errors in our derivative data sets, we used the program DSCALE-AD (M. Rould) to local-scale each derivative data set to the native data set (Matthews and Czerwinski, 1975). The locally-scaled derivative data was used in the calculation of Cross R-values (see Table 1, Chapter 4), and all subsequent phasing procedures.

4. Phasing, Model Building and Refinement

Isomorphous difference Patterson and anomolous difference Patterson maps (Figures 3, 4 and 5) were calculated for each derivative crystal using the program PROTEIN (Steigemann, 1974). Figure 3 shows the Harker sections from a representative isomorphous difference Patterson map calculated from native data

and a 12+12' derivative (second crystal) dataset. The periodicity of peaks in the Harker sections show that a few terms bias the map. After removing the reflexions with large intensity differences (>7σ) between the native and derivative data sets (6 reflexions), we used local-scaling again on the derivative data set, and then calculated new difference Patterson maps. Figure 4a shows that removing the reflexions with the large intensity differences significantly improved the interpretability of the maps as viewed from the Harker sections; we treated all subsequent Patterson maps for the other derivatives in a similar manner. Even with this map improvement, however, it was not straightforward to identify all of the heavy atom sites from the Harker sections alone. From our calculations of solvent content, we expected that we would have two copies of a MyoD dimer bound to DNA in each asymmetric unit. Given that the 12+12' derivative was doubly substituted with iodine, we expected a total of four heavy atom sites in each asymmetric unit. Using the Harker sections as a guide, we picked potential heavy atom sites. We used the program HASSP (\ et al., 1987) as an independent method to pick potential heavy atom positions. HASSP works by systematically searching the difference Patterson function for heavy atom sites which have large values for both the self- and cross-vector positions. In the case of the second crystal for the 12+12' derivative, HASSP picked out four potential sites. These sites were also found in our Harker sections of the isomorphous difference Patterson map. There were also corresponding peaks in the anomalous difference Patterson maps for

this derivative, but these peaks were not as clear as in the isomorphous difference Patterson maps (Figure 5a).

For the other derivative crystals, we used similar procedures to the one described above for identifying the positions of the heavy atom sites (Figures 4 and 5). Once we had identified candidate sites for the heavy atoms by using the results from HASSP and the difference Patterson maps, we refined the heavy atom parameters for the derivative crystal against the centric reflexions using the program REFINE from the CCP4 package (S.E.R.C, 1979). After refining the heavy atom parameters (occupancy, position, and thermal parameter) for the 11+11', and 12+12' derivatives, we used the refined positions of the heavy atom sites to fix the origin of the unit cell with respect to the positions of the heavy atom sites from the two derivatives. Since the 11+11', and 12+12' derivatives have iodine substitutions on adjacent bases, we expected that the positions of the heavy atom sites from the two derivatives should differ by between 3.5 to 4.5 Å for each site. We used these expected differences in positions to fix the origin of the unit cell. With the major heavy atom sites identified and initially refined for each derivative, we used difference Fourier methods to check the correctness of the heavy atom sites and to identify any additional sites. For the 11+11' derivative (first crystal), we had originally identified and refined three heavy atom sites. Using these three sites, we calculated a difference map for the residuals (Henderson and Moffat, 1971) and were able to identify a fourth heavy atom site. Thus, both the 12+12' and 11+11' derivatives contained four

heavy atom sites, as expected. The final heavy atom parameters

from REFINE for each derivative are shown in Table 8.

After refining the heavy atom parameters for each derivative

crystal individually with REFINE, we used the program PHARE

(S.E.R.C., 1979) for phase-refinement using data from all of the

derivative crystals. Since we had data sets from four derivative

crystals (two crystals each of derivatives 11+11' and 12+12'), we

used cross-phased refinement in order to reduce bias (Blow and

Matthews, 1973). With this procedure, the parameters of one

derivative are refined while the other derivatives are used to

calculate phases. After each derivative was refined twice by

cross-phased refinement (8 rounds of refinement in total; see Table

9), we generated an initial M.I.R. map. At this stage, the map was of

reasonable quality; we were able to identify alpha-helical regions,

as well as much of the phosphate backbone of the DNA. In order to

improve the quality of our phases, we used a procedure developed by

Mark Rould (Rould et al., 1992). The procedure decouples heavy atom

parameter refinement from the calculation of parent phases by first

solvent flattening (Wang, 1985) the initial M.I.R. map in order to

generate new solvent-flattened phases. These new phases, in turn,

are used in the re-refinement of the heavy atom parameters; during

this re-refinement, the phases are not updated. After convergence

of the refinement, new M.I.R. phases are calculated and the

procedure is repeated, if desired. The parameters from this heavy-

atom refinement against solvent-flattened phases are shown in

Table 10. This additional refinement step improved the figure of

merit from 0.63 to 0.65, and more importantly, improved the interpretability of the M.I.R. map.

Using this solvent-flattened M.I.R. map, we modeled the alpha-helical regions as polyalanine helices and the loop region as an extended polypeptide chain. We placed idealized B-form DNA (generated by INSIGHT) in the density and then adjusted each nucleotide to improve the fit. After completing this initial model, we used the program O (Jones et al., 1991) to superimpose each monomer of MyoD and half of each DNA duplex on a reference MyoD monomer and DNA half-site. We used the rotation and translation matrices generated by this superposition to find the non-crystallographic symmetry axes within the asymmetric unit. By using the non-crystallographic symmetry axes to superpose parts of the M.I.R. map (each part corresponding to an individual monomer and DNA half-site) on a region of the M.I.R. map corresponding to a reference MyoD monomer and DNA half-site, we generated a four-fold non-crystallographic symmetry averaged M.I.R. map. We note that after generating this map, we did not use iterative real-space averaging (Bricogne et al., 1976) to improve the quality of the maps since we felt that the non-crystallographic symmetry averaged M.I.R. map (which had also undergone heavy atom refinement against solvent-flattened phases) was already of high quality. The main chains of the alpha-helices and the loops were visible, as were most of the side chains. The DNA was also well-resolved, with clear density for both the phosphate backbone and the base pairs. Model building was done primarily with the program FRODO (Jones,

1978) written for use on the Evans and Sutherland's PS390. A more recent version of FRODO, written for Silicon Graphics machines, contains a data base of protein structures from which the modeler can fit fragments of a protein chain to the electron density (Jones and Thirup, 1986). We found the use of this DGLP feature to be very useful in building the loop regions of our model. Figure 6 shows various parts of our model with the corresponding M.I.R. density. We refined our model using the programs XPLOR (Brunger, 1987, 1992) and TNT (Tronrud, 1987), monitoring the course of our refinement with the R_{free} and R values (See 'Experimental Procedures' section of Chapter 4 for detailed discussion about refinement). Figure 7 shows a Ramachandran plot for the final refined structure. The Ramachandran plot shows that only one residue (out of 248) is in a region that is not favorable (Morris et al., 1992). This residue, Asn 343, is in one of the four loops in our asymmetric unit (we have two bHLH dimers per asymmetric unit) and is in a region where the exact structure of the chain is more difficult to determine.

References

Aggarwal, A.K., Rogers, D.W., Drottar, M., Ptashne, M., and Harrison, S.C. (1988). Recognition of a DNA operator by the repressor of phage 434: A view at high resolution. Science *242*, 899-907.

Anderson, J.E., Ptashne, M., and Harrison, S.C. (1987). Structure of the repressor-operator complex of bacteriophage 434. Nature *326*, 846-852.

Blow, D.M., and Matthews, B.W. (1973). Parameter refinement in the multiple isomorphous-replacement method. Acta Cryst. *A29*, 56-62.

Bricogne, G. (1976). Methods and Programs for Direct-Space Exploration of Geometric Redundancies. Acta Cryst. *A32*, 832-847.

Brunger, A.T. (1992). X-PLOR v3.1 Manual (New Haven, Connecticut: Yale University Press).

Brunger, A.T., Kuriyan, J., and Karplus, M. (1987). Crystallographic R-factor refinement by molecular dynamics. Science 235, 458-460.

Harrison, S.C., Winkler, F.K., Schutt, C.E., and Durbin, R.M. (1985). Oscillation method with large unit cells. Methods in Enzymology *114*, 211-237.

Henderson, R., and Moffat, J.K. (1971). The difference fourier technique in protein crystallography: Errors and their treatment. Acta Cryst. *B27*, 1414-1420.

Jones, T.A. (1978). A graphics model building and refinement system for macromolecules. J. Appl. Cryst. *11*, 268-272.

Jones, T.A., and Thirup, S. (1986). Using known substructures in protein model building and crystallography. EMBO J. *5*, 819-822.

Jones, T.A., Zou, J.-Y., Cowan, S.W., and Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Cryst. *A47*, 110-119.

Jordan, S.R., and Pabo, C.O. (1988). Structure of the lambda complex at 2.5 Å resolution: Details of the repressor-operator interactions. Science *242*, 893-899.

Matthews, B.W. (1968). Solvent content of protein crystals. J. Mol. Biol. *33*, 491-497.

Matthews, B.W., and Czerwinski, E.W. (1975). Local scaling: A method to reduce systematic errors in isomorphous replacement and anomalous scattering measurements. Acta Cryst. A31, 480-487.

Morris, A.L., MacArthur, M.W., Hutchinson, E.G., and Thornton, J.M. (1992). Stereochemical quality of protein structure coordinates. Proteins *12*, 345-364.

Rould, M.A., Perona, J.J., and Steitz, T.A. (1992). Improving multiple isomorphous replacement phasing by heavy-atom refinement using solvent-flattened phases. Acta Cryst. *A48*, 751-756.

S.E.R.C. [U.K.] (1979). Collaborative Computing Project No. 4, a suite of programs for protein crystallography (Distributed from the Daresbury Laboratory, Warrington, U.K.).

Steigemann, W. (1974). Thesis, TU Muenchen.

Terwilliger, T.C., Kim, S.C., and Eisenberg, D. (1987). Generalized method of determining heavy-atom positions using the difference patterson function. Acta Cryst. *A43*, 1-5.

Tronrud, D.E., TenEyck, L.F., and Matthews, B.W. (1987). An efficient general-purpose least-squares refinement program for macromolecular structures. Acta Cryst. *A43*, 489-501.

Wang, B.C. (1985). Resolution of phase ambiguity in macromolecular crystallography. Methods in Enzymology *115*, 90-112.

Wolberger, C., Dong, Y., Ptashne, J., and Harrison, S.C. (1988). Structure of a phage 434 cro/DNA complex. Nature *335*, 789-795.

Wonacott, A.J. (1977). Geometry of the rotation method. In The Rotation Method in Crystallography, U.W. Arndt and A.J. Wonacott, eds. (Amsterdam, The Netherlands: Elsevier/North-Holland Biomedical Press), pp. 75-101.

Table Legends

Table 1: Iodine-substituted DNA sites.

Sequences of DNA oligonucleotides which had thymine bases replaced by 5-iodouracil and/or cytosine bases replaced by 5-iodocytosine. We summarize the results of cocrystallization trials with these DNA oligomers and MyoD peptide MD-bHLH_C135S (see Figure 2, Chapter 2 for primary sequence of this peptide). The column "Diffraction" refers to the diffraction limit of the crystal tested.

Table 2: Summary of data reduction statistics for Native Data Set 1.

We show the summary of output from the program SCALEPACK (Z. Otwinowski), which was used for the scaling of the diffraction data (reduced with the program DENZO (Z. Otwinowski)). This native data set contains the merged data from two separate native crystals.

Table 3: Summary of data reduction statistics for Native Data Set 2.

This native data set contains the merged data from two separate crystals. These two crystals diffracted to a higher resolution (2.6 Å versus 2.8 Å) than the two native crystals used for Native Data

Set 1. Attempts to merge Native Data Set 1 with Native Data Set 2 resulted in higher values for R_{sym} than the values for the individual native data sets (R_{sym} = 8.1% for the merged Native Data Set 1 plus Native Data Set 2; R_{sym} = 7.9 % for Native Data Set 1; R_{sym} = 6.5% for Native Data Set 2); thus, we have kept the two native data sets separate. The higher R_{sym} value that resulted from merging the two native data sets could have been due in part to the fact that the two crystals used in Native Data Set 1 had a different crystal alignment (a axis parallel to the spindle axis) than the crystal alignment for the two crystals used in Native Data Set 2 (a axis perpendicular to the spindle axis).

Table 4: Summary of data reduction statistics for 12+12' derivative (First crystal).

The data is from a cocrystal containing DNA which had thymine replaced by 5-iodouracil at the 12 and 12' positions (for numbering scheme, see Table 1). As explained in the text, we treated data from each derivative crystals separately. Thus, this data set is from one derivative crystal only.

Table 5: Summary of data reduction statistics for 12+12' derivative (Second crystal).

The data is from a single derivative crystal.

**Table 6: Summary of data reduction statistics for 11+11'
derivative (First crystal).**

The data is from a cocrystal containing DNA which had thymine
replaced by 5-iodouracil at the 11 and 11' positions. The data is
from a single derivative crystal.

**Table 7: Summary of data reduction statistics for 11+11'
derivative (Second crystal).**

The data is from a single derivative crystal.

**Table 8: Heavy atom parameters after refinement on
centric reflexions from each individual derivative crystal.**

The heavy atom parameters (an overall scale factor for each
crystal, and the occupancy, isotropic temperature factor (B factor),
and coordinates for each iodine site) from each individual
derivative crystal were refined separately against centric
reflexions using the program REFINE from the CCP4 package
(S.E.R.C., 1979). The heavy atom parameters shown are from after
two rounds of refinement with REFINE.

Table 9: Heavy atom parameters after cross-phased refinement.

Heavy atom parameters (an overall scale factor for each crystal, an overall isotropic temperature factor (B factor) for each crystal, and the real occupancy, anomalous occupancy, isotropic temperature factor, and coordinates for each iodine site) from all of the derivative crystals were refined using cross-phased refinement (see text for details) with the program PHARE from the CCP4 package (S.E.R.C., 1979). The heavy atom parameters shown are from after two rounds of refinement with PHARE.

Table 10: Heavy Atom Parameters after cross-phased refinement against solvent-flattened phases.

The procedure follows that first presented in Rould et al., 1992, and is discussed further in the text. The heavy atom parameters are from after one round of refinement with PHARE.

Figure Legends

Figure 1: Reverse-phase HPLC chromatography on iodine-substituted DNA.

We show a chromatograph from a trityl-off reverse-phase HPLC run on DNA which had 5-iodouracil substituted for thymine at the 12 and 12' positions (see Chapter 2 for further discussion of the DNA purification protocol). The DNA elutes as a single peak with a shoulder at the front end of the peak. The gradient is represented by sloping lines running from right to left and is shown at the bottom of the graph. The scale shown records the percentage of eluant (acetonitrile in 50 mM triethyl ammonium acetate, pH 6.0) at points within the elution gradient. On the DNA peak, we have marked our cut-off points for collection of the eluting peak.

Figure 2: Ion-exchange chromatography on iodine-substituted DNA.

DNA with 5-iodouracils substituted for thymine at the 12 and 12' positions were purified by two rounds of reverse-phase HPLC chromatography (the chromatograph from the final round of reverse-phase HPLC chromatography is shown in Figure 1). After annealing the HPLC-purified DNAs, we ran them through a Mono Q ion-exchange column (Pharmacia LKB, Piscataway, New Jersey; see text for details of protocol). Figure 2a shows the ion-exchange run of DNA oligomer M14-12 (the DNA used in the native crystals; it

was purifed by reverse-phase HPLC and does not contain iodine substitutions). The DNA elutes as a single peak. The gradient of the chromatography run is represented by a sloping line running from left to right. The scale shown records the percentage of eluant (1 M NaCl in 10 mM NaOH, pH 12) at points within the elution gradient. Figure 2b shows the ion-exchange run of DNA with 12+12' iodine substitutions (purified by reverse-phase HPLC). Notice that in contrast to the DNA in Figure 2a, this DNA elutes as two peaks, with the left-most peak eluting at a position in the gradient similar to that of the M14-12 DNA peak shown in Figure 2a. Since the left-most peak is the smaller of the two, we assume that it is an impurity. In Figure 2c, we had mixed the M14-12 DNA with the 12+12' iodine-substituted DNA before running the mixture through the ion-exchange column. Notice that the height of the left-most peak is larger for the mixture than for the 12+12' iodine-substituted DNA alone (compare Figure 2b with Figure 2c), suggesting that the impurity in Figure 2b has similar elution characteristics to that of the unsubstituted M14-12 DNA. Figure 2d shows the ion-exchange run of 12+12'-substituted DNA which had been synthesized using FOD phosphoramidites and purified by reverse-phase HPLC (see text for details). Notice that in contrast to Figure 2b, the height of the left-most peak is much smaller.

Figure 3: Isomorphous difference Patterson map.

Harker sections from the 12+12' derivative (second crystal). We show three sections of the isomorphous difference Patterson map,

corresponding to u=1/2, v=1/2, and w =0, calculated from Native

Data Set 1 and data from the 12+12' derivative (second crystal).

The map was generated by the program PROTEIN (Steigemann, 1975)

and covers data from 20 to 3.0 Å. The contours of the maps start at

1 sigma and are in increments of 1 sigma. Notice that the maps

show a strong periodicity of peaks. The map was thus, difficult to

interpret. Removing the terms with large intensity differences

between the native and derivative sets resulted in more

interpretable maps (Figure 4).

**Figure 4: Isomorphous difference Patterson maps after
removal of reflexions with large intensity differences
between the native and derivative data sets.**

We removed reflexions with a greater than 7 sigma difference in

intensity between the native and derivative data sets. Notice that

the maps are no longer dominated by a periodicity of peaks as in

Figure 3. Figure 4a shows Harker sections from the 12+12'

derivative (second crystal). Figure 4b shows Harker sections from

the 12+12' derivative (first crystal). Figure 4c shows Harker

sections from the 11+11' derivative (first crystal). Figure 4d

shows Harker sections from the 11+11' derivative (second crystal).

In the maps, peaks which correspond to the heavy atom sites of the

derivative are circled.

Figure 5: Anomalous difference Patterson maps.

We show Harker sections calculated from anomalous differences for each derivative data set. Before generating the maps, we removed the reflexions with large intensity differences (> 6 sigma) between the primary and friedal mates of an anomalous pair. Figure 5a shows Harker sections from the 12+12' derivative (second crystal). Figure 5b shows Harker sections from the 12+12' derivative (first crystal). Figure 5c shows Harker sections from 11+11' derivative (first crystal). Figure 5d shows Harker sections from the 11+11' derivative (second crystal). Notice that the peaks in the anomalous difference Patterson maps are generally not as clear as in the isomorphous difference Patterson maps (Figure 4). Peaks or regions of the maps which correspond to the heavy atom sites of the derivative are circled.

Figure 6: Electron density for the MyoD/DNA structure.

M.I.R. maps were made after heavy-atom parameter refinement of solvent-flattened phases and subsequent recalculation of the phases. As a final step, we averaged the M.I.R. map about the two separate non-crystallographic symmetry axes within each asymmetric unit. This figure shows the solvent-flattened M.I.R electron density map at 3.0 Å resolution after four-fold non-crystallographic symmetry averaging. The maps are contoured at 1 sigma and are superimposed on stick figures showing the final model. a) shows the M.I.R. electron density around the DNA. b)

shows the M.I.R. electron density around the loop region of the MyoD peptide. Starting at the top center of the figure and moving clockwise, we show residues 136 to 146, which has the sequence TSSNPNQRLPK. c) shows the M.I.R. electron density around helices 1 and 2. We show a region at the dimer interface (corresponding to the region A discussed in Chapter 5). Helix 1 is shown in yellow and Helix 2' (from the other monomer within the MyoD dimer) is shown in magenta. Starting from the bottom center of the figure and moving upwards, we show residues 120 to 135 from helix 1 (which has the sequence RRLSKVNEAFETLKRS) and residues 146 to 158 from helix 2 (which has the sequence KVEILRNAIRYIE). d) shows the M.I.R. electron density at the carboxy terminal ends of the MyoD peptide. We show the dimer interface between helix 2 and helix 2'. Residues 152 to 166 (with the sequence NAIRYIEGLQALLRD) are shown for each helix.

Figure 7: Ramachandran plot.

This figure shows the location in Phi,Psi space of each amino acid residue from the final model of the MyoD/DNA complex. The angle phi (which measures the dihedral angle about the N-Ca bond) is shown on the abscissa, and the angle psi (which measures the dihedral angle about the Ca-C bond) is shown on the ordinate.

Derivative	Crystals	Diffraction
1) 1+1' 5' **U**CAACAGCTGTTGA 3' 3' AGTTGTCGACAAC**U** 5'	no crystals	n/a
2) 12+12' 5' TCAACAGCTGT**U**GA 3' 3' AG**U**TGTCGACAACT 5'	crystals	diffraction (~ 2.8 Å)
3) 11+11' 5' TCAACAGCTG**U**TGA 3' 3' AGT**U**GTCGACAACT 5'	crystals	diffraction (~ 2.8 Å)
4) 9+9' 5' TCAACAGC**U**GTTGA 3' 3' AGTTG**U**CGACAACT 5'	no crystals	n/a
5) 11,12 + 11',12' 5' TCAACAGCTG**UU**GA 3' 3' AG**UU**GTCGACAACT 5'	crystals	diffraction (~ 4 Å)
6) 2,12 + 2',12' 5' T**C**AACAGCTGT**U**GA 3' 3' AG**U**TGTCGACAA**C**T 5'	crystals	diffraction (~ 4 Å)
7) 2+2' 5' T**C**AACAGCTGTTGA 3' 3' AGTTGTCGACAA**C**T 5'	crystals	diffraction (~ 4 Å)
8) 5+5' 5' TCAA**C**AGCTGTTGA 3' 3' AGTTGTCGA**C**AACT 5'	no crystals	n/a
9) 8+8' 5' TCAACAG**C**TGTTGA 3' 3' AGTTGT**C**GACAACT 5'	no crystals	n/a

U = 5-iodouracil   **C** = 5-iodocytosine

Table 1

Native 1

Shell				Summary of observation redundancies by shells:								
Lower	Upper	No. of reflections with given No. of observations										
limit	limit	0	1	2	3	4	5-6	7-8	9-12	13-19	>19	total
20.00	5.76	192	177	239	426	352	148	7	0	0	0	1349
5.76	4.59	41	93	375	390	323	224	0	0	0	0	1405
4.59	4.01	0	96	365	455	282	206	0	0	0	0	1404
4.01	3.65	2	104	371	395	352	165	0	0	0	0	1387
3.65	3.39	1	93	328	426	324	197	0	0	0	0	1368
3.39	3.19	3	104	336	388	383	187	0	0	0	0	1398
3.19	3.03	1	98	334	323	414	180	0	0	0	0	1349
3.03	2.90	3	118	354	405	372	123	0	0	0	0	1372
All hkl		243	883	2702	3208	2802	1430	7	0	0	0	11032

Summary of reflections intensities and R-factors by shells

R linear = SUM (ABS(I - <I>)) / SUM (I))
R square = SUM ((I - <I>) ** 2) / SUM (I ** 2))
Chi**2   = SUM ((I - <I>) ** 2) / (Error ** 2)) * N / (N-1)
In all sums single measurements are excluded

Shell	Lower	Upper	Average	Average	Norm.	Linear	Square
limit		Angstrom	I	error	stat. Chi**2	R-fac	R-fac
	20.00	5.76	2513.6	107.7	72.2	0.808	0.031	0.029
	5.76	4.59	1108.5	59.6	47.1	0.978	0.056	0.055
	4.59	4.01	1149.4	66.4	52.8	1.162	0.070	0.065
	4.01	3.65	727.2	57.5	50.5	1.152	0.102	0.094
	3.65	3.39	634.7	57.1	51.2	1.231	0.129	0.104
	3.39	3.19	415.7	51.0	46.9	1.118	0.185	0.124
	3.19	3.03	160.4	40.5	39.8	1.052	0.340	0.310
	3.03	2.90	89.8	36.2	35.8	1.004	0.500	0.526
All reflections			848.4	59.4	49.5	1.065	0.079	0.048

Table 2

Native 2

```
    Shell            Summary of observation redundancies by shells:
Lower  Upper       No. of reflections with given No. of observations
limit  limit    0     1     2     3     4    5-6   7-8  9-12 13-19   >19   total
20.00  5.17     462   438   475   378   169   175    0     0     0     0    1638
 5.17  4.12     345   386   539   393   179   157    0     0     0     0    1654
 4.12  3.60     249   400   558   443   161   138    0     0     0     0    1700
 3.60  3.27     207   356   601   432   200   121    0     0     0     0    1710
 3.27  3.04     178   371   615   406   222   110    0     0     0     0    1724
 3.04  2.86     291   365   604   372   192    86    0     0     0     0    1619
 3.86  2.72     553   537   470   224    77    12    0     0     0     0    1320
 2.72  2.60    1141   386   309    96     0     0    0     0     0     0     791
All hkl        3426  3239  4171  2744  1200   799    0     0     0     0   12156
```

Summary of reflections intensities and R-factors by shells

R linear = SUM (ABS(I - <I>)) / SUM (I))
R square = SUM ((I - <I>) ** 2) / SUM (I ** 2))
Chi**2   = SUM ((I - <I>) ** 2) / (Error ** 2)) * N / (N-1)
In all sums single measurements are excluded

```
Shell Lower Upper Average     Average    Norm.  Linear Square
limit     Angstrom      I    error  stat. Chi**2 R-fac  R-fac
       20.00   5.17  9646.3  430.4  275.5  1.640  0.040  0.043
        5.17   4.12  4867.1  219.5  147.4  1.875  0.054  0.061
        4.12   3.60  3066.1  170.1  130.3  1.751  0.067  0.068
        3.60   3.27  2371.6  156.2  127.3  1.555  0.075  0.067
        3.27   3.04   705.3  109.3  103.8  1.322  0.175  0.178
        3.04   2.86   320.3  100.2   98.5  1.131  0.343  0.418
        2.86   2.72   175.7   86.6   85.9  1.167  0.464  0.571
        2.72   2.60   132.8   71.6   71.2  1.204  0.383  0.470
All reflections       2894.9  176.5  135.1  1.513  0.065  0.053
```

Table 3

12+12' (First crystal)

```
        Shell              Summary of observation redundancies by shells:
     Lower Upper         No. of reflections with given No. of observations
     limit limit     0     1     2     3     4   5-6   7-8  9-12 13-19   >19   total
     20.00  5.56   706   260   361   100   262     0     0     0     0     0    983
      5.56  4.43   521   226   429   117   270     0     0     0     0     0   1042
      4.43  3.88   523   189   474   105   257     0     0     0     0     0   1025
      3.88  3.52   504   212   440   115   257     0     0     0     0     0   1024
      3.52  3.27   474   220   454   112   242     0     0     0     0     0   1028
      3.27  3.08   515   172   475   108   224     0     0     0     0     0    979
      3.08  2.93   544   208   460    84   211     0     0     0     0     0    963
      2.93  2.80   681   167   447   105   108     0     0     0     0     0    827
  All hkl        4468  1654  3540   846  1831     0     0     0     0     0   7871
```

```
            Summary of reflections intensities and R-factors by shells
   R linear = SUM ( ABS(I - <I>)) / SUM (I) )
   R square = SUM ( (I - <I>) ** 2) / SUM (I ** 2) )
   Chi**2   = SUM ( (I - <I>) ** 2) / (Error ** 2) ) * N / (N-1)
   In all sums single measurements are excluded
```

```
Shell Lower Upper Average       Average     Norm.  Linear Square
limit      Angstrom      I     error   stat. Chi**2 R-fac  R-fac
      20.00  5.56  2894.5     147.5   107.1  1.065  0.029  0.031
       5.56  4.43  1556.0      88.3    69.1  0.948  0.040  0.040
       4.43  3.88  1514.2      86.4    69.4  1.016  0.045  0.045
       3.88  3.52  1005.7      68.1    58.0  1.016  0.060  0.057
       3.52  3.27   956.1      70.1    60.6  0.944  0.059  0.050
       3.27  3.08   361.0      56.7    54.6  1.007  0.134  0.111
       3.08  2.93   163.7      54.8    54.1  0.981  0.287  0.243
       2.93  2.80   100.1      51.9    51.7  0.985  0.429  0.381
  All reflections  1095.8      78.6    65.9  0.995  0.057  0.043
```

Table 4

12+12' (Second crystal)

Summary of observation redundancies by shells:
No. of reflections with given No. of observations

Shell Lower limit	Upper limit	0	1	2	3	4	5-6	7-8	9-12	13-19	>19	total
20.00	5.76	688	180	364	60	245	16	0	0	0	0	865
5.76	4.59	568	134	388	68	262	27	0	0	0	0	879
4.59	4.01	539	130	414	57	248	31	0	0	0	0	880
4.01	3.65	523	121	424	62	240	20	0	0	0	0	867
3.65	3.39	437	134	494	58	231	13	0	0	0	0	930
3.39	3.19	532	125	442	62	225	25	0	0	0	0	879
3.19	3.03	426	151	497	62	204	18	0	0	0	0	932
3.03	2.90	518	164	458	58	171	22	0	0	0	0	873
All hkl		4231	1139	3481	487	1826	172	0	0	0	0	7105

Summary of reflections intensities and R-factors by shells

R linear = SUM (ABS(I - <I>)) / SUM (I))
R square = SUM ((I - <I>) ** 2) / SUM (I ** 2))
Chi**2   = SUM ((I - <I>) ** 2) / (Error ** 2)) * N / (N-1)
In all sums single measurements are excluded

Shell Lower limit	Upper Angstrom	Average I	Average error	stat.	Norm. Chi**2	Linear R-fac	Square R-fac
20.00	5.76	2700.5	114.7	69.2	1.684	0.037	0.036
5.76	4.59	952.4	50.6	38.8	1.401	0.056	0.051
4.59	4.01	1054.3	61.7	48.8	1.142	0.056	0.052
4.01	3.65	735.3	58.3	50.5	1.082	0.078	0.066
3.65	3.39	678.5	59.8	53.4	1.214	0.085	0.070
3.39	3.19	466.1	59.6	56.0	1.053	0.120	0.090
3.19	3.03	163.0	51.2	50.6	1.004	0.293	0.255
3.03	2.90	101.5	53.0	52.8	0.930	0.424	0.415
All reflections		847.2	63.4	52.5	1.192	0.070	0.048

Table 5

11+11' (First crystal)

```
   Shell              Summary of observation redundancies by shells:
Lower Upper        No. of reflections with given No. of observations
limit limit      0     1     2     3     4  5-6  7-8 9-12 13-19  >19 total
20.00  5.95    617   171   341    21   233    0    0    0     0    0   766
 5.95  4.75    511   103   406    59   233    0    0    0     0    0   801
 4.75  4.15    471    84   445    57   217    0    0    0     0    0   803
 4.15  3.78    443    96   432    22   249    0    0    0     0    0   799
 3.78  3.51    472   112   412    50   219    0    0    0     0    0   793
 3.51  3.30    430   126   409    50   204    0    0    0     0    0   789
 3.30  3.14    490   133   399    58   185    0    0    0     0    0   775
 3.14  3.00    455   127   365    49   195    0    0    0     0    0   736
All hkl       3889   952  3209   366  1735    0    0    0     0    0  6262
```

```
           Summary of reflections intensities and R-factors by shells
   R linear = SUM ( ABS(I - <I>)) / SUM (I) )
   R square = SUM ( (I - <I>) ** 2) / SUM (I ** 2) )
   Chi**2   = SUM ( (I - <I>) ** 2) / (Error ** 2) ) * N / (N-1)
   In all sums single measurements are excluded
```

```
Shell Lower Upper Average       Average      Norm.  Linear Square
limit      Angstrom       I    error  stat. Chi**2  R-fac  R-fac
      20.00  5.95   1567.3    79.6   57.4  1.515  0.045  0.043
       5.95  4.75    644.0    40.4   33.5  1.424  0.070  0.069
       4.75  4.15    665.7    45.9   39.3  1.545  0.078  0.078
       4.15  3.78    521.5    45.1   40.8  1.269  0.086  0.078
       3.78  3.51    398.4    44.8   41.8  1.105  0.103  0.092
       3.51  3.30    399.2    47.5   44.3  1.102  0.104  0.086
       3.30  3.14    196.5    43.3   42.3  1.014  0.184  0.146
       3.14  3.00     86.3    41.2   41.0  0.942  0.392  0.341
All reflections      561.2    48.4   42.5  1.247  0.085  0.064
```

Table 6

11+11' (Second crystal)

Summary of observation redundancies by shells:

Shell Lower limit	Upper limit	No. of reflections with given No. of observations										total
		0	1	2	3	4	5-6	7-8	9-12	13-19	>19	
20.00	5.56	251	245	477	174	546	22	0	0	0	0	1464
5.56	4.43	166	165	483	134	597	44	0	0	0	0	1423
4.43	3.88	140	160	503	142	583	41	0	0	0	0	1429
3.88	3.52	127	164	518	133	572	48	0	0	0	0	1435
3.52	3.27	133	119	518	150	573	36	0	0	0	0	1396
3.27	3.08	71	178	552	144	543	36	0	0	0	0	1453
3.08	2.93	163	161	508	158	511	39	0	0	0	0	1377
2.93	2.80	285	291	431	193	295	23	0	0	0	0	1233
All hkl		1336	1483	3990	1228	4220	289	0	0	0	0	11210

Summary of reflections intensities and R-factors by shells

R linear = SUM (ABS(I - <I>)) / SUM (I))
R square = SUM ((I - <I>) ** 2) / SUM (I ** 2))
Chi**2 = SUM ((I - <I>) ** 2) / (Error ** 2)) * N / (N-1)
In all sums single measurements are excluded

Shell Lower limit	Upper Angstrom	Average I	Average error	stat.	Norm. Chi**2	Linear R-fac	Square R-fac
20.00	5.56	2414.9	95.9	51.5	1.837	0.042	0.042
5.56	4.43	1050.8	48.5	34.3	1.527	0.054	0.055
4.43	3.88	936.5	48.9	37.8	1.273	0.060	0.058
3.88	3.52	621.2	43.9	38.1	1.146	0.073	0.065
3.52	3.27	561.8	45.0	39.6	1.103	0.083	0.064
3.27	3.08	210.0	37.1	36.1	1.012	0.168	0.136
3.08	2.93	110.9	37.1	36.7	0.997	0.314	0.259
2.93	2.80	70.2	39.1	38.9	0.995	0.473	0.456
All reflections		766.2	49.8	39.2	1.246	0.070	0.052

Table 7

Derivative 12+12'

First crystal	Second crystal
Scale factor = 0.98319	Scale factor = 1.02899

First crystal

Scale factor = 0.98319

Site 1 Iodine

Occupancy B factor Coordinates
3.613   28.48   0.0859,0.4621,0.1046

Site 2 Iodine

Occupancy B factor Coordinates
4.095   33.83   0.2838,0.5554,0.3440

Site 3 Iodine

Occupancy B factor Coordinates
4.490   48.71   0.1687,0.3964,-0.0126

Site 4 Iodine

Occupancy B factor Coordinates
4.447   34.78   0.4528,0.2328,0.1989

Second crystal

Scale factor = 1.02899

Site 1 Iodine

Occupancy B factor Coordinates
4.407   29.84   0.0857,0.4590,0.0643

Site 2 Iodine

Occupancy B factor Coordinates
5.341   30.95   0.2843,0.5571,0.3494

Site 3 Iodine

Occupancy B factor Coordinates
5.153   34.01   0.1690,0.3938,-0.0352

Site 4 Iodine

Occupancy B factor Coordinates
5.300   28.51   0.4522,0.2326,0.2135

Table 8

Derivative 11+11'

First crystal	Second crystal

First crystal

Scale factor = 1.01053

Site 1 Iodine

Occupancy	B factor	Coordinates
4.975	27.97	0.0703,0.4660,-0.0087

Site 2 Iodine

Occupancy	B factor	Coordinates
6.534	30.70	0.2795,0.5288,0.4719

Site 3 Iodine

Occupancy	B factor	Coordinates
6.091	26.88	0.1791,0.3887,-0.1686

Site 4 Iodine

Occupancy	B factor	Coordinates
5.277	27.24	0.4621,0.2045,0.1318

Second crystal

Scale factor = 1.01435

Site 1 Iodine

Occupancy	B factor	Coordinates
4.185	27.82	0.0708,0.4654,-0.0024

Site 2 Iodine

Occupancy	B factor	Coordinates
5.351	15.55	0.2793,0.5290,0.4761

Site 3 Iodine

Occupancy	B factor	Coordinates
5.242	25.70	0.1791,0.3889,-0.1673

Site 4 Iodine

Occupancy	B factor	Coordinates
5.227	29.34	0.4622,0.2040,0.0914

Table 8 (cont.)

