Protein Science (1995), *4*:2168–2178. Cambridge University Press. Printed in the USA.

Homology modeling of an immunoglobulin-like domain in the *Saccharomyces cerevisiae* adhesion protein α-agglutinin

PETER N. LIPKE,[1] MIN-HAO CHEN,[1,4] HANS DE NOBEL,[2] JANET KURJAN,[2] AND PETER C. KAHN[3]

[1] Department of Biological Sciences and The Institute for Biomolecular Structure and Function, Hunter College of the City University of New York, New York, New York 10021
[2] Department of Microbiology and Molecular Genetics, University of Vermont, College of Medicine and College of Agriculture and Life Sciences, Burlington, Vermont 05405
[3] Department of Biochemistry and Microbiology, Cook College, Rutgers University, New Brunswick, New Jersey 08903

(RECEIVED June 2, 1995; ACCEPTED July 26, 1995)

Abstract

The *Saccharomyces cerevisiae* adhesion protein α-agglutinin is expressed by cells of α mating type. On the basis of sequence similarities, α-agglutinin has been proposed to contain variable-type immunoglobulin-like (IgV) domains. The low level of sequence similarity to IgV domains of known structure made homology modeling using standard sequence-based alignment algorithms impossible. We have therefore developed a secondary structure-based method that allowed homology modeling of α-agglutinin domain III, the domain most similar to IgV domains. The model was assessed and where necessary refined to accommodate information obtained by biochemical and molecular genetic approaches, including the positions of a disulfide bond, glycosylation sites, and proteolytic sites. The model successfully predicted surface exposure of glycosylation and proteolytic sites, as well as identifying residues essential for binding activity. One side of the domain was predicted to be covered by carbohydrate residues. Surface accessibility and volume packing analyses showed that the regions of the model that have greatest sequence dissimilarity from the IgV consensus sequence are poorly structured in the biophysical sense. Nonetheless, the utility of the model suggests that these alignment and testing techniques should be of general use for building and testing of models of proteins that share limited sequence similarity with known structures.

Keywords: cell adhesion molecule; glycoprotein; immunoglobulin variable domain; homology modeling; secondary structure prediction; sequence alignment; surface accessibility; volume packing

The cell adhesion protein α-agglutinin, which is expressed by *Saccharomyces cerevisiae* cells of α mating type, contains a single binding site for its ligand **a**-agglutinin, which is expressed by cells of **a** mating type (Lipke & Kurjan, 1992; Cappellaro et al., 1994). The α-agglutinin gene *AGα1* encodes a 650-amino acid protein with an N-terminal secretion signal and a C-terminal signal for addition of a glycosyl phosphatidylinositol anchor (Lipke et al., 1989; Wojciechowicz et al., 1993). The C-terminal half of α-agglutinin (Agα1p) is serine/threonine rich, is highly glycosylated, and is predicted to form an extended stalk (Jentoft, 1990; Chen et al., 1995). A soluble, truncated protein consisting of the N-terminal half (350 amino acids) of Agα1p is fully active in binding **a**-agglutinin. The α-agglutinin sequence between residues 200 and 320 (called domain III) has been aligned with the consensus sequence for immunoglobulin variable (IgV) domains (Williams & Barclay, 1988), with an alignment score 6.6 SDs above the mean for random sequences of the same composition. Two additional domains (domains I and II) also are predicted to form IgV fold-like structures (Chen et al., 1995). The N-terminal half of α-agglutinin has a CD spectrum consistent with a structure containing a high proportion of antiparallel β-sheet and <7% α-helix, as predicted for a protein consisting of three IgV domains.

Structural determination for glycoproteins is particularly difficult, because sugars interfere with crystallization and the carbohydrate ring protons absorb in the same region of NMR spectra as do protons important for structural determination in polypeptides. Homology modeling has been applied with some success to members of the Ig superfamily, including ICAM-1,

Reprint requests to: Peter Lipke, Department of Biological Sciences, Hunter College, 695 Park Avenue, New York, New York 10021; e-mail: lipke@genectr.hunter.cuny.edu, or Peter Kahn, Department of Biochemistry and Microbiology, Cook College, Rutgers University, New Brunswick, New Jersey 08903; e-mail: kahn@biovax.rutgers.edu.

[4] Present address: Department of Pathology, Harvard Medical School, 200 Longwood Avenue, Boston, Massachusetts 02115.

CD4, Po, and CEA (Bates et al., 1989, 1992; Berendt et al., 1992; Wells et al., 1993). However, the degree of sequence similarity of α-agglutinin domain III to IgV domains of known structure is significantly lower than for the Ig domains that have been modeled, making the alignment of sequences using commercially available programs inaccurate. We have developed an approach to model α-agglutinin domain III based on known IgV structures and have tested the validity of the model. This approach provides useful algorithms for the prediction and testing of structural models for other distantly related proteins.

Modeling procedure

Construction of the α-agglutinin domain III model involved four steps: (1) identification of homologs of known structure and of conserved regions of these reference proteins, (2) alignment of the α-agglutinin domain III sequence with the reference proteins, (3) assignment of initial coordinates to the domain III sequence, and (4) knowledge-based refinement of the model. This sequence of steps is standard (Bajorath et al., 1993; Sudarsanam et al., 1994), but we modified the alignment procedure to accommodate the low degree of similarity of the α-agglutinin domain III sequence with the reference structures. In addition, we have also applied a series of tests for the accuracy and physicochemical quality of the resulting model, which, to the best of our knowledge, have not been applied previously to homology modeled structures.

