parate eukaryotic expres-
es for the heavy and light
nd V_L expression vectors
ATCC CCL10) using the
methotrexate and G418 as
lones that produce up to
er bottle cultures without

ied CEA and to CEA on
the original mouse MAb
ies, although it has not yet
t of the mouse MAb, which

ıe murine BW431/26 MAb
ersion. Surprisingly, 17 out
ı as well as the mouse MAb;
ill detectable. This finding
tween the variable domains
. Bosslet *et al.*, unpublished

cid sequences of the donor
th the acceptor V_H (human
egree of sequence similarity
s are clustered in the regions
ı domains. Even in positions
each other the amino acid
ıave contributed to the suc-
ody. The crystal structures of
ıd we use their respective V
any murine MAb. Computer
potential framework amino
ıct with the CDRs or directly
sferred to the human frame-
ing scheme described here we
y particular monoclonal anti-

ne **73,** 419 (1988).

Queen *et al.*[31] have reshaped an antibody by selecting human acceptor V domains from the Kabat database to match the framework sequences of the murine donor V domains as closely as possible. This alternative reshaping strategy raises the possibility that any human antibody framework can be used as acceptor for the murine CDRs. Humanization of murine MAbs is a technique that is now widely applied to MAbs of potential use in diagnosis and therapy in humans. Whether one of the reshaping schemes will offer any particular advantage in successful humanization of MAbs will become clearer as the number of reshaped antibodies increases.

Acknowledgments

Initial work was done by D. G. at the MRC/LMB, CB2 2QH Cambridge, England. We thank Dr. Greg Winter and P. T. Jones for their advice and help in establishing the humanization technique in our laboratory, and for their technical help with the computer analyses, Mrs. K. Müller and R. Bier for excellent technical assistance, Dr. K. Bosslet for helpful discussion, and Dr. H.-H. Sedlacek for support.

[31] C. Queen, W. P. Schneider, H. E. Selick, P. W. Payne, N. F. Landolfi, J. F. Duncan, N. M. Avdalovitc, M. Levitt, R. P. Junghans, and T. A. Waldmann, *Proc. Natl. Acad. Sci. U.S.A.* **86,** 10029 (1989).

[6] Molecular Modeling of Antibody Combining Sites

By ANDREW C. R. MARTIN, JANET C. CHEETHAM, and ANTHONY R. REES

Introduction

Both the variable and constant domains of an antibody Fab consist of two twisted antiparallel β sheets which form a β-sandwich structure. The constant regions have three- and four-stranded β sheets arranged in a Greek key-like motif,[1] while variable regions have a further two short β strands producing a five-stranded β sheet.

The two β sheets of the variable domain are inclined at 30° to one another,[2] with a conserved disulfide bridge in each domain linking the two β sheets. Lesk and Chothia[3] have shown the relative orientation of the two

[1] J. S. Richardson, *Adv. Protein Chem.* **34,** 168 (1981).
[2] C. Chothia and J. Janin, *Proc. Natl. Acad. Sci. U.S.A.* **78,** 4146 (1981).
[3] A. M. Lesk and C. Chothia, *J. Mol. Biol.* **160,** 325 (1982).

β sheets to vary by up to 18° in variable domains and up to 10° in constant domains. This characteristic "immunoglobulin fold"[4-6] is also seen in a number of molecules of related and unrelated function,[7] one of the most recent examples being a bacterial protein, PapD, which mediates assembly of pili in *Escherichia coli*.[8]

The V_L and V_H domains interact via the five-stranded β sheets to form a nine-stranded β barrel of about 8.4 Å radius, with the strands at the domain interface inclined at approximately 50° to one another.[9] The loops linking the β strands form the CDRs with the domain pairing bringing these loops from both V_L and V_H into close proximity. The CDRs themselves form some 25% of the V_L/V_H domain interface.[10]

The six hypervariable loops [complementarity determining regions (CDRs)] that are supported on the β-barrel framework form the antigen combining site (ACS). While their sequence is hypervariable in comparison with the rest of the immunoglobulin structure,[11] some of the loops show a relatively high degree of both sequence and structural conservation. In particular, CDR-L2 and CDR-H1 are highly conserved in conformation.[12] Analysis of conserved key residues has led Chothia and co-workers to define canonical ensembles into which the CDRs may be grouped.[13,14]

Requirement for Modeling

Since the first X-ray crystallographic structure determinations, sequences of immunoglobulin light and heavy chain variable regions have accumulated at an ever increasing rate. X-Ray structures, however, are accruing at a relatively slow pace with no more than three or four immunoglobulin structures being published each year and fewer than that being deposited in the Brookhaven Protein Databank.[15] Protein crystallography

[4] R. J. Poljak, L. M. Amzel, H. P. Avey, B. L. Chen, R. P. Phizackerley, and F. Saul, *Proc. Natl. Acad. Sci. U.S.A.* **70**, 3305 (1973).
[5] M. Schiffer, R. L. Girling, K. R. Ely, and A. B. Edmunds, *Biochemistry* **12**, 4620 (1973).
[6] C. Chothia, J. Novotný, R. E. Bruccoleri, and M. Karplus, *J. Mol. Biol.* **186**, 651 (1985).
[7] A. F. Williams and A. N. Barclay, *Annu. Rev. Immunol.* **6**, 381 (1988).
[8] A. Holmgren and C.-I. Brändén, *Nature (London)* **342**, 248 (1989).
[9] J. Novotný, R. E. Bruccoleri, J. Newell, D. Murphy, E. Haber, and M. Karplus, *J. Biol. Chem.* **258**, 14433 (1983).
[10] P. M. Colman, *Adv. Immunol.* **43**, 99 (1988).
[11] T. T. Wu and E. A. Kabat, *J. Exp. Med.* **132**, 211 (1970).
[12] P. de la Paz, B. J. Sutton, M. J. Darsley, and A. R. Rees, *EMBO J.* **5**, 415 (1986).
[13] C. Chothia and A. M. Lesk, *J. Mol. Biol.* **196**, 901 (1987).
[14] C. Chothia, A. M. Lesk, A. Tramontano, M. Levitt, S. J. Smith-Gill, G. Air, S. Sheriff, E. A. Padlan, D. Davies, W. R. Tulip, P. M. Colman, S. Spinelli, P. M. Alzari, and R. J. Poljak, *Nature (London)* **342**, 877 (1989).
[15] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.* **112**, 535 (1977).

is limited by two maj
refine an X-ray data s
lographic analysis (for
tallize, or do not diff
addition to the techni
ing biochemistry can
available in limited qu

Currently, the onl
nuclear magnetic res
technique is applicab
(<20 kDa).[18] Thus, w
importance of antibo
attempts to develop m
and thus circumvent e
and reliable modeling
effects of site-directed
intelligent application
bining site (CDR rep
metal binding sites) a
antigens.

Antibody Modeling–

Various workers h
using, in the main, k
attempts have been m
mational space availab
procedure. Generally t
tion calculated using a

Modeling of antib
classes:

1. modeling panels
   either by crystall
   antibody

[16] R. M. Cooke and I. D. Car
[17] K. Wüthrich, "NMR of Pr
[18] G. M. Clore and A. M. Gr
[19] M. J. Darsley, P. de la Paz
and Exploitation of Antibo
and Analysis Vol. 15 (E. R
1985.
[20] C. R. Mainhart, M. Potter,
[21] M. E. Snow and L. M. Am

is limited by two major factors: the time required to collect, process, and refine an X-ray data set and the intractability of certain proteins to crystallographic analysis (for example, certain proteins prove impossible to crystallize, or do not diffract to acceptable resolutions once crystallized). In addition to the technical complexity of crystallography itself, the supporting biochemistry can be difficult. The protein may be difficult to purify, available in limited quantities, or unstable.

Currently, the only experimental alternative to X-ray techniques is nuclear magnetic resonance (NMR).[16,17] However, at the moment, the technique is applicable only to the solution of relatively small proteins (<20 kDa).[18] Thus, with the experimental limitations and the extreme importance of antibodies in therapy and research, there have been many attempts to develop methods by which to model antibody combining sites and thus circumvent experimental procedures. The availability of accurate and reliable modeling procedures would also allow the prediction of the effects of site-directed mutagenesis (SDM) experiments and allow the intelligent application of SDM as well as larger modifications to the combining site (CDR replacement, introduction of catalytic activity, and metal binding sites) and, eventually, tailoring of combining sites to new antigens.

Antibody Modeling—Approaches to Date

Various workers have attempted to model antibody combining sites using, in the main, knowledge-based approaches.[12,19-24] More recently, attempts have been made to use *ab initio* methods[25-27] where the conformational space available to a loop is saturated, followed by a screening procedure. Generally this involves selecting the lowest energy conformation calculated using an empirically derived potential function.[28,29]

Modeling of antibody combining sites may be divided into three classes:

1. modeling panels of mutant antibodies given a structure (determined either by crystallography or by molecular modeling) for the parent antibody

[16] R. M. Cooke and I. D. Campbell, *BioEssays* **8**, 52 (1988).
[17] K. Wüthrich, "NMR of Proteins and Nucleic Acids." Wiley, New York, 1986.
[18] G. M. Clore and A. M. Gronenborn, *Protein Eng.* **1**, 275 (1987).
[19] M. J. Darsley, P. de la Paz, D. C. Phillips, A. R. Rees, and B. J. Sutton, *in* "Investigation and Exploitation of Antibody Combining Sites" Methodological Surveys in Biochemistry and Analysis Vol. 15 (E. Reid, G. M. W. Cook, and D. J. Morré, eds.), Plenum, New York, 1985.
[20] C. R. Mainhart, M. Potter, and R. J. Feldmann, *Mol. Immunol.* **21**, 469 (1984).
[21] M. E. Snow and L. M. Amzel, *Proteins: Struct. Funct. Genet.* **1**, 276 (1986).

2. modeling insertions, deletions, and CDR replacements given a structure for the parent antibody
3. the larger problem of modeling an unknown antibody structure from its amino acid sequence alone. This involves two stages—building the framework region and building the CDRs

Modeling Single-Site Mutations

This involves the replacement of one or more amino acid residues by others. In modeling such replacements, it is necessary to assess the possibility of backbone conformational changes in addition to side chain placement. The work of Sibanda and Thornton,[30] Thornton *et al.*,[31] and of Greer[32] indicates that changes in loops tend to be accommodated locally and one can thus be reasonably confident that the overall structure of the antibody will not be modified. Support for this assumption also comes from X-ray structures of mutant hemoglobins,[33] T4 lysozyme[34] and crambins,[35] the isomorphous crystallization of mutant proteins,[36] and solution data for phage λ-repressor mutants.[37,38] If the residue replacements are conservative (especially if they are surface residues not making inter-CDR interactions), it may only be necessary to consider the placement of the side chain, but if nonhomologous changes are being made, changes to

[22] E. A. Padlan, D. R. Davies, I. Pecht, D. Givol, and C. Wright, *Cold Spring Harbor Symp. Quant. Biol.* **41,** 627 (1976).
[23] S. J. Smith-Gill, C. R. Mainhart, T. B. Lavoie, R. J. Feldmann, W. Drohan, and B. R. Brooks, *J. Mol. Biol.* **194,** 713 (1987).
[24] E. A. Padlan and E. A. Kabat, *Proc. Natl. Acad. Sci.* **85,** 6885 (1988).
[25] R. M. Fine, H. Wang, P. S. Shenkin, D. L. Yarmush, and C. Levinthal, *Proteins: Struct. Funct. Genet.* **1,** 342 (1986).
[26] R. E. Bruccoleri, E. Haber, and J. Novotný, *Nature (London)* **335,** 564 (1988).
[27] J. Moult and M. N. G. James, *Proteins: Struct. Funct. Genet.* **1,** 146 (1986).
[28] B. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, *J. Comput. Chem.* **4,** 187 (1983).
[29] J. Åqvist, W. F. van Gunsteren, M. Leifonmark, and O. Tapia, *J. Mol. Biol.* **183,** 461 (1985).
[30] B. L. Sibanda and J. M. Thornton, *Nature (London)* **316,** 170 (1985).
[31] J. M. Thornton, B. L. Sibanda, M. S. Edwards, and D. J. Barlow, *BioEssays* **8,** 63 (1988).
[32] J. Greer, *J. Mol. Biol.* **153,** 1027 (1981).
[33] G. Fermi and M. Perutz, *in* "Atlas of Molecular Structures in Biology" (D. C. Phillips and F. M. Richards, eds.). Oxford Univ. Press (Clarendon) London and New York, 1981.
[34] M. G. Grütter, R. B. Hawkes, and B. W. Matthews, *Nature (London)* **277,** 667 (1979).
[35] W. A. Hendrickson and M. M. Teeter, *Nature (London)* **290,** 107 (1981).
[36] M. Knossow, R. S. Daniels, A. R. Douglas, J. J. Skehel, and D. C. Wiley, *Nature (London)* **311,** 678 (1984).
[37] M. H. Hecht, H. C. M. Nelson, and R. T. Sauer, *Proc. Natl. Acad. Sci.* **80,** 2676 (1983).
[38] M. A. Weiss, M. Karplus, D. J. Patel, and R. T. Sauer, *J. Biomol. Struct. Dynamics* **1,** 151 (1983).

backbone conformation
philic residue to a hydr
bury the hydrophobic g

A number of appro
tions. One of the s
(MOP).[12,20,23,38a,39–41] Ar
residue is constructed w
to the atoms in the par
dihedrals are inherited b
subjected to energy mi
protein.