122

Derivative 12+12'

First crystal

Scale factor = 1.00936   Overall B = -0.11297

Site 1 Iodine
Real occupancy   3.57198
Anomalous occupancy   3.13250
Isotropic temperature factor   45.25246
Atomic coordinates
        0.08628   0.46213   0.08014

Site 2 Iodine
Real occupancy   4.47473
Anomalous occupancy   4.70482
Isotropic temperature factor   47.46457
Atomic coordinates
        0.28407   0.55610   0.34426

Site 3 Iodine
Real occupancy   4.33637
Anomalous occupancy   4.39871
Isotropic temperature factor   49.12519
Atomic coordinates
        0.16867   0.39557   -0.02842

Site 4 Iodine
Real occupancy   4.52842
Anomalous occupancy   5.15625
Isotropic temperature factor   43.24976
Atomic coordinates
        0.45240   0.23429   0.21126

Phasing power to 3.0 Å = 1.68[a]

Second crystal

Scale factor = 0.97342   Overall B = -0.38405

Site 1 Iodine
Real occupancy   5.11988
Anomalous occupancy   4.84505
Isotropic temperature factor   47.83429
Atomic coordinates
        0.08604   0.46162   0.07695

Site 2 Iodine
Real occupancy   6.11132
Anomalous occupancy   6.87854
Isotropic temperature factor   45.07910
Atomic coordinates
        0.28413   0.55724   0.34709

Site 3 Iodine
Real occupancy   6.26975
Anomalous occupancy   6.88597
Isotropic temperature factor   48.31748
Atomic coordinates
        0.16867   0.39571   -0.03068

Site 4 Iodine
Real occupancy   5.68761
Anomalous occupancy   5.70042
Isotropic temperature factor   43.03117
Atomic coordinates
        0.45241   0.23443   0.20942

Phasing power to 3.0 Å = 2.07[a]

[a] rms F_H/E, where E = residual lack of closure

Table 9

Derivative 11+11'

	First crystal	Second crystal
	Scale factor = 0.99079 Overall B = -0.33002	Scale factor = 0.96290 Overall B = 0.13387
Site 1 Iodine		
Real occupancy	5.22575	5.59465
Anomalous occupancy	5.54299	4.34584
Isotropic temperature factor	36.98766	42.10398
Atomic coordinates	0.07070 0.46437 -0.01701	0.07063 0.46407 -0.01758
Site 2 Iodine		
Real occupancy	6.99931	7.17804
Anomalous occupancy	6.67583	5.87884
Isotropic temperature factor	30.18929	27.55417
Atomic coordinates	0.27941 0.52741 0.47214	0.27935 0.52668 0.47556
Site 3 Iodine		
Real occupancy	6.45164	6.92796
Anomalous occupancy	6.23974	5.75457
Isotropic temperature factor	34.36960	36.63137
Atomic coordinates	0.17894 0.38793 -0.15972	0.17891 0.38792 -0.16062
Site 4 Iodine		
Real occupancy	5.54428	5.82248
Anomalous occupancy	6.61023	4.72151
Isotropic temperature factor	31.85756	35.72726
Atomic coordinates	0.46224 0.20379 0.10515	0.46223 0.20374 0.10309
Phasing power to 3.0 Å = 2.09[a]		Phasing power to 3.0 Å = 2.54[a]

Mean figure of merit to 3.0 Å for all derivatives = 0.63

Table 9 (cont.)

Derivative 12+12'

	First crystal	Second crystal
	Scale factor = 1.00887 Overall B = -0.11297	Scale factor = 0.97033 Overall B = -0.38405

Site 1 Iodine

	First crystal	Second crystal
Real occupancy	3.72854	5.43781
Anomalous occupancy	3.87528	6.12636
Isotropic temperature factor	46.30278	50.59309
Atomic coordinates	0.08633 0.46155 0.08112	0.08607 0.46122 0.07716

Site 2 Iodine

	First crystal	Second crystal
Real occupancy	4.55377	6.53778
Anomalous occupancy	5.62542	8.20054
Isotropic temperature factor	49.01730	46.72416
Atomic coordinates	0.28399 0.55639 0.34450	0.28420 0.55750 0.34758

Site 3 Iodine

	First crystal	Second crystal
Real occupancy	4.26164	6.46658
Anomalous occupancy	5.19894	8.28199
Isotropic temperature factor	48.21379	48.66072
Atomic coordinates	0.16870 0.39563 -0.02731	0.16870 0.39567 -0.03076

Site 4 Iodine

	First crystal	Second crystal
Real occupancy	4.79627	6.03464
Anomalous occupancy	6.23052	6.86735
Isotropic temperature factor	45.11608	44.69779
Atomic coordinates	0.45245 0.23427 0.21158	0.45245 0.23515 0.20966

Phasing power to 3.0 Å = 1.69[a]	Phasing power to 3.0 Å = 2.12[a]

[a]rms F_H/E, where E = residual lack of closure

Table 10

Derivative 11+11'

First crystal

Scale factor = 0.99004 Overall B = -0.33002

Site 1 Iodine
Real occupancy 5.44029
Anomalous occupancy 6.32572
Isotropic temperature factor 40.44254
Atomic coordinates
-0.01740 0.07071 0.46477

Site 2 Iodine
Real occupancy 7.07708
Anomalous occupancy 8.09575
Isotropic temperature factor 30.93017
Atomic coordinates
0.47144 0.27948 0.52739

Site 3 Iodine
Real occupancy 6.50727
Anomalous occupancy 7.57448
Isotropic temperature factor 36.30181
Atomic coordinates
-0.15784 0.17895 0.38799

Site 4 Iodine
Real occupancy 5.63924
Anomalous occupancy 7.68985
Isotropic temperature factor 32.03269
Atomic coordinates
0.10475 0.46228 0.20383

Phasing power to 3.0 Å = 2.05[a]

Second crystal

Scale factor = 0.96272 Overall B = 0.13387

Site 1 Iodine
Real occupancy 5.68596
Anomalous occupancy 5.30385
Isotropic temperature factor 48.28655
Atomic coordinates
-0.01743 0.07070 0.46398

Site 2 Iodine
Real occupancy 7.14976
Anomalous occupancy 7.65180
Isotropic temperature factor 29.90097
Atomic coordinates
0.47530 0.27940 0.52703

Site 3 Iodine
Real occupancy 6.93636
Anomalous occupancy 7.31050
Isotropic temperature factor 39.87103
Atomic coordinates
-0.15953 0.17893 0.38812

Site 4 Iodine
Real occupancy 5.67871
Anomalous occupancy 5.44083
Isotropic temperature factor 37.79224
Atomic coordinates
0.10249 0.46234 0.20385

Phasing power to 3.0 Å = 2.55[a]

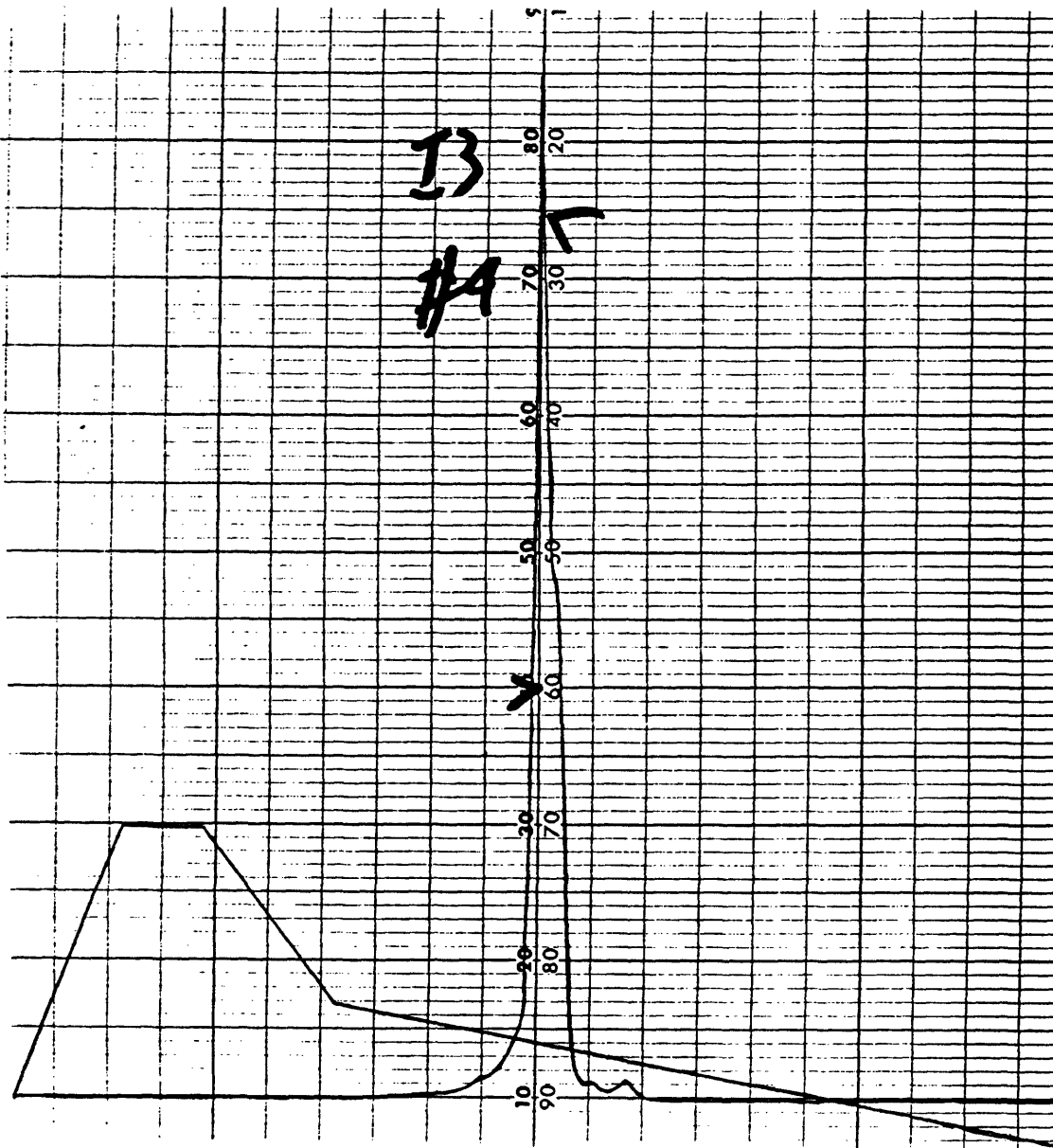Mean figure of merit to 3.0 Å for all derivatives = 0.65
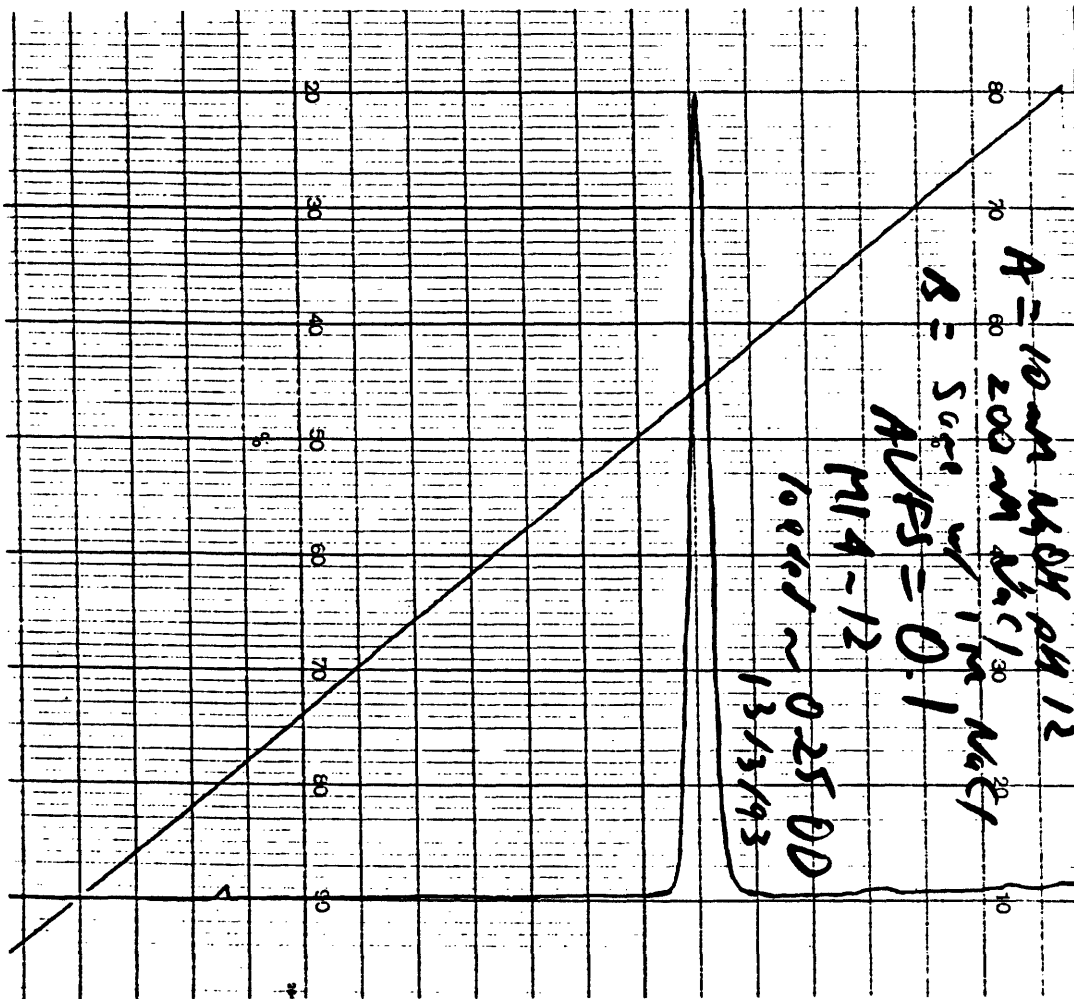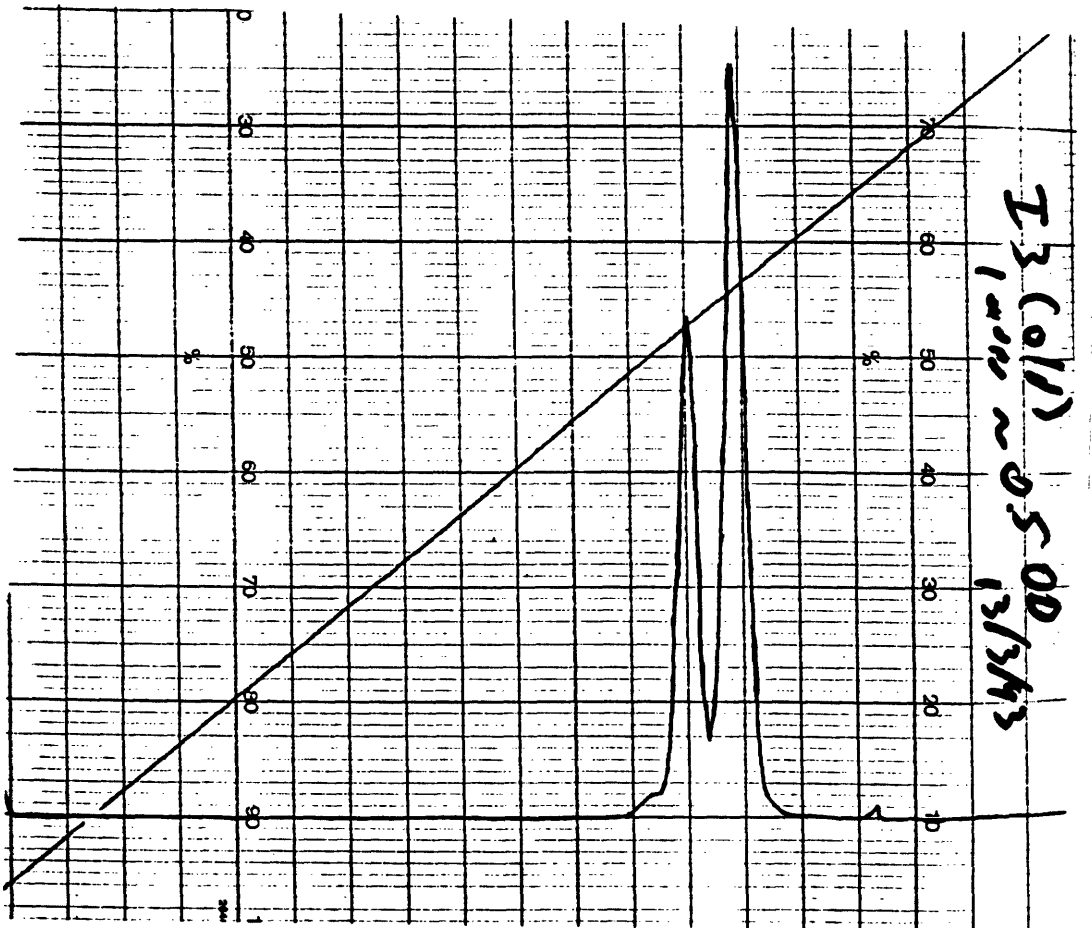
Table 10 (cont.)
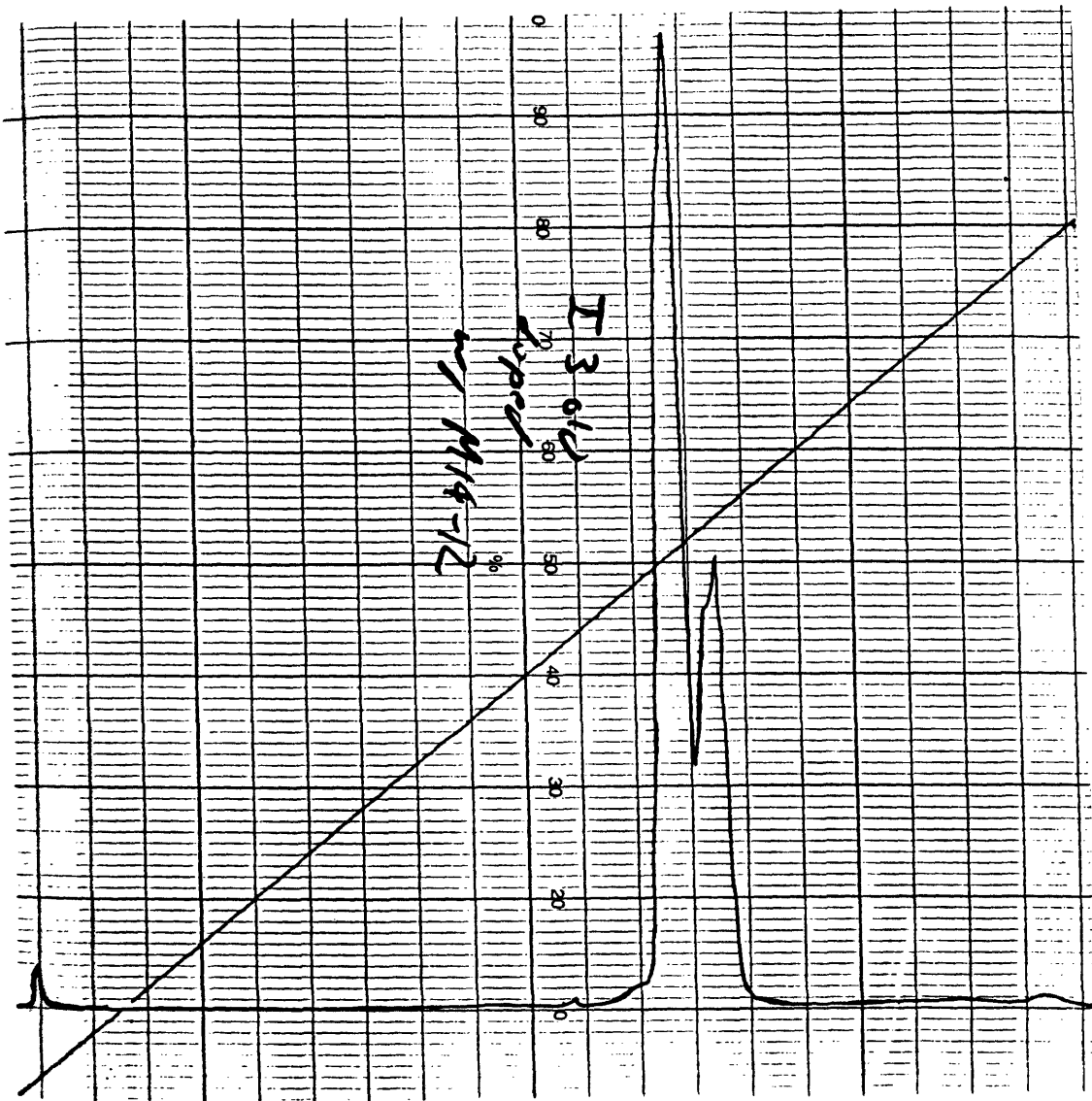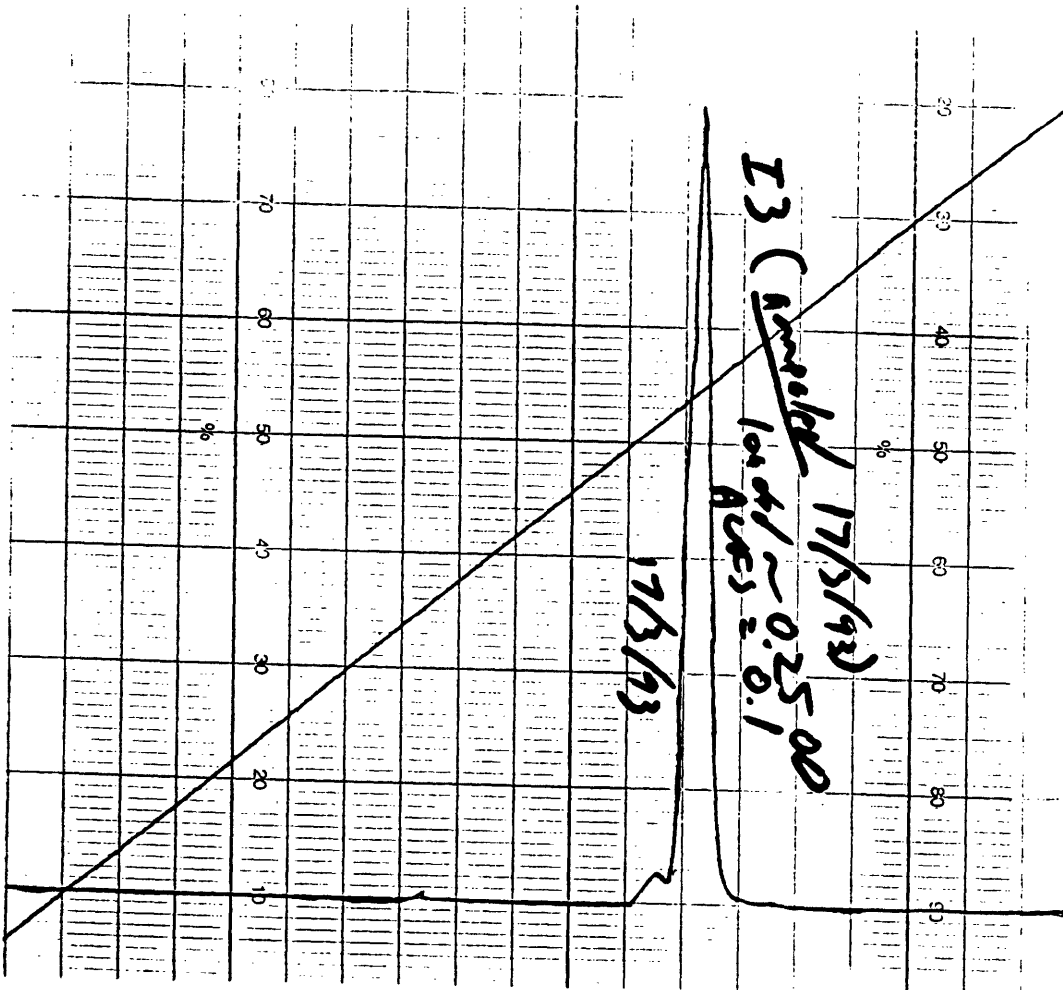
Figure 1

Figure 2a

Figure 2b

Figure 2c

Figure 2d

12+12' (Second Crystal)
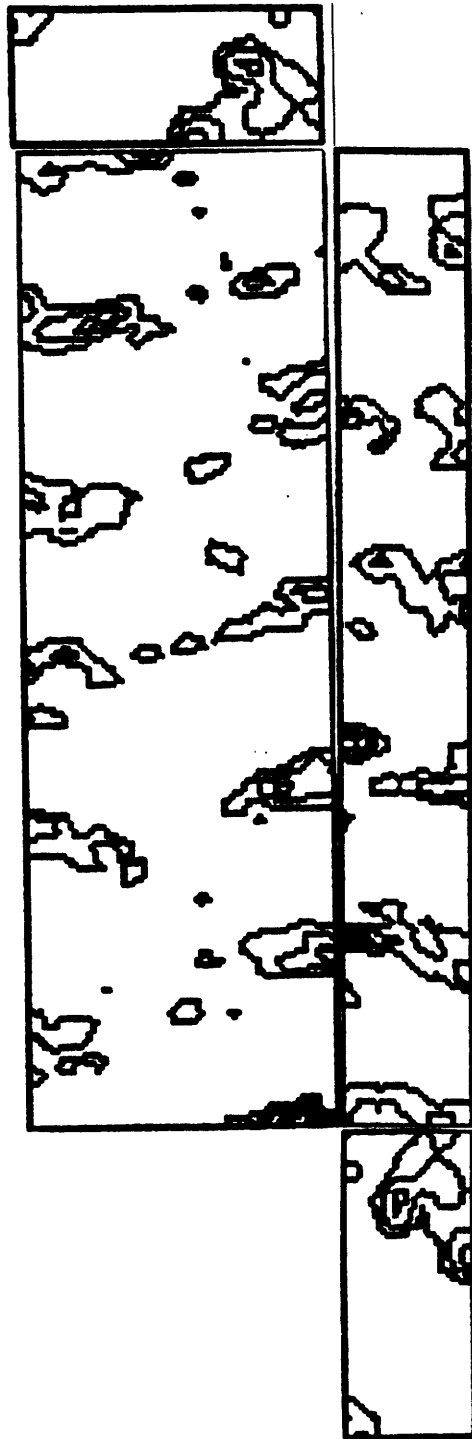

Figure 3

12+12' (Second Crystal)

Figure 4a

12+12' (First Crystal)

Figure 4b

11+11' (First Crystal)

Figure 4c

11+11' (Second Crystal)