Identification of homologous structures

Sequence-based searches of Genbank did not identify α-agglutinin domain III as a homolog of any other gene (Lipke et al., 1989). The initial identification of domain III as a homolog of IgV domains was carried out visually and confirmed statistically by alignment of a 110-residue sequence fragment to an Ig consensus (Wojciechowicz et al., 1993). IgV domains consist of nine β-strands (called A, B, C, C′, C″, D, E, F, and G) that form two β-sheets and often contain a disulfide between Cys residues in strands B and F (Williams & Barclay, 1988). The alignment of α-agglutinin domain III to the consensus IgV domain sequence was strongest near the conserved cysteine residues. As a check for similarity to other structures, this region of α-agglutinin was also aligned with consensus sequences for other common extracellular domains. Similarity scores were compiled for alignments with domains rich in β-sheets, including fibronectin type II and type III, epidermal growth factor, Ig constant, and Ig I set (Bork, 1991; Harpaz & Chothia, 1994). Only the alignment with IgV domains gave a statistically significant score.

Homology modeling depends on identification of structurally conserved regions, SCRs, within a gene family (Bajorath et al., 1993; Wells et al., 1993). The homologous regions of a new member of the family are then assigned to the coordinates of the corresponding regions of the reference proteins. The first step in the modeling of α-agglutinin domain III was therefore the identification of SCRs for reference proteins. IgV domains of limited sequence similarity were chosen for this comparison to define those structural regions that are most strongly conserved. Three reference proteins, the heavy chain variable region of Ig KOL (Huber et al., 1976; PDB file 2FB4), Ig domain 1 of CD4 (Wang et al., 1990; PDB file 2CD4), and the IgV domain of CD8α (Leahy et al., 1992; PDB file 1CD8) were aligned based on their homologous positions in the three-dimensional structures (Bajorath et al., 1993). SCRs for each of the β-strands B–F were defined as those regions with <1 Å RMS deviation (RMSD) of α-carbons between each pair. The α-carbons in the regions of these domains encompassing β-strands B–F aligned with an RMSD of <1.4 Å between each pair of structures, despite low sequence identities (Table 1). β-Strands A and G had different coordinates in the three structures and did not constitute SCRs.

Table 1. *Comparisons of reference proteins and α-agglutinin*[a]

	α-Agglutinin domain III	Ig KOL heavy chain V	CD4 domain I	CD8α
α-Agglutinin domain III	–	13 (43)	14 (33)	15 (41)
Ig KOL heavy chain V	**2.55 Å**	–	22 (46)	21 (48)
CD4 domain I	n.d.[b]	**1.37 Å**	–	15 (34)
CD8α	n.d.[b]	**0.96 Å**	**1.11 Å**	–

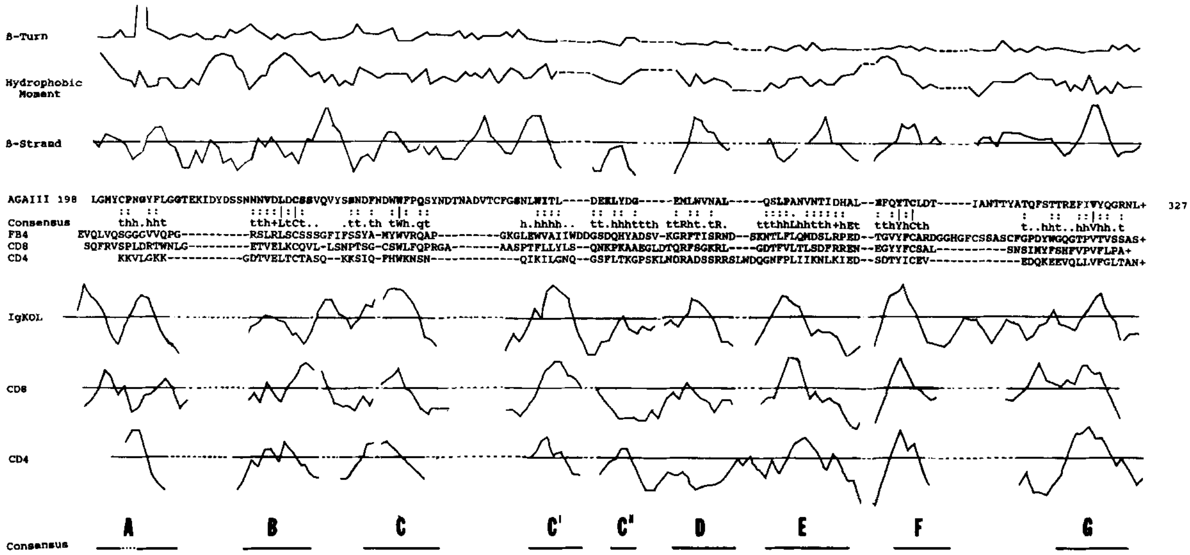
[a] The data are based on the alignment in Figure 1, the crystallographic structures of the reference proteins, and the model of α-agglutinin. The SCRs for β-strands B–F were compared, and RMSD of the three-dimensional structures are listed in bold in the bottom left half of the table. Percent sequence identities and similarities (in parentheses) are shown in the top right half of the table. Similarity sets were: AFILMPVWY, DEKR, ND, NQ, EQ