Sidechain replacem
PLACE and REFI com
gram FRODO.[42] REFI
side chain conformatio
"molten atom" or "loc
time such that each m
the atom by decreasing
angles, and dihedral an
lated and applied to ea
moving one atom will a
50 cycles are required t
procedures may be imp
Waltham, MA).

However, when usin
in a local energy minim
perturbation procedure
making the appropriate
performed about the di
conformations are ident
tion.[28] Restraints are a
sumed to be local to the

[38a] The terms MOP (maxim
dure), and CPP (coupled pe
[39] R. J. Feldmann, M. Potter,
[40] M. K. Swenson, A. W. Bu
Biology" (B. Pulman, ed.),
[41] P. Warme, F. A. Momany,
13, 768 (1974).
[42] T. A. Jones, this series, Vol
[43] J. Hermans and J. E. McQu
[44] R. E. Hubard, *Proc. Compu*
[45] H. L. Shih, J. Brady, and M

backbone conformation are likely. For example, changing a surface hydrophilic residue to a hydrophobic one may result in a refolding of the loop to bury the hydrophobic group away from the solvent.

A number of approaches have been used to model single-site mutations. One of the simplest is the maximum overlap procedure (MOP).[12,20,23,38a,39-41] Amino acid replacements are made such that the new residue is constructed with as many atoms as possible in identical positions to the atoms in the parent structure; that is, wherever possible, the parent dihedrals are inherited by the mutated residue. The entire structure is then subjected to energy minimization to obtain a prediction for the mutant protein.

Sidechain replacement by MOP is readily achieved using the REPLACE and REFI commands of the interactive molecular graphics program FRODO.[42] REFI uses an optimization technique to fit a template side chain conformation to the atoms present in the parent structure. This "molten atom" or "local change" approach[43] moves only one atom at a time such that each movement improves the immediate environment of the atom by decreasing the differences from ideality of bond lengths, bond angles, and dihedral angles near this atom. Small displacements are calculated and applied to each atom in turn. The process is thus iterative, as moving one atom will affect the environment of another. Typically, up to 50 cycles are required to achieve convergence to less than 0.01 Å. Similar procedures may be implemented using HYDRA[44] or QUANTA (Polygen, Waltham, MA).

However, when using MOP, it is possible for the structure to be trapped in a local energy minimum resulting in an incorrect model. The minimum perturbation procedure (MPP)[45] helps to overcome this problem. After making the appropriate residue replacement(s), a conformational search is performed about the dihedral angles of the new side chain. Low-energy conformations are identified and subjected to restrained energy minimization.[28] Restraints are applied such that conformational changes are assumed to be local to the area of the residue replacement(s).

[38a] The terms MOP (maximum overlap procedure), MPP (minimum perturbation procedure), and CPP (coupled perturbation procedure) were introduced by Snow and Amzel.[21]
[39] R. J. Feldmann, M. Potter, and C. P. J. Glaudemans, *Mol. Immunol.* **18,** 683 (1981).
[40] M. K. Swenson, A. W. Burgess, and H. A. Scheraga, *in* "Frontiers in Physicochemical Biology" (B. Pulman, ed.), p. 115. Academic Press, New York, 1978.
[41] P. Warme, F. A. Momany, S. V. Rumball, R. W. Tuttle, and H. A. Scheraga, *Biochemistry* **13,** 768 (1974).
[42] T. A. Jones, this series, Vol. 115, p. 157.
[43] J. Hermans and J. E. McQueen, *Acta Crystallogr. Sect. A* **30,** 730 (1974).
[44] R. E. Hubard, *Proc. Comput.-Aided Mol. Des. Conf.* p. 99 (1984).
[45] H. L. Shih, J. Brady, and M. Karplus, *Proc. Natl. Acad. Sci.* **82,** 1697 (1985).

Simple approaches such as MOP and MPP can prove extremely successful, especially when the replacement is conservative and the residue makes few interactions with neighboring residues. Karplus's group[45] cites an example where they have modeled a Gly → Asp mutation in influenza virus hemagglutinin by MPP. They accurately positioned the aspartate side chain in the region indicated by a difference density peak from crystallographic techniques.[36]

The coupled perturbation procedure (CPP)[21] extends the MPP approach by searching the conformational space not only of the replaced residue, but also of a "dependency set." The dependency set for each replaced residue constitutes those residues whose conformations are likely to be affected by the replacement, i.e., those which are capable of making interactions with any conformation of the parent or replacement residue. As in MPP, the structure is then subjected to restrained energy minimization. By using more comprehensive conformational search procedures such as CONGEN,[46] CPP can search side chain conformations throughout the whole structure, effectively extending the dependency set to all residues in the molecule. However, the computer time required for such a search is likely to become prohibitive.

All three methods (MOP, MPP, CPP), however, are likely to prove inadequate if a drastic change in the nature of a residue is made, since none accounts for the possibility of backbone conformational change.

Modeling Insertions, Deletions, and CDR Replacements

As with modeling single-site mutations, such changes are likely to be accommodated locally within the CDR. However, the magnitude of these local changes may be large and it is thus always necessary to consider conformational changes in the backbone of the loop. CDR replacement is thus an extension of the insertion/deletion problem as it is obviously necessary to consider the whole loop conformation. However, in contrast to modeling the entire combining site, the environment in which the loop is constructed is largely defined by the presence of the other five loops of the combining site.

To date, most modeling in this category has relied on the presence of CDRs of the required length in the database of known antibody structures. When these are not present, the closest available CDR has been modified manually using computer graphics.[12,23,24] When loops of the correct length are available, such methods are likely to be quite successful. However, when manual insertions or deletions need to be made, the results are highly

[46] R. E. Bruccoleri and M. Karplus, *Biopolymers* **26**, 137 (1987).

dependent on user int
natively, it has bee
searches,[27,46] not only
vicinity of the replace
sive and again require
ods, below).

Modeling the Framew

When modeling a
alone, it is first nece
straightforward as it is
packing of the V_L and
degree of variability is
analysis of this effect i
unpublished, 1991).

A single known hig
homology to the struct
for maximum overlap
Alternatively, light and
ing structures.[14] V_L anc
sequence homology. If
ies composed of $V_L1/$
formed by least squares
to leave the V_L1/V_H2
V_L1/V_H1.

Problems have beer
influencing the constru
the heavy chain packs
antibody HyHEL-5 the
the equivalent residues
The temperature factor
higher than the mean
identified five residues
distribution: PCA-1H,
suggests these residues
the crystal structure. Th
atoms (N, Cα, C, O)
identified: PCA-1H, Va
poorly defined in the cr

[47] P. M. Alzari, M.-B. Lascon
[48] A. C. R. Martin, D.Phil. Th

dependent on user interaction and their reproducibility is thus low. Alternatively, it has been possible to employ complete conformational searches,[27,46] not only of the side chains, but also of the backbone in the vicinity of the replacement. Such methods are extremely computer intensive and again require user interaction (see Conformational Search Methods, below).

Modeling the Framework

When modeling a completely unknown antibody from its sequence alone, it is first necessary to build the framework. This is relatively straightforward as it is highly conserved,[47] although differences do occur in packing of the V_L and V_H domains with respect to one another. A higher degree of variability is seen around the takeoff region of CDR-H3 and an analysis of this effect is currently under way (J. Pedersen and A. R. Rees, unpublished, 1991).

A single known high-resolution antibody structure with high sequence homology to the structure being modeled may be used as a starting point for maximum overlap (MOP)-type correction of sequence differences.[12] Alternatively, light and heavy chains may be modeled from separate starting structures.[14] V_L and V_H domains are selected separately on the basis of sequence homology. If the two domains are chosen from different antibodies composed of V_L1/V_H1 and V_L2/V_H2, respectively, the composite is formed by least squares fitting V_L1 onto V_L2 before removing V_H1 and V_L2 to leave the V_L1/V_H2 composite which inherits the V_H/V_L packing of V_L1/V_H1.

Problems have been identified with poorly defined framework regions influencing the construction of CDRs.[48] For example, the N terminus of the heavy chain packs against CDR-H3 and in the crystal structure of the antibody HyHEL-5 the first two residues of V_H are placed differently from the equivalent residues in Gloop2 and other antibody crystal structures. The temperature factors of the HyHEL-5 atoms were examined and those higher than the mean plus three standard deviations ($\bar{x} + 3\sigma$) noted. This identified five residues with temperature factors falling outside the normal distribution: PCA-1H, Val-2H, Gln-3H, Arg-40H, and Asp-102H. This suggests these residues are either extremely mobile, or difficult to place in the crystal structure. The analysis was repeated examining just main chain atoms (N, Cα, C, O) and, again, the N-terminal residues of V_H were identified: PCA-1H, Val-2H, and Gln-3H. Thus, these residues seem to be poorly defined in the crystal structure, with the side chains of the first two

[47] P. M. Alzari, M.-B. Lascombe, and R. J. Poljak, *Annu. Rev. Immunol.* **6**, 555 (1987).
[48] A. C. R. Martin, D.Phil. Thesis University of Oxford, Oxford (1990).

residues and the main chain of the third being particularly poor.[49] There is, however, no evidence for any other conformation being more correct.

Main-chain temperature factors should thus be examined to identify poorly defined regions of the structure and such regions should be replaced with consensus conformations from other known crystal structures with lower temperature factors in these regions. An automated framework-building procedure has now been developed (J. Pedersen and A. R. Rees, unpublished, 1991).

Modeling the CDRs of an Entire Antibody Fv

Modeling the CDRs themselves is a more difficult task. By their very nature, these loops are hypervariable, showing, in some cases, extreme variability in sequence, length, and conformation.

Most approaches to the problem so far have used the available CDRs from antibody crystal structures as starting models for those in the antibody to be modeled.[12,13,20,23,24,39,50]

MOP-Based Methods. Procedures based on the MOP method for single-residue replacements have been used independently by at least two groups.[12,20,23] Their approaches differ in the way in which they select loops as starting points. The groups of Feldmann and Smith-Gill[20,23] choose a single antibody on which to model all the hypervariable loops, selecting loops from other antibodies only when the required loop length is not present in that starting antibody structure. In contrast, the group of Rees[12] has used a single, high-resolution framework chosen by sequence homology. Individual loops are selected from a database of the available antibody crystal structures in the Brookhaven Protein Databank,[15] first on length and, if two or more loops of equally suitable length exist, one loop is selected on sequence homology. These loops are then attached to the framework model.

The selected loop is fitted onto the model framework using interactive computer graphics and the sequence of the loop is then corrected using the MOP protocol. In both methods, if no loop of the required length is available, the loop of closest length to the unknown is used and insertions or deletions are made using interactive molecular graphics while attempting to maintain the overall shape of the loop and intraloop hydrogen bonding.

When this procedure has been repeated for each of the six loops and the

[49] S. Sheriff, personal communication.

[50] C. Chothia, A. M. Lesk, M. Levitt, A. G. Amit, R. A. Mariuzza, S. E. V. Phillips, and R. J. Poljak, *Science* **233**, 755 (1986).

side chains of the framev
the whole modeled Fv i
strained[23] or unrestrained
tential.