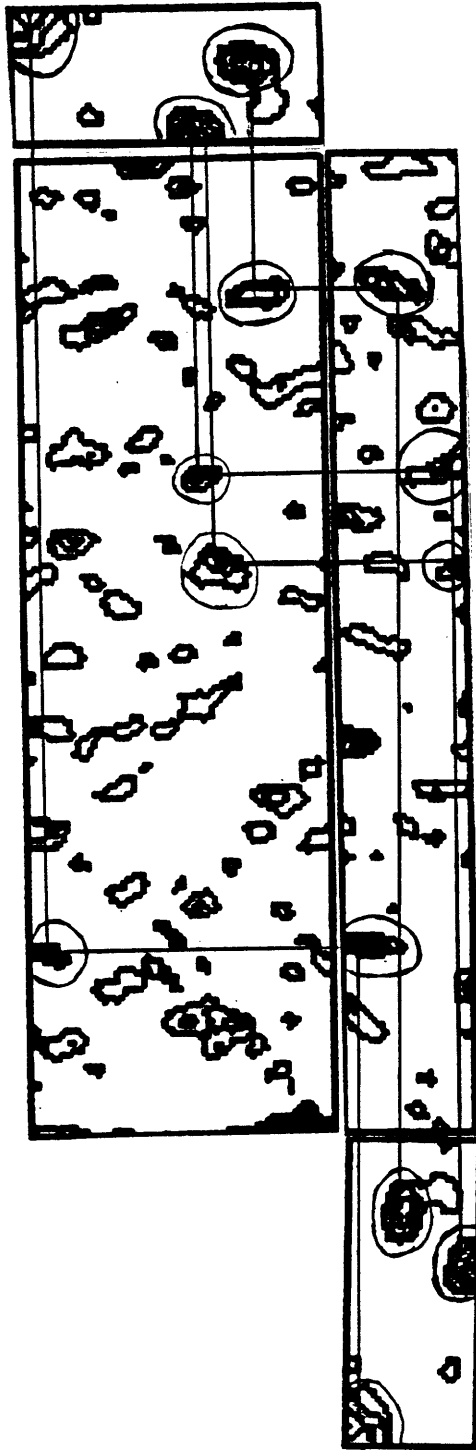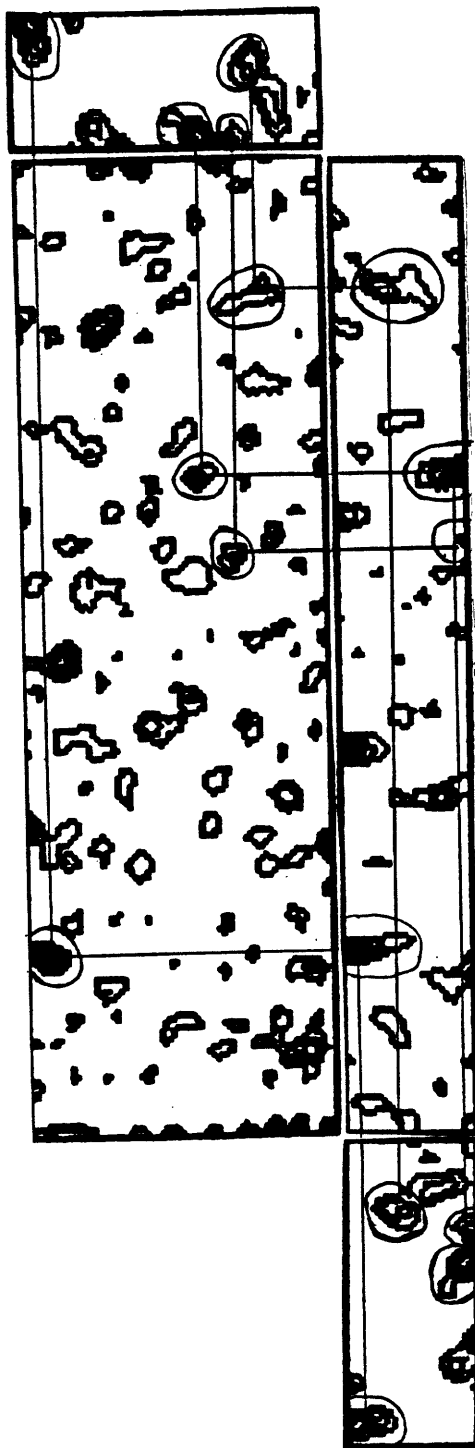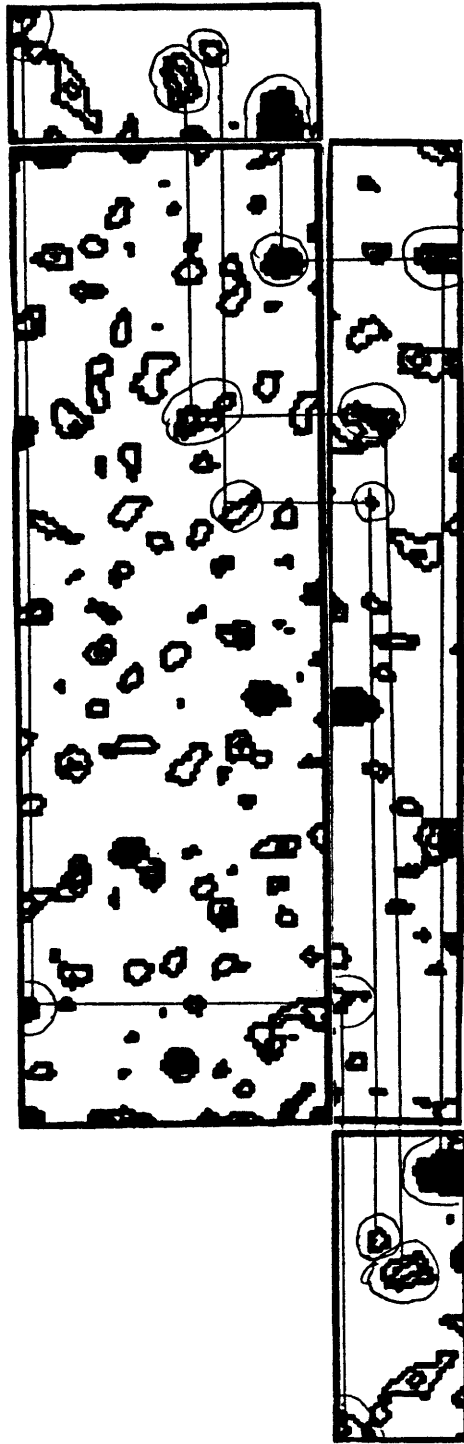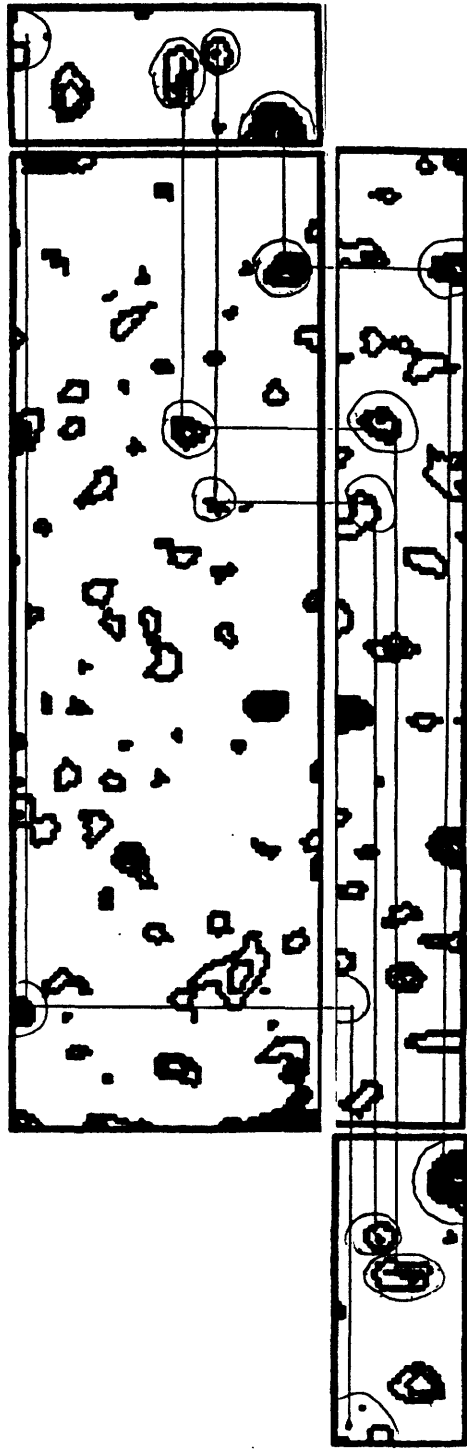Figure 4d

12+12' (Second Crystal)
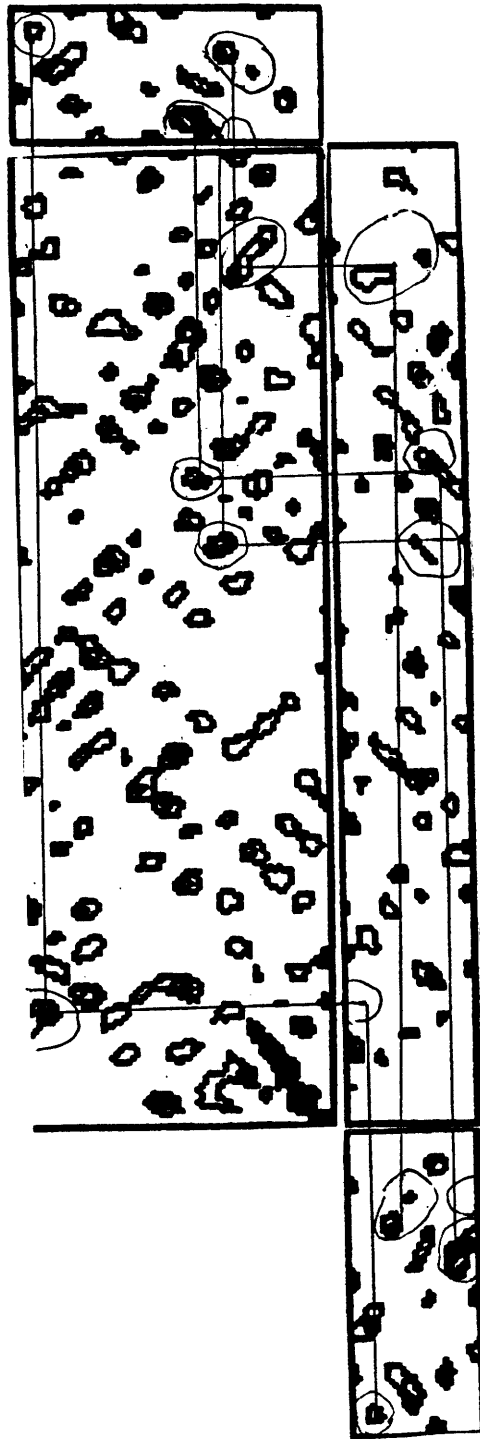
Figure 5a

12+12' (First Crystal)
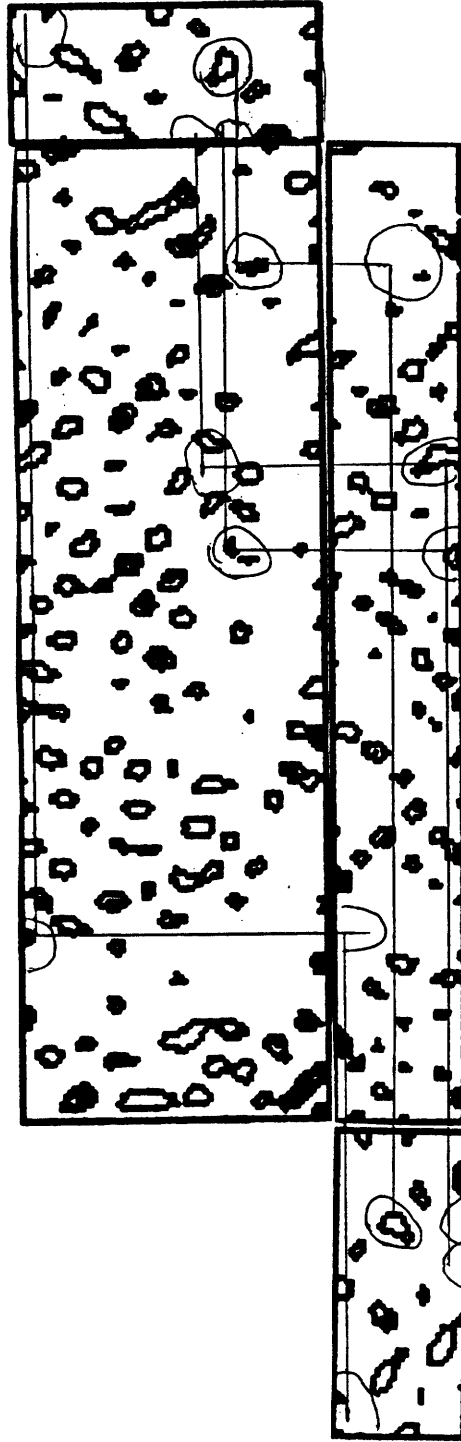
Figure 5b

11+11' (First Crystal)

Figure 5c

11+11' (Second Crystal)
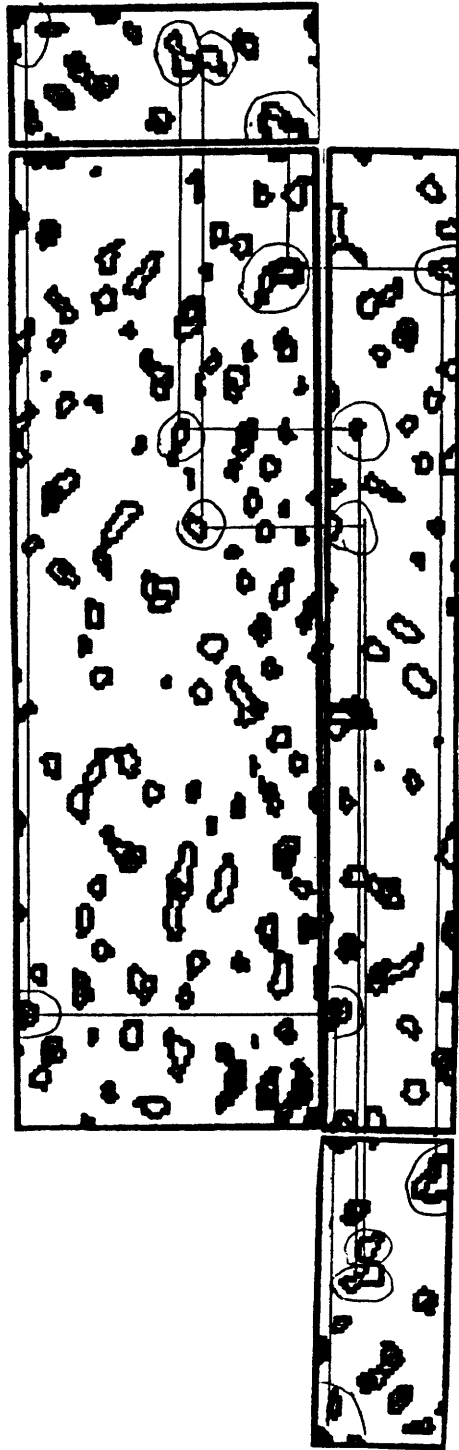
Figure 5d

a)
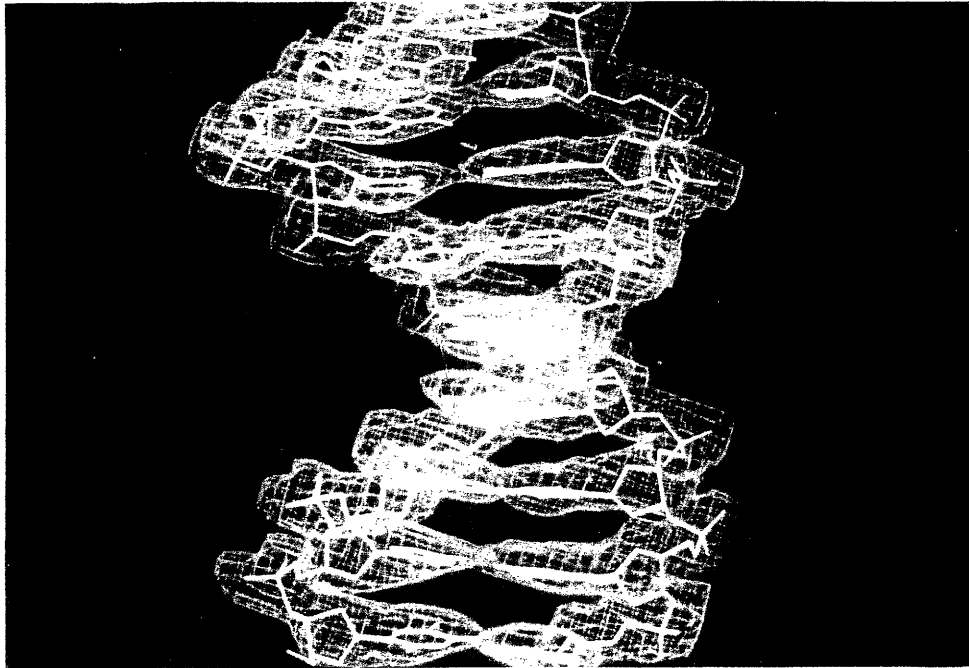
b)
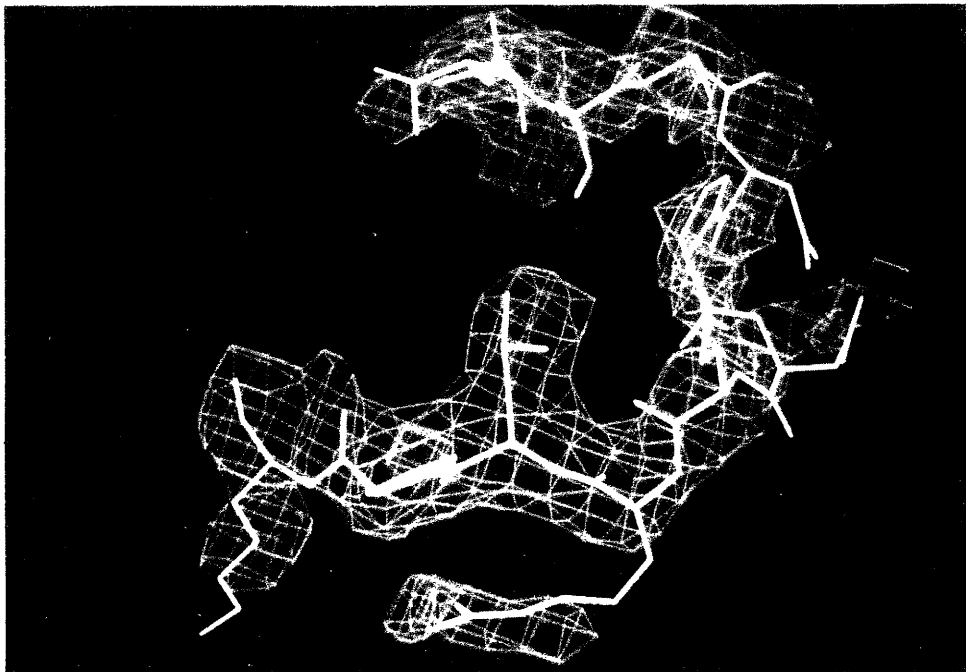
Figure 6

c)
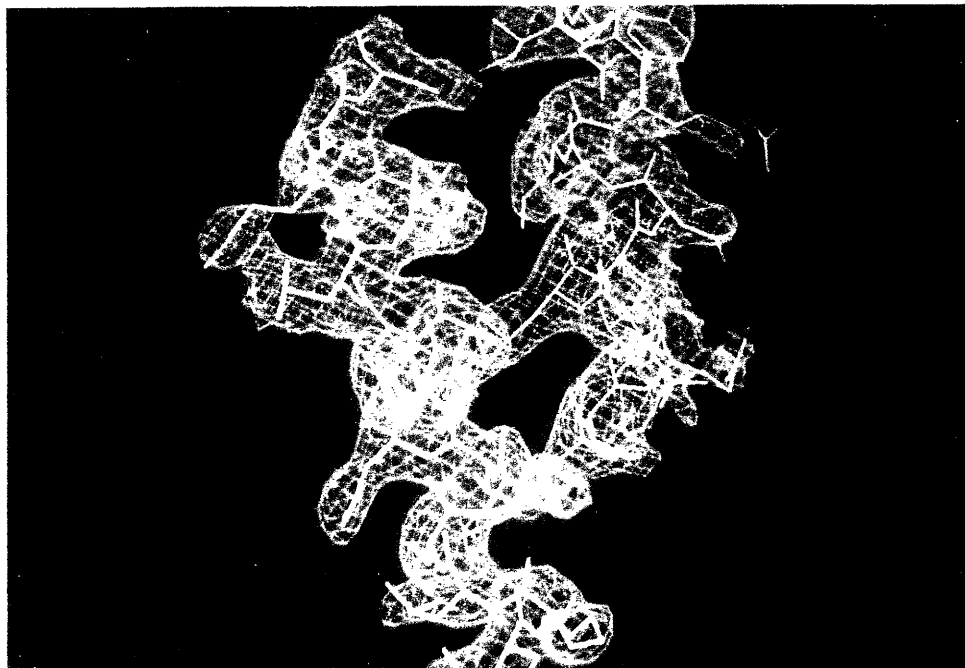
d)
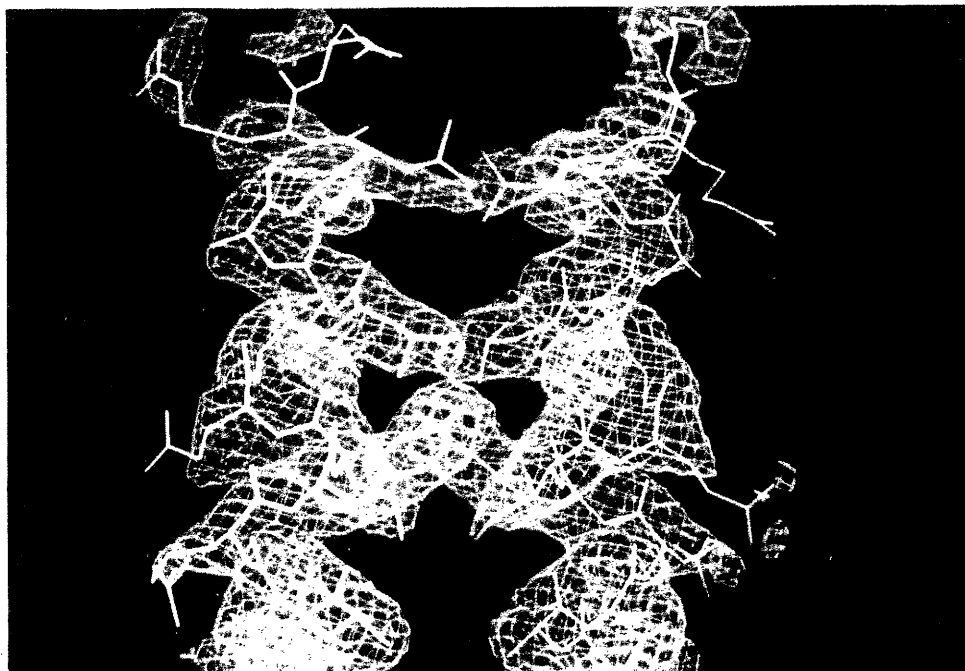
Figure 6

Ramachandran Plot
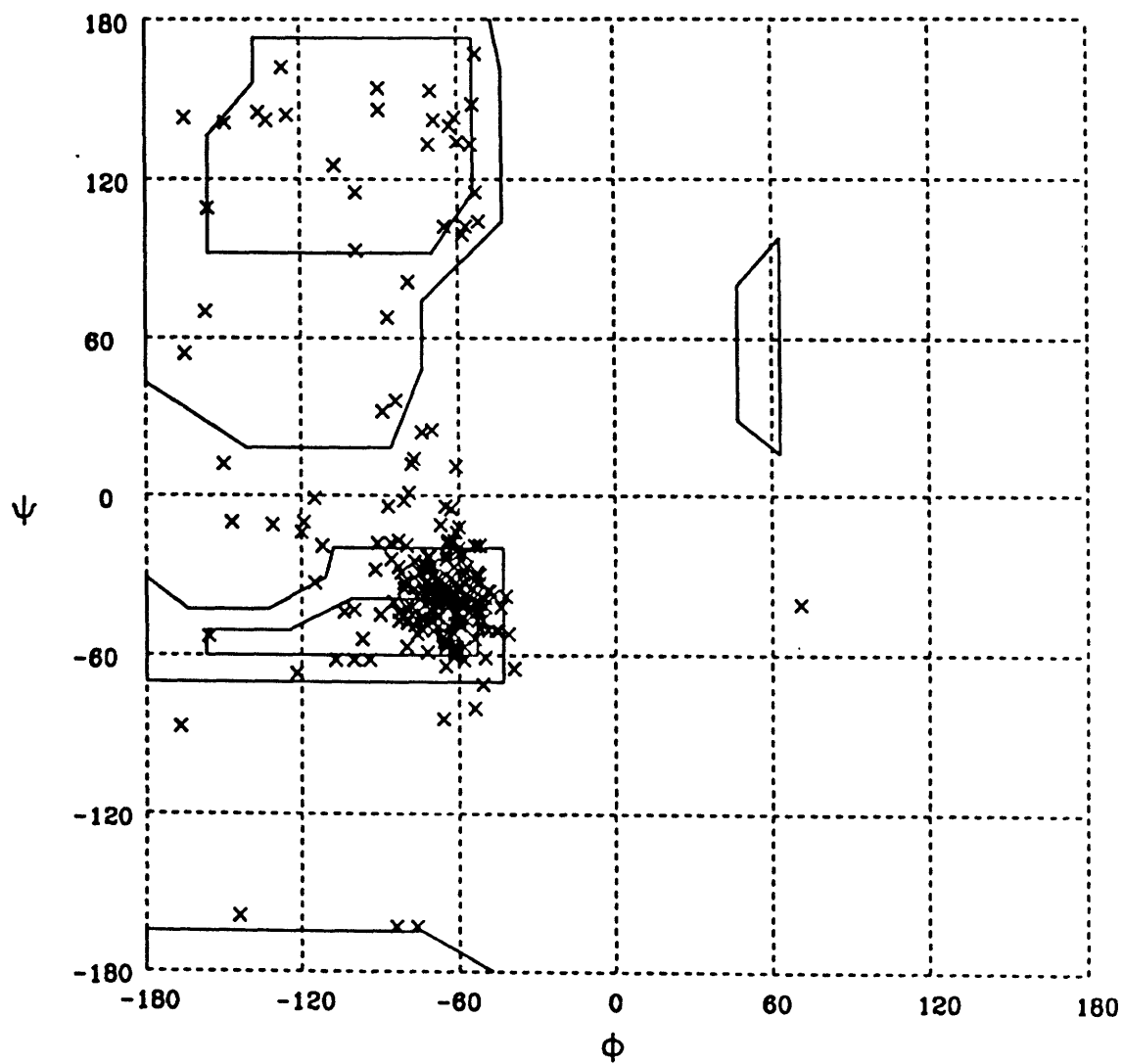
Figure 7

CHAPTER 4

Crystal Structure of MyoD bHLH Domain Bound
to DNA: New Perspectives on DNA Recognition
and Transcriptional Activation

Abstract

The crystal structure of a MyoD basic-helix-loop-helix (bHLH) domain-DNA complex has been solved and refined at 2.8 Å resolution. This structure: proves that bHLH and bHLH-leucine zipper (bHLH-ZIP) proteins are remarkably similar; helps us understand subtle differences in binding preferences for these proteins; and has surprising implications for our understanding of transcription. Specifically, Ala 114 and Thr 115, which are required for positive control in the myogenic proteins, are buried at the protein-DNA interface. These residues are not available for direct protein-protein contacts but they may determine the conformation of Arg 111. Comparisons with Max suggest that the conformation of this arginine - which is different in the two structures - may play an important role in myogenic transcription.

Introduction

Members of the MyoD family of bHLH proteins (MyoD, Myf-5, MRF4, and myogenin) are transcriptional activators that play a central role in determining muscle cell identity (Weintraub, H., 1993; Rudnicki et al., 1993). The basic region mediates DNA binding, and MyoD - like other members of the bHLH superfamily - binds to DNA sites with the consensus sequence CANNTG (Blackwell and Weintraub, 1990). The helix-loop-helix domain mediates dimerization: MyoD forms homodimers and also forms heterodimers with members of the E2 class of bHLH proteins. The crystal structure of the Max bHLH-leucine zipper-DNA (bHLH-ZIP-DNA) complex (Ferré-D'Amaré et al., 1993) provided an initial framework for understanding the helix-loop-helix proteins. The crystal structure of a truncated upstream stimulatory factor bHLH-DNA complex (the zipper region which is normally present in this protein was deleted for crystallization) has provided further information (Ferré-D'Amaré et al., 1994). However, there still are important questions about family/subfamily relationships among the bHLH proteins and about subtle differences in binding site preferences. A particularly intriguing aspect of MyoD biology that has not been understood is the observation that certain mutations in the basic region ("positive control" mutations) bind normally to DNA, but fail to activate transcription.

We have determined the crystal structure of a complex that contains the bHLH domain of MyoD with its preferred binding site.

Even though MyoD (bHLH) and Max (bHLH-ZIP) belong to different subfamilies and have limited sequence identity (25%) in their bHLH regions, we find that these complexes are remarkably similar. Comparing and contrasting these structures: 1) emphasizes the remarkable overall conservation of the bHLH fold, 2) helps us understand subtle differences in the dimerization and binding site preferences of the bHLH and bHLH-ZIP proteins, and 3) suggests how critical residues in the basic region of MyoD may contribute to transcriptional activation.

Results and Discussion

Overall Structure of the MyoD-DNA Complex

We have solved the structure of a complex that contains the bHLH domain of MyoD and a DNA fragment with an optimized binding site (Blackwell and Weintraub, 1990; Figure 1). MyoD binds to this site as a dimer, and the asymmetric unit of the crystal contains two copies of this MyoD-DNA complex. We find that the overall fold of the bHLH domain is strikingly similar to that observed in Max and USF. Each MyoD monomer, which includes residues 102 to 166, forms two long alpha-helices connected by an eight residue loop. The first helix contains the basic region and the region originally identified as H1; the second helix corresponds to the region identified as H2 (Figure 2). As observed in the Max complex, the basic region fits into the major groove, and each monomer of MyoD makes identical DNA contacts. The helical regions H1 and H2 are at

the dimer interface and participate in forming a parallel, left-handed, four-helix bundle (Figure 2). Even though Max and MyoD belong to different subfamilies (Max is a bHLH-ZIP protein), we find that the bHLH regions of these proteins are very similar: if we exclude the loops, these structures can be superimposed with a 1.2 Å rms distance between corresponding alpha carbons. The two copies of the MyoD complex that are present in the asymmetric unit of our crystal provide a useful control for our structural analysis and comparison. Superimposing these two MyoD complexes shows that they have very similar structures and provides an internal check on the consistency of our models.

The Basic Region and DNA Recognition

Most of the known bHLH proteins recognize sites of the form CANNTG, and studies of the Max-DNA and upstream stimulatory factor-DNA complexes have provided an initial description of these interactions. As observed with Max, residues in the basic region of MyoD make most of the DNA contacts (Figures 2, 3 and 4). Glu 118 plays an especially important role in recognition. One of the side chain oxygens accepts a hydrogen bond from the N4 of base C5 (with distances ranging from 2.8 - 3.1 Å in the four different monomers in our crystal). The other oxygen makes a weak hydrogen bond to the N6 of base A6 (3.5 - 3.8 Å). Finally, the carboxylate group also makes a water-mediated contact with the N4 of base C8' and a water-mediated contact with the N7 of base A4. The side chain conformation of Glu 118 appears to be stabilized by salt

bridge interactions with Arg 121, which is located in the next turn of the alpha-helix. Arg 121, in turn, contacts the phosphate of C5 (Figures 3 and 4).

Residues in the basic region of MyoD also contact the conserved T and G of the bHLH binding site, and Arg 111 [which will have a central role in our later discussion of positive control] plays a very important role in making these contacts. Specifically, it 1) donates a hydrogen bond to the N7 of base G10' (2.6 - 3.1 Å), 2) makes a water-mediated contact to the O6 of base G10', and 3) contacts a phosphodiester oxygen at position G10'. Finally, the Arg 111 side chain hydrogen bonds to the hydroxyl group of Thr 115, which, in turn, uses its gamma carbon to make hydrophobic contacts with the methyl group of T9' (the methylene groups of the critical Glu 118 side chain also help to form a hydrophobic pocket at this position). In addition, Thr 115 donates a hydrogen bond to a phosphodiester oxygen at position T9'.

A number of other residues makes important phosphate contacts (Figures 3 and 4). There are three critical residues in the basic region: Arg 117 contacts phosphates at positions C2 and A3 (Arg 117 may also have an electrostatic interaction with the N7 of base A3, but it is too far away (4.0 - 4.4 Å) for a direct hydrogen bond). Arg 110 contacts a phosphate at position C2 and also contacts the terminal -OH at position T1. [We presume it could contact a phosphate at this position, but our synthetic DNA fragment has a terminal -OH.] Arg 119 makes phosphate contacts at

positions G7' and C8'. Several residues outside of the basic region also make phosphate contacts: Asn 126 (from helix 1) contacts position A6', Arg 143 (in the loop) contacts position C5', and Lys 146 (from helix 2) contacts positions C5' and A6'. In all, each MyoD monomer contacts a total of ten positions on the DNA backbone.

Our structure of the MyoD complex raises a number of interesting questions about recognition of the central base pairs. Selection experiments reveal that MyoD and myogenin have limited specificity for the central two base pairs, although they can discriminate against the sequences CACATG and CACGTG (Figure 1c; Blackwell and Weintraub, 1990; Wright et al., 1991). MyoD does not make any direct contacts with the central bases, but the crystal structure shows a water mediated hydrogen bond between Glu 118 and the N4 of base C8'. It seems possible that MyoD's discrimination of the central bases may involve the indirect readout of subtle changes in the DNA conformation and/or additional water-mediated interactions (Otwinowski et al., 1988) not readily identifiable at this level of resolution. There may be fundamental differences in the way that bHLH and bHLH-ZIP proteins contact the central base pairs. Max has an arginine (corresponding to Leu 122 of MyoD) that contacts the N7 atom of the guanine CACGTG and helps explain the sequence preference. In contrast, Leu 122 of MyoD is not in a position to make any base-specific contact, and it is interesting that bHLH proteins typically have a hydrophobic amino acid at this position, while bHLH-ZIP proteins usually have an

arginine. Moreover, genetic studies have shown that changing Leu 122 in MyoD to an arginine changes the preferred binding site from CA GC TG to CA CG TG (Blackwell et al., 1993).