Key Residue Metho
workers[13,14,50] also exploit
mation is predicted on th
These are residues which
Trp, Tyr, or Phe), can for
Asn, Gln, Asp, Glu, Arg,
tions (e.g., Gly or Pro). As
manual although it is a pr

All Protein Database
restricted size of the anti
only around a dozen str
known immunoglobulin
tions in CDRs when cont
able among the known str
procedures. Such change
computer graphics.[12,23,24]
its repeatability is low. Th
critical in determining th
structures become availat
CDR of identical length a
eled and of identifying a
crease. At the present time
the Brookhaven Protein I
two per year.

One way of expandin
loops from all known pro
database to antibodies. Th
protein loops was propos
developed a method for
COMPOSER) which utili
tures. They first define a st
structural comparison of
body structures). Loop re
structed. If the loop is a ti

[51] A. R. Rees and P. de la Paz, *T*

[52] T. A. Jones and S. Thirup, *EM*

[53] M. J. Sutcliffe, I. Haneef, D. C

side chains of the framework region have been resequenced using MOP, the whole modeled Fv is subjected to energy minimization, either restrained[23] or unrestrained,[12,51] using the GROMOS[29] or CHARMM[28] potential.

Key Residue Method. The approach devised by Chothia and coworkers[13,14,50] also exploits the antibody structure database. Loop conformation is predicted on the basis of the presence of critical or key residues. These are residues which affect loop packing (e.g., bulky residues such as Trp, Tyr, or Phe), can form hydrogen bonds or salt bridges (e.g., Ser, Thr, Asn, Gln, Asp, Glu, Arg, or Lys), or are able to adopt unusual conformations (e.g., Gly or Pro). As used by Chothia *et al.* the method is completely manual although it is a prime candidate for automation.

All Protein Database Method. The above methods are limited by the restricted size of the antibody crystal structure database, which contains only around a dozen structures. When the database is restricted to the known immunoglobulin crystal structures, making insertions and deletions in CDRs when conformations of the required lengths are not available among the known structures is one of the most unreliable parts of the procedures. Such changes are generally made by hand, using interactive computer graphics.[12,23,24] Clearly, such a procedure is highly subjective and its repeatability is low. The site at which insertions or deletions are made is critical in determining the resulting conformation.[48] When more crystal structures become available, the probability of the database containing a CDR of identical length and high sequence homology to that being modeled and of identifying all key residues (as defined by Chothia) will increase. At the present time, however, the number of antibody structures in the Brookhaven Protein Databank is increasing only at the rate of one or two per year.

One way of expanding the database of loop structures is to examine loops from all known protein crystal structures rather than restricting the database to antibodies. The use of nonhomologous structures in modeling protein loops was proposed by Jones and Thirup.[52] Sutcliffe *et al.*[53] have developed a method for modeling loops (implemented in the program COMPOSER) which utilizes a database of all high-resolution crystal structures. They first define a structurally conserved framework region (SCR) by structural comparison of all homologous known structures (here the antibody structures). Loop regions not defined within this SCR are then constructed. If the loop is a tight turn, it is built *ab initio*. Otherwise, geometri-

[51] A. R. Rees and P. de la Paz, *Trends Biochem. Sci.* **11**, 144 (1986).
[52] T. A. Jones and S. Thirup, *EMBO J.* **5**, 819 (1986).
[53] M. J. Sutcliffe, I. Haneef, D. Carney, and T. L. Blundell, *Protein Eng.* **1**, 377 (1987).

cal constraints for the adjoining section of the SCR are determined and loops of appropriate length which satisfy the same constraints in the framework to which they originally were attached are selected from the database. Selection from this group is performed on the basis of structurally determining residues (SDRs). When attaching a loop to the SCR, a sliding weighting scale is used to match the SCR and framework onto which the chosen loop was originally attached.

From analysis of residue replacements in homologous proteins, they also provide a rule for each possible side-chain replacement in sheet, helix, or loop (i.e., a total of 1200 rules).[54] The side-chain χ torsion angles within the SCRs of globins and immunoglobulins were calculated. Their distributions were analyzed and the most probable torsion angles were calculated with their standard deviations. Correlations between side-chain atom positions where substitutions have occurred were examined and these form the basis of the rules. Side chains are built by matching the parent side-chain conformation against the correlated parent positions for the required replacement. If all correlated conformations score equally, the side chain is built up in the most probable conformation. A similar set of rules has been suggested by Karplus's group,[55] although they do not split the rules into the three groups for sheet, helix, and loop. They state that for structurally and functionally similar proteins, there is a high probability that side-chain orientation of the parent can be transferred to the mutant side chain and that the basic rule for this transfer is that maximal occupancy of atom positions in the two structures is maintained. By implication this supports the use of MOP in side-chain replacement. For mutated residues, the transferability is in the range 60–89% for γ atoms and 56–84% for δ atoms. This compares with transferabilities of 60–97% (γ) and 50–90% (δ) for identical residues. Higher degrees of homology between proteins result in higher degrees of transferability, although highly accessible residues appear to have more random orientations and this is important in surface loops.

Stanford and Wu Method. Stanford and Wu[56] proposed a method for generating backbone conformation of the CDRs by analysis of backbone torsion angles for tripeptides observed in other β-sheet proteins. Since several torsion angles are generally possible for each amino acid, 1000–2000 conformations were generated for each CDR. However, this number is reduced by the requirement to link the two takeoff points of the framework onto which the loops must be fitted. The presence of the framework

[54] M. J. Sutcliffe, F. R. F. Hayes, and T. L. Blundell, *Protein Eng.* **1**, 385 (1987).
[55] N. L. Summers, W. D. Carlson, and M. Karplus, *J. Mol. Biol.* **196**, 175 (1987).
[56] J. M. Stanford and T. T. Wu, *J. Theor. Biol.* **88**, 421 (1981).

also applies van der W
loops. An iterative m
end-points of the loop
best fit onto the frame

Random Search
which combines mol
conformations for the
with van der Waals o
molecule. Energy mi
ergy minimization, is
number of random s
space, selecting possib
set of the randomly ge
The results obtained
phosphorylcholine an
well for the shortest lo
residues) especially in
structure. When cons
CDRs the prediction
CDR-L2 in the region
For these shorter loops
is the primary factor
longer loops studied (C
energy minima were f
predicted loops showin
the crystal structure. N
the method means th
selected. At the time c
available to them was i
and CDR-H2, 17 and

Conformational Se
conformational space
tree search.[27,46] A large
ated by rotation about
chain together with sic
lems with this type of
exponentially as the nu
increases. Thus, the co
for longer loops. Bruc

[57] S. Rudikoff, Y. Satow, E.
(1981).

also applies van der Waals constraints on the possible conformations of the loops. An iterative method is applied to adjust the torsion angles to fit the end-points of the loops onto the framework and the conformation with the best fit onto the framework is selected.

Random Search Method. Levinthal's group has suggested a method which combines molecular dynamics and generation of random sets of conformations for the CDRs.[25] These are screened to eliminate structures with van der Waals overlap either within the loop or with the rest of the molecule. Energy minimization, or molecular dynamics followed by energy minimization, is then applied to the system. By generating a large number of random structures, they attempt to saturate conformational space, selecting possible structures for the loop to be modeled by refining a set of the randomly generated structures and selecting loops of low energy. The results obtained by applying this method to modeling the anti-phosphorylcholine antibody McPC603[57] are mixed. The method works well for the shortest loops studied (CDR-H1, five residues; CDR-L2, seven residues) especially in the presence of the other CDRs from the crystal structure. When constructions were made in the absence of the other CDRs the prediction for CDR-H1 was still good, but the prediction for CDR-L2 in the region where it interacts with the other CDRs was poor. For these shorter loops, they find that the length of the loop to be modeled is the primary factor in determining the structure of the loop. For the longer loops studied (CDR-L3, 9 residues; CDR-H3, 11 residues) multiple energy minima were found, some significantly lower than the energy of predicted loops showing minimal root mean square (RMS) deviation from the crystal structure. No actual RMS deviations are cited as the nature of the method means that clusters of similar possible conformations are selected. At the time of publication of these results, the computer power available to them was insufficient to tackle the two longest loops (CDR-L1 and CDR-H2, 17 and 19 residues, respectively) in McPC603.

Conformational Search Methods. An alternative method to saturate conformational space is to generate all possible loop conformations by a tree search.[27,46] A large number of alternative loop conformations is generated by rotation about the dihedral angles, ϕ and ψ, of the peptide main chain together with side-chain χ dihedral angles. One of the major problems with this type of approach is that the size of the tree search grows exponentially as the number of degrees of freedom included in the search increases. Thus, the conformational search procedure becomes impractical for longer loops. Bruccoleri *et al.*[26] have employed methods such as "real

[57] S. Rudikoff, Y. Satow, E. A. Padlan, D. R. Davies, and M. Potter, *Mol. Immunol.* **18,** 705 (1981).

space renormalization"[58-60] to reduce the search time for longer loops. Here, loops are built from both points at which they leave the framework simultaneously. The bottom residues are constructed first and the conformations generated are written into a conformations file or "CG" file. Around 10 low-energy conformations for this pair of residues are read back from the CG file and used on which to build the next pair of residues. The 10 lowest energy conformations of these four residues are, in turn, used to construct the next pair of residues until the size of the loop remaining for construction is small enough for a full conformational search. Alternatively, rather than constructing one residue at a time, fragments of two or three residues may be built. Numerous construction protocols thus become possible in order to construct a long loop. In Bruccoleri and Karplus's procedure, CONGEN,[46] final "chain closure" of the loop over three residues is performed analytically using a modification[61] of Gō and Scheraga's chain closure algorithm.[62] Additionally, cycles of CHARMM energy minimization[28] can be incorporated within the conformational search; this frees the resulting model from being restricted to the step size (generally 30°) used for the conformational search, thus permitting structures which might otherwise be rejected on the basis of van der Waals overlap.

The ϕ and ψ angles for an amino acid are combined into a single degree of freedom and the amount of conformational space explored is restricted by looking up ϕ/ψ combinations in a tabulated form of the Ramachandran plot[63] and examining only those combinations with energies below a specified cutoff. Three such Ramachandran maps are used: one for glycine, one for proline and one for alanine, which is taken as being representative of all the other amino acids. Three residues of the section being constructed are generated using the analytical chain closure algorithm of Gō and Scheraga[62] (as modified by Bruccoleri and Karplus[61]). Thus, for a five-residue section, two residues are constructed by full conformational search while three are constructed by chain closure. Residues constructed by full conformational search will be termed *backbone* while those constructed by chain closure will be termed *chain.*

In order to improve the efficiency of the algorithm, after each combination of *backbone* torsion angles is generated, the distance between the endpoints is checked to ensure that it is possible to span the remaining gap with the number of residues left to construct (when these are in all-trans

[58] H. A. Scheraga, *Biopolymers* **22,** 1 (1983).
[59] M. R. Pincus and R. D. Klausner, *Proc. Natl. Acad. Sci.* **79,** 3413 (1982).
[60] M. R. Pincus, R. D. Klausner, and H. A. Scheraga, *Proc. Natl. Acad. Sci.* **79,** 5107 (1982).
[61] R. E. Bruccoleri and M. Karplus, *Macromolecules* **18,** 2767 (1985).
[62] N. Gō and H. A. Scheraga, *Macromolecules* **3,** 178 (1970).
[63] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, *J. Mol. Biol.* **7,** 95 (1963).

configuration with the
Only if this condition
followed by the *chain*
conformations are rejec
cutoff (typically 20 kcal/
kcal/mol for side chains

Two major problem
apart from the extreme
very good at reproducir
chain orientations[26]—t
achieved for each loop, i
structures is generated
the conformation obser
lowest potential energy s
of solvent accessibility a
in energy by less than 2
ing both conformations
conformation with the
However, this is freque
generated by CONGEN.
ployed, the order in wh
come. In a real modelin
put together will be relat

Results of Antibody Dat

So far, all these meth
structures have become
large deviations betwee
obtained, especially in
the modeled structure[20]
tions of between 1.1 Å (
seen for C_αs and betwe
atoms. Although it is m
quacies in the modeling
crystal packing leads to
of the antibody D1.3[50]
(backbone—N, C_α, C,
RMS of 2.07 Å. (All a

[64] B. K. Lee and F. M. Richar
[65] S. W. Suh, T. N. Bhat, M.
Davies, *Proteins: Struct. Fu*

configuration with the bond angles stretched by 5° from the optimum). Only if this condition is met are further *backbone* residues constructed followed by the *chain* residues and, finally, the side chains. In addition, conformations are rejected with van der Waals energy above a specified cutoff (typically 20 kcal/mol for *backbone,* 100 kcal/mol for *chain,* and 20 kcal/mol for side chains).