Selection studies also show that the myogenic proteins prefer purines on the 5' flanking side of the site (RA CANNTG) and pyrimidines on the 3' flanking side (CANNTG TY). The crystal structure shows some contacts with the 5' flanking residues that may - directly or indirectly - be responsible for these preferences. As mentioned previously: Glu 118 makes a water mediated contact to the N7 of base A4; Arg 117 makes phosphate contacts at positions C2 and A3, and Arg 117 may have a weak interaction with the N7 of base A3. The Arg residue at position 117 is highly conserved in the bHLH proteins and clearly is important for binding. Changing this residue to leucine (which appears at the corresponding position in Max) knocks out DNA binding (Van Antwerp et al., 1992). Max is rather different in this region: There are no contacts with the flanking bases, binding site selections show no sequence preferences at these positions (Blackwell et al., 1993), and the Leu at position 117 does not make any DNA contacts.

Structure of the DNA

The crystal structure shows that the binding site is essentially B-DNA, but there are a few distinctive features. First, we note that base pairs A6:T9' and A6':T9 - which make critical contacts with Glu 118 - are buckled. There also are two distinctive

global features of the MyoD DNA. First, we note that the base pairs of MyoD have an inclination of about 8 degrees, whereas the base pairs are essentially perpendicular to the helical axis in standard B-DNA. Second, there are interesting differences in groove width. The major groove is rather narrow (10.8 Å in MyoD vs. 12.3 Å in B-DNA) and the minor groove is rather wide (8.1 Å in MyoD vs. 4.8 Å in B-DNA). It is interesting that the Max DNA has similar characteristics. Although these features were not discussed in the original report, the Max DNA has buckled base pairs at positions A6:T9' and A6':T9, a wide minor groove, and a significant inclination of the base pairs. In spite of these distinctive features of the MyoD DNA, the overall structure is essentially B-form DNA, and we find that the DNA is relatively straight. This is consistent with biochemical experiments which had indicated that binding of MyoD does not significantly bend the DNA (Fisher, D., personal communication).

Dimer Formation in bHLH Proteins

Comparing the Max and MyoD structures helps us understand dimer formation in the bHLH and bHLH-ZIP proteins. As mentioned before, the structures of the bHLH domains of Max and MyoD are remarkably similar, and a more detailed comparison shows that they have conserved packing arrangements for the hydrophobic residues in the core of the four-helix bundle. All of the hydrophobic residues in Max have structural counterparts in MyoD, and mutagenesis experiments confirm the importance of these residues

for dimer stability (Davis et al., 1990; Winter et al., 1992; Voronova and Baltimore, 1990). These similarities are particularly noteworthy given the distinctive dimerization specificities of the bHLH and bHLH-ZIP proteins. For most bHLH-ZIP proteins (including Max and *myc*), the entire bHLH-ZIP region is needed for efficient dimerization and DNA binding (Kato et al., 1992; Smith et al., 1990), and these proteins usually fail to dimerize with bHLH proteins. To understand why the bHLH-ZIP proteins need a leucine zipper for effective dimerization and DNA binding, we looked for differences in the helix-loop-helix dimer interfaces of Max and MyoD. We first calculated the surface areas buried upon dimer formation and found that these are essentially equal (the buried surface area is 2% greater for the MyoD bHLH domain than for the Max bHLH domain). Comparing the dimer interfaces in more detail shows interesting differences that might help us understand differences in dimer stability, but mutagenic analysis may be needed to rigorously determine the energetic significance of individual contacts.

Although the four-helix bundles are very similar, there are significant structural differences in the loop regions of Max and MyoD. More generally, comparing sequences of bHLH proteins shows that the loop regions vary in composition and length (Figure 1) and swap experiments show that these loops are not always functionally interchangeable among bHLH proteins (Pesce and Benezra, 1993). Finally, in comparing the helix-loop-helix regions of Max and MyoD, we noted a curious cavity between the four helix bundle and the DNA. This cavity, which is just above the central

base pairs in the MyoD and Max structures, is large enough to accommodate a small molecule about the size of a nucleotide, but we do not yet know whether the cavity has any special structural or functional significance.

On the whole, a comparison of Max and MyoD complexes shows that the bHLH domain is a very well conserved motif, and the leucine zipper provides a simple way of extending the dimer interface. In fact, superimposing the Max and MyoD structures indicates that the C-termini of the H2 helices of MyoD could readily be extended with a leucine zipper. As proposed in the original report on the Max complex (Ferré-D'Amaré et al., 1993), the overall conservation of the bHLH fold (and the corresponding dimer interface) reflects the inherent stability of this structure and allows for regulatory networks that can be controlled via a set of competing homodimer and heterodimer interactions. The Max and MyoD structures may provide a reasonable basis for modeling other bHLH domains and for modeling heterodimer formation. (Since this fold is so well conserved in different bHLH proteins, it seems plausible to assume that it will be conserved as the heterodimers form). In this regard, we note that Ellenberger et al. (personal communication; Ellenberger et al., 1994) have recently solved the structure of the E47 homodimer. Combining our structures should provide a plausible basis for modeling the MyoD-E47 heterodimer, which may be the biologically active species.

Myogenic Transcription

Previous studies of MyoD have shown that the bHLH domain is necessary and sufficient for DNA binding and muscle cell conversion. In addition, a region of about 50 residues at the N-terminus can function as a transcriptional activation domain (Weintraub et al., 1991). However, there also are several positions in the basic region where mutations can alter transcriptional activation without affecting DNA binding. Ala 114 is one of the critical positions where mutations can give such a "positive control" phenotype (Bengal et al., 1994). "Swap experiments" which mix and match segments from MyoD and E12 also have been used to determine which residues are required for transcriptional activation. These experiments highlight the roles of Ala 114, Thr 115 and Lys 124, which are conserved in all of the myogenic proteins. Remarkably, introducing just these three residues into comparable positions in E12 allows that protein to function as a myogenic activator (Davis and Weintraub, 1992).

Examining the MyoD-DNA complex shows that Lys 124 is exposed on the surface and thus could readily participate in protein-protein interactions (with the activation domain or with the transcriptional apparatus). However, Ala 114 and Thr 115 are buried in the major groove (Figure 5a), and this raises an interesting question: How can these two residues participate in transcriptional activation when they are not exposed at the

surface? Could they have subtle, indirect effects on the overall structure of the complex?

Although other scenarios are possible, comparing the Max and MyoD structures suggests a plausible hypothesis about how residues 114 and 115 might exert critical - but indirect - effects that could be "relayed" to the surface of the complex. In particular, we note that Arg 111 has very different conformations in these two complexes, and it appears that residues 114 and 115 help to determine the conformation of the arginine (Figure 5b). In the MyoD complex, Arg 111 is buried in the major groove and contacts the N7 of base G10. (Thr 115 makes a base contact and a backbone contact; the small size of Ala 114 helps provide room for the arginine to fit in the groove.) In the Max complex, Arg 111 has a very different conformation: It swings out of the major groove and makes a phosphate contact. (His 114 of Max contacts the N7 of base G10 and appears to displace the arginine side chain).

Sequence comparisons are consistent with the general notion that Arg 111 may need to be buried for myogenic transcriptional activation: Most bHLH proteins - with the notable exception of those involved in myogenesis - have a His or Asn residue at position 114. As in Max, this residue could contact G10 and force Arg 111 out of the major groove (in a way that might interfere with activation of muscle-specific genes). We are currently undertaking mutagenesis experiments to test for interactions among the residues at positions 111, 114 and 115. In particular, we are looking for

second-site revertants to the positive control mutants, and we predict that at least some of these revertants will be found at position 111. Preliminary results are consistent with this model: Thus, our model predicts that substituting a small amino acid, such as glycine, at position 114 would still accommodate Arg 111 in the major groove and give wild type MyoD activity. This has recently been confirmed (J. Huang, P.C.M., M.A.R., C.O.P., H.W., unpublished results). Finally, we note that the structure of the E47-DNA homodimer complex (which does not activate myogenesis) is consistent with our hypothesis. In the recently solved structure of the complex, Arg 111 has a conformation similar to that seen in Max (Ellenberger et al., personal communication; Ellenberger et al., 1994).

Conclusions

This crystal structure of the MyoD-DNA complex gives important new insights about the bHLH proteins in general and about the myogenic proteins in particular. Specifically, we note that: 1) the structure of the bHLH domain of MyoD and of the corresponding region of Max (a bHLH-ZIP protein) are remarkably similar. Thus, it appears that the bHLH structure is very well conserved in spite of the modest sequence identities between the Max and MyoD domains (25% identity) and that the presence of the leucine zipper does not significantly affect the structure of the bHLH domain. 2) Our crystal structure helps us understand subtle

differences in the site specificities of the bHLH proteins. 3) The structure shows that Ala 114 and Thr 115, two of the key residues implicated in transcriptional activation, are buried in the major groove and thus, unlikely to be involved in any direct protein-protein interactions. These residues may be required because of indirect effects on the conformation of Arg 111 (introducing other residues at positions 114 and 115 may force Arg 111 out of the major groove and thereby interfere with transcriptional activation).

Experimental Procedures

The peptide used in our studies contains residues 102-166 of
the mouse MyoD protein (which is identical to human MyoD in this
region) and three extraneous residues (MEL) that were added at the
N-terminus during cloning (Figure 1a). We also replaced Cys 135
with Ser to simplify crystallization attempts: earlier experiments
had shown that the Cys residue can readily form intermolecular
disulfide bonds which knock out DNA binding activity (Starovasnik
et al., 1992; Ma and Pabo, unpublished data). [NMR studies also have
shown that this oxidized peptide (Starovasnik et al., 1992) forms an
antiparallel four-helix bundle, but the biological relevance of this
form is unclear.] We find that replacing the Cys residue by Ser did
not noticeably affect the biological properties of the peptide. The
peptide was expressed in E. coli and purified to homogeneity by
reverse phase chromatography using Vydac C_4 and C_{18} columns.
The DNA used in the crystallization attempts was synthesized by
standard phosphoramidite chemistry and purified by reverse-phase
HPLC and ion exchange chromatography. A series of different DNA
duplexes was tested during our crystallization attempts (Jordan et
al. , 1985) using the vapor diffusion method. The best crystals used
a 14 bp DNA duplex (Figure 1b) and grew at 22 °C from drops
containing 10%-15% PEG 3500, 100 mM Tris-HCl, pH 8.5, 100mM
NaCitrate, 20 mM $BaCl_2$. Crystals grew over the course of a few
days with a final size of approximately 0.3 mm x 0.3 mm x 0.4 mm.
The crystals belong to the space group $P2_12_12$ and have unit cell
dimensions of a=222.8 Å, b=70.8 Å, c=30.0 Å with two protein

dimer-DNA complexes in the asymmetric unit (55% solvent

content); each complex has a non-crystallographic two-fold axes.

Isomorphous derivatives of the complex were obtained by using DNA

duplexes with 5-iodouracil substituted for thymine: 1) at positions

11 and 11' and 2) at positions 12 and 12' (Figure 1b). Two crystals

were used for each derivative, but each crystal was treated

separately during heavy atom parameter refinement and phasing.

Diffraction data were collected at 22 °C using a Rigaku

RAXIS-IIC imaging plate area detector. Diffraction from the

crystals was anisotropic, with the best diffraction (about 2.6 Å)

along the DNA axis. Data were processed with DENZO/SCALEPACK

(Z. Otwinowski, personal communication), and DSCALEAD (M. Rould)

was used for local scaling of the derivative data sets to the native

data. Heavy atom positions were determined from a combination of

isomorphous and anomalous difference Patterson maps; the program

HASSP (Terwilliger et al., 1987) was used to help determine the

initial positions. The positions and occupancies of the heavy atoms

were refined using the program REFINE (S.E.R.C., 1979) on centric

reflections for each of the individual derivatives, and then by

cross-phased refinement with PHARE (S.E.R.C., 1979) in which each

derivative is refined using parental phases determined solely from

the native and other derivatives (Table 1). An initial MIR map at

3.0 Å revealed clear density for the DNA, the alpha-helices and part

of the loop region. One round of solvent flattening (Wang, B.C.,

1985) significantly reduced the noise level and improved the

continuity of the electron density in the loop region. The heavy

atom parameters were then refined again using the solvent-flattened phases as the parent phases (Rould et al., 1992), and another MIR map was calculated. Finally, this map was averaged about the two non-crystallographic symmetry axes to give a solvent-flattened, four-fold non-crystallographic symmetry-averaged MIR map that allowed us to see most of the side chains (Figure 4b).

Idealized B-form DNA (generated by INSIGHT) was placed in the clear MIR density and each nucleotide adjusted to improve the fit. The alpha-helical regions were modeled as polyalanine helices and the loop region was fit with an extended polypeptide chain. Using this initial framework, a model with side chains was built into the MIR map using the program FRODO (Jones, 1978) and this model was refined with XPLOR (Brunger et. al, 1987; Brunger, 1992a) using positional refinement and simulated annealing. Before any refinement, 10% of the reflections were removed for cross-validation (free R-factor). We monitored the free R-factor during all stages of rebuilding and refinement (Brunger, 1992b) and changes in the free R factor were used to evaluate the effects of symmetry restraints and to make other basic decisions about the refinement strategy. Strict noncrystallographic symmetry was applied during the first rounds of refinement at 3.0 Å and 2.9 Å; tightly restrained noncrystallographic symmetry was used in the later stages. In order to correct for anisotropic diffraction, data were locally scaled against calculated structure factors using the program DSCALEAD. The conformations of the side chains and bases

were checked using simulated annealing omit maps (Brunger, 1992a), as well as positionally refined omit maps. During the last rounds of refinement, data up to 2.6 Å were included, although the limited completeness and anisotropy of diffraction of the higher resolution shells effectively limits the resolution of our structure to about 2.8 Å. The residues N-terminal to 105 are disordered in all but one monomer (where it happens to be ordered because of crystal packing contacts). During the last stages of model rebuilding, 25 water molecules were added and accepted because they met the following strict criteria: 1) they were found at the interface of the basic region and the DNA, and hence were of potential biological interest; 2) they appeared with strong density in F_o-F_c maps in similar positions in at least two of the four monomers; 3) they participated in at least one hydrogen bonded interaction (most participated in at least two) and 4) they refined to an isotropic B-factor of no more than 60 $Å^2$ with unit occupancy. Tightly restrained individual B-factors were used in the final stages of refinement. In addition to refinement with XPLOR, we also used positional refinement with the program TNT (Tronrud et al., 1987). We believe that TNT has a more accurate stereochemical dictionary for DNA and that XPLOR has a more accurate stereochemical dictionary for proteins (Engh and Huber, 1991). Hence, our final model of the MyoD-DNA complex is a chimera which uses the DNA from TNT and the protein from XPLOR. The crystallographic R factor of the final model is 25.3% using all of the data and 22.4% using data with I > 2σ. The corresponding free R values are 33.0% and 30.4% respectively. The rms deviation in bond lengths and bond

angles for the protein are 0.008 Å and 1.289 degrees and 0.013 Å and 3.270 degrees for the DNA.

Acknowledgments

References

Bengal, E., Flores, O., Rangarajan, P.N., Chen, A., Weintraub, H., and Verma, I. (1994). Positive control mutations in MyoD basic region fail to show cooperative DNA binding and transcriptional activation *in vitro*. Proc. Natl. Acad. Sci. USA. In press.

Blackwell, T.K., and Weintraub, H. (1990). Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. Science *250*, 1104-1110.

Blackwell, T.K., Huang, J., Ma, A., Kretzner, L., Alt, F.W., Eisenman, R.N., and Weintraub, H. (1993). Binding of myc proteins to canonical and noncanonical DNA sequences. Mol. Cell Biol. *13*, 5216-5224.

Brunger, A.T., Kuriyan, J., and Karplus, M. (1987). Crystallographic R-factor refinement by molecular dynamics. Science *235*, 458-460.

Brunger, A.T. (1992a). X-PLOR v3.1 Manual (New Haven, Connecticut: Yale University Press).

Brunger, A.T. (1992b). The free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature *355*, 472-474.

Davis, R.L., Cheng, P.F., Lassar, A.B., and Weintraub, H. (1990). The MyoD DNA binding domain contains a recognition code for muscle-specific gene activation. Cell *60*, 733-746.

Davis, R.L., and Weintraub, H. (1992). Acquisition of myogenic specificity by replacement of three amino acid residues from MyoD into E12. Science *256*, 1014-1018.

Ellenberger, T., Fass, D., Arnaud, M., and Harrison, S.C. (1994). Crystal structure of transcription factor E47: E-box recognition by a basic region helix-loop-helix dimer. Genes Dev. In press.

Engh, R.A., and Huber, R. (1991). Accurate bond and angle parameters for X-ray protein-structure refinement. Acta Cryst. *A47*, 392-400.

Ferré-D'Amaré, A.R., Prendergast, G.C., Ziff, E.B., and Burley, S.K. (1993). Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain. Nature *363*, 38-45.

Ferre-D'Amare, A.R., Pognonec, P., Roeder, R.G., and Burley, S.K. (1994). Structure and function of the b/HLH/Z domain of USF. EMBO J. *13*, 180-189.

Jones, T.A. (1978). A graphics model building and refinement system for macromolecules. J. Appl. Cryst. *11*, 268-272.

Jordan, S.R., Whitcombe, T.V., Berg, J.M., and Pabo, C.O. (1985).
Systematic variation in DNA length yields highly ordered repressor-
operator cocrystals. Science *230*, 1383-1385.

Kato, G.J., Lee, W.M.F., Chen, L. and Dang, C.V. (1992). Max: functional
domains and interactions with myc. Genes Dev. *6*, 81-92.

Otwinowski, Z., Schevitz, R.W., Zhang, R.G., Lawson, C.L., Joachimiak,
A., Marmostein, R.Q., Luisi, B.F. and Sigler, P.B. (1988). Crystal
structure of *trp* repressor/operator complex at atomic resolution.
Nature *335*, 321-329.

Pesce, S., and Benezra, R. (1993). The loop region of the helix-loop-
helix protein Id1 is critical for its dominant negative activity. Mol.
Cell. Biol. *13*, 7874-7880.

Rould, M.A., Perona, J.J., and Steitz, T.A. (1992). Improving multiple
isomorphous replacement phasing by heavy-atom refinement using
solvent-flattened phases. Acta Cryst. *A48*, 751-756.

Rudnicki, M.A., Schnegelsberg, P.N.J., Stead, R.H., Braun, T., Arnold,
H.H., and Jaenisch, R. (1993). MyoD or Myf-5 is required for the
formation of skeletal muscle. Cell *75*, 1351-1359.

S.E.R.C. [U.K.] (1979). Collaborative Computing Project No. 4, a suite
of programs for protein crystallography (Distributed from the
Daresbury Laboratory, Warrington, U.K.).

Smith, M.J., Charron-Prochownik, D.C., and Prowchownik, E.V. (1990). The leucine zipper of c-myc is required for full inhibition of erythroleukemia differentiation. Mol. Cell Biol. *10*, 5333-5339.

Starovasnik M.A., Blackwell, T.K., Laue, T.M., Weintraub, H., and Klevit, R.E. (1992). Folding topology of the disulfide-bonded dimeric DNA-binding domain of the myogenic determination factor MyoD. Biochemistry *31*, 9891-9903.

Sun, X.H., and Baltimore, D. (1991). An inhibitory domain of E12 transcription factor prevents DNA binding in E12 homodimers but not in E12 heterodimers. Cell *64*, 459-470.

Terwilliger, T.C., Kim, S.C., and Eisenberg, D. (1987). Generalized method of determining heavy-atom positions using the difference patterson function. Acta Cryst. *A43*, 1-5.

Tronrud, D.E., TenEyck, L.F., and Matthews, B.W. (1987). An efficient general-purpose least-squares refinement program for macromolecular structures. Acta Cryst. *A43*, 489-501.

Van Antwerp, M.E., Chen, D.G., Chang, C., and Prochownik, E.V. (1992). A point mutation in the MyoD basic domain imparts c-myc-like properties. Proc. Natl. Acad. Sci. USA *89*, 9010-9014.

Voronova, A., and Baltimore, D. (1990). Mutations that disrupt DNA binding and dimer formation in the E47 helix-loop-helix protein map to distinct domains. Proc. Natl. Acad. Sci. USA *87*, 4722-4726.

Wang, B.C. (1985). Resolution of phase ambiguity in macromolecular crystallography. Meth. Enzymol. *115*, 90-112.

Weintraub, H., Dwarki, V.J., Verma, I., Davis, R., Hollenberg, S., Snider, L., Lassar, A., and Tapscott, S.J. (1991). Muscle-specific transcriptional activation by MyoD. Genes Dev. *5*, 1377-1386.

Weintraub, H. (1993). The MyoD family and myogenesis: redundancy, networks, and thresholds. Cell *75*, 1241-1244.

Winter, B., Braun, T., and Arnold, H.H. (1992). Co-operativity of functional domains in the muscle-specific transcription factor Myf-5. EMBO J. *11*, 1843-1855.

Wright, W.E., Binder, M., and Funk, W. (1991). Cyclic amplification and selection of targets (CASTing) for the myogenin consensus binding site. Mol. Cell Biol. *11*, 4104-4110.

Table Legends

Table 1: Data collection, phasing, and refinement statistics.

These statistics group together data from a number of computer programs, used at different stages of the structure determination. The data collection and reduction statistics are from DENZO and SCALEPACK (Z. Otwinowski). The phasing statistics are from REFINE and PHARE from the CCP4 package (S.E.R.C., 1979). The refinement statistics are from XPLOR (Brunger, 1987, 1992a) and TNT (Tronrud et al., 1987).

Figure Legends

Figure 1: Sequences of MyoD bHLH domain and DNA fragment used in cocrystallization

Figure 1a

Sequence alignment of MyoD with other members of the bHLH subfamily, along with sequence of the bHLH-ZIP protein Max. Amino acids are given in one-letter code and conserved residues are denoted by stippled boxes. The MyoD numbering is used throughout the alignment and the text.

Figure 1b

Sequence of the symmetric DNA duplex used for crystallization, showing the numbering scheme used in this paper. The thymine bases which were replaced by 5-iodo-uracil in making isomorphous derivatives are highlighted.

Figure 1c

Summary of selection data from optimizing the DNA binding sites for MyoD and myc (Blackwell and Weintraub, 1990; Blackwell et al., 1993). The residues CAxxTG, which were not randomized in these experiments, are highlighted.

Figure 2: Overview of the MyoD-DNA complex

Figure 2a

In this figure, the 14 bp DNA site is shown in light blue, with the nucleotides CA--TG highlighted in violet. The basic region is yellow, as are the side chains which make base-specific and phosphate contacts to the DNA. The helix-loop-helix region of MyoD is shown in red.

Figure was generated using Insight II software (BIOSYM Technologies, Inc.).

Figure 2b

Sketch of the MyoD-DNA complex in the same orientation as in Figure 2a. The alpha-helices are represented as cylinders and a few residue numbers are given for reference. The sugar-phosphate backbones of the DNA are represented with ribbons.

Figure 3: Summary of base and phosphate contacts

Sketch summarizing the base and phosphate contacts made by one monomer of MyoD. The contacts made by the other monomer in the dimer are not shown, but they involve identical contacts at symmetry-related positions on the opposite strand of the DNA (the symmetry axis is shown as a black oval between bases G7 and G7').

The DNA is represented as a cylindrical projection with phosphates indicated by circles. Critical CA--TG base pairs are indicated with heavier lines; bases and phosphates contacted by one monomer are shaded. As noted in the text, there are no 5' terminal phosphates on our DNA, and Arg 110 makes a contact to the O5' of T1.

Figure 4: Interactions of the MyoD basic region with DNA

Figure 4a

Stereo diagram showing the key contacts made by one of the basic helices with its binding site. For reference, we have also shown the alpha-carbon trace of the other basic helix. The view is similar to that in Figure 2, but the complex has been rotated slightly in order to better show the side chain-base contacts. Base-specific contacts are made by Arg 111, Thr 115, and Glu 118. Phosphate contacts are made by Arg 110, Arg 111, Thr 115, Arg 117, Arg 119, and Arg 121. Water-mediated contacts are made by Arg 111 and Glu 118. Some of the critical hydrogen bond contacts are represented as dashed lines (111R-G10', 118E-C5, 118E-A6, 121R-C5, 121R-118E).

Figure 4b

Solvent-flattened MIR electron density map at 3.0 Å, showing the interface between the basic region and the DNA (looking down the

axis of the DNA). The sidechain residues of Arg 111, Glu 118, and Arg 121 are highlighted in magenta, while the peptide backbone and side chains for the other residues in the region 111 to 121 are shown, as are the DNA nucleotides C5 and G10'.

Figure 5: Conformation of residues involved in positive control

Figure 5a

Figure highlights critical residues implicated in positive control. For clarity, only one basic helix and one half of the DNA site (i.e. 7 base pairs) are shown, and the view is looking down the axis of the basic helix. Residues 108 to 124 are shown in light blue and the side chains of the positive control residues - Ala 114, Thr 115, and Lys 124 - are shown in red. The DNA is shown in yellow, and the sugar-phosphate backbones are highlighted with ribbons.

Figure 5b

Superposition of the alpha carbon trace of the basic regions from MyoD and Max showing the side chains of residues Arg 111, Ala 114, and Thr 115 from MyoD and the corresponding residues from Max: Arg 25, His 28 and Asn 29. We also show the C5 and G10' nucleotide pair from MyoD and the corresponding nucleotides from Max. The view is looking down the axis of the DNA, as in Figure 4b.

The residues and nucleotides from MyoD are shown in red; those from Max are shown in light blue.

Figures were generated using Insight II software (BIOSYM Technologies, Inc.).

Statistics for Data and Derivatives

	Native1	Native2	Derivative 12+12' First Crystal	Derivative 12+12' Second Crystal	Derivative 11+11' First Crystal	Derivative 11+11' Second Crystal
Resolution (A)	2.9	2.6	2.8	2.9	3.0	2.8
Measured reflections	37,898	30,065	20,456	16,094	19,412	31,193
Unique reflections	11,032	12,156	7871	7105	6262	11,210
Rsym [a]	7.9%	6.5%	5.7%	7.0%	8.5%	7.0%
Data completeness	97.8%	78.0%	63.8%	62.7%	61.7%	89.4%
Mean isomorphous difference [b]			17.8%	23.4%	26.8%	26.2%

Phasing statistics: 20 to 3.0 A

Cullis R factor [c]			.52	.52	.51	.52
Phasing power [d]			1.69	2.12	2.05	2.55

Mean figure of merit = 0.65 [e]

Refinement: 20 to 2.8 A

	All data	$|I| > 2\sigma_I$
Rfactor (%)	25.3%	22.4%
Free Rfactor	33.0%	30.4%
Reflections	10585	8963
Total number of atoms	3984	
R.m.s. bond length (A)	0.008 (protein)[f]	0.013 (DNA)[g]
R.m.s. bond angle (degrees)	1.3 (protein)[f]	3.3 (DNA)[g]
R.m.s difference in B between bonded atoms	2.38 A^2	

[a] $\Sigma_h\Sigma_i |I_{h,i} - I_h| / \Sigma_h\Sigma_i I_{h,i}$ where $I_{h,i}$ is mean intensity of the i observations of reflection h

[b] $\Sigma |I_{PH} - I_P| / \Sigma <I>$ where I_{PH} and I_P are the derivative and native intensities, and $<I>$ is $[I_{PH} + I_P] / 2$

[c] $\Sigma ||F_{der} - F_{nat}| - F_{H(calc)}| / \Sigma |F_{der} - F_{nat}|$ for centric reflections, where $F_{H(calc)}$ is the calculated heavy structure factor

[d] rms F_H/E, where E = residual lack of closure error

[e] After solvent flattening, Mean figure of merit=0.75

[f] Engh and Huber (1991) dictionary for protein: parhcsdx.pro in XPLOR

[g] TNT dictionary for DNA

Table 1

1a

activation ⟍⟍ basic helix1 loop helix2 ⟍⟍

1   53   108   125   137   146   166   319

* = residues conserved among myogenic proteins
\# = residues conserved among Class A proteins

MyoD

```
          110      120        130           140                    150        160
           |        |          |             |                      |          |
          * *  #    |    #  *  #     |                   |      |               # #
MyoD         MELKRKTTNADRRKAATMRERRRLSKVNEAFETLKRCTS----SNP-----------NQRLPKVEILRNAIRYIEGLQALLRD
myogenin     VDRRRAATLREKRRLKKVNEAFEALKRSTL----LNP-----------NQRLPKVEILRSAIQYIERLQALLS
Myf5         MDRRKAATMRERRRLKKVNQAFETLKRCTT----TNP-----------NQRLPKVEILRNAIRYIESLQELLR
E12          KERRVANNARERLRVRDINEAFKELGRMCQ----LHLN---------SEKPQTKLLILHQAVSVILNLEQQVR
E47          RERRMANNARERVRVRDINEAFRELGRMCQ----MHLK---------SDKAQTKLLILQQAVQVILGLEQQVR
daughterless KERRQANNARERIRIRDINEALKELGRMCM----THLK---------SDKPQTKLGILNMAVEVIMTLEQQVR
twist        SNQRVMANVRERQRTQSLNDAFKSLQQIIP----TLP----------SDKLSKIQTLKLATRYIDFLCRMLS
AS-C sc (T4) VQRR---NARERNRVKQVNNSFARLRQHIPQSIITDLTKG--G-GRGPHKKISKVDTLRIAVEYIRSLQDLVD

consensus      RR     RER RΘ  ΘN Φ  LK                      K  K  ΘL AΘ YΘ Θ  ΘΘ
                \ /      \ /
             variable  variable      Θ = L,I,V,M        Φ = F,L,I,Y
             region 1  region 2
```

Max   ADKRAHHNALERKRRDHIKDSFHSLRDSVP---------SLQG------KRASRAQILDKATEYIQYMRRKND

1b

```
 1  2  3  4  5  6  7  8  9 10 11 12 13 14
```

TCAACAGCTGTTGA
AGTTGTCGACAACT

14'13'12'11'10' 9' 8' 7' 6' 5' 4' 3' 2' 1'

1c

A A   C   T
  CAG TGT
G G   G   C

MyoD specificity
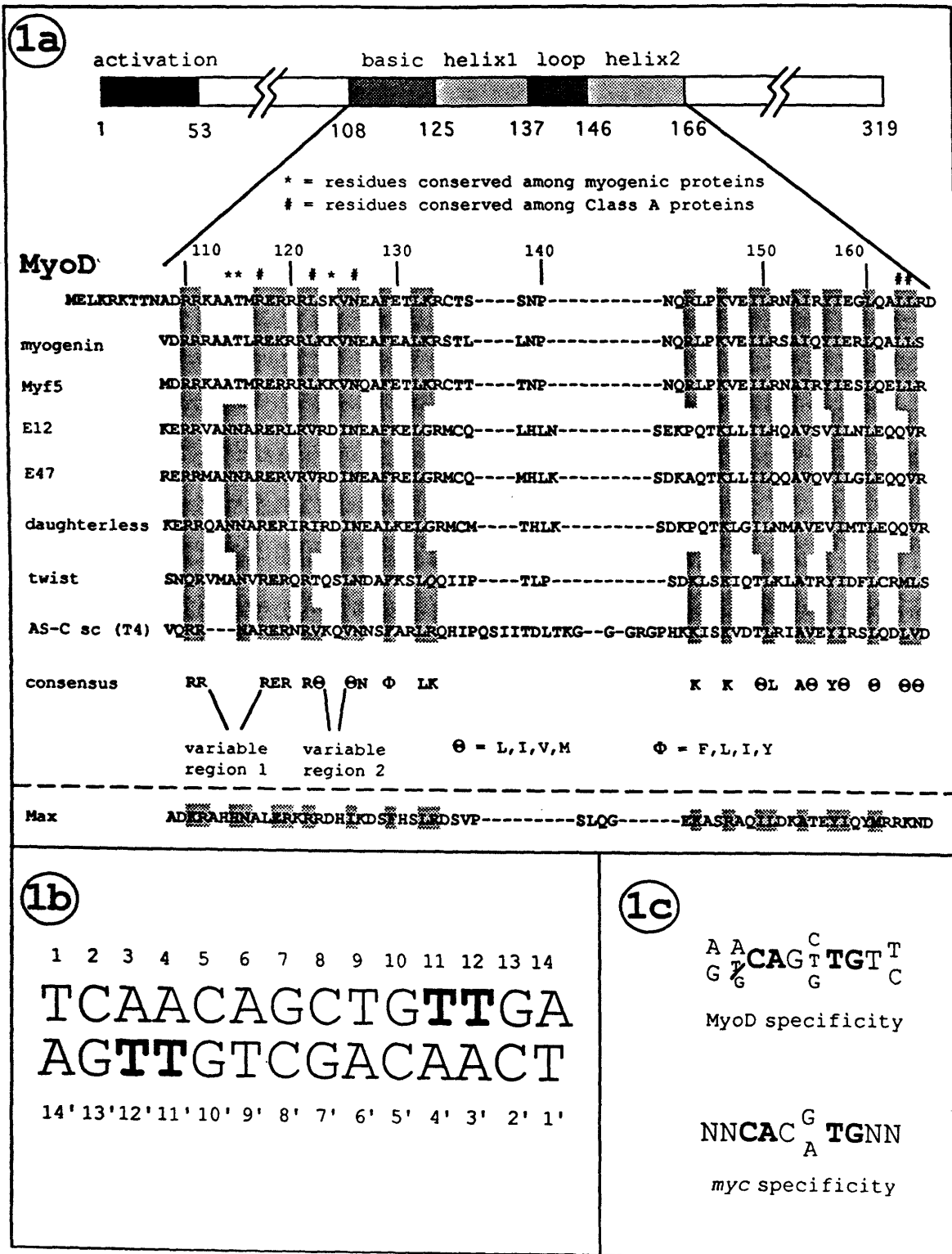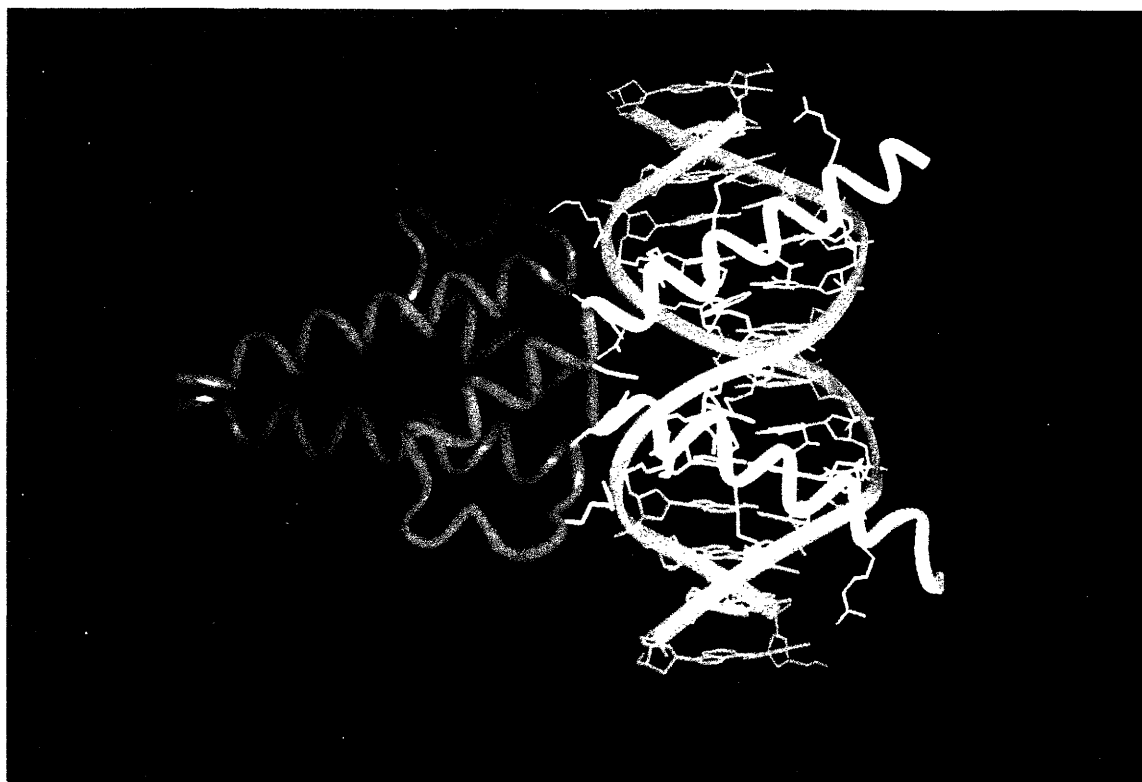
                    G
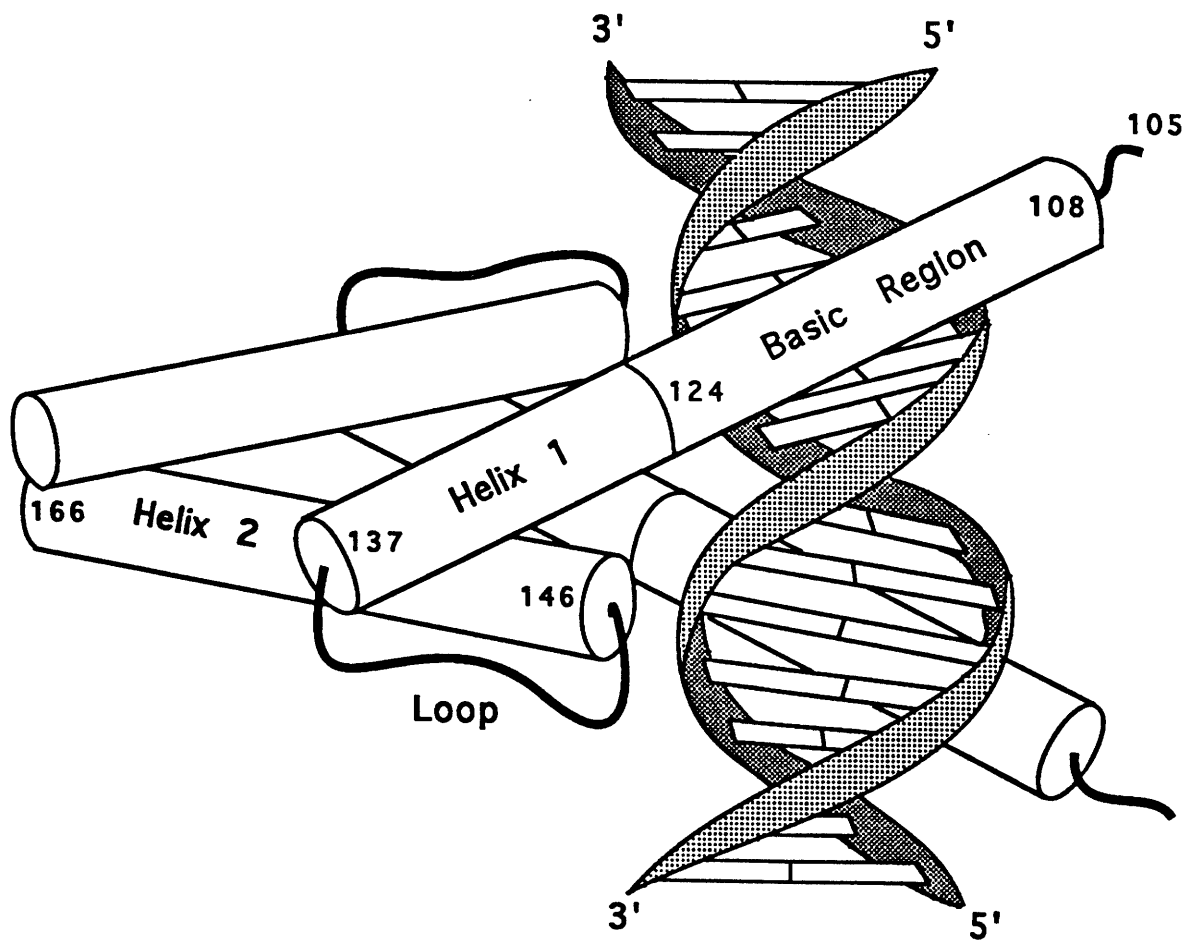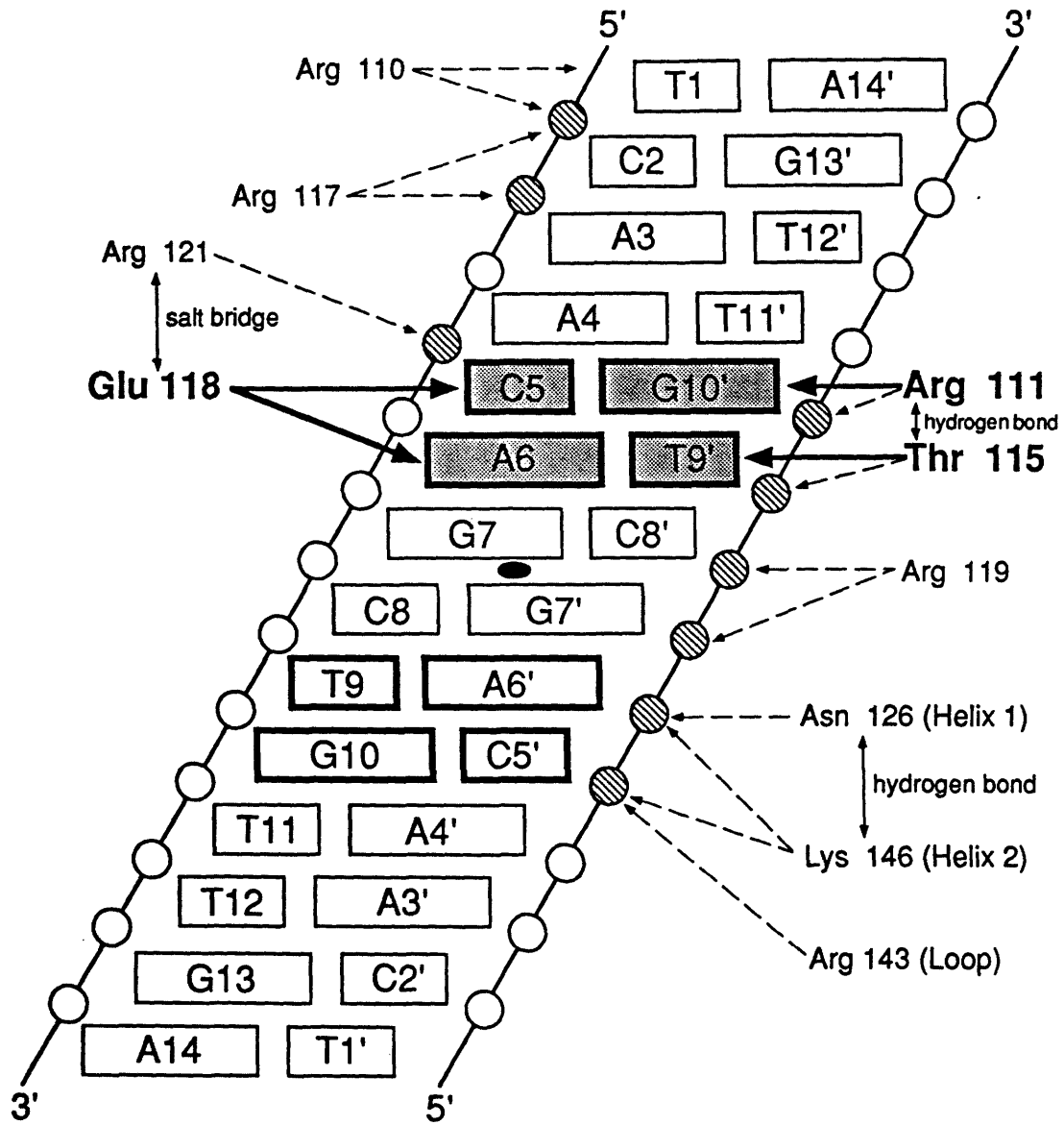NNCAC   TGNN
                    A
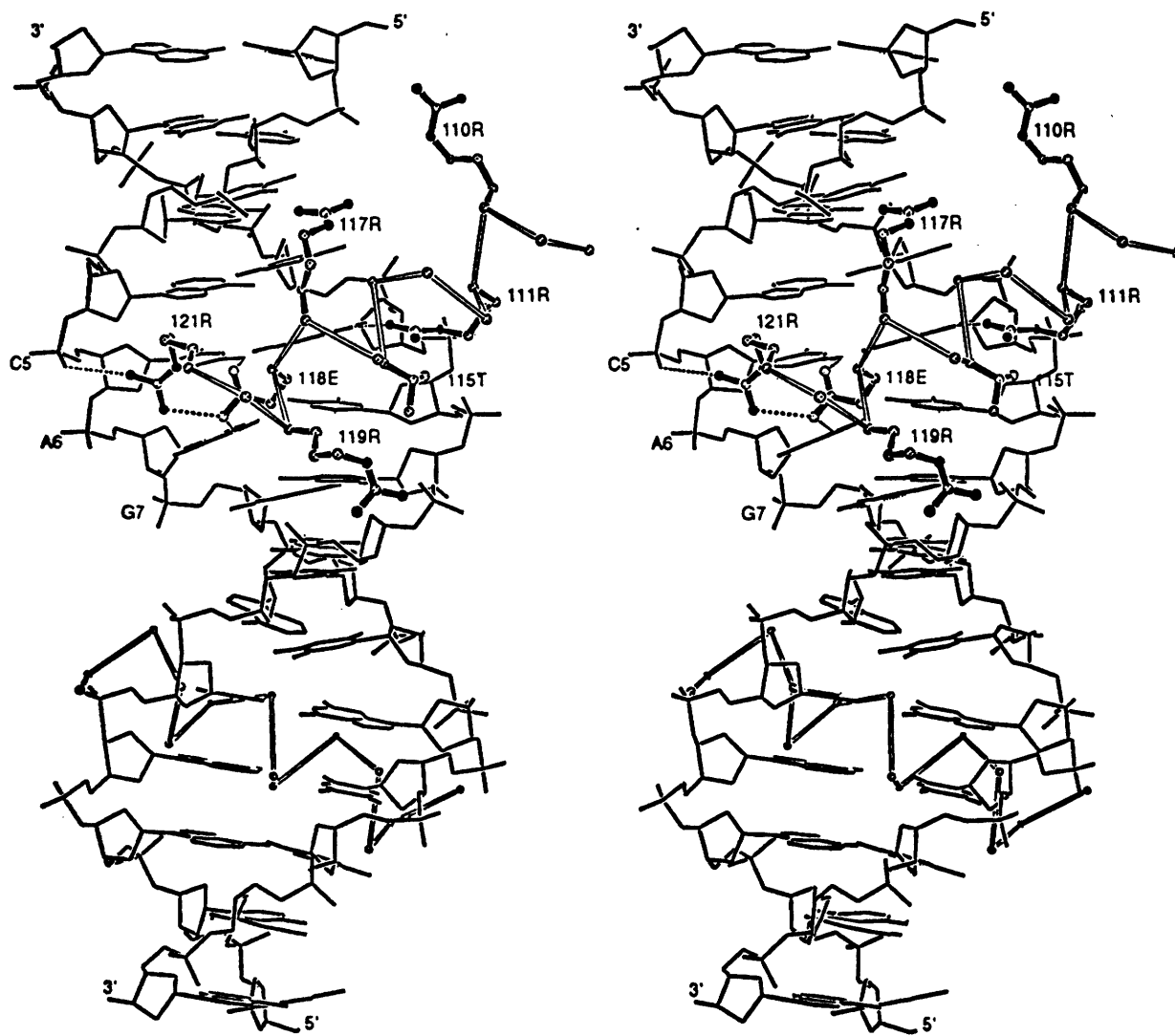
myc specificity

Figure 1
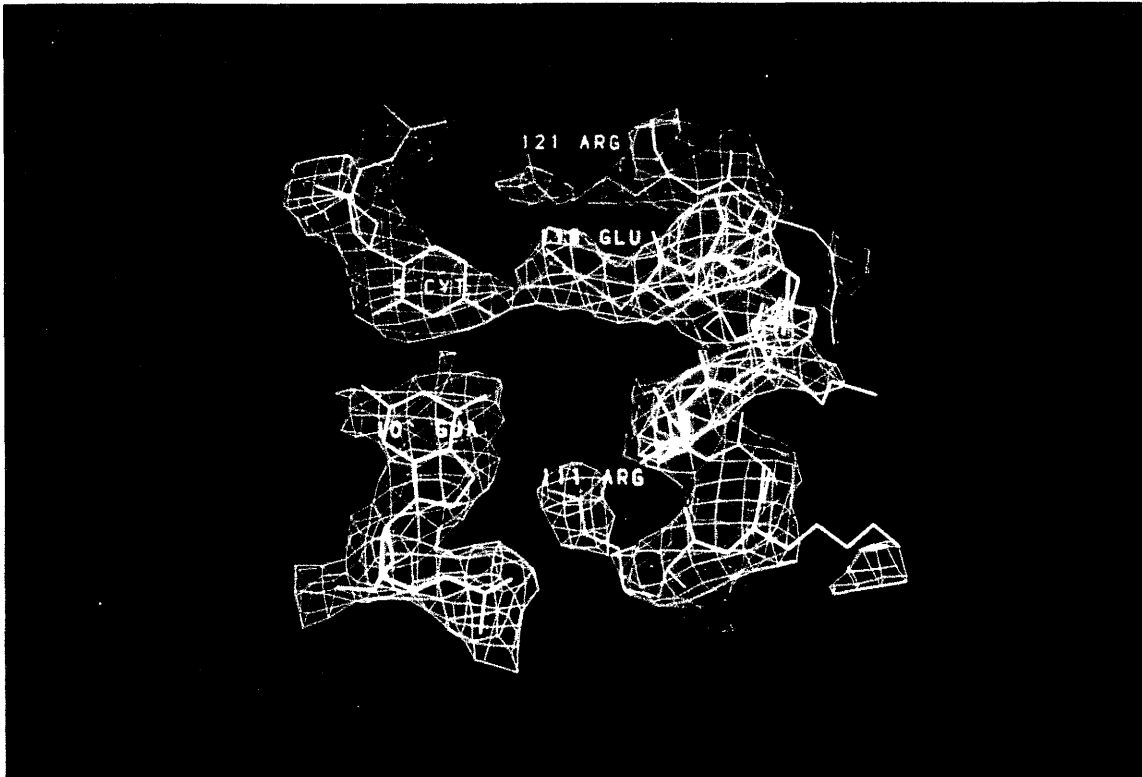
Figure 2a

Figure 2b

Figure 3
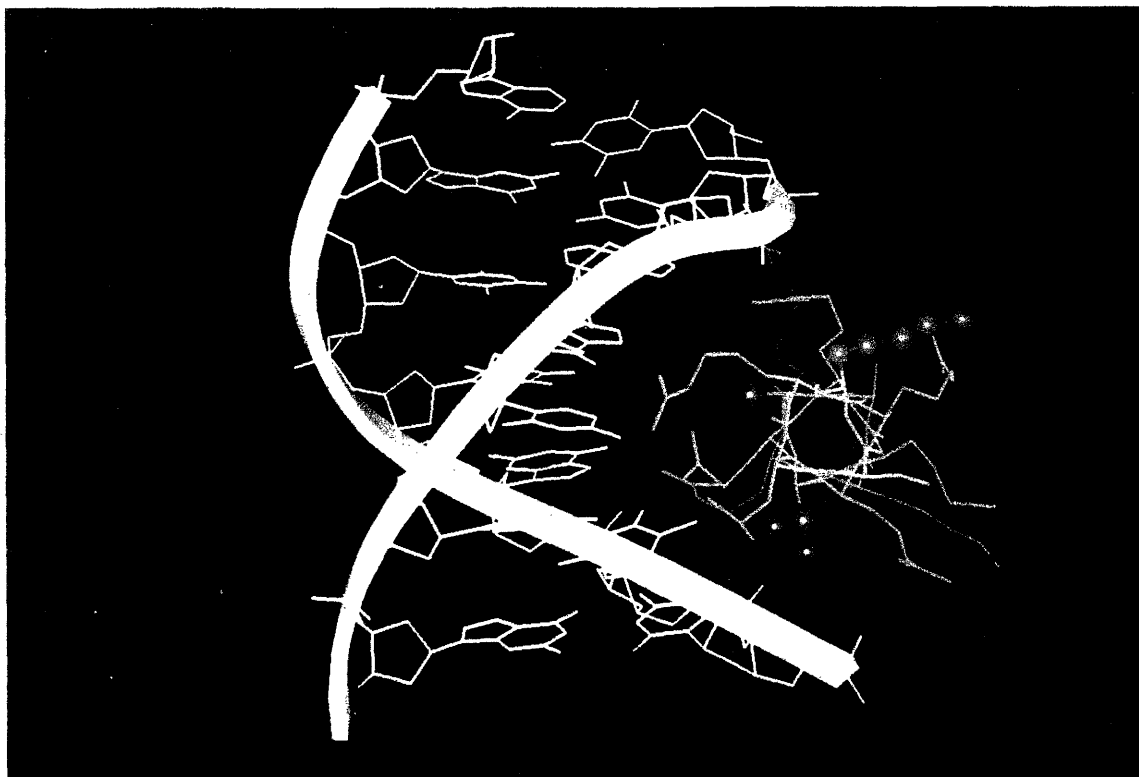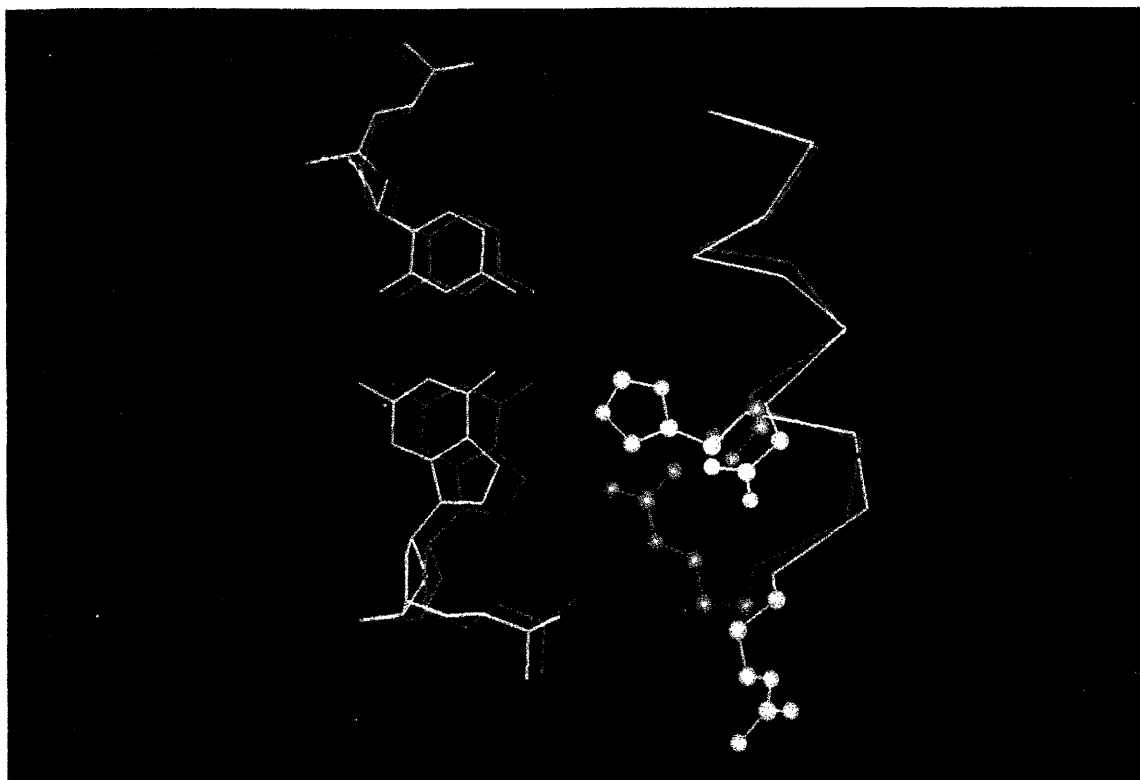
Figure 4a

Figure 4b

Figure 5a

Figure 5b

CHAPTER 5

Structural Basis for Dimer Stabilization
and Dimer Specificity for MyoD
and other bHLH Proteins

Like the bZIP proteins, the bHLH and bHLH-ZIP proteins bind
DNA as dimers, with each monomer binding to half of the DNA site
(Ellenberger, T., 1994). Since the bZIP, bHLH, and bHLH-ZIP proteins
can form homodimers as well as heterodimers, and since these have
different affinities and specificities for the DNA, dimerization
plays an important role in determining the DNA binding activity of
these proteins. With the bHLH and bHLH-ZIP proteins, we are
fortunate in having four structures of homodimers bound to DNA: 1)
the Max bHLH-ZIP domain (Ferré-D'Amaré et al., 1993); 2) the USF
bHLH domain in which the leucine zipper which is normally present
in the protein has been deleted for crystallization purposes (Ferré-
D'Amaré et al., 1994); 3) the E47 bHLH domain (Ellenberger et al.,
1994); and 4) the MyoD bHLH domain. In this chapter, the structural
basis of dimerization for the bHLH proteins is addressed. We first
describe the dimer interface for the MyoD bHLH domain bound to
DNA, concentrating on aspects of the structure that were not
discussed in detail in Chapter 4. Next, we examine the structure of
a model of the MyoD-E47 heterodimer in which the MyoD homodimer
coordinates and the E47 homodimer coordinates (kindly provided by
T. Ellenberger and S.C. Harrison) were combined to generate a model
of the heterodimer. In this section, we address the question of why
the MyoD-E47 heterodimer may be a more stable structure than the
MyoD homodimer. In the last section of this chapter, we will look
at other members of the bHLH family of proteins and use what we
have learned from the Max, USF, E47 and MyoD structures to suggest
plausible determinants for dimerization specificity within this
group of proteins.

A. The MyoD bHLH Homodimer

We have grouped the contacts at the dimer interface of the MyoD homodimer into four major regions: first, a hydrophobic core in the interior of the four helix bundle; second, region A at the interface between helix 1 of one monomer and the apposing helix 2 of the dimeric partner; third, region B where the C-terminus of helix 1, the N-terminus of the loop, and helix 2 residues from both dimeric partners come together; and fourth, region C, near the part of the hydrophobic core closest to the DNA, where helix 1 first crosses helix 2 of the same monomer. We describe each one of these four regions in turn.

1. The Core

In MyoD, this region consists of nine residues from each monomer (Figures 1a and 2a). Most of these residues are highly conserved over all bHLH and bHLH-ZIP proteins, and almost all of these residues are hydrophobic in character. Isoleucine and/or leucine residues are particularly well-represented and well-conserved, both within the HLH region itself, as well as in the zipper region of bHLH-ZIP (Beckmann and Kadesch, 1991; Dang et al., 1991). As mentioned earlier, our discussion will focus on the bHLH region (see (Ellenberger, 1994) for a discussion of zipper regions in bZIP and bHLH-ZIP proteins). The parallel, left-handed, four-helix bundle, formed by helix 1 and helix 2 of each monomer, was first described in the report on the Max structure (Ferré-D'Amaré et al.,

1993). The structure of the Max-DNA complex showed that all of the conserved hydrophobic residues from the bHLH region are found within the core of the four-helix bundle. The residues within the hydrophobic core of MyoD form structural counterparts to those residues in Max. Val 125 and Phe 129 from helix 1 and Ile 149 and Leu 150 from helix 2 are in the region of the core closest to the DNA, near a cavity located directly above the central base pairs of the DNA (see Chapter 4 for more on the cavity). Leu 132 of helix 1, together with Ala 153 and Ile 154 from helix 2, form the middle of the core. The carboxy terminus of the core is composed entirely of residues from helix 2, Ile 157 and Leu 160, which pack in an arrangement similar to the side-by-side packing of coiled-coils (Ellenberger, 1994). Outside of the core, there are two more hydrophobic residues in helix 2, Leu 163 and Leu 164, which also pack in a side-by-side arrangement. These residues contribute to the stability of the C-terminus of the four-helix bundle in the MyoD structure (Figure 3), but since these residues are not well-conserved within the bHLH and bHLH-ZIP proteins, we have not included them with the other residues in the hydrophobic core.

The functional importance of the residues in the core region has been shown in a number of mutagenesis studies with the myogenic and E-proteins (which include E12 and E47). In myogenin, the Val 125 residue was changed to an Ala, which resulted in a reduction in heterodimerization with E12 or E47 (Winter et al., 1992). The Phe residue at position 129 is highly conserved throughout all bHLH and bHLH-ZIP proteins, and where it is not

conserved, it is replaced with the hydrophobic amino acids Ile, Leu,
or Tyr. Mutagenesis studies of E47 and myogenin have demonstrated
that this Phe residue, along with the residues corresponding to Leu
132, Ile 149, Leu 150, Ile 157, and Leu 160, are absolutely essential
for dimerization (Voronova and Baltimore, 1990; Winter et al.,
1992).

Region A

The region sits at an interface between helix 1 and the
apposing helix 2' (Figure 2b) and includes some residues which are
not within the hydrophobic core (Figure 1b). There are a number of
charged and polar residues within region A, or close to it, and it has
been proposed that these residues may form salt bridge or hydrogen
bond interactions which may stabilize or destabilize the dimer
interface (Shirakata et al., 1993). In the structure of the MyoD,
homodimer bound to DNA, we do not see any strong charge-charge or
hydrogen bond interactions in region A. The residues which are
closest to the interface between helix 1 and helix 2' and which we
have grouped together as belonging to region A are shown in Figures
1b and 4a. Val 125 in helix 1 is a hydrophobic core residue that
packs against Val 147' of helix 2', and thus, helps to stabilize the
region A interface. The guanidinium group of Arg 151', from helix
2', sits above Val 125 and Val 147', and may help to seclude the
Val-Val interaction from external solvent (the conformation of Arg
151' may be partially stabilized by an intra-helical salt bridge
formed between Arg 151' and Glu 148'). Ala 128, Thr 131, and Ser

135 (which was mutated from the original cysteine because of concerns about oxidation; see Chapters 2 and 4) from helix 1, pack respectively against residues Leu 150', Ile 154', and Ile 157' from the hydrophobic core.

Region B

Many parts of the four-helix bundle structure come together in this region (Figures 1c and 2c). Analysis of the MyoD homodimer structure shows that a key role is played by the tyrosine residue at position 156 in helix 2. This residue is involved in two important interactions (Figure 4b). First, it sits above the carboxy terminus of helix 1, over Thr 136, preventing further extension of helix 1 at the carboxy terminus, and also stabilizing the amino terminus of the loop. Second, the OH of Tyr 156 makes a hydrogen bond with the OE2 of Gln 161' from helix 2', helping to stabilize the dimer interface in region B. The Tyr residue at position 156 is well-conserved over the bHLH and bHLH-ZIP proteins (Figure 1c); the E-proteins are unique in having a Val residue instead of Tyr residue at the position equivalent to 156 in MyoD (the structural significance of the Val residue will be discussed in more detail later). Mutagenesis results support the importance of Tyr 156 in stabilizing the MyoD homodimer structure. A myogenin mutant where Tyr was substituted by Val, had a reduced ability to oligomerize and bind DNA (Winter et al., 1992).

The other residues in region B help to stabilize the loop and the carboxy terminus of helix 1. Lys 133 projects into the interior of the loop and makes a contact with the carbonyl of residue 142 from the loop; the conformation of Lys 133 is also stabilized by a intrahelical salt bridge between Lys 133 and Glu 130 of helix 1. Thr 136 and Ser 137 at the C-terminus of helix 1 pack against helix 2. Thr 136 packs against Ala 153 from the hydrophobic core, while Ser 137 points toward Asn 152. The Asn 152 residue is within range to make a hydrogen bond with Ser 137 of helix 1, but does not do so in the structure, possibly because it is involved in an intrahelical salt bridge with Arg 155.

Region C

This region comprises three well-conserved residues (Figure 1d), located near the portion of the hydrophobic core closest to the DNA, where helix 1 first crosses helix 2 (Figure 2d). The crossing angle between helices 1 and 2 in MyoD may be partly determined by the interactions between the residues in this region. Asn 126 is at the amino terminus of helix 1, near the junction with the basic region, while Lys 146, the first residue in helix 2, is at the junction with the loop region. The NZ of Lys 146 contacts the phosphate of A6, while the main chain amide contacts the phosphate at C5 (Figure 4c). The A6 phosphate is also contacted by ND2 of Asn 126. Besides the phosphate contact, the OD1 of Asn 126 makes a hydrogen bond with the NZ of Lys 146. Thus, the two residues Asn 126 and Lys 146 help 'anchor' the four-helix bundle to the DNA,

while also stabilizing the conformation of the bundle itself. Phe 129 from helix 1 helps to keep Lys 146 in place as it sits over the carbon atoms of the Lys side chain. Phe 129 contains an aromatic ring with an extensive surface for Van der Waals interactions, both within the core, and with the Lys 146 residue. The structural importance of the Lys residue at position 146 in MyoD is supported by mutagenesis experiments with E47 and myogenin. In E47, a Lys to Ala substitution knocks out oligomerization and DNA binding (Voronova and Baltimore, 1990), while in myogenin, a Lys to Met substitution reduces oligomerization and DNA binding (Winter et al., 1992). The Lys residue at position 146 is highly conserved over all bHLH and bHLH-ZIP proteins (where a Lys residue is not present, there is an Arg, as in Max). Moreover, the X-ray structures of the Max (Ferré-D'Amaré et al., 1993), USF (Ferré-D'Amaré et al., 1994) and E47 (Ellenberger et al., 1994) complexes all show the Lys (or Arg in Max) residue in a similar conformation to that seen in MyoD.