Two major problems exist with the conformational search algorithms apart from the extreme expense in CPU time. First, although the method is very good at reproducing crystal structures, including quite precise side-chain orientations[26]—typically, RMS deviations below 2 Å may be achieved for each loop, including side-chains, a large number of low-energy structures is generated from which only one must be selected. Generally the conformation observed in the crystal structure does not represent the lowest potential energy structure. Bruccoleri *et al.*[26] have suggested the use of solvent accessibility as a filter. If the bottom two energy structures differ in energy by less than 2 kcal (three times the Boltzmann factor kT, indicating both conformations are well populated at room temperature) then the conformation with the smaller solvent accessible surface[64] is chosen.[26] However, this is frequently still unable to select the best conformations generated by CONGEN. Second, where real space renormalization is employed, the order in which the segments are constructed affects the outcome. In a real modeling situation any such decisions about how a loop is put together will be relatively arbitrary.

Results of Antibody Database Methods

So far, all these methods have had only limited success; where crystal structures have become available after modeling has been performed, quite large deviations between predicted and observed structures have been obtained, especially in side-chain conformations.[50,65] In a comparison of the modeled structure[20] of J539 with the crystal structure,[65] RMS deviations of between 1.1 Å (CDR-H1 and CDR-L2) and 4.0 Å (CDR-H3) are seen for C_αs and between 2.0 Å (CDR-L1) and 6.5 Å (CDR-H3) for all atoms. Although it is most likely that these differences result from inadequacies in the modeling, the authors cannot rule out the possibility that crystal packing leads to distortions in loop conformation. Chothia's model of the antibody D1.3[50] is better, with RMS deviations of 0.50–0.97 Å (backbone—N, C_α, C, C_β) for five of the six loops, CDR-H1 having an RMS of 2.07 Å. (All atom RMS deviations have not been published; a

[64] B. K. Lee and F. M. Richards, *J. Mol. Biol.* **55,** 379 (1971).
[65] S. W. Suh, T. N. Bhat, M. A. Navia, G. H. Cohen, D. N. Rao, S. Rudikoff, and D. R. Davies, *Proteins: Struct. Funct. Genet.* **1,** 74 (1986).

higher resolution map for D1.3, the map used in the comparison is at 2.8-Å resolution,[66] is required to make a more critical evaluation of the accuracy of side-chain predictions. However, examination of Fig. 1 in Chothia *et al.*[50] suggests that the side-chain predictions in some of the loops may be poor as the C_β positions inferred from peptide plane orientations appear to be badly misplaced. However, it should be noted that, at 2.8-Å resolution, the exact orientation of peptide planes outside regions of defined secondary structure may not be well defined in the electron density map.

In general, when the key residues defined by Chothia are present and loops of identical length are available in the database, the predictions are good, a least on backbone, as indicated above. However, for loops of lengths not represented in the database or lacking the critical residues required for this type of analysis, the predictions tend to be poor. In addition, relatively unconserved residues can have a profound effect on loop packing.[67] A variable residue at the base of CDR-H2 in Gloop2 (an antibody against the loop region of lysozyme), when changed from Glu to Ser, resulted in abolition of binding. Roberts *et al.*[67] argued that, since this residue is relatively buried in the model of the antibody,[12] it is most likely that the Glu50H → Ser mutation disrupts the packing of CDRs in the ACS, thus drastically altering the topology of the combining site surface and hence abolishing antigen binding. Chothia's approach would not identify this residue as being critical in loop packing. The recent crystal structure of Gloop2[68,69] supports this prediction, as Glu-50H is not accessible on the surface of the ACS and is involved in inter-CDR contacts.

A Combined Algorithm

Clearly, none of the methods presented above is wholly satisfactory, although all show a limited amount of success. The ideal modeling procedure may thus involve a combination of the best features of a number of these approaches. Database methods can be used to reduce the search times of conformational search procedures for long loops, thus eliminating the need for real-space renormalization.[60] A number of conformations can be extracted from a database, either by using methods such as that of Sutcliffe *et al.*[53] (see All Protein Database Method, above) or, alternatively, by using distance constraints within the loops themselves defined from the known antibody structures. A conformational search program such as

[66] A. G. Amit, R. A. Mariuzza, S. E. V. Phillips, and R. J. Poljak, *Science* **233**, 747 (1986).
[67] S. Roberts, J. C. Cheetham, and A. R. Rees, *Nature (London)* **328**, 731 (1987).
[68] P. D. Jeffrey, R. E. Griest, G. L. Taylor, and A. R. Rees, manuscript in preparation (1991).
[69] P. D. Jeffrey, D.Phil. Thesis University of Oxford, Oxford (1989).

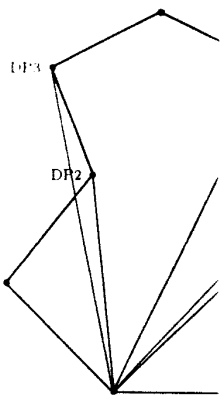


FIG. 1. C_α distances
antibody and Bence-Jon
then used to search the d
mation.

CONGEN[46] may the
best backbone confo
ing only the top of
searching. By remov
from the knowledge
renormalization fro
minimally dependen
Such a method h

Database Searching

In order to ident
seen in antibodies a
antibody being mod
loops are used as sh
from an analysis of k
CDRs, both antibod
body structures were
search the database is
virtually 100% of a n
was chosen in order
distribution and to p

[70] A. C. R. Martin, J. C. C

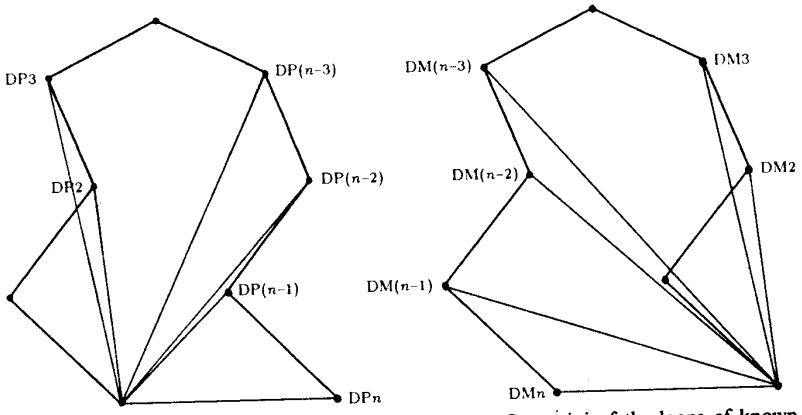


FIG. 1. C_α distances are measured from the N and C termini of the loops of known antibody and Bence-Jones protein structures as shown in the diagram. These distances are then used to search the database of all known protein structures for loops of similar conformation.

CONGEN[46] may then be used to select, from this set of database loops, the best backbone conformations for the bases of the loops, while reconstructing only the top of the loops and the side chains by conformational searching. By removing the requirement for manual insertions or deletions from the knowledge-based methods and the requirement for real space renormalization from the *ab initio* method, a protocol results which is minimally dependent on user choice.

Such a method has been developed[48,70] and is described below.

Database Searching

In order to identify protein loops of the same general shape as loops seen in antibodies and, most importantly, of the required length for the antibody being modeled, sets of inter-C_α distance constraints within the loops are used as shown in Fig. 1. These distance constraints are derived from an analysis of known antibody crystal structures. For the light-chain CDRs, both antibody and Bence-Jones structures were used, while antibody structures were used for the heavy chain CDRs. The range used to search the database is the mean $\pm$ 3.5σ (σ = standard deviation). 3σ covers virtually 100% of a normal distribution; since the sample size is small, 3.5σ was chosen in order not to exclude values falling just outside the current distribution and to prevent the search from being overrestrictive.

[70] A. C. R. Martin, J. C. Cheetham, and A. R. Rees, *Proc. Natl. Acad. Sci.* **86**, 9268 (1989).

A database has been created containing inter-C_α distances for every pair of residues in the range $C_{\alpha_i} \rightarrow C_{\alpha_{i+n}}$ $(-20 \leq n \leq +20)$ within each protein. The database is searched by indexing a column from the database and performing a binary search on the sorted column.

Loop Processing

Having identified a set of loops matching the distance constraints applied at both the N and C termini, redundancies resulting from updated entries in the protein databank and from crystal structures containing multiple copies of a protein fragment are removed. The loops are extracted from the protein databank and are fitted onto the framework (the known crystal structure in the case of single-loop replacements, or a modeled framework where a complete antibody is being modeled). In order to orientate the loops correctly, they are overlapped onto the original loops present on the framework. Overlapping is performed in four stages:

1. The N terminus of the database loop (N_D) is moved onto the N terminus of the framework loop being replaced (N_F).
2. The C terminus of the database loop (C_D) is rotated onto the vector between N_F and the C terminus of the framework loop (C_F).
3. The database loop is translated along the vector between N_F and C_F such that the distance $(N_F - N_D) = (C_F - C_D)$.
4. The database loop is rotated about the $N_D - C_D$ vector such that the plane formed by the center of gravity of its backbone atoms, N_D and C_D, is coplanar with that formed by the center of gravity of the framework loop backbone atoms, N_F and C_F.

This procedure relies on the assumption that the loop takeoff angles defined by the planes through the backbone center of gravity, and the N and C terminal takeoff points, do not vary greatly between different antibodies with loops of different lengths. This has indeed been shown to be the case.[48]

The sequences of the database loops are then corrected to match the loop being modeled by placing side chains using template conformations which are least squares fitted onto the backbone atoms (N, C_α, C, O) of the residue being replaced. Since the side chains are later repositioned using the conformational search program, CONGEN, the actual positioning of side-chain atoms with the exception of the C_β is not critical. CONGEN treats the C_β as a backbone atom, as its position is defined by the orientation of the backbone atoms, N, C_α, C, and O. Thus a side-chain conformational search which first rotates the χ_1 torsion angle ($C_\alpha - C_\beta$) is unable to position the C_β. The position of the C_β is taken from the parent amino acid except when this is a Gly in which case it is taken from the template

conformation. More sop
been proposed[54,55]; how
they were not considered

As with the majority
GEN reduces the compu
implicitly by using "ext
the "heavy" atom to
modified to account for
gens potentially involve
Since coordinates for hy
techniques, such "polar
standard orientations.

The atom order is c
which residues will be c
GEN conformations file

In some cases, the nu
large to process—the co
such cases, the number
manageable number by
using the algorithm of Su
assigned as being struct
moved relative to the po
and the structurally dete
being modeled using th
Sutcliffe[72])

Conformational Searchi

Conformational sea
GEN,[46] which saturates
fragment by rotation ab
sidechain χ angles.

Side chains are cons
starts with an energetic
possible conformations
conformation with the

[71] J. A. McCammon, P. G. W
[71a] J. A. McCammon and S.
182. Cambridge Univ. Press
[72] M. J. Sutcliffe, Ph.D. Thesis
[73] W. C. Barker and M. O. D
Dayhoff, ed.), Vol. 5. Natio
1972.

conformation. More sophisticated side-chain replacement techniques have been proposed[54,55]; however, since only the C_β position is important here, they were not considered worthwhile.

As with the majority of molecular mechanics-based programs, CONGEN reduces the computational burden by treating most hydrogen atoms implicitly by using "extended" (or "united") atoms[71] (i.e., parameters for the "heavy" atom to which one or more hydrogens are attached are modified to account for the presence of the hydrogens). Only those hydrogens potentially involved in hydrogen bonding are treated explicitly.[71a] Since coordinates for hydrogens are not provided by X-ray crystallographic techniques, such "polar" hydrogens are added to the heavy atoms in standard orientations.

The atom order is corrected to a standard order and, having specified which residues will be constructed by conformational searching, a CONGEN conformations file is generated for further processing by CONGEN.

In some cases, the number of loops extracted from the database is too large to process—the computer time required would become excessive. In such cases, the number of database fragments may be reduced to a more manageable number by using a structurally determining residue filter, using the algorithm of Sutcliffe.[72] Each residue in each database fragment is assigned as being structurally determining if it causes the next C_α to be moved relative to the position of that C_α in any of the other database loops and the structurally determining residues are scored against the sequence being modeled using the Dayhoff mutation matrix.[73] (For details, see Sutcliffe[72])

Conformational Searching

Conformational searching is performed using the program CONGEN,[46] which saturates the conformational space available to a peptide fragment by rotation about the backbone ϕ and ψ torsion angles and the sidechain χ angles.