B. Heterodimer formation

Several lines of evidence suggest that a heterodimer formed between MyoD and a member of the E2 proteins (E12 and E47) is the biologically active species. First, it appears that *in vitro*, MyoD forms heterodimers more efficiently than it forms homodimers (Sun and Baltimore, 1991). Second, examination of whole cell extracts with antibodies specific for MyoD and the E2 proteins reveals that MyoD can associate with E12/E47 *in vivo* (Lassar et al., 1991). Third expression of antisense E2 RNA lowers the level of gene

expression induced by MyoD (Lassar et al., 1991). In addition to its interactions with E47 and E12, MyoD can also form heterodimers with the Id protein. Since Id does not have a functional basic region, formation of the MyoD-Id heterodimer may be a means by which the DNA binding activity of MyoD is negatively regulated in *trans* (Benezra et al., 1990). The formation of heterodimers between MyoD and other bHLH proteins thus appears to be an important mechanism where the activity of MyoD may be regulated. In this regard, we note that selection experiments which optimize the DNA binding site for MyoD and E47 homodimers, and for MyoD-E47 heterodimers have shown that the three forms of dimer show subtle difference in DNA binding site preference (Blackwell and Weintraub, 1990).

We do not yet have a structure of MyoD as part of a heterodimer, but we are fortunate in having the two homodimer structures of the MyoD and E47 bHLH domains bound to DNA. A superposition of the alpha carbon atoms from the two homodimer structures shows that outside of the loop region, the two structures are very similar (1.0 Å rms difference between the alpha carbon atoms). The high degree of similarity between the bHLH domains of MyoD and E47 suggested to us that a combination of the coordinates from the MyoD and E47 homodimer structures would be a good starting model for the structure of the MyoD-E47 heterodimer. With the assistance of Tom Ellenberger, who kindly provided us with the E47 homodimer coordinates, we have constructed such a heterodimer model by a simple substitution of

the coordinates from one E47 bHLH monomer for the coordinates of a MyoD bHLH monomer, after the E47 and MyoD homodimer coordinates were superimposed. The resulting MyoD-E47 heterodimer model shows the familiar four-helix bundle structure (Figure 5). We note that there are no obvious steric clashes between residues in this MyoD/E47 heterodimer model. Thus, we believe that the model is a reasonable approximation for a functional heterodimer.

In the heterodimer model, the E47 half of the heterodimer has an extended helix 1, which is one turn longer than helix 1 of MyoD (Ellenberger et al., 1994). Thus, the substitution of an E47 monomer for a MyoD monomer in the heterodimer model may create additional buried surface area for the dimer interface compared to that of the MyoD homodimer. Aside from this increase in the surface area of the dimer interface due to the extended helix 1 of the E47 monomer, there may be other contributions to heterodimer stability which are significantly more subtle. The core regions of the MyoD homodimer, the E47 homodimer, and our model of the MyoD-E47 heterodimer are very similar, with only two out of nine residues in the core being different between MyoD and E47 (Figure 1a). It is in region A that the differences between the MyoD homodimer, the E47 homodimer, and the MyoD-E47 heterodimer model become more pronounced (Figure 1b). In our model of the heterodimer, there are two distinct region A interfaces: one which has helix 1 from MyoD and helix 2' from E47, and another which has the reverse arrangement of helices. At the region A interface with helix 1 from MyoD and helix

2' from E47, there are several possible interactions between side chains which could affect the stability of the heterodimer (Figure 6a): 1) In the model, Val 125 from MyoD helix 1 packs against Leu 147' from E47 helix 2' (for ease of comparison, the number scheme follows that used to describe the MyoD homodimer). This Val-Leu interaction has a greater potential for Van der Waals interactions than the Val-Val interaction seen in the MyoD homodimer does, and could be an important factor in heterodimer stabilization (for comparison, the E47 homodimer has a Ile-Leu interaction at this part of the region A interface). 2) The Lys 124 from MyoD helix 1 has the potential to form a hydrogen bond with Gln 151' from E47 helix 2'. [We note that the formation of hydrogen bonds and/or salt bridges between different subunits has been hypothesized to play an important role in determining dimer stability and specificity in bZIP proteins (Schuermann et al., 1991; O'Shea et al., 1992; Hu et al., 1993; Vinson et al., 1993). Initial mutagenesis studies with the bHLH proteins MyoD and E47 also suggest that salt bridges and hydrogen bonds are important for the dimerization properties of bHLH proteins (Shirakata et al., 1993).] We note that neither the MyoD homodimer nor the E47 homodimer have residues in positions 124 and 151' which could form as good a hydrogen bond pair as Lys 124 and Gln 151'. The MyoD homodimer has a Lys and an Arg at these positions, respectively, which do not interact in the structure, possibly because of electrostatic repulsion between the two residues. The E47 homodimer has an Asp and a Gln at the corresponding residues, which are too far apart to form a good hydrogen bond pair (Ellenberger et al., 1994). 3) At the carboxy

terminus of this region A interface, the residues Ser 135 from MyoD helix 1 and Leu 158' from E47 helix 2' could interact. It is difficult to predict *a priori* the energetics of such a potential interaction, especially in comparison to the energetics of the interactions between Ser 135 and Glu 158' from the MyoD homodimer structure, and between Met 135 and Leu 158' from the E47 homodimer structure.

At the other region A interface formed between helix 1 of E47 and helix 2' of MyoD, there are also a number of interesting interactions which are possible (Figure 6b): 1) In the model, Ile 125 from E47 helix 1 packs against Val 147' from MyoD helix 2', and as with the Val-Leu pair at the other region A interface of the heterodimer, the Ile-Val pair here would likely have a more favorable energy of interaction than the Val-Val pair found in the MyoD homodimer structure. 2) Asp 124 from E47 helix 1 could form a salt bridge with Arg 151' from MyoD helix 2'. 3) In lieu of a potential salt bridge with Asp 124, the Arg 151' residue could form a salt bridge with Glu 131 of E47 helix 1. We cannot predict *a priori* whether the Asp 124 - Arg 151' interaction would contribute more to the heterodimer stability than the Glu 131 - Arg 151' interaction would. We do note, though, that neither the MyoD homodimer nor the E47 homodimer structures have residues which could form a stable salt bridge or hydrogen bond between residues at positions 131 and 151'. 4) At the carboxy terminus of this region A interface, the Met 135 from the E47 helix 1 and the Glu 158' from the MyoD helix 2' could interact. As with Ser 135 and Leu

158' from the other region A interface, however, we cannot easily predict the energetics of the potential Met 135 - Glu 158' interaction.

Outside of region A, the potential interactions at the dimer interface of the heterodimer model do not differ so much from the interactions seen at the dimer interfaces of the MyoD and E47 homodimer structures. In region C, for example, the interactions between Asn 126 from helix 1, Phe 129 from helix 1, and Lys 146 from helix 2 are nearly identical in the MyoD and E47 homodimer structures, and as a result, are predicted to be very similar in the MyoD-E47 heterodimer model. For region B, the individual dimer interfaces for each half of the heterodimer look similar to the corresponding homodimer interfaces. Thus, in the heterodimer model, the region B interface formed by the MyoD helix 1, loop, and helix 2 and the E47 helix 2' (Figure 6c) looks similar to the same region in the MyoD homodimer (Figure 4b). At the other region B interface for the heterodimer, formed by the E47 helix 1, loop, and helix 2 and the MyoD helix 2', the configuration of the amino acids (Figure 6d) is nearly identical to that seen in the E47 homodimer structure (Figure 5 in Ellenberger et al., 1994). Thus, the carboxy terminus of helix 1 is extended by one more turn than is seen in the MyoD and Max homodimer structures because replacing the Tyr residue at position 158 with a smaller Val residue removes the steric constraint on the extension of helix 1 (compare Figure 6d with Figure 4b).

In summary, the MyoD-E47 heterodimer model predicts that the major determinants for the greater stability of the heterodimer vis-à-vis the MyoD homodimer are to be found in two regions of the dimer interface. First, in region B, the extended helix 1 in the E47 monomer provides additional buried surface area at the dimer interface without losing any potential hydrogen bonds: the Tyr 156 - Gln 161' hydrogen bond interaction seen in the MyoD homodimer structure (Figure 4b) is replaced by a His 139 - Gln 161' hydrogen bond interaction in the MyoD-E47 heterodimer model (Figure 6d). Second, in region A, the potential hydrogen bonds and/or salt bridges which can form in the MyoD-E47 heterodimer, but not in the MyoD or E47 homodimers, will be expected to stabilize the heterodimer interface. A number of recent mutagenesis studies involving the Id, MyoD, E12 and E47 proteins provide some support for our model of dimer stabilization. The importance of residues in region B for dimer stability is supported by recent studies on the Id protein (Pesce and Benezra, 1993). The Id protein is unique among bHLH proteins in that it does not form homodimers with itself, but does form heterodimers with MyoD and E47 (Benezra et al., 1990; Jen et al., 1992; Sun et al., 1991). Replacing four residues in the helix 1/loop junction in region B of E12 or E47 (QMHL) with the corresponding residues from Id (PTLP) conferred the Id dimerization specificity on the E-proteins, that is, the mutant E-proteins were unable to form homodimers, but could form heterodimers with MyoD (Pesce and Benezra, 1993). As discussed in the report on the E47 homodimer structure, the Id residues introduced in region B could significantly alter the structure of this region and thus, interfere

with E47 homodimer formation (Ellenberger et al., 1994). As for region A, a recent mutagenesis study on MyoD showed that substituting the region A residues Thr 131 and Cys 135 from helix 1 with Ile and Arg, respectively, could significantly decrease the ability of the mutant to form heterodimers with E12 (Shirakata et al., 1993). More mutagenesis studies need to be done in order to fully address the role played by residues in regions A and B, but the data, thus far, supports the importance of these residues in influencing homodimer and heterodimer stability.

C. Dimerization in the bHLH family of proteins

For MyoD and E47, differences in regions A and B of the dimer interface may contribute to dimer specificity. In region C, the residues are identical in MyoD and E47, but there are other bHLH proteins which have significant differences in this region. These differences in sequence, especially at position 129 where a Phe residue is usually found, might have significant effects on the oligomerization properties of the bHLH proteins. The structure of the USF bHLH domain bound to DNA (in which the normally present zipper was deleted for crystallization purposes) has recently been solved (Ferré-D'Amaré et al., 1994). The USF bHLH domain shows a fold that is similar, but not identical, to the parallel four-helix bundle structure seen in Max, E47, and MyoD. In particular, the USF bHLH fold shows a significantly different crossing angle between helices 1 and 2 at the part of the hydrophobic core closest to the DNA. This difference in crossing angle causes the USF bHLH

structure to have helices 2 and 2' splayed away from the dimer axis, in comparison to Max, E47 and MyoD. The USF protein normally has a leucine zipper extension, as does Max, and indeed, the bHLH domain by itself (in contrast to the full bHLH-ZIP domain) dimerizes poorly and binds DNA poorly (Ferré-D'Amaré et al., 1994). Although, it may thus be difficult to draw any firm conclusions about dimerization based on the USF bHLH structure, it is nonetheless interesting to note that the difference in helix crossing angle between USF and the other three bHLH(/Z) structures may be linked to a difference in structure in region C. Instead of a Phe residue at position 129, which is found in Max, MyoD, and E47, USF has a Ile residue (Figure 1d). In the MyoD structure, the Phe 129 residue packs tightly against Ile 149 and Leu 150, as well as against residues from the other dimeric partner, Phe 129', Ile 149' and Leu 150' (Figure 7). Substituting a Ile residue for the Phe might be expected to significantly affect the amount of surface area available for Van der Waals interactions within the hydrophobic core, if the orientation of the helices within the core were unchanged. If the helices adopted different crossing angles, however, the residue change from Phe to Ile might be better accommodated. A recent mutagenesis study, which tested the importance of the sequence identity of residue 129 in MyoD, finds that substituting Phe with Ala, Val, Ile, Leu, or Trp significantly reduces the protein's dimerization potential (Shirakata et al., 1993). One mutation, a Tyr substitution, retained more of the original dimerization potential, presumably because Tyr is quite close in structure to Phe.

Aside from the different hydrophobic residues which can be found at position 129, the differences in residue identity within the hydrophobic core are quite subtle, and a comparison of the structures of Max, USF, E47, and MyoD does not suggest any obvious way in which this residue variability may affect the structure of the four-helix bundle. More structures of bHLH and bHLH-ZIP proteins with different sequences in the core region will need to be solved before we can more fully address this question. In this regard, we note that it is still not clear how important the sequence identity in hydrophobic core regions is for determining the structure of proteins in general. Earlier studies have shown that the pattern of hydrophobic and polar residues, rather than the actual sequence identity for each residue, is the more important factor in determining protein structure (Lim and Sauer, 1991; Dao-pin et al., 1991; Bowie et al., 1992). More recently, a detailed study on GCN4 zipper mutants has shown that the sequence identity of the residues at the **a** and **d** positions of a coiled coil (which can be thought of as constituting the hydrophobic core for this type of structure), has a key role to play in determining whether the alpha-helices of the GCN4 mutants form two-, three-, or four-helical bundles (Harbury et al., 1993). Given that all of the bHLH and bHLH-ZIP structures solved thus far show a parallel, four-helix bundle conformation, it will be interesting to see if mutations in the core region will significantly affect the packing arrangements of the helices.

We note that all of our analysis has been based upon the structures of bHLH and bHLH-ZIP proteins bound to DNA as

homodimers. In the absence of DNA, various studies have shown that bHLH-ZIP and bHLH proteins can interact as tetramers as well as dimers (Dang et al., 1989; Hu et al., 1990; Fisher et al., 1991; Anthony-Cahill et al., 1992; Starovasnik et al., 1992; Farmer et al., 1992). Recent studies with the USF protein also suggest that under certain conditions, the protein may associate with DNA as a homotetramer (Ferré-D'Amaré et al., 1994). For the myogenic and E-proteins, at least, the evidence suggests that the proteins bind to DNA as dimers, even though they may associate in different oligomeric forms in the absence of DNA (Farmer et al., 1992; Lassar et al., 1991; Fairman et al., 1993). Thus, for the purpose of understanding the structural basis for the bHLH proteins' involvement in gene regulation, we feel justified in focusing our discussion on dimerization. As we discuss in the next section, a detailed understanding of dimerization preferences and the role of dimerization in regulating biological activity will require the solution of heterodimer structures, as well as careful study of dimer stabilities and DNA binding constants. We also note that our structural analysis has focused on the bHLH regions of these proteins. Other regions of the proteins may influence the dimerization states. Further mutational and deletional analysis of the bHLH(-ZIP) proteins may help to address this possibility.

References

Anthony-Cahill, S.J., Benfield, P.A., Fairman, R., Wasserman, Z.R., Brenner, S.L., Stafford, W.F., III, Altenbach, C., Hubbell, W.L., and DeGrado, W.F. (1992). Molecular characterization of helix-loop-helix peptides. Science *255*, 979-983.

Beckmann, H., and Kadesch, T. (1991). The leucine zipper of TFE3 dictates helix-loop-helix dimerization specificity. Genes Dev. *5*, 1057-1066.

Benezra, R., Davis, R.L., Lockshon, D., Turner, D.L., and Weintraub, H. (1990). The protein Id: a negative regulator of helix-loop-helix DNA binding proteins. Cell *61*, 49-59.

Blackwell, T.K., and Weintraub, H. (1990). Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. Science *250*, 1104-1110.

Bowie, J.U., Luthy, R., and Eisenberg, D. (1992). A method to identify protein sequences that fold into a known three-dimensional structure. Science *253*, 164-170.

Dang, C.V., Barrett, J., Villa-Garcia, M., Resar, L.M.S., Kato, G.J., and Fearon, E.R. (1991). Intracellular leucine zipper interactions suggest c-myc hetero-oligomerization. Mol. Cell. Biol. *11*, 954-962.

Dang, C.V., McGuire, M., Buckmire, M., and Lee, W.M.F. (1989). Involvement of the 'leucine zipper' region in the oligomerization and transforming activity of human c-myc protein. Nature *337*, 664-666.

Dao-pin, S., Anderson, D.E., Baase, W.A., Dahlquist, F.W., and Matthews, B.W. (1991). Structural and thermodynamic consequences of burying a charged residue within the hydrophobic core of T4 lysozyme. Biochemistry *30*, 11521-11529.

Ellenberger, T. (1994). Getting a grip on DNA recognition: Structures of the bZIP and the bHLH DNA-binding domains. Curr. Opin. Struc. Biol. In press.

Ellenberger, T., Fass, D., Arnaud, M., and Harrison, S.C. (1994). Crystal structure of transcription factor E47: E-box recognition by a basic region helix-loop-helix dimer. Genes Dev. In press.

Fairman, R., Beran-Steed, R.K., Anthony-Cahill, S.J., Lear, J.D., Stafford, W.F., DeGrado, W.F., Benfield, P.A., and Brenner, S.L. (1993). Multiple oligomeric states regulate the DNA binding of helix-loop-helix peptides. Proc. Natl. Acad. Sci. USA *90*, 10429-10433.

Farmer, K., Catala, F., and Wright, W.E. (1992). Alternative multimeric structures affect myogenin DNA binding activity. J. Biol. Chem. *267*, 5631-5636.

Ferré-D'Amaré, A.R., Pognonec, P., Roeder, R.G., and Burley, S.K. (1994). Structure and function ot the b/HLH/Z domain of USF. EMBO J. *13*, 180-189.

Ferré-D'Amaré, A.R., Prendergast, G.C., Ziff, E.B., and Burley, S.K. (1993). Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain. Nature *363*, 38-45.

Fisher, D.E., Carr, C.S., Parent, L.A., and Sharp, P.A. (1991). TFEB has DNA-binding and oligomerization properties of a unique helix-loop-helix/leucine zipper family. Genes Dev. *5*, 2342-2352.

Harbury, P.B., Zhang, T., Kim, P.S., and Alber, T. (1993). A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. Science *262*, 1401-1407.

Hu, J.C., Newell, N.E., Tidor, B., and Sauer, R.T. (1993). Probing the role of residues at the e and g positions of the GCN4 leucine zipper by combinatorial mutagenesis. Protein Sci. *2*, 1072-1084.

Hu, Y.-F., Lüscher, B., Admon, A., Mermod, N., and Tjian, R. (1990). Transcription factor AP-4 contains multiple dimerization domains that regulate dimer specificity. Genes Dev. 4, 1741-1752.

Jen, Y.H., Weintraub, H., and Benezra, R. (1992). Overexpression of Id protein inhibits the muscle differentiation program: In vivo association of Id with E2A protein. Genes Dev. *6*, 1466-1479.

Jones, T.A., Zou, J.-Y., Cowan, S.W., and Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Cryst. *A47*, 110-119.

Lassar, A.B., Davis, R.L., Wright, W.E., Kadesch, T., Murre, C., Voronova, A., and Baltimore, D. (1991). Functional activity of myogenic HLH proteins requires hetero-oligomerization with E12/E47-like proteins in vivo. Cell *66*, 305-315.

Lim, W.A., and Sauer, R.T. (1991). The role of internal packing in the structure and stability of a protein. J. Mol. Biol. *219*, 359-376.

O'Shea, E.K., Rutkowski, R., and Kim, P.S. (1992). Mechanism of specificity in the fos-jun oncoprotein heterodimer. Cell *68*, 699-708.

Pesce, S., and Benezra, R. (1993). The loop region of the helix-loop-helix protein Id1 is critical for its dominant negative activity. Mol. Cell. Biol. *13*, 7874-7880.

Schuermann, M., Hunter, J.B., Hennig, G., and Mueller, R. (1991). Non-leucine residues in the leucine repeats of fos and jun contribute to the stability and determine the specificity of dimerization. Nuc. Acids Res. *19*, 739-746.

Shirakata, M., Friedman, F.K., Wei, Q., and Paterson, B.M. (1993). Dimerization specificity of myogenic helix-loop-helix DNA-binding factors directed by nonconserved hydrophilic residues. Genes Dev. 7, 2456-2470.

Starovasnik, M.A., Blackwell, T.K., Laue, T.M., Weintraub, H., and Klevit, R.E. (1992). Folding topology of the disulfide-bonded dimeric DNA-binding domain of the myogenic determination factor MyoD. Biochemistry 31, 9891-9903.

Sun, X.-H., Copeland, N.G., Jenkins, N.A., and Baltimore, D. (1991). Id proteins Id1 and Id2 selectively inhibit DNA binding by one class of helix-loop-helix proteins. Mol. Cell. Biol. 11, 5603-5611.

Sun, X.H., and Baltimore, D. (1991). An inhibitory domain of E12 transcription factor prevents DNA binding in E12 homodimers but not in E12 heterodimers. Cell 64, 459-470.

Vinson, C.R., Hai, T., and Boyd, S.M. (1993). Dimerization specificity of the leucine zipper-containing bZIP motif on DNA binding: Prediction and rational design. Genes Dev. 7, 1047-1058.

Voronova, A., and Baltimore, D. (1990). Mutations that disrupt DNA binding and dimer formation in the E47 helix-loop-helix protein map to distinct domains. Proc. Natl. Acad. Sci. USA 87, 4722-4726.

Winter, B., Braun, T., and Arnold, H.H. (1992). Co-operativity of functional domains in the muscle-specific transcription factor Myf-5. EMBO J. *11*, 1843-1855.

Figure Legends

Figure 1: Amino acid identities of residues at the dimer interface for bHLH and bHLH-ZIP proteins.

Sequence alignment of the helix-loop-helix region of members of the bHLH (above dashed line) and bHLH-ZIP families (below dashed line). Key residue positions were identified by looking for conserved residues within the primary sequences and/or by using the structure of the MyoD/DNA complex as a template. Figure 1a highlights the residues within the hydrophobic core. The hydrophobic core residues are shown in stippled boxes. Figure 1b also shows the hydrophobic core residues within stippled boxes. In addition, it shows residues at the dimer interface corresponding to region A (between helix 1 and helix 2'). These residues from region A are underlined and highlighted in bold. Figure 1c shows the hydrophobic core residues and residues from the dimer interface corresponding to region B (where the carboxy terminus of helix 1, the loop, and parts of helices 2 and 2' meet). These residues are underlined and highlighted in bold. For E47, the residues from the homodimer structure (Ellenberger et al., 1994) which were identified as being important for dimerization are italicized, underlined, and highlighted in bold. Figure 1d shows the hydrophobic core residues and the residues from region C of the dimer interface (where helix 1 first crosses helix 2). As in Figures 1b and 1c, the key residues are highlighted in bold and underlined.

Figure 2: Overview of the MyoD helix-loop-helix dimerization motif, highlighting residues at the dimer interface.

We show the alpha carbon trace of the helix-loop-helix regions of MyoD in blue. Key residues which are important for dimerization stability and/or specificity are shown in red. Figure 2a shows the position of residues within the hydrophobic core in red. Figure 2b shows the positions and side chain residues from region A in red. Figure 2c shows the positions and side chain residues from region B in red. Figure 2d shows the positions and side chain residues from region C in red.

Figure 3: A tetrad of leucines at the carboxy termini of a MyoD dimer help to stabilize the dimer interface.

The alpha carbon traces of helices 2 and 2' are shown in red. The side chains of leucine residues 163 and 164, and the symmetry related residues (Leu 163' and Leu 164') are represented by ball-and-stick models in blue.

Figure 4: Key residues at the dimer interface of the MyoD homodimer.

Ball-and-stick models of important residues at each of the three key regions of the dimer interface. Figure 4a shows residues from region A. We have labeled important residues that are involved or

may be involved in dimerization stability and/or specification for the MyoD homodimer: from helix 1, these are Lys 124, Val 125, Ala 128, Thr 131, and Ser 135 (we had replaced the normally present cysteine at this position with serine because of concerns about oxidation during the cocrystallization experiments); from helix 2' (from the symmetry-related monomer), these residues are Val 147', Arg 151', and Glu 158'. Figure 4b shows residues from region B. We show the following key residues: from the carboxy terminal end of helix 1, Lys 133, Thr 136, and Ser 137; from helix 2, Asn 152 and Tyr 156; and from helix 2', Gln 161'. Figure 4c shows residues from region C, where helices 1 and 2 cross each other for the first time, and where the four helix bundle is 'anchored' to the DNA phosphate backbone. The important residues shown are Asn 126 and Phe 129 from helix 1, and Lys 146 from helix 2. The nucleotides from C5 and A6 are also shown. Figures made with PREMA (M. Rould).

Figure 5: Model of the MyoD-E47 heterodimer.

We used the .LSQ command from the graphics program O (Jones et al., 1991) to superimpose the coordinates of the E47 homodimer (provided by T. Ellenberger and S.C. Harrison, Harvard University) on the MyoD homodimer coordinates. The MyoD-E47 heterodimer model consists of one monomer from the MyoD dimer combined with one monomer from the superimposed E47 dimer. The heterodimer model is shown 'bound' to the DNA site from the MyoD homodimer structure. Residues from MyoD are shown in magenta, those from

E47 are shown in red. Notice that the carboxy terminus of helix 1 is extended in the E47 monomer compared to that of MyoD.

Figure 6: Key residues from the dimer interface of the MyoD-E47 heterodimer model.

We show residues from the heterodimer model which correspond to the regions 1 and 2 shown in Figure 4. For purposes of exposition, we use the MyoD numbering scheme for residues. Thus, the E47 residues are numbered according to the scheme used for MyoD. Figure 6a shows the region A interface where helix 1 is from MyoD and helix 2' is from E47. We show side chains from key residues that may be important for heterodimer stabilization and/or specificity: from MyoD, Lys 124, Val 125, Ala 128, Thr 131, and Ser 135; from E47, Leu (147'), Gln (151'), and Leu (158'). Figure 6b shows region A where helix 1 is from E47 and helix 2' from MyoD. The key residues are: from E47, Asp (124), Ile (125), Ala (128), Glu (131), and Met (135); from MyoD, Val 147', Arg 151', and Glu 158'. Figure 6c shows the region B interface where the carboxy terminal end of helix 1 and residues from helix 2 of MyoD meet helix 2' from E47. The important residues are similar to those shown in Figure 4b for the MyoD homodimer, except that a Glu residue from E47 replaces that of a Gln from MyoD at position 161'. Figure 6d shows region B from the other half of the heterodimer. Here, the carboxy terminus of helix 1, along with parts of loop region and helix 2 from E47 meet helix 2' from MyoD. We show E47 residues which were identified from the E47 homodimer structure to be important for

dimer stabilization (Ellenberger et al., 1994): Gln (137), His (139), Leu (140), Gln (144), Gln (152) and Val (156). Notice that the carboxy terminus of helix 1 from E47 is extended by an additional turn compared to that of helix 1 from MyoD. In the E47 homodimer structure, a Glu at position 161' from helix 2' makes a hydrogen bond contact with the His at position 139. Here, in the heterodimer model, helix 2' is from MyoD, and a Gln replaces the E47 Glu at position 161'. Figures made with PREMA (M. Rould)

Figure 7: The interior of the core around the residues of region C.

We show a view looking down the dimer axis of the four-helix bundle from the MyoD homodimer. Hence we look down from the carboxy termini towards the amino termini of helices 1 and 1' and helices 2 and 2'. We show key residues involved in dimer stabilization at the region where helix 1 first crosses helix 2: from helices 1 and 1', Phe 129 and 129'; from helices 2 and 2', Ile 149, Leu 150, Ile 149', and Leu 150'. Figure made with PREMA (M. Rould).

```
                 helix 1                loop              helix 2

MyoD      KVNEAFETLKRCTS----SNP-----------NQRLPKVEILRNAIRYIEGLQAL

E47       DINEAFRELGRMCQ----MHLK---------SDKAQTKLLILQQAVQVILGLEQQ

Twist     SLNDAFKSLQQIIP----TLP-----------SDKLSKIQTLKLATRYIDFLCRM

AS-C sc   QVNNSFARLRQHIPQSIITDLTKG--G-GRGPHKKISKVDTLRIAVEYIRSLQDL

Id        DMNGCYSRLKELVP----TLP----------QNRKVSKVEILQHVIDYIRDLQLE

Tall      NVNGAFAELRKLIP----THPP----------DKKLSKNEILRLAMKYINFLAKL

Hairy     RINNCLNELKTLIL----DATKKDPA-----RHSKLEKADILEKTVKHLQELQRQ

----------------------------------------------------------------

N-Myc     DLRSSFLTLRDHVP---------EL-----VKNEKAAKVVILKKATEYVHSLQAE

c-Myc     ELKRSFFALRDQIP---------EL-----ENNEKAPKVVILKKATAYILSVQAE

Max       HIKDSFHSLRDSVP---------SLQG------QKASRAQILDKATEYIQYMRRK

USF       KINNWIVQLSKIIP----DC-SMEST------KSGQSKGGILSKACDYIQELRQS

TFE3      NINDRIKELGTLIP------KSSDP-------QMRWNKGTILKASVDYIRKLQKE

AP4       SINAGFQSLKTLIP--------HTD--------GEKLSKAAILQQTAEYIFSLEQE

PHO4      RLAVALHELASLIP---------EWKQ--QNVSAAPSKATTVEAACRYIRHLQQN
```