Side chains are constructed using the ITERATIVE algorithm,[46] which starts with an energetically acceptable position for all the side chains. All possible conformations for the first side chain are then regenerated and the conformation with the lowest energy is selected and its energy recorded.

[71] J. A. McCammon, P. G. Wolynes, and M. Karplus, *Biochemistry* **18**, 927 (1979).
[71a] J. A. McCammon and S. C. Harvey, *in* "Dynamics of Proteins and Nucleic Acids," p. 182. Cambridge Univ. Press, Cambridge, 1987.
[72] M. J. Sutcliffe, Ph.D. Thesis University of London Birkbeck College, London (1988).
[73] W. C. Barker and M. O. Dayhoff, *in* "Atlas of Protein Sequence and Structure" (M. O. Dayhoff, ed.), Vol. 5. National Biomedical Research Foundation, Silver Spring, Maryland, 1972.

The process is repeated for each side chain in turn and, when all side chains have been searched, the process starts again from the first side chain until the energy does not change, or an iteration limit is reached. This procedure generates only one energetically feasible side-chain conformation per backbone conformation and has been shown to be the most effective[46] in generating an accurate conformation in a minimum amount of computer time.

In practice, it is rarely practical to construct more than five residues (occasionally six or seven if the conformational flexibility of the region is restricted). As discussed previously, longer regions may be constructed by "real space renormalization"[60] CONGEN treats the extracted set of database loops which have been processed into the form of a CONGEN conformations file as it would the base residues of a peptide being constructed by real space renormalization (see Conformational Search Methods, above).

A five-residue fragment in the middle of each database loop is reconstructed using a 30° search of the torsion angles of the outer two *backbone* residues and the Gō and Scheraga chain closure on the middle three *chain* residues. If this fails to find more than 100 conformations, the torsional search is reduced to 15° or, if necessary, 5°. If the search still fails, additional *backbone* residues are searched until conformations are generated.

It has been shown[48] that when building a complete combining site, the constructions are best performed in the absence of the other loops. If a loop is built between two other previously built loops, cumulative errors can cause the current loop to be built badly. In the case of short loops, however, where the FILTER algorithm (see below) cannot be used, energy is the only criterion on which the screening is performed and these loops must thus be built last in the presence of the other loops.

Energy Screening

CONGEN is built around the CHARMM[28] energy minimization and dynamics package. The work of Bruccoleri in using CONGEN has selected a conformation from those generated by selecting the lowest energy conformation as evaluated using the CHARMM potential *in vacuo*. The *in vacuo* CHARMM potential, however, does not appear to rank the conformations very well—conformations of low RMS deviation compared with the crystal structure frequently do not fall among the lowest energy structures, as shown in Fig. 2. An examination of the low-energy conformations of CDR-H2 of HyHEL-5 using molecular graphics shows the low energy to result from an optimization of the van der Waals packing. In solvent, this attractive effect would be counterbalanced by an attraction of the loop

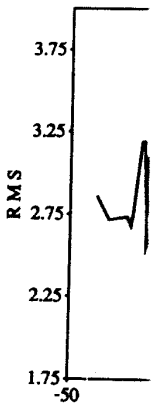


FIG. 2. Root mean sc
energy for each conform
lated using the CHARM
low in energy are not tho

toward the solvent.
simulation of T1 rib
collapse, while in t
retained.

Use of a static rar
tions, by chance, hav
do not. In order to
simulate the waters d
large when it is neces

As an alternative,
packing which occu
Lennard–Jones pote
der Waals radius is l
whatever the distanc
drops very rapidly wit
dynamically (whether
attempting to minimi
into account. Since at
is simply being used
necessary. In addition
a high dielectric const

[74] A. D. Mackerell, Jr., L.

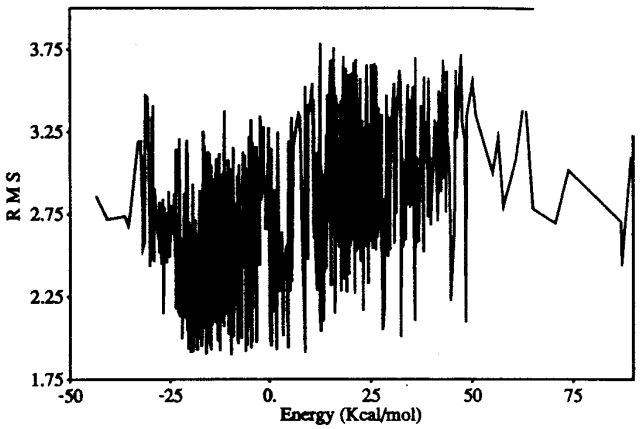


FIG. 2. Root mean square (RMS) deviation from the crystal structure is plotted against energy for each conformation generated for CDR-H2 of HyHEL-5. The energies are calculated using the CHARMM potential *in vacuo*. It is clear that those conformations which rank low in energy are not those of most similar structure to that observed in the crystal structure.

toward the solvent. This is analagous to the situation seen in a dynamics simulation of T1 ribonuclease,[74] where *in vacuo* the active site is seen to collapse, while in the presence of solvent, the native conformation is retained.

Use of a static random water box is not practical since some conformations, by chance, have favorable interactions with the solvent while others do not. In order to overcome this problem, it would be necessary to simulate the waters dynamically and computer time becomes unacceptably large when it is necessary to screen many hundreds of conformations.

As an alternative, in order to prevent the optimization of van der Waals packing which occurs *in vacuo*, the attractive ($-B/r_{ij}^6$) part of the 12-6 Lennard–Jones potential is removed. By doing this, the concept of a van der Waals radius is lost, since all atoms repel each other to some extent, whatever the distance between them [even though this repulsive effect drops very rapidly with distance (A/r_{ij}^{12})]. If the atoms were being simulated dynamically (whether by molecular dynamics, or simply by movements in attempting to minimize the potential energy), this would have to be taken into account. Since atoms are not being allowed to move, but the potential is simply being used to calculate comparative energies, this step is not necessary. In addition, the presence of solvent around the loops will lead to a high dielectric constant (around 50), reducing the importance of electro-

[74] A. D. Mackerell, Jr., L. Nilsson, and R. Rigler, *Biochemistry* **27**, 4547 (1988).

statics. Thus the electrostatic potential is also removed. Similar "solvent-modified" potentials have been used in molecular dynamics of L-arabinose-binding protein,[75] in simulations of carbohydrates[76] and in analysis of misfolded proteins.[77] In the latter case, Novotný *et al.* omit the attractive part of the Lennard–Jones potential for all sulfur atoms and carbon atoms other than those in carbonyl groups, where either of the atoms in the pair being considered has a solvent accessibility greater than 0.1 Å^2. Since the loops of interest in antibody combining sites are on the protein surface and thus the majority of their atoms are solvent exposed, removing the attractive part of the potential for *all* atoms is effective and saves the computation time required to calculate the solvent accessibility. Novotný *et al.* calculate the solvent-modified electrostatic energy by multiplying the atomic charges by a factor dependent on the ratio of the distance of the atom from the center of the closest surface atom. The resulting "dielectric screening factor," first described by Northrup *et al.*[78] decreases the effective charge of an atom linearly from 1.0 in the center of the protein to 0.3 on the surface. For our purposes of modeling surface loops, we have found that removal of the electrostatic interaction altogether results in a greatly improved simulation and the added complexity of accounting more accurately for the center of the protein is unnecessary. This solvent-modified potential has been implemented in a modified version of GROMOS.

The rankings obtained by this solvent-modified potential are subsequently better than those from the *in vacuo* CHARMM potential, with low RMS conformations falling among the low-energy conformations (Fig. 3).

Filtering

In general, however, the lowest energy conformation is still not the lowest RMS conformation observed amongst the bottom five energies. Selection from this group is performed using one of two filtering procedures described below.

Solvent Accessibility. The solvent-modified potential used to rank the conformations considers only the enthalpic term of the free energy—entropy is ignored. In considering the local folding of small fragments of the structure of a protein, the major entropic component is that relating to the hydrophobic effect[79]—the tendency for hydrophobic atoms to pack away from the solvent. By calculating the solvent-accessible surface area of

[75] B. Mao, M. R. Pear, J. A. McCammon, and F. A. Quicho, *J. Biol. Chem.* **257,** 1131 (1982).
[76] K. Bock, M. Meldal, D. R. Bundle, T. Iversen, P. J. Garegg, T. Norberg, A. A. Lindberg, and S. B. Svenson, *Carbohydr. Res.* **130,** 23 (1984).
[77] J. Novotný, A. A. Rashin, and R. E. Bruccoleri, *Proteins: Struct. Funct. Genet.* **4,** 19 (1988).
[78] S. H. Northrup, M. R. Pear, J. D. Morgan, J. A. McCammon, and M. Karplus, *J. Mol. Biol.* **153,** 1087 (1981).
[79] C. Tanford, "The Hydrophobic Effect," 2nd Ed. Wiley, New York, 1980.

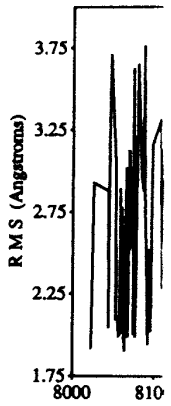


FIG. 3. RMS deviatic
conformation generated
solvent-modified GROM
much improved, with tw
energies.

hydrophobic atoms,
phobics to the solver
for this purpose are
number of possible
all side-chain carbon
ized groups or adjace

However, screeni
ing of atoms which

ATOMS DEFIN
CONFORMATI

Ala-C_α,C_β
Cys-C_α,C_β,S_γ
Asp-C_α,C_β
Glu-C_α,C_β,C_γ
Phe-C_α,C_β,C_γ,$C_{\delta 1}$,
His-C_α,C_β,C_γ,$C_{\delta 2}$,
Ile-C_α,C_β,$C_{\gamma 1}$,$C_{\gamma 2}$,
Lys-C_α,C_β,C_γ,C_δ
Leu-C_α,C_β,C_γ,$C_{\delta 1}$,

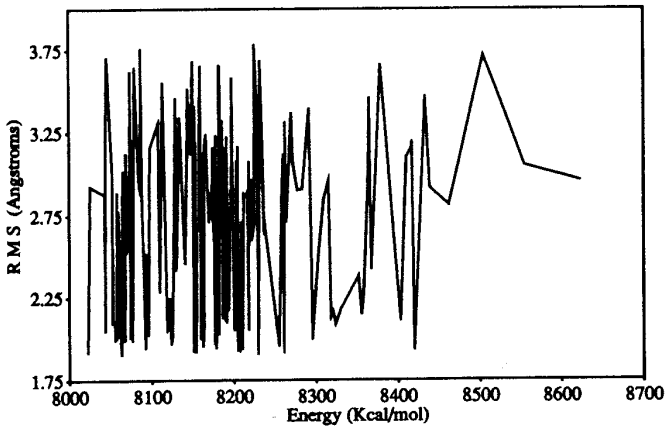


FIG. 3. RMS deviation from the crystal structure is plotted against energy for each conformation generated for CDR-H2 of HyHEL-5. The energies are calculated using the solvent-modified GROMOS potential described in the text. The ranking of conformations is much improved, with two of the lowest RMS conformations appearing in the bottom five energies.

hydrophobic atoms, that conformation with the lowest exposure of hydrophobics to the solvent can be selected. The atoms defined as hydrophobic for this purpose are listed in Table I. This list was compiled by testing a number of possible combinations to find the optimum set and represents all side-chain carbons (from C_α out), excluding those in charged, delocalized groups or adjacent to hydrophilic groups.

However, screening by solvent accessibility is oversensitive to positioning of atoms which are not part of the protein fragment being modeled.