Figure 1a

```
            helix 1              loop          helix 2

MyoD     KVNEAFETLKRCTS----SNP-----------NQRLPKVEILRNAIRYIEGLQAL

E47      DINEAFRELGRMCQ----MHLK---------SDKAQTKLLILQQAVQVILGLEQQ

Twist    SINDAFKSLQQIIP----TLP-----------SDKLSKIQTLKLATRYIDFLCRM

AS-C sc  QVNNSFARLRQHIPQSIITDLTKG--G-GRGPHKKISKVDTLRIAVEYIRSLQDL

Id       DMNGCYSRLKELVP----TLP-----------QNRKVSKVEILQHVIDYIRDLQLE

Tal1     NVNGAFAELRKLIP----THPP----------DKKLSKNEILRLAMKYINFLAKL

Hairy    RINNCLNELKTLIL----DATKKDPA------RHSKLEKADILEKTVKHLQELQRQ


---------------------------------------------------------


N-Myc    DLRSSFLTLRDHVP---------EL-----VKNEKAAKVVILKKATEYVHSLQAE

c-Myc    ELKRSFFALRDQIP---------EL-----ENNEKAPKVVILKKATAYILSVQAE

Max      HIKDSFHSLRDSVP---------SLQG------QKASRAQILDKATEYIQYMRRK

USF      KINNWIVQLSKIIP----DC-SMEST------KSGQSKGGILSKACDYIQELRQS

TFE3     NINDRIKELGTLIP-------KSSDP-------QMRWNKGTILKASVDYIRKLQKE

AP4      SINAGFQSLKTLIP--------HTD--------GEKLSKAAILQQTAEYIFSLEQE

PHO4     RLAVALHELASLIP---------EWKQ--QNVSAAPSKATTVEAACRYIRHLQQN
```
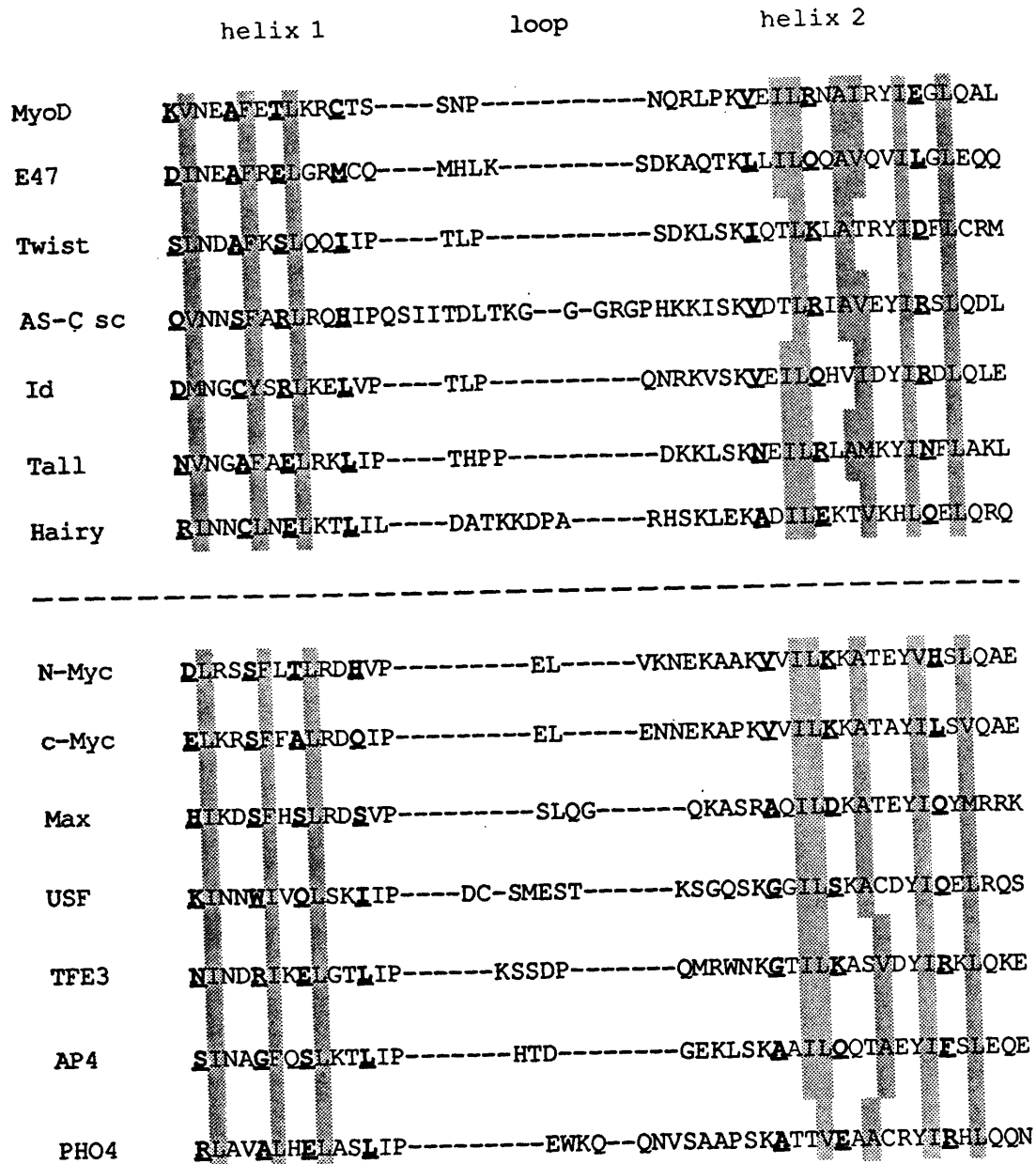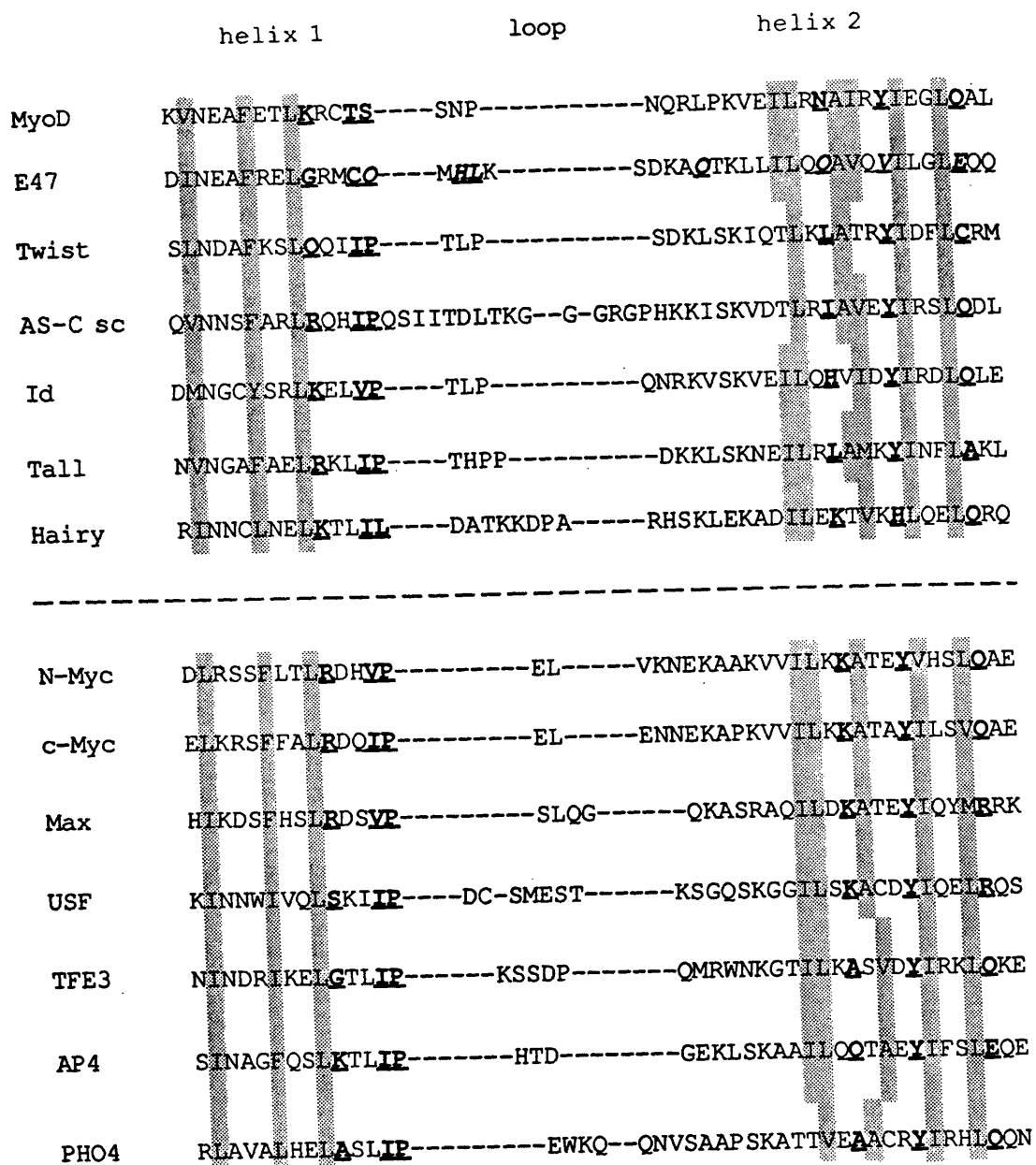
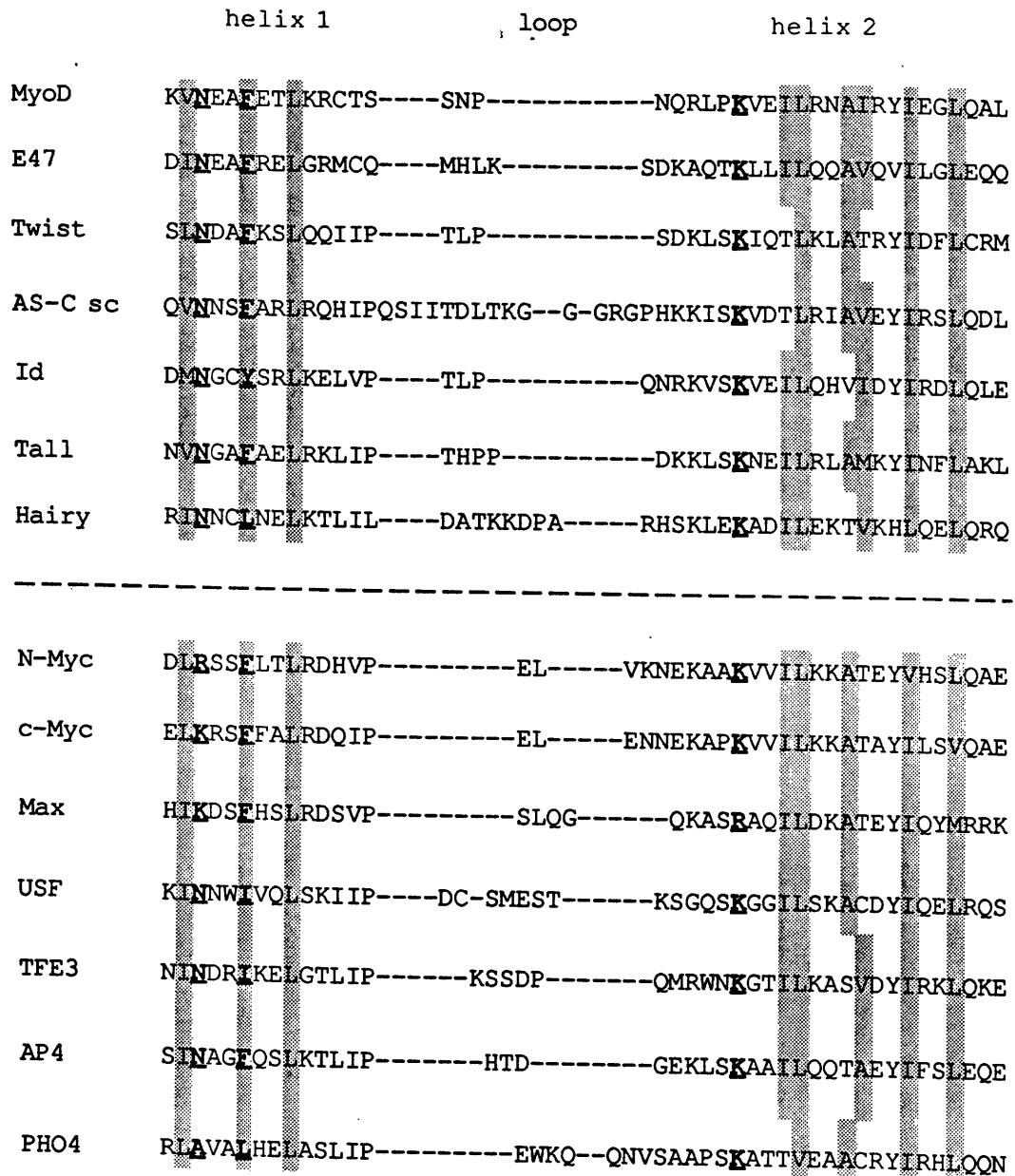# Figure 1b

```
            helix 1              loop             helix 2

MyoD     KVNEAFETLKRCTS----SNP-----------NQRLPKVEILRNAIRYIEGLQAL

E47      DINEAFRELGRMCQ----MHLK---------SDKAQTKLLILQQAVQVILGLEQQ

Twist    SLNDAFKSLQQIIP----TLP-----------SDKLSKIQTLKLATRYIDFLCRM

AS-C sc  QVNNSFARLRQHIPQSIITDLTKG--G-GRGPHKKISKVDTLRIAVEYIRSLQDL

Id       DMNGCYSRLKELVP----TLP-----------QNRKVSKVEILQHVIDYIRDLQLE

Tal1     NVNGAFAELRKLIP----THPP----------DKKLSKNEILRLAMKYINFLAKL

Hairy    RINNCLNELKTLIL----DATKKDPA-----RHSKLEKADILEKTVKHLQELQRQ

         ------------------------------------------------------

N-Myc    DLRSSFLTLRDHVP----------EL-----VKNEKAAKVVILKKATEYVHSLQAE

c-Myc    ELKRSFFALRDQIP----------EL-----ENNEKAPKVVILKKATAYILSVQAE

Max      HIKDSFHSLRDSVP----------SLQG------QKASRAQILDKATEYIQYMRRK

USF      KINNWIVQLSKIIP----DC-SMEST------KSGQSKGGILSKACDYIQELRQS

TFE3     NINDRIKELGTLIP------KSSDP--------QMRWNKGTILKASVDYIRKLQKE

AP4      SINAGFQSLKTLIP-------HTD--------GEKLSKAAILQQTAEYIFSLEQE

PHO4     RLAVALHELASLIP---------EWKQ--QNVSAAPSKATTVEAACRYIRHLQQN
```

Figure 1c

```
              helix 1              loop              helix 2

MyoD     KVNEAFETLKRCTS----SNP------------NQRLPKVEILRNAIRYIEGLQAL

E47      DINEAFRELGRMCQ----MHLK---------SDKAQTKLLILQQAVQVILGLEQQ

Twist    SINDAFKSLQQIIP----TLP------------SDKLSKIQTLKLATRYIDFLCRM

AS-C sc  QVNNSFARLRQHIPQSIITDLTKG--G-GRGPHKKISKVDTLRIAVEYIRSLQDL

Id       DMNGCYSRLKELVP----TLP-----------QNRKVSKVEILQHVIDYIRDLQLE

Tall     NVNGAFAELRKLIP----THPP----------DKKLSKNEILRLAMKYINFLAKL

Hairy    RINNCLNELKTLIL----DATKKDPA-----RHSKLEKADILEKTVKHLQELQRQ


N-Myc    DLRSSELTLRDHVP---------EL-----VKNEKAAKVVILKKATEYVHSLQAE

c-Myc    ELKRSFFALRDQIP---------EL-----ENNEKAPKVVILKKATAYILSVQAE

Max      HIKDSFHSLRDSVP---------SLQG------QKASRAQILDKATEYIQYMRRK

USF      KINNWIVQLSKIIP----DC-SMEST------KSGQSKGGILSKACDYIQELRQS

TFE3     NINDRIKELGTLIP------KSSDP-------QMRWNKGTILKASVDYIRKLQKE

AP4      SINAGFQSLKTLIP-------HTD--------GEKLSKAAILQQTAEYIFSLEQE

PHO4     RLAVALHELASLIP---------EWKQ--QNVSAAPSKATTVEAACRYIRHLQQN
```
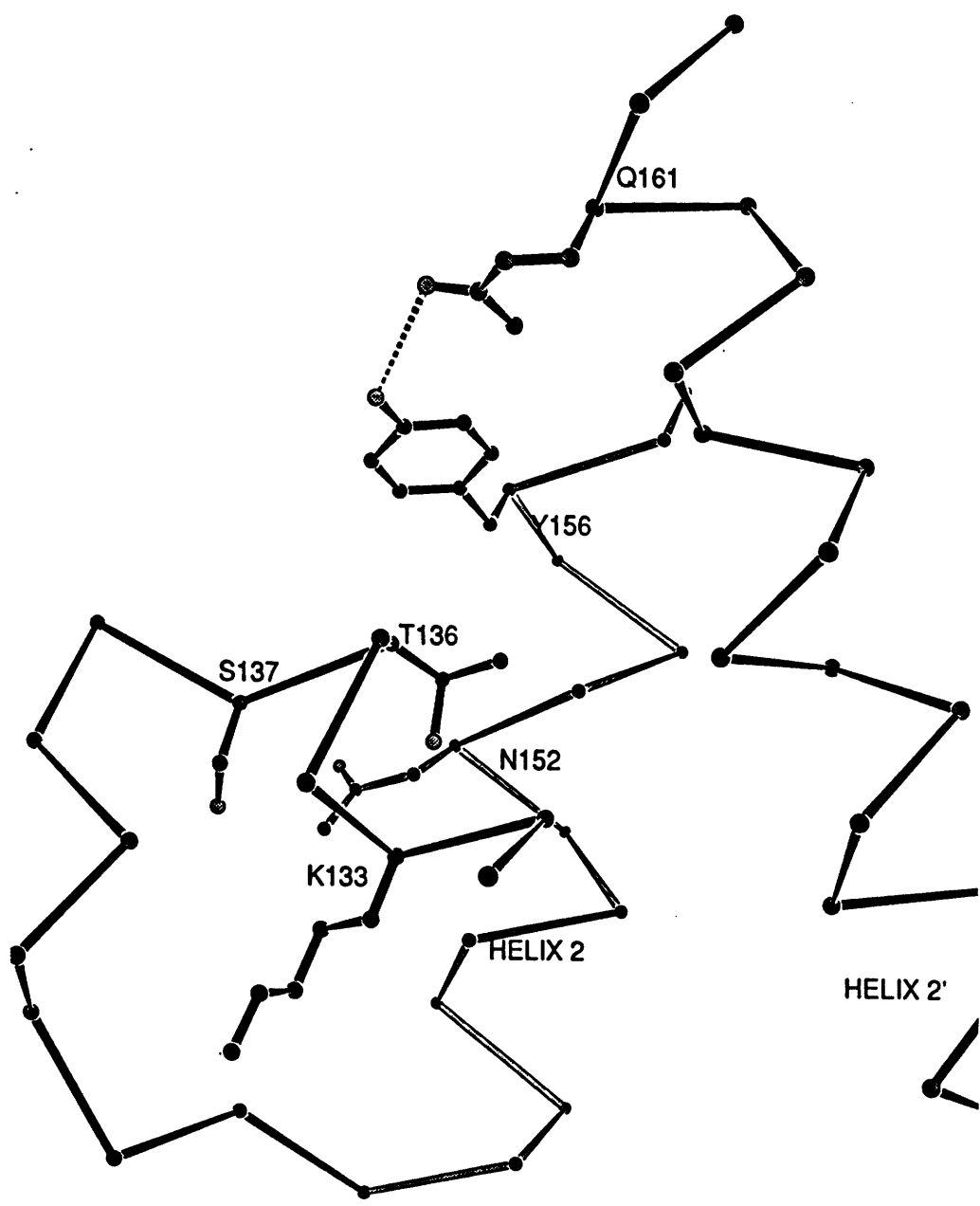
Figure 1d

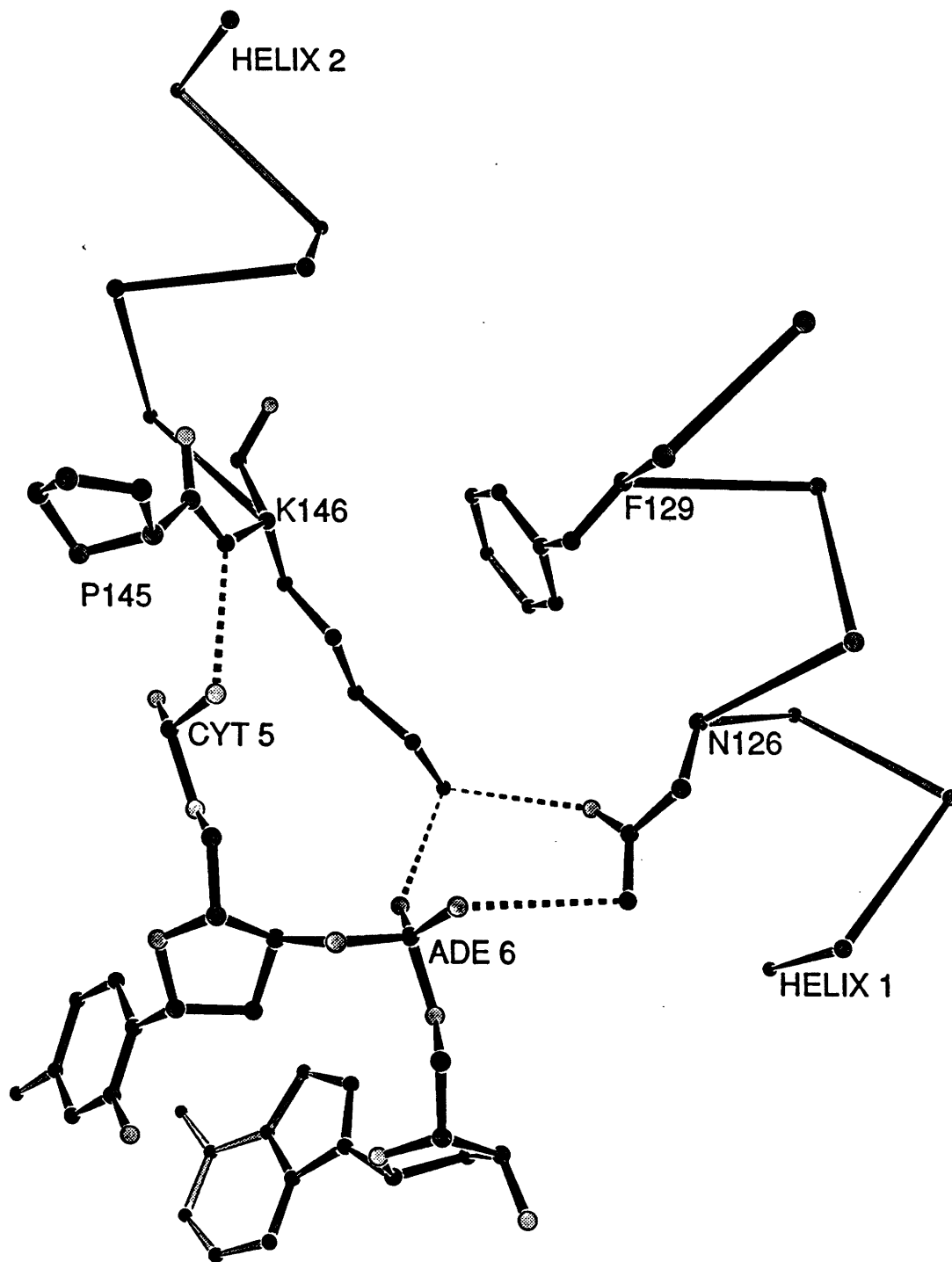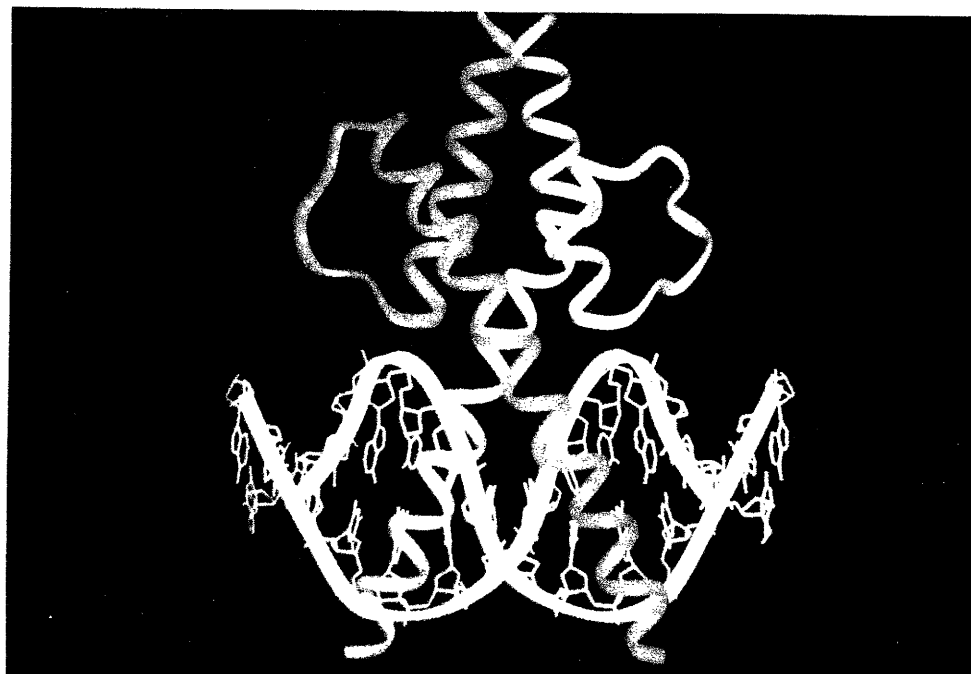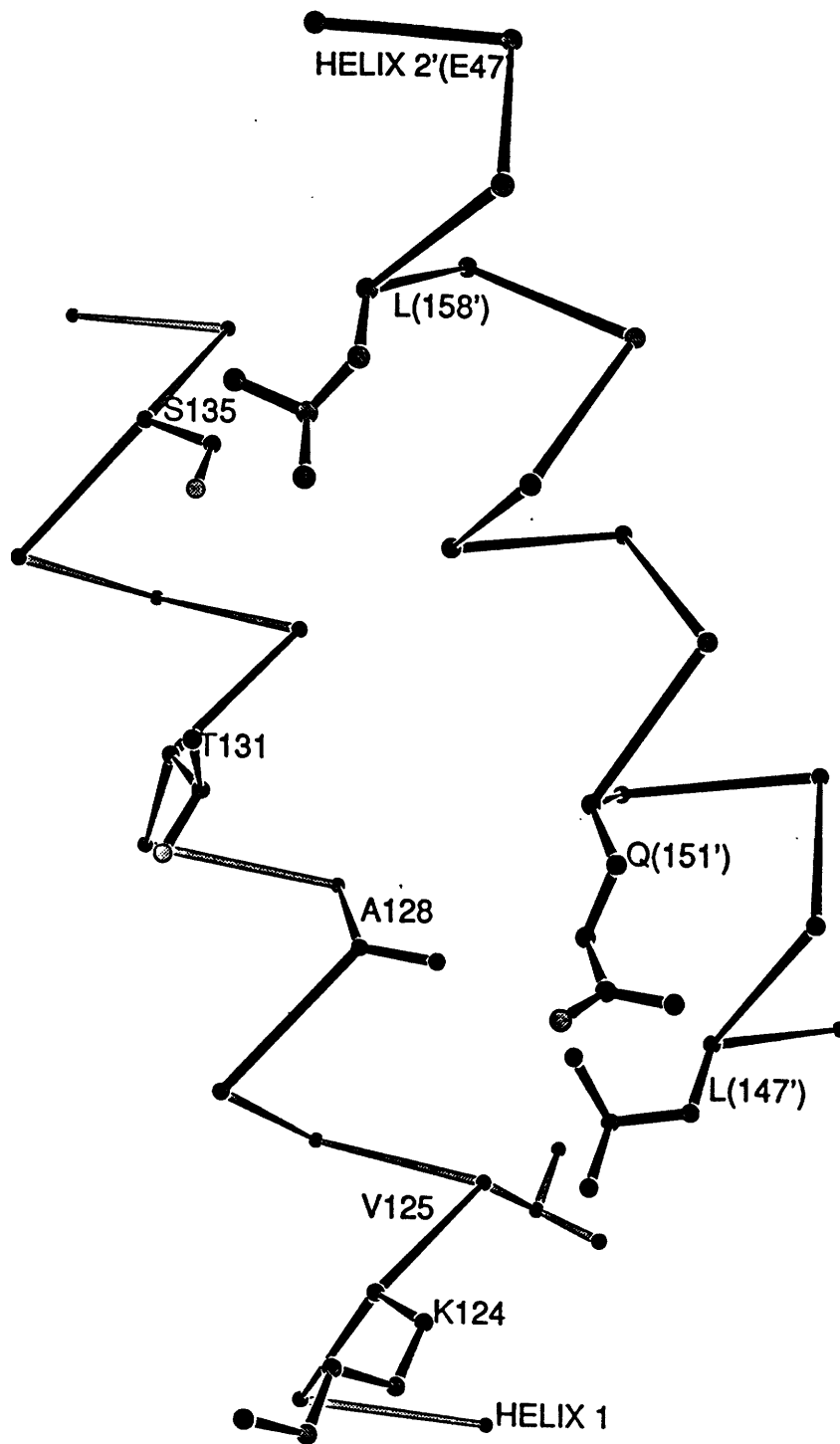a)

b)

Figure 2

c)

d)

Figure 2
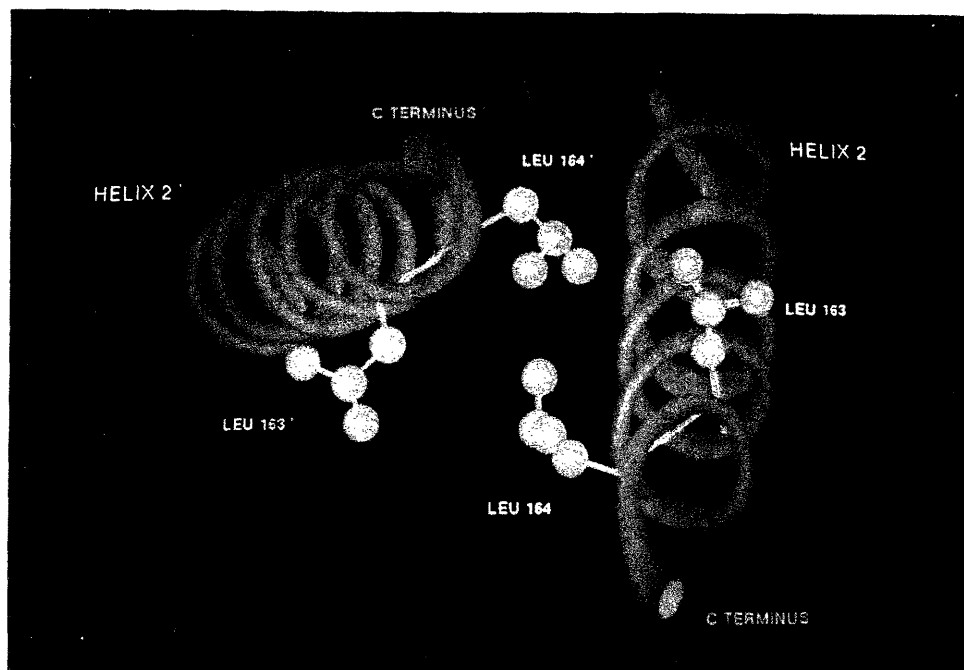
Figure 3
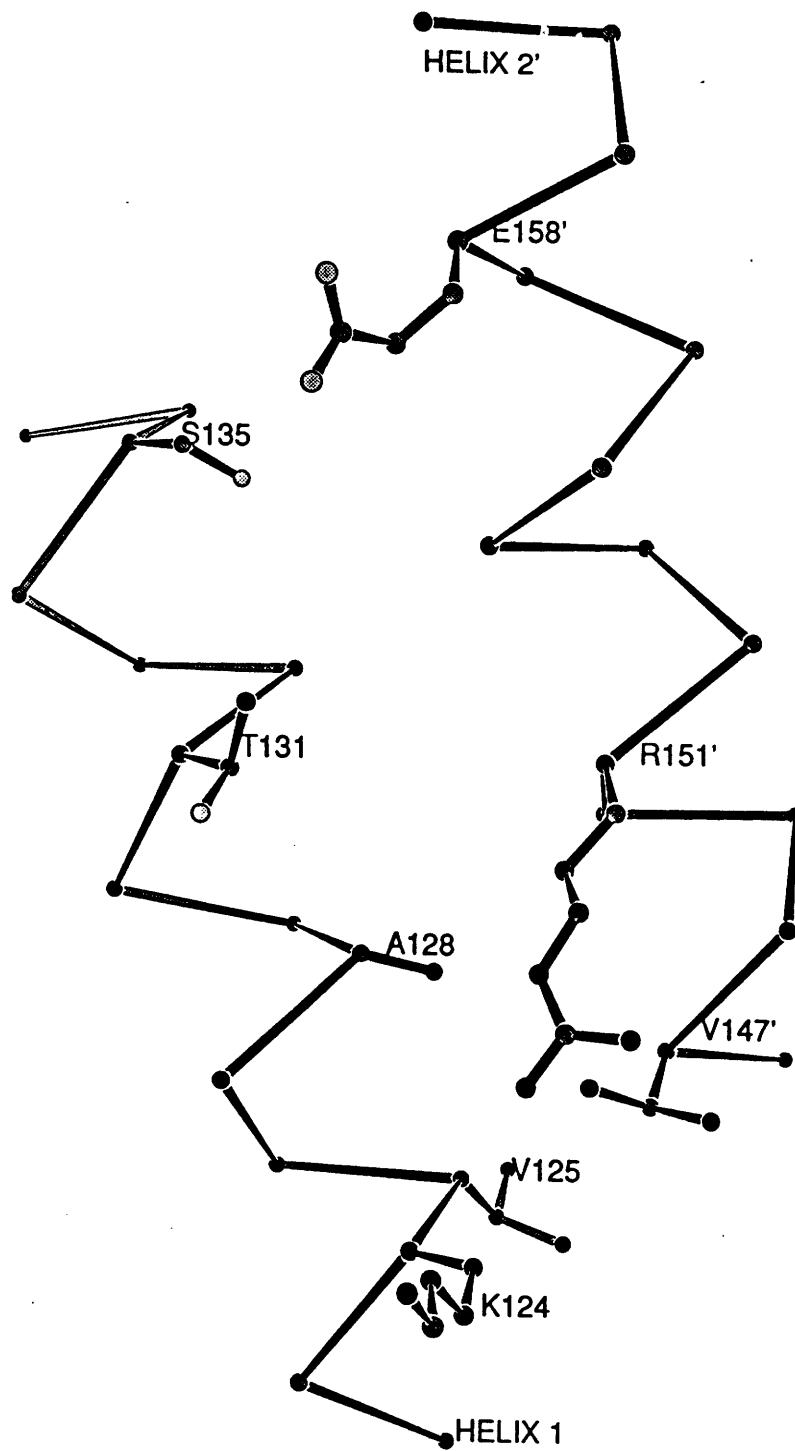
Figure 4a
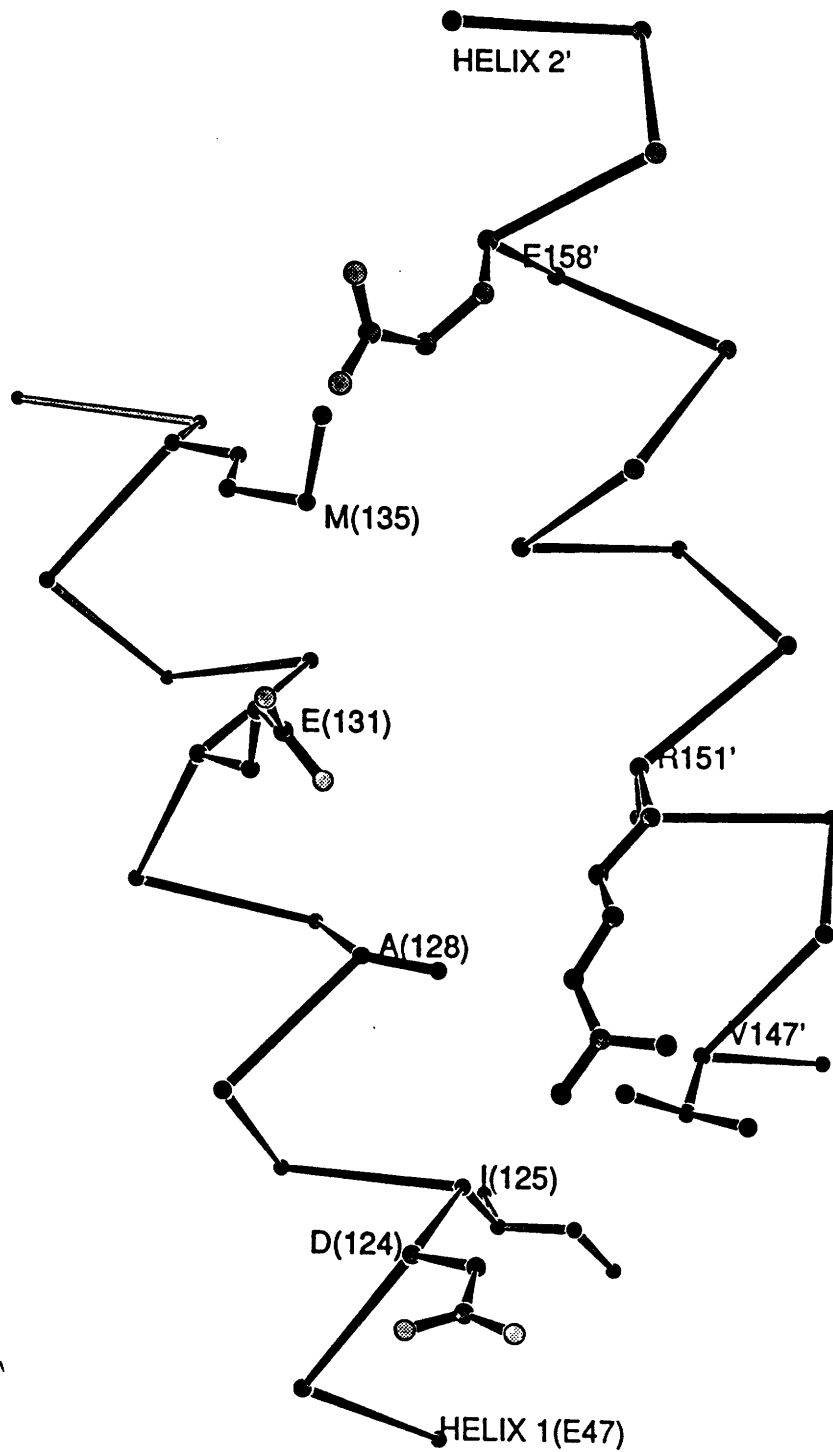
Figure 4b