TABLE I

ATOMS DEFINED AS HYDROPHOBIC FOR THE PURPOSE OF SCREENING CONFORMATIONS ON THE BASIS OF THE SOLVENT ACCESSIBILITY OF HYDROPHOBIC ATOMS

Ala-C_α,C_β	Met-C_α,C_β,C_γ,S_δ,C_ϵ
Cys-C_α,C_β,S_γ	Asn-C_α,C_β
Asp-C_α,C_β	Pro-C_α,C_β,C_γ,C_δ
Glu-C_α,C_β,C_γ	Glu-C_α,C_β,C_γ
Phe-C_α,C_β,C_γ,$C_{\delta 1}$,$C_{\delta 2}$,$C_{\epsilon 1}$,$C_{\epsilon 2}$,C_ζ	Arg-C_α,C_β,C_γ
His-C_α,C_β,C_γ,$C_{\delta 2}$,$C_{\epsilon 1}$	Thr-C_α,C_γ
Ile-C_α,C_β,$C_{\gamma 1}$,$C_{\gamma 2}$,$C_{\delta 1}$	Val-C_α,C_β,C_γ,$C_{\delta 1}$,$C_{\delta 2}$
Lys-C_α,C_β,C_γ,C_δ	Trp-C_α,C_β,C_γ,$C_{\delta 1}$,$C_{\delta 2}$,$C_{\epsilon 2}$,$C_{\epsilon 3}$,$C_{\zeta 2}$,$C_{\zeta 3}$,$C_{\eta 2}$
Leu-C_α,C_β,C_γ,$C_{\delta 1}$,$C_{\delta 2}$	Tyr-C_α,C_β,C_γ,$C_{\delta 1}$,$C_{\delta 2}$,$C_{\epsilon 1}$,$C_{\epsilon 2}$

Thus, while it performs well in testing the modeling procedure by reconstructing individual CDRs of a known crystal structure with the other CDRs present[70] and should perform equally well in modeling CDR replacements or mutations, it performs poorly when framework residues have also been modeled and the other CDRs are absent.

In addition, solvent accessibility cannot be used as a screen when loops are being constructed in a combining site with no prior knowledge of the structure of the other loops, since the solvent accessibility is meaningless with no other loops present. These problems led to the development of an alternative filtering method.

Structurally Determining Residues—FILTER. As an alternative to solvent accessibility, a structurally determining residue algorithm has been developed. This is applicable only to loops of six residues or longer, as it requires information from the database search (see later). For each residue of the loop in turn, the database loops are searched for the required residue type and the ϕ and ψ torsion angles are recorded. If the required residue type is not identified in any of the database loops, then the most similar residue types are identified from the Dayhoff mutation matrix and the confidence for the prediction at this residue is reduced by 5%. This procedure is repeated until residues are identified among the database loops.

The ϕ and ψ angles for each of the conformations being screened are then calculated and scored against the torsion angles observed in the database. The terminal residues start with a confidence of 50% since one of the two torsion angles is undefined. Two scoring schemes are implemented. In the first scheme, torsion angles are scored as $1/\theta$, where θ is the difference between model and database torsion angles (in radians) and θ is less than a specified cutoff. This value is then multiplied by the confidence. The second scheme simply scores a 1 for a residue if it has ϕ/ψ angles within the cutoff of any ϕ/ψ angles observed in the database conformations and thus represents the distribution of scores across the loops. Both scoring schemes are examined. If the first scoring scheme is unable to distinguish between the model conformations, the distribution of the scores is also examined—conformations which have their score distributed across the amino acids are chosen in preference to ones which score highly on only a few residues. In addition, if the conformation scoring highest in scheme 1 does not have a high distribution, this selection is rejected and a new angular cutoff is used.

The angular cutoff is initially chosen at 15°. If this is unable to select a conformation, the cutoff is increased to 30° and if this is still unable to distinguish between the models, the cutoff is further increased to 45°. This protocol is based on the ability of the different angular cutoffs to distinguish correctly between the low-energy conformations. Figure 4 shows the

ability of different angu
algorithm still fails to s
that all the low-energy
energy conformation n

Variations to the Proce

For loops of six or
constructed from the c
extracted from the data
backbone of a loop app
crystal structures. Conf
sary and is used only to
ing of the side chains gi
of the MOP protocol.

For loops of five re
constraints means that
is selected from the da
impractical and thus C
database is not searche
cases, it has been showr

While the combine
computer intensive. Cl

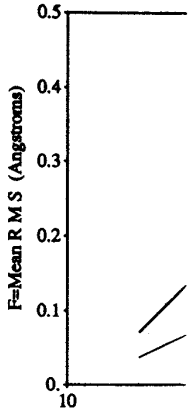


FIG. 4. Ability of the FI
conformations of Gloop2 an
$F = \sum(R_{sel} - R_{min})/n$, the m
mation and the conformatio
FILTER is unable to select a
the distribution selects agains

ability of different angular cutoffs to select the correct conformation. If the algorithm still fails to select a unique conformation, it is a fair indication that all the low-energy conformations are extremely similar and the lowest energy conformation may be selected.

Variations to the Procedure

For loops of six or seven residues, only one or two residues will be constructed from the database. In these cases, several hundred loops are extracted from the database and the conformational space available to the backbone of a loop appears to be well saturated in the current database of crystal structures. Conformational search of the backbone is thus unnecessary and is used only to construct the side chains. Conformational searching of the side chains gives better fits to the crystal structure than does use of the MOP protocol.

For loops of five residues or shorter, the limited number of distance constraints means that an enormous number of conformations (> 10,000) is selected from the database. Processing this number of loops becomes impractical and thus CONGEN is used alone in these cases. Because the database is not searched, the FILTER algorithm cannot be used. In these cases, it has been shown that the lowest energy structure is acceptable.

While the combined algorithm performs very well, it is extremely computer intensive. Chothia's approach of defining canonical ensembles

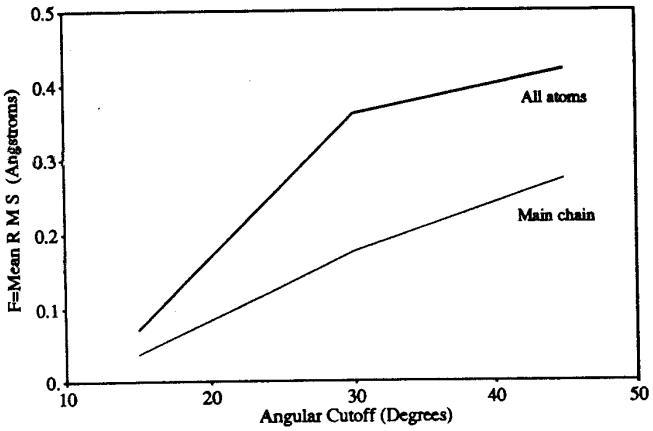


FIG. 4. Ability of the FILTER algorithm to distinguish correctly from the low-energy conformations of Gloop2 and HyHEL-5 CDRs at different angular cutoffs. For each cutoff, $F = \sum(R_{sel} - R_{min})/n$, the mean difference in RMS deviation between the selected conformation and the conformation with the minimum RMS deviation, is plotted. Cases where FILTER is unable to select a conformation are excluded, as are those where consideration of the distribution selects against the highest scoring conformation.

into which the CDRs fit requires very little computer time, but is limited to those CDRs which fit the currently defined set of canonicals. Where these structures do exist, the method works well (especially on backbone conformation) and the two approaches may successfully be combined. Thus a model may be produced by constructing those loops which fit canonicals using Chothia's approach, the remaining loops being built by the combined algorithm.

Summary

The overall protocol may be summarized as follows:

1. Construct the framework on the basis of sequence homology, rejecting sections with unusually high temperature factors and selecting them from other structures.
2. For loops of five residues or shorter, construct the loop using CONGEN. Go to step 8.
3. For loops of six residues or longer, search the distance database for backbone conformations.
4. Overlap the database loops onto the framework.
5. Correct the sequences.
6. Add explicit hydrogens.
7. For loops of eight or more residues, delete the midsection of the loop and reconstruct with CONGEN.
8. Reconstruct side chains with CONGEN.
9. Calculate the energy for each conformation with the solvent-modified potential.
10. Screen the bottom five energy conformations with the FILTER algorithm.

The procedures described for modeling by the combined algorithm (independent of both CHARMM and GROMOS) will be made available in an integrated modeling package from Oxford Molecular Limited (Oxford, England).

Gloop2—A Case Study

Gloop2[80] is a monoclonal antibody raised against the loop peptide region of lysozyme and selected to cross-react with the native protein. This antibody has been the target of numerous site-directed mutagenesis experiments[67,81,82] (S. Roberts, *et al.* unpublished, 1991) and modeling exercises[12,48,70] (C. Chothia *et al.*, unpublished). The crystal structure has be-

[80] M. J. Darsley and A. R. Rees, *EMBO J.* **4**, 393 (1985).
[81] S. Roberts, D.Phil. Thesis University of Oxford, Oxford (1988).
[82] K. Hilyard, D.Phil. Thesis University of Oxford, Oxford (1991).

	Se
LFR1	D I Q M T Q
L1	R A S Q E I
LFR2	W L Q Q K P
L2	A A S T L D
LFR3	G V P K R F
L3	L Q Y L S Y
LFR4	F G A G T K
HFR1	Q V Q L Q Q
H1	T F G I T
HFR2	W V K Q R T
H2	E I F P G N
HFR3	K A T L T A
H3	E I R Y
HFR4	W G Q G T T

[a] Light-chain CDRs ar
the light-chain frame
those from the heavy
[b] The section of this C
modeling exercise.

come available,[68,69] al
be tested.

The Gloop2 struct
described above and
used. The Gloop2 V_L
model each loop is sh
loops are all built on
the two short CDRs,
other four loops. The
tions encloses the re
region used for chain
come from database
are constructed by C(

Framework

The V_L and V_H d
built separately from

[83] O. Epp, P. Colman, H.
J. Biochem. **45**, 513 (19
[84] S. Sheriff, E. W. Silverto
D. R. Davies, *Proc. Natl*

TABLE II
SEQUENCE OF THE V_L REGION OF GLOOP2[a]

LFR1	D I Q M T Q S P S S L S A S L G E R V S L T C
L1	R A S Q E I S G Y L S
LFR2	W L Q Q K P D G T I K R L I Y
L2	A A S T L D S
LFR3	G V P K R F S G R R S G S D Y S L T I S S L E S E D F A D Y Y C
L3	L Q Y L S Y P L T
LFR4	F G A G T K L E
HFR1	Q V Q L Q Q S G T E L A R P G A S V R L S C K A S G Y T F T
H1	T F G I T
HFR2	W V K Q R T G Q G L E W I G
H2	E I F P G N S K T Y(Y A E R F K G)[b]
HFR3	K A T L T A D K S S T T A Y M Q L S S L T S E D S A V Y F C A R
H3	E I R Y
HFR4	W G Q G T T L T V

[a] Light-chain CDRs are named L1, L2, and L3; heavy-chain CDRs are H1, H2, and H3; the light-chain framework regions are named LFR1, LFR2, LFR3, and LFR4 while those from the heavy chain are named HFR1, HFR2, HFR3, and HFR4.
[b] The section of this CDR enclosed in parentheses was considered as framework for the modeling exercise.

come available,[68,69] allowing some of the predictions from the modeling to be tested.

The Gloop2 structure has been modeled using the combined algorithm described above and will be used as a case study to describe the approach used. The Gloop2 V_L sequence is shown in Table II. The protocol used to model each loop is shown in Table III and will be described in turn. The loops are all built onto an "empty" combining site with the exception of the two short CDRs, H1 and H3, which are built last in the presence of the other four loops. The notation used in Table III to describe the constructions encloses the region built by CONGEN in square brackets and the region used for chain closure in parentheses. Residues not thus enclosed come from database structures. In all cases, the side chains of all residues are constructed by CONGEN.

Framework

The V_L and V_H domains of the framework for the Gloop2 model were built separately from REI[83] and HyHEL-5,[84] respectively. The β strands of

[83] O. Epp, P. Colman, H. Fehlhammer, W. Bode, M. Schiffer, R. Huber, and W. Palm, *Eur. J. Biochem.* **45**, 513 (1974).
[84] S. Sheriff, E. W. Silverton, E. A. Padlan, G. H. Cohen, S. J. Smith-Gill, B. C. Finzel, and D. R. Davies, *Proc. Natl. Acad. Sci.* **84**, 8075 (1987).

TABLE III
CONSTRUCTION OF GLOOP2 USING THE COMBINED ALGORITHM

Loop	Sequence	Number of conformations	CPU time (hr)	Global fit RMS Deviation			
				$P2_1$		$P1$	
				All (Å)	C_α (Å)	All (Å)	C_α (Å)
L1	RAS[Q(EIS)G]YLS	1368	59 + 25	1.90	0.92	2.09	0.86
L2	AASTLDS	157	4 + 1	1.25	0.68	1.10	0.66
L3	LQ[Y(LSY)P]LT	721	24 + 14	2.08	0.76	2.00	0.75
H1	[T(FGI)T]	375	18 + 8	2.06	0.79	2.04	1.03
H2	EI[F(PGN)S]KTY	3066	123 + 57	2.51	1.35	2.23	1.20
H3	[R(EIR)Y]	144	7 + 3	2.17	1.00	1.76	0.76
Overall			235 + 108	2.11	1.23	1.96	1.25

[a] For each of the six CDRs of Gloop2, the construction protocol is indicated together with the number of conformations generated, the computer time required for the conformational search and the energy screening (on a VAX 3200), and global RMS deviations compared with the $P2_1$ and $P1$ crystal structures. These RMS deviations were calculated by fitting the structures on the framework regions and calculating the deviation over the CDRs on all atoms and C_αs. The all-atom RMS deviations drop between 0.02 and 0.55 Å while the C_α deviations drop between 0.03 and 0.44 Å when a local RMS fit is calculated.