Figure 4c

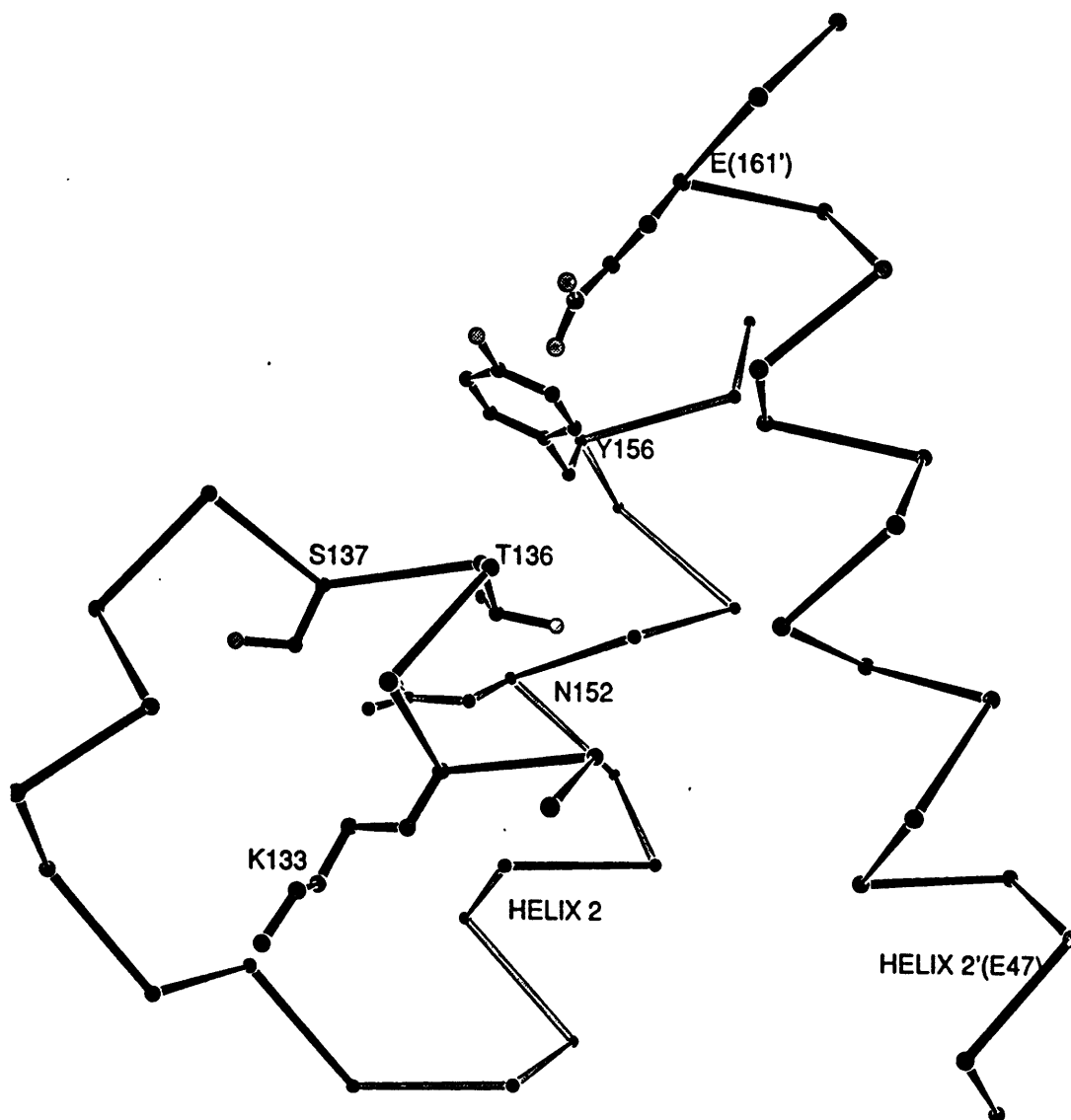Figure 5

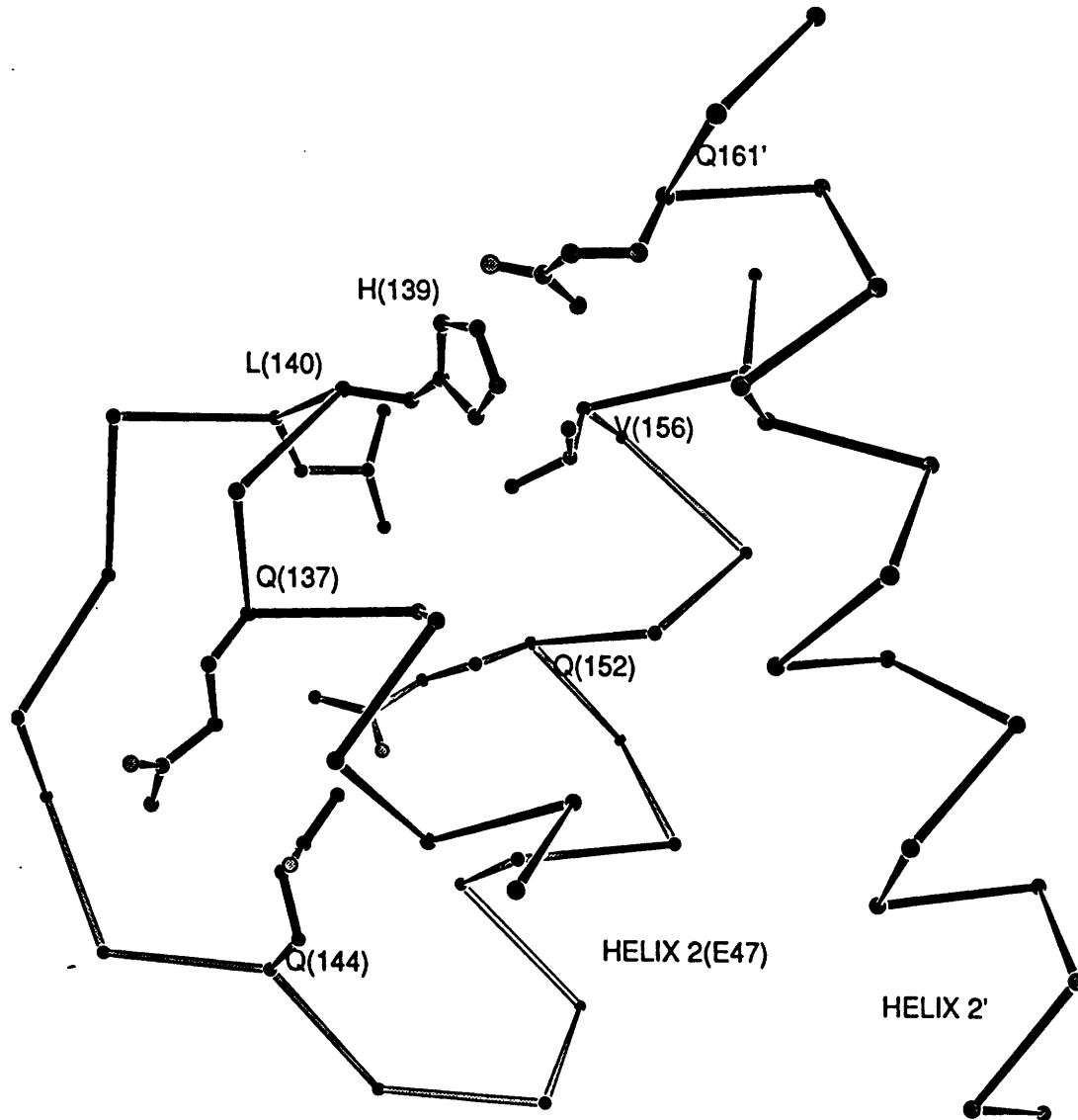Figure 6a
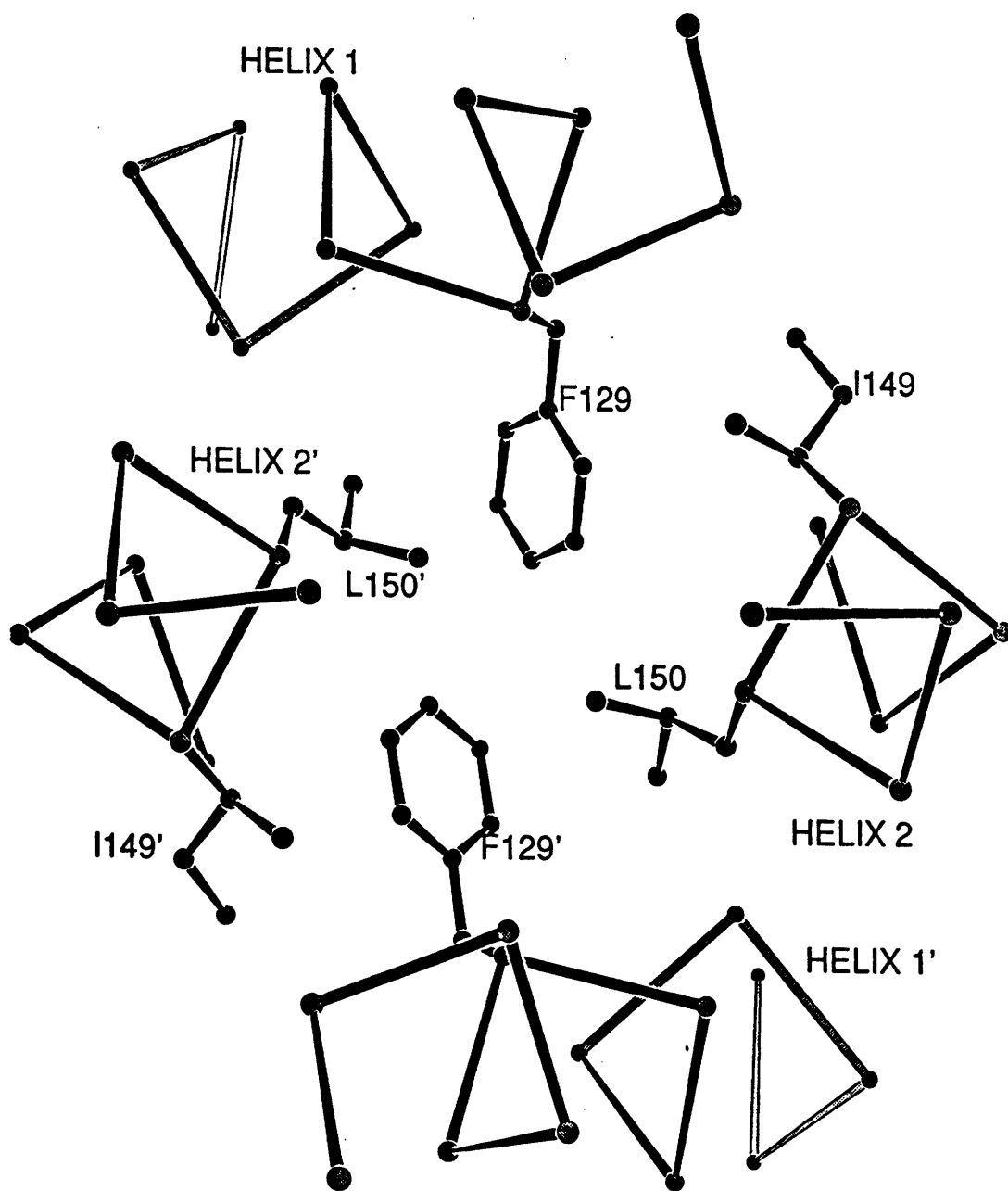
Figure 6b

Figure 6c

Figure 6d

Figure 7

CHAPTER 6

The Structural Basis for the Control of
Gene Regulation by MyoD:
Outlook for Future Experiments

At this stage, we can answer many of the structural and biological questions that motivated this work. In summary, we have: 1) Determined that the fold of the MyoD bHLH domain is very similar to that of the Max bHLH domain, and the E47 bHLH domain, but is less similar to the USF bHLH domain (which may be a result of the deletion of the leucine zipper from the normal bHLH/Z domain of USF). It thus appears that the bHLH domain is a fairly well-conserved fold. 2) Determined the basis of specificity for the conserved four bases in the central E-box DNA binding site (CANNTG), as well as for a few of the flanking base pairs. 3) Shown that a conserved core of hydrophobic residues within the four helix bundle forms the bulk of the dimer interface for the MyoD homo-dimer. We have also identified a few regions of the dimer interface which contain residues that may be important in influencing homo-dimer and heterodimer stability. 4) Examined the positions of the postive control mutants in the context of the MyoD structure. Two of these residues, Ala 114 and Thr 115, are found to be buried within the major groove of the DNA, and hence, are not likely to interact directly with other proteins. We have found, though, that Ala 114 and Thr 115 may determine the structure of Arg 111, and we hypothesize that the position of Arg 111 could have an effect on the transcriptional activity of MyoD.

Even as we have met some of our initial goals in our study of MyoD, there are questions about the structure and function of the MyoD bHLH domain that remain to be answered. Indeed, a number of these questions have arisen only after the determination of the

crystal structure. In the following section, we outline a few of these outstanding questions and suggest possible directions for future research.

A. Tertiary fold

The four crystal structures of bHLH and bHLH/Z proteins which have been solved recently show that the parallel four-helix bundle fold is very well conserved. In particular, three of the four structures, those of Max, E47, and MyoD, have very similar structures in the four-helix bundle region (Figure 1). [Conver-sations with Adrian Ferré-D'Amaré and Stephen Burley support our observation that the structure of the USF bHLH region is signifi-cantly different from that of Max and MyoD.] From sequence comparisons of the four proteins (Max, USF, E47 and MyoD), it appears that the hydrophobic core accounts for much of the conser-vation of structure in the four-helix bundle fold. It would be interesting to learn how residue changes within the hydrophobic core change the structure of the four-helix bundle. As discussed in the previous chapter, there is still some uncertainty about how the identity of hydrophobic residues affects the structure of proteins, in general. Mutagenesis experiments on the hydrophobic core region of MyoD may help determine how the structure of the four-helix bundle is affected by different hydrophobic residues, and may also add to our understanding of the importance of hydrophobic residues for protein structures, in general. One could make random changes in the core region of MyoD by means of cassette mutagenesis (Reidhaar-Olson

et al., 1991), screen the mutants for DNA binding activity and oligomerization potential, and then determine the structures of the most interesting mutants. Alternatively, one could use a more directed approach to mutagenesis, such as that used in one study on MyoD (Shirakata et al., 1993), and make systematic residue changes at one or more of the positions in the core.

B. DNA recognition

There is one important question about DNA recognition by MyoD that was not fully answered by the structure of the MyoD-DNA complex, and that question has to do with the specificity for the central base pairs of the MyoD binding site. MyoD, along with most other bHLH proteins, has a DNA binding preference for CAGCTG, and against CACGTG (Blackwell and Weintraub, 1990). In contrast, Max, myc, and most other bHLH/Z proteins seem to preferentially bind CACGTG over CAGCTG (Dang et al., 1992; Blackwell et al., 1993). The structure of the Max complex showed that an Arg, at a position corresponding to Leu 122 in MyoD, made a hydrogen bond contact to the N7 atom of the central G base in CACGTG (Ferré-D'Amaré et al., 1993). A similar arginine-guanine contact was seen in the structure of the USF complex (Ferré-D'Amaré et al., 1994). In contrast to this defined arginine-guanine contact in the Max and USF structures, the structure of the MyoD complex did not show any direct contacts at the central base pairs. Indeed, the leucine residue at position 122 was not in a good position to make any sort of base-specific contact at the central base pairs. [The structure of the E47 complex

showed that the Val residue at position 122 does not make any contact to the central base pairs either (Ellen-berger et al., 1994)]. It appears, then, that the bHLH proteins may have different mechanisms for recognizing the central base pairs than do the bHLH/Z proteins. The mechanisms used by the bHLH proteins may depend upon the sequence being recognized. Selection experiments had revealed that while the MyoD homodimer prefers the symmetric DNA sequence CAGCTG, the E47 homodimer or the MyoD-E47 heterodimer prefers the asymmetric DNA sequence CAGGTG (Blackwell et al., 1990). In the E47 structure, where the protein is bound to DNA containing the sequence CAGGTG, one of the basic regions has an Arg (corresponding to Arg 119 in MyoD) that makes a contact to the N7 of the central G base in CAGGTG (Ellen-berger et al., 1994). The other basic region has the Arg residue making a phosphate contact (both of the Arg 119 residues in the MyoD homodimer make phosphate contacts). In the MyoD homodimer structure, which contains the sequence CAGCTG, there are no direct contacts made to the central base pairs, although there is a water-mediated contact made by Glu 118 to the central C base in CAGCTG. Given the resolution limit of 2.8 Å in the MyoD structure, however, one cannot make a definitive conclusion about the importance of water-mediated contacts.

We have seen, then, that for the MyoD homodimer, the basis of specificity for the central bases in CAGCTG does not appear to be due to direct base contacts, but may rather be due to water-mediated contacts and/or an indirect readout mechanism that

senses sequence-dependent changes in the phosphate backbone. In
order to clarify this question of specificity for the central base
pairs, we can take several approaches. The first is to obtain a
higher resolution structure of the MyoD homodimer bound to DNA
site containing the sequence CAGCTG. The MyoD cocrystals which
we used in our structure determination have an effective resolution
limit of 2.8 Å. It might be possible to get better diffraction by
finding new crystallization conditions, possibly using a slightly
different DNA site. Flash-freezing the crystals with liquid nitrogen
after treating the crystals with an appropriate cryoprotectant
(Petsko, 1975), may extend the resolution of diffraction by limiting
radiation damage during data collection. We had earlier exper-
imented with flash-freezing, but since our crystals were very
sensitive to mechanical stress and changes in mother liquor (see
Chapter 3), we were unable to find a cryoprotectant that did not
damage the crystals. At this stage, then, it appears, then, finding
new crystallization conditions may be the best means of getting
better diffracting crystals. Aside from obtaining a higher resol-
ution structure of the MyoD complex, we might get more
information about specificity for the central base pairs by
cocrystallizing MyoD with DNA sequences that differ at the center.
A crystal structure of MyoD bound to a DNA site containing CACGTG
might reveal why this site is not favored by MyoD. [Although MyoD
binds poorly to the CACGTG site in gel shift experiments (Blackwell
et al., 1993), the millimolar concentrations of protein and DNA used
in crystallization experiments may be high enough to compensate
for the poor DNA binding.]

C. Dimerization

Dimerization has been shown to be an important determinant of the biological activities for the bHLH(-ZIP) and bZIP families of transcriptional regulators (for a review, see Ellenberger, 1994). In Chapter 5, we examined some of the details of the dimer interface in the MyoD homodimer structure. In particular, we identified three regions of the dimer interface that may be especially important for determining the specificity of dimer formation. We also presented a model of the MyoD-E47 heterodimer, based on the coordinates of the MyoD homodimer and E47 homodimer (Ellenberger et al., 1994). Although our model has identified a few residues that may be important for heterodimer specification, definitive conclusions about how the MyoD-E47 heterodimer is stabilized will have to await the solution of a heterodimer structure. The biggest barriers to obtaining a structure of the heterodimer will likely be at the purification and crystallization stages. Since a mixture of MyoD and E47 will result in an equilibrium of MyoD homodimers, E47 homodimers and MyoD-E47 heterodimers (Sun and Baltimore, 1991), it is likely that there would be significant problems of microheterogeneities during purification and crystallization. One way around this problem of homodimer-heterodimer equilibrium, would be to use a covalently-linked heterodimer. Heterodimers of MyoD tethered to E47 by cysteine disulfide bonds (Anthony-Cahill et al., 1992; Fairman et al., 1993), or by flexible polypeptide linkers (Neuhold and Wold, 1993) have been constructed and shown to bind DNA effectively. Moreover, once these tethered heterodimers are

formed, they are difficult to break apart, as they seem to be resistant to negative inhibition by the Id protein (Fairman et al., 1993; Neuhold and Wold, 1993). Thus, a tethered heterodimer is likely to be stably folded and a good candidate for crystallization studies. We would probably need to generate a number of different tethered heterodimers, joined by linkers of various length and composition, before finding one that produced suitable crystals.

Short of solving a crystal structure of a MyoD-E47 heterodimer (tethered or not), we could get some information about the determinants for heterodimer formation by means of mutagenesis experiments. In Chapter 5, we identified a few residues in regions A and B of the dimer interface that might be important for heterodimer stabilization. We could test our theories by making directed mutations at these residues, and then checking to see how these mutations affected homodimerization and heterodimerization. A recent study on MyoD used such an approach to generate a MyoD mutant that had an increased preference for homodimerization (Shirakata et al., 1993; the change was in a residue in region A: Thr 131 to Glu). By using directed mutagenesis at other residues at the dimer interface to generate MyoD and E47 mutants that had different preferences for homodimeration and heterodimerization, we could learn more about the rules for heterodimer and homodimer formation. A detailed study of dimer formation among the myogenic proteins (which would include structural studies, careful measurements of dimer stability and DNA binding affinities of different homo- and heterodimers, as well

as mutagenesis experiments) may add to our understanding of how dimerization controls the activities of bHLH and bZIP proteins, in general.

D. Transcriptional Activation

The structure of the MyoD complex revealed that the positive control mutants may influence transcriptional activation indirectly through determining the position of Arg 111 in the basic region. We can test this hypothesis by generating mutants in the positive control residues and testing for their ability to bind DNA and activate transcription. Previous studies showed that a Thr to Asn substitution at position 115 generated a mutant that was able to bind DNA as well as activate transcription (Bengal et al., 1994). This result is consistent with our model since in the MyoD structure, Thr 115 makes a phosphate and base specific contact. An Asn substituted for the Thr would still be able to make a contact to the DNA without significantly affecting the position of Arg 111. In contrast to the mutations at position 115, mutations at Ala 114 would be expected to have an effect on the position of Arg 111. Thus, an Ala to Asp substitution at position 114 resulted in a mutant that still bound DNA, but was no longer able to activate transcription (Bengal et al., 1994). It will be interesting to test if other mutations at position 114 can generate the same phenotype as the Ala to Asp substitution. Such mutagenesis studies are currently under way (Weintraub et al., personal communication). In addition to further mutagenesis studies at positions 114 and 115, our

collaborators are currently attempting to generate second site

reverants to the positive control mutants by changing the identity

of the residue at position 111 (Weintraub et al., personal

communication). Once these revertants are identified, it would be

quite interesting to determine their structures.

The work presented in this thesis has concentrated on

describing the bHLH region of MyoD. There are other areas of the

protein which are important for biological activity. For example,

the first 53 residues of MyoD have been shown to function as a

transcriptional activation region when fused to the GAL4 DNA

binding domain (Weintraub et al., 1991). It will be very interesting

to see how this N-terminal activation region interacts with other

regions of MyoD, especially the bHLH region. A deeper understanding

of how the positive control residues affect transcription may come

from solving the structure of the intact MyoD protein, or else, a

fragment of the protein which contains both the N-terminal

activation and bHLH regions. Proteolytic dissection of the lambda

repressor protein revealed that an amino-terminal domain binds

DNA and that a carboxy-terminal domain mediates oligomerization

(Pabo et al., 1979). Proteolytic digestion experiments with MyoD, in

the presence and absence of DNA, may uncover a protease-resistant

domain of the protein that maintains both DNA binding and

transcriptional activities. If such a stable domain can be found, it

would be a good candidate for further structural studies.

References

Anthony-Cahill, S.J., Benfield, P.A., Fairman, R., Wasserman, Z.R., Brenner, S.L., Stafford, W.F., III, Altenbach, C., Hubbell, W.L., and DeGrado, W.F. (1992). Molecular characterization of helix-loop-helix peptides. Science *255*, 979-983.

Bengal, E., Flores, O., Rangarajan, P.N., Chen, A., Weintraub, H., and Verma, I. (1994). Positive control mutations in MyoD basic region fail to show cooperative DNA binding and transcriptional activation *in vitro*. Prc. Natl. Acad. Sci. USA. In press.

Blackwell, T.K., and Weintraub, H. (1990). Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. Science *250*, 1104-1110.

Blackwell, T.K., Huang, J., Ma, A., Kretzner, L., Alt, F.W., Eisenman, R.N., and Weintraub, H. (1993). Binding of myc proteins to canonical and noncanonical DNA sequences. Mol. Cell Biol. *13*, 5216-5224.

Dang, C.V., Dolde, C., Gillison, M.L., and Kato, G.J. (1992). Discrimination between related DNA sites by a single amino acid residue of Myc-related basic-helix-loop-helix proteins. Proc. Natl. Acad. Sci. USA *89*, 599-602.

Ellenberger, T., Fass, D., Arnaud, M., and Harrison, S.C. (1994). Crystal structure of transcription factor E47: E-box recognition by a basic region helix-loop-helix dimer. Genes Dev. In press.

Fairman, R., Beran-Steed, R.K., Anthony-Cahill, S.J., Lear, J.D., Stafford, W.F., DeGrado, W.F., Benfield, P.A., and Brenner, S.L. (1993). Multiple oligomeric states regulate the DNA binding of helix-loop-helix peptides. Proc. Natl. Acad. Sci. USA *90*, 10429-10433.

Ferré-D'Amaré, A.R., Pognonec, P., Roeder, R.G., and Burley, S.K. (1994). Structure and function of the b/HLH/Z domain of USF. EMBO J. *13*, 180-189.

Ferré-D'Amaré, A.R., Prendergast, G.C., Ziff, E.B., and Burley, S.K. (1993). Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain. Nature *363*, 38-45.

Jones, T.A., Zou, J.-Y., Cowan, S.W., and Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Cryst. *A47*, 110-119.

Neuhold, L.A., and Wold, B. (1993). HLH forced dimers: Tethering MyoD to E47 generates a dominant positive myogenic factor insulated from negative regulation by Id. Cell *74*, 1033-1042.

Pabo, C.O., Sauer, R.T., Sturtevant, J.M., and Ptashne, M. (1979). The λ repressor contains two domains. Proc. Natl. Acad. Sci. USA *76*, 1608-1612.

Petsko, G. (1975). Protein crystallography at sub-zero temperatures: Cryo-protective mother liquors for protein crystals. J. Mol. Biol. *96*, 381-392.

Reidhaar-Olson, J.F., Bowie, J.U., Breyer, R.M., Hu, J.C., Knight, K.L., Lim, W.A., Mossing, M.C., Parsell, D.A., Shoemaker, K.R., and Sauer, R.T. (1991). Random mutagenesis of protein sequences using oligonucleotide cassettes. Methods in Enzymology *208*, 564-586.

Shirakata, M., Friedman, F.K., Wei, Q., and Paterson, B.M. (1993). Dimerization specificity of myogenic helix-loop-helix DNA-binding factors directed by nonconserved hydrophilic residues. Genes Dev. *7*, 2456-2470.

Sun, X.H., and Baltimore, D. (1991). An inhibitory domain of E12 transcription factor prevents DNA binding in E12 homodimers but not in E12 heterodimers. Cell 64, 459-470.

Weintraub, H., Dwarki, V.J., Verma, I., Davis, R., Hollenberg, S., Snider, L., Lassar, A., and Tapscott, S.J. (1991). Muscle-specific transcriptional activation by MyoD. Genes Dev. *5*, 1377-1386.

Figure Legends

Figure 1: Superposition of the basic-helix-loop-helix regions of Max, E47, and MyoD reveal a strongly conserved tertiary fold.

We used the program O (Jones et al., 1991) to superimpose the coordinates of the bHLH regions of the Max homodimer structure (provided by Stephen Burley, Rockefeller University), and the E47 homodimer structure (provided by Tom Ellenberger and S.C. Harrison, Harvard University) on the coordinates of the MyoD homodimer structure. As first suggested in the report on the Max structure (Ferré-D'Amaré et al., 1993), the superposition reveals that the four-helix bundle domain is remarkably well conserved between the three structures. The three bHLH domains are shown 'bound' to DNA from the MyoD homodimer structure. Residues from MyoD are shown in magenta, those from E47 are in red, and those from Max are in green.
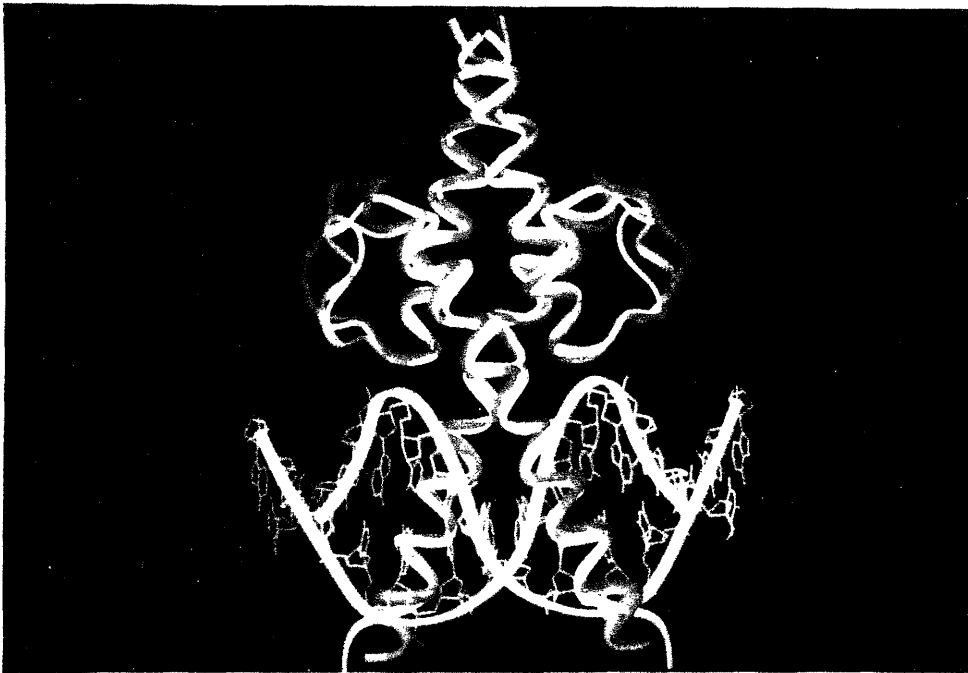
Figure 1