REI were fitted to the β strands of the HyHEL-5 V_L, using the following residue ranges (numbered consecutively):

REI	HyHEL-5
34–39	33–38
43–47	42–46
84–90	83–89
100–104	98–102

The V_L domain of HyHEL-5 was then deleted to leave a hybrid of REI V_L and HyHEL-5 V_H. Temperature factors of the main-chain atoms (N, C_α, C, O) of the REI and HyHEL-5 domains were examined as described above. The N-terminal residues of HyHEL-5 V_H were identified as having high-temperature factors (greater than the mean plus three standard deviations): PCA-1H, Val-2H, and Gln-3H. These residues were replaced with the equivalent conformation from J539.[65] Finally, side-chain replacements were performed with the REPLACE and REFI functions of the interactive molecular graphics program FRODO.[42]

CDR-L1

CDR-L1 is 11 residues long. The middle five residues were constructed using CONGEN, the remaining residues coming from database loops. The

FILTER algorithm (T
mations at any angul
mations are very sim
lected.

CDR-L2

CDR-L2 is seven r
this length, the main-
rated by the available
side chains were built
were ranked with the
rithm was used to sele
conformation 9 scores
(distribution = 1); sim
but ranks lowest on di
both score and distribu

CDR-L3

CDR-L3 is nine r
structed using CONG
loops. Conformation 8
the seventh position w
cis conformation.

CDR-H1

CDR-H1 is only fi
constraints is thus toc
sands of conformatio
loop was constructed
means the FILTER al
only available criterio
that the other loops
CDR-H1 (and CDR-H
the presence of, CDH
construction was per
expected, a worse con
1.42 Å for N, C_α,
the presence of the ot
C_α, C).

ALGORITHM

Global fit RMS Deviation			
$P2_1$		$P1$	
Å)	C_α (Å)	All (Å)	C_α (Å)
0	0.92	2.09	0.86
5	0.68	1.10	0.66
8	0.76	2.00	0.75
6	0.79	2.04	1.03
1	1.35	2.23	1.20
7	1.00	1.76	0.76
11	1.23	1.96	1.25

ndicated together with the number
ıe conformational search and the
ıs compared with the $P2_1$ and $P1$
ng the structures on the framework
oms and C_αs. The all-atom RMS
ons drop between 0.03 and 0.44 Å

V_L, using the following

eave a hybrid of REI V_L
ain-chain atoms (N, C_α,
examined as described
were identified as having
lus three standard devia-
dues were replaced with
side-chain replacements
nctions of the interactive

esidues were constructed
from database loops. The

FILTER algorithm (Table IV) is unable to distinguish between the conformations at any angular cutoff, suggesting that all the low-energy conformations are very similar. Thus the lowest energy conformation was selected.

CDR-L2

CDR-L2 is seven residues long. It has been shown[48] that, for loops of this length, the main-chain conformational space is sufficiently well saturated by the available database of protein crystal structures. Thus only the side chains were built using the conformational search; the conformations were ranked with the solvent-modified potential and the FILTER algorithm was used to select a final conformation (Table IV). At the 15° cutoff, conformation 9 scores highest, but this score is achieved on a single residue (distribution = 1); similarly at 30°, conformation 9 has the highest score, but ranks lowest on distribution; at 45°, conformation 93 ranks highest on both score and distribution and is thus the conformation selected.

CDR-L3

CDR-L3 is nine residues long. The middle five residues were constructed using CONGEN, the remaining residues coming from database loops. Conformation 82 was selected using FILTER at 45°. The proline at the seventh position within the loop is correctly predicted in the unusual cis conformation.

CDR-H1

CDR-H1 is only five residues long. The number of available distance constraints is thus too small to perform a database search (tens of thousands of conformations are extracted from the database) and the whole loop was constructed using CONGEN. The absence of the database search means the FILTER algorithm cannot be used and, since energy is thus the only available criterion on which to select a conformation, it is important that the other loops are present in order to calculate the energy. Thus CDR-H1 (and CDR-H3, which is also short) was constructed after, and in the presence of, CDR-L1, CDR-L2, CDR-L3, and CDR-H2. A repeat construction was performed in the absence of the other loops and, as expected, a worse conformation was selected (RMS: 3.21 Å for all atoms, 1.42 Å for N, C_α, C) compared with the conformation selected in the presence of the other loops (RMS: 2.18 Å for all atoms, 1.24 Å for N, C_α, C).

TABLE IV
LOW-ENERGY CONFORMATIONS FOR THE SIX CDRS OF GLOOP2 CONSTRUCTED ONTO THE EMPTY COMBINING SITE[a]

CDR	Conformation	Energy (kcal/mol)	RMS (Å) All	RMS (Å) B/B	FILTER score 15°	FILTER score 30°	FILTER score 45°	Identifier
L1	247*	5975.5	2.94	2.04	1.36(3)	2.60(6)	2.56(6)	1REI-A-0024
	319	5975.7	2.97	2.11	1.36(3)	2.22(5)	2.62(7)	1REI-A-0024
	320	5978.2	2.96	2.09	1.36(3)	2.51(6)	2.62(7)	1REI-A-0024
	241	5985.9	2.64	1.88	1.36(3)	2.60(6)	2.56(6)	1REI-A-0024
	253	5986.0	2.99	2.11	1.36(3)	2.60(6)	2.56(6)	1REI-A-0024
L2	43	5273.0	4.22	2.56	1.92(2)	2.97(2)	9.48(4)	1PYP-0-0109
	93*	5297.9	1.51	1.08	0.27(1)	9.25(5)	26.23(7)	2RHE-0-0051
	11	5300.6	1.86	1.33	3.52(5)	5.81(5)	21.58(7)	1FB4-L-0049
	109	5358.7	2.26	1.33	2.22(3)	5.08(4)	7.81(5)	2YHX-0-0383
	9	5366.7	2.06	1.33	6.64(1)	12.45(2)	26.06(6)	1CN1-A-0143
L3	98	5816.3	3.67	1.95	1.61(3)	2.82(4)	3.40(4)	1REI-A-0089
	84	5816.7	3.68	1.88	1.61(3)	2.82(4)	3.40(4)	1REI-A-0089
	107	5816.9	3.23	2.04	1.61(3)	2.82(4)	3.71(5)	1REI-A-0089
	82*	5818.4	2.82	1.99	1.61(3)	2.82(4)	3.87(6)	1REI-A-0089
	96	5818.9	3.70	1.94	1.61(3)	2.82(4)	3.40(4)	1REI-A-0089
H1	14*	6228.4	2.18	1.24	—	—	—	—
	12	6233.0	2.30	1.40	—	—	—	—
	16	6238.7	2.21	1.54	—	—	—	—
	4	6239.4	2.37	1.40	—	—	—	—
	11	6239.8	2.28	1.53	—	—	—	—
H2	240	5559.3	3.81	1.87	0.63(2)	2.28(5)	2.72(5)	1FB4-H-0050
	463*	5559.7	3.00	1.06	2.20(5)	3.42(6)	4.51(7)	1FBJ-H-0050
	1657	5566.7	2.50	1.10	0.96(3)	2.71(5)	4.10(6)	2HFL-H-0050
	471	5567.9	3.04	1.07	1.41(4)	2.84(6)	4.26(7)	1FBJ-H-0050
	722	5568.8	3.78	1.63	0.43(1)	1.47(3)	2.66(5)	1IG2-H-0050
H3	128*	7495.3	2.36	1.45	—	—	—	—
	125	7191.4	2.38	1.59	—	—	—	—
	41	7495.9	4.01	3.03	—	—	—	—
	120	7500.6	2.26	1.52	—	—	—	—
	138	7509.7	2.85	1.49	—	—	—	—

[a] FILTER scores show the normal score with the distribution (i.e., the number of residues over which the score is achieved) in parentheses. RMS deviations are calculated against the $P2_1$ crystal structure by least squares fitting over the framework and are quoted over the regions constructed. The identifier refers to the parent database loop from which the conformation was built. The first four characters are the PDB code, followed by the chain identifier (0 if no identifier) and the residue number of the first residue of the loop.

CDR-H2

CDR-H2, as define
EIFPGNSKTYYAER
definition[13] of just fou
The latter definition
affect antigen binding
excessive since it inclu
sheet and part of the l
here, and recommend
bined algorithm, cons
minal residue defined
loop. For Gloop2, C
EIFPGNSKTY). Conf
15°. The all-atom RM
the Phe at position 3 i
120° when compared
solved in two separate
the *P1* structure, the
atom RMS drops to 2.

This concerted side
effects of crystallizatio
though surface side ch
a result of crystal pack
high. In the Gloop2 *P*
for the Fv domain is
residues of H2 show n
Phe at position 3 and
they are relatively imm

CDR-H3

CDR-H3 is only fo
short length precludes
construct five residue
possible to construct fc
residue was built at th
algorithm could not be

Summary

Each of the six CD
Fig. 5.

)OP2 CONSTRUCTED

score		
	45°	Identifier
)(6)	2.56(6)	1REI-A-0024
':(5)	2.62(7)	1REI-A-0024
(6)	2.62(7)	1REI-A-0024
)(6)	2.56(6)	1REI-A-0024
)(6)	2.56(6)	1REI-A-0024
7(2)	9.48(4)	1PYP-0-0109
5(5)	26.23(7)	2RHE-0-0051
1(5)	21.58(7)	1FB4-L-0049
8(4)	7.81(5)	2YHX-0-0383
5(2)	26.06(6)	1CN1-A-0143
2(4)	3.40(4)	1REI-A-0089
2(4)	3.40(4)	1REI-A-0089
2(4)	3.71(5)	1REI-A-0089
2(4)	3.87(6)	1REI-A-0089
2(4)	3.40(4)	1REI-A-0089
–	—	—
–	—	—
–	—	—
–	—	—
–	—	—
28(5)	2.72(5)	1FB4-H-0050
42(6)	4.51(7)	1FBJ-H-0050
71(5)	4.10(6)	2HFL-H-0050
84(6)	4.26(7)	1FBJ-H-0050
17(3)	2.66(5)	1IG2-H-0050
—	—	—
—	—	—
—	—	—
—	—	—
—	—	—

, the number of residues over which
ated against the $P2_1$ crystal structure
over the regions constructed. The
nformation was built. The first four
(0 if no identifier) and the residue

CDR-H2

CDR-H2, as defined by Wu and Kabat,[11] is 17 residues long (sequence: EIFPGNSKTYYAERFKG). This contrasts with the Chothia and Lesk definition[13] of just four residues at the top of the loop (sequence: PGNS). The latter definition would ignore residues such as Glu-50H, known to affect antigen binding (see above). The Wu and Kabat definition seems excessive since it includes, on the C-terminal side, a complete strand of β sheet and part of the loop at the back of the β barrel. The definition used here, and recommended for all future modeling exercises using the combined algorithm, consists of the Kabat and Wu loop, but with the C-terminal residue defined as the strand partner of the N-terminal residue of the loop. For Gloop2, CDR-H2 is thus defined as 10 residues (sequence: EIFPGNSKTY). Conformation 463 was selected using a FILTER cutoff of 15°. The all-atom RMS deviation is poor (3.00 Å), owing to rotations of the Phe at position 3 in the loop and Tyr at position 10 by approximately 120° when compared with the $P2_1$ crystal structure. Gloop2 has been solved in two separate crystal forms, $P2_1$ and *P1*.[68,69] When compared with the *P1* structure, the side chains are placed almost perfectly and the all-atom RMS drops to 2.23 Å (global fit).

This concerted side-chain motion between crystal forms illustrates the effects of crystallization conditions on surface side-chain placement. Even though surface side chains may show relatively low temperature factors as a result of crystal packing interactions, their mobility in solution may be high. In the Gloop2 *P1* structure, the mean side-chain temperature factor for the Fv domain is 13.46 ($\sigma = 8.20$) while the side chains of these two residues of H2 show mean temperature factors of 5.56 ($\sigma = 0.68$) for the Phe at position 3 and 710 ($\sigma = 1.73$) for the Tyr at position 10, suggesting they are relatively immobile.

CDR-H3

CDR-H3 is only four residues long. As was the case with CDR-H1, this short length precludes the use of the database. It was actually necessary to construct five residues rather than four, using CONGEN, as it was not possible to construct four residues by conformational search. Thus an extra residue was built at the N terminus of the loop. Once again, the FILTER algorithm could not be used and conformation 128 was selected on energy.

Summary

Each of the six CDRs of Gloop2 is shown with the modeled structure in Fig. 5.

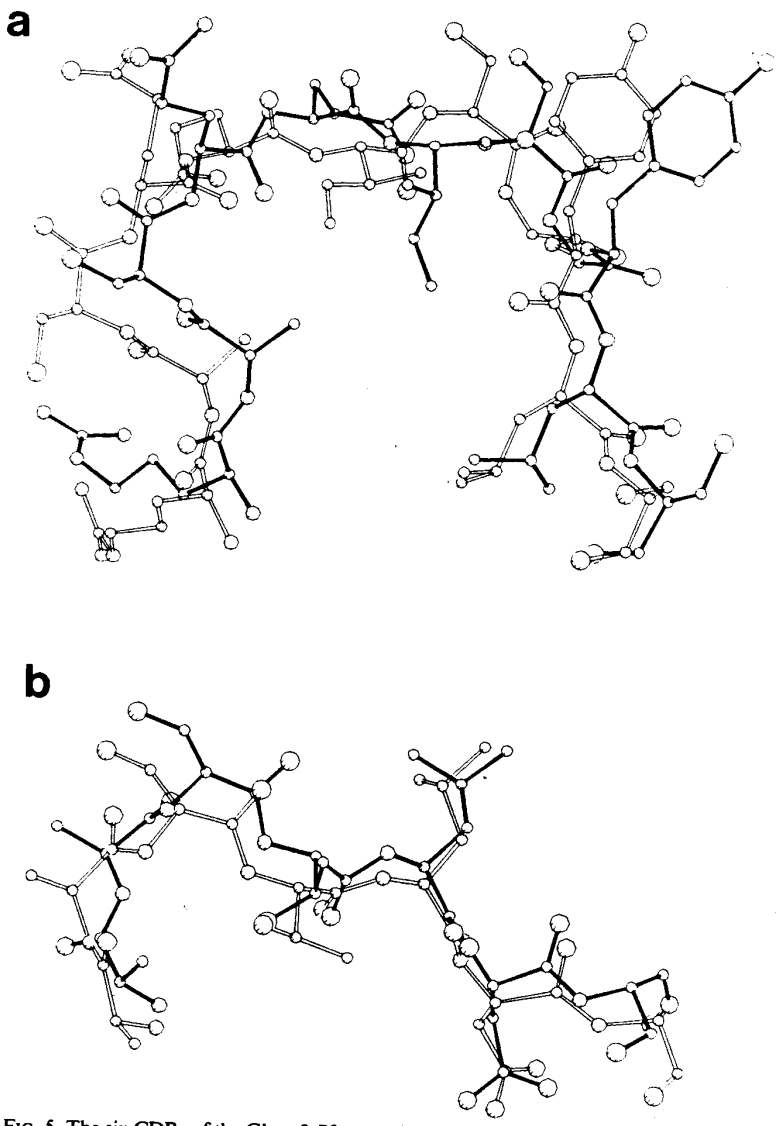
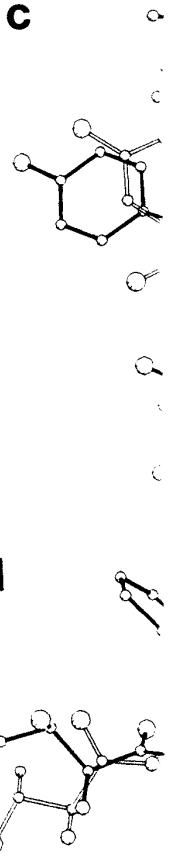
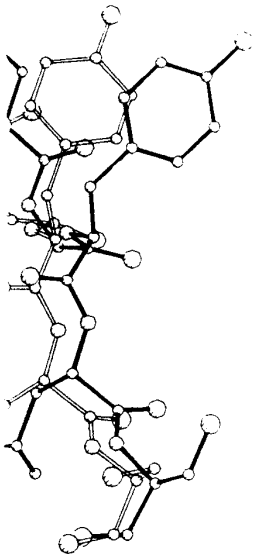
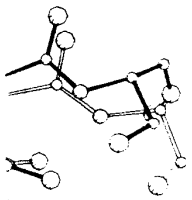


FIG. 5. The six CDRs of the Gloop2 $P2_1$ crystal structure are shown in filled lines with the modeled loops generated by the combined algorithm shown in open lines. All the fits are global (i.e., the framework regions are fitted rather than a local fit being performed on the CDRs themselves). (a) CDR-L1; (b) CDR-L2; (c) CDR-L3; (d) CDR-H1; (e) CDR-H2; (f) CDR-H3.



Overall, the results
in accuracy to those a
However, the canonic
residues identified by
structures currently a
canonical ensembles.
remaining CDRs do
the protein engineer r
the use of Chothia's n





re shown in filled lines with the
in open lines. All the fits are
cal fit being performed on the
(d) CDR-H1; (e) CDR-H2; (f)

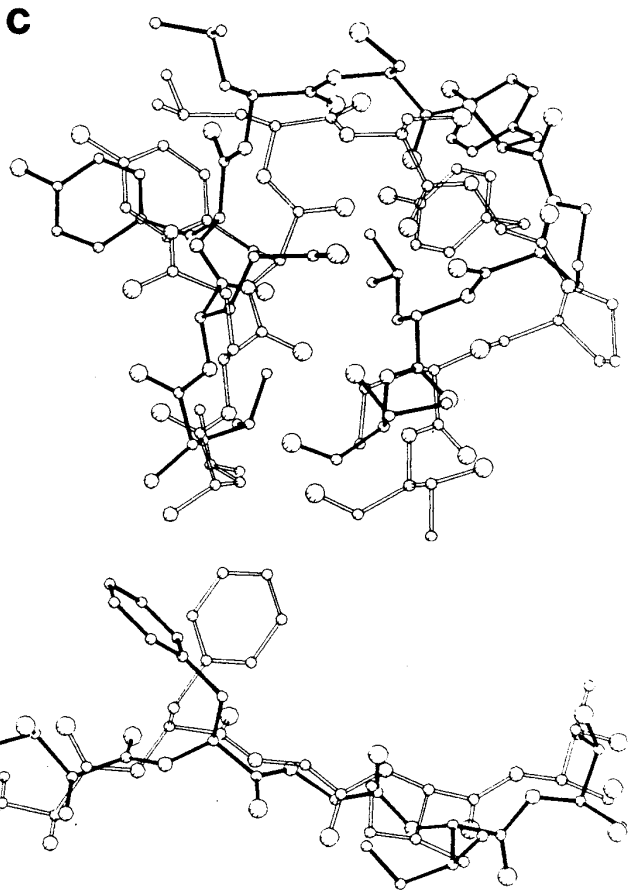


FIG. 5c, d

Overall, the results obtained using the combined algorithm are similar in accuracy to those achieved using the canonical method of Chothia *et al.* However, the canonical method is limited to those loops where the key residues identified by Chothia are present. With the number of antibody structures currently available, it is not possible to classify CDR-H3 into canonical ensembles. Additionally, a small percentage of examples in the remaining CDRs do not match the current canonical classifications and the protein engineer may well wish to mutate the key residues, precluding the use of Chothia's method for modeling the resulting conformation.

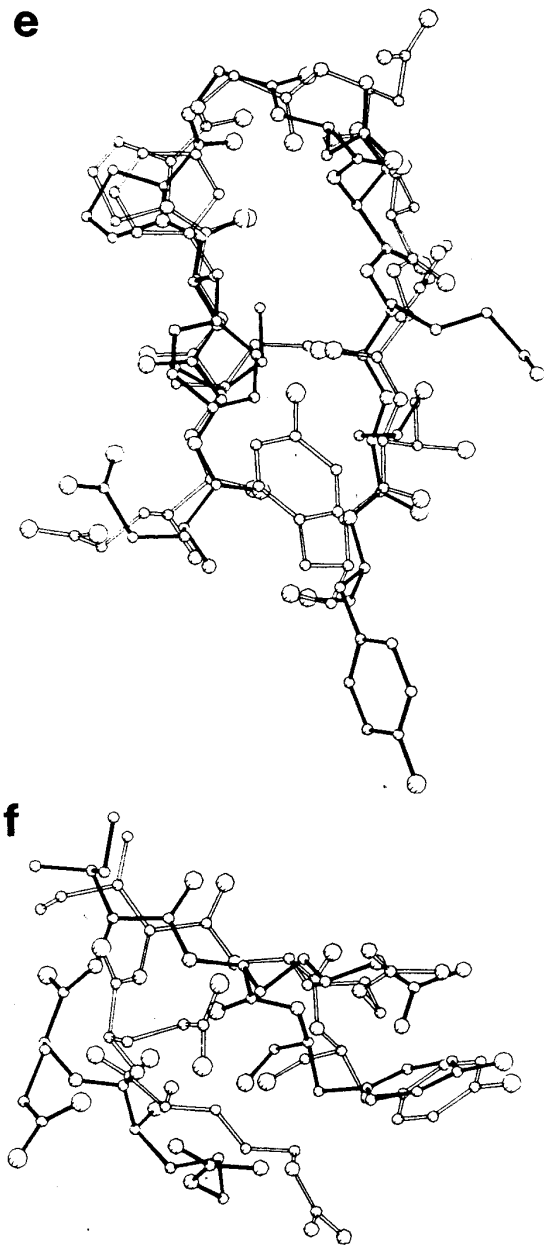


FIG. 5e, f

Thus the best appro
to model the backbon
represented in the data
unrepresented among
where mutations have
eled by the combined a

Acknowledgments

The authors would like t
tions were performed, Phil J
and Gloop2, and Bob Bruce
SERC, UK, for financial sup

[7] X-Ray Cr
Antigen-Comp
Structura

By IAN A. WIL
GAIL

Introduction

The development
duce unlimited numb
given antigen. This op
detailed structural infc
and has driven the pre
recognition. The struct
solved both as free Fat
determinations, comb
from multiple myelon
edge of antibody–anti
The rapid increase

[1] G. Kohler and C. Milstein
[1a] Throughout this review,
for Fab or Fab' fragments
distinct from Fab will be

Thus the best approach appears to be to use Chothia's method (at least to model the backbone conformation) when the loop to be modeled is represented in the database of canonical structures. Any other loops, either unrepresented among the known canonicals (including CDR-H3), or where mutations have been made to the key residues, may then be modeled by the combined algorithm presented here.

Acknowledgments

The authors would like to thank Celltech for providing a V53000 on which the computations were performed, Phil Jeffrey and Steven Sheriff for providing coordinates of HyHEL-5 and Gloop2, and Bob Bruccoleri for providing CONGEN and its source code. We also thank SERC, UK, for financial support.

[7] X-Ray Crystallographic Analysis of Free and Antigen-Complexed Fab Fragments to Investigate Structural Basis of Immune Recognition

By IAN A. WILSON, JAMES M. RINI, DAVED H. FREMONT, GAIL G. FIESER, and ENRICO A. STURA

Introduction

The development of hybridoma technology[1] made it possible to produce unlimited numbers of antibodies of known specificity against any given antigen. This opportunity led to an upsurge of interest in obtaining detailed structural information concerning antibody–antigen recognition and has driven the present investigations of the structural basis of immune recognition. The structures of many monoclonal antibodies have now been solved both as free Fabs[1a] and as Fab–antigen complexes. These structure determinations, combined with four crystal structures of Fabs derived from multiple myeloma antibodies, have markedly increased our knowledge of antibody–antigen interactions.

The rapid increase in the number of free and antigen-bound antibody

[1] G. Kohler and C. Milstein, *Nature (London)* **256,** 495 (1975).

[1a] Throughout this review, the term Fab will frequently be used as the general terminology for Fab or Fab′ fragments of antibodies. In specific instances, where appropriate, Fab′ as distinct from Fab will be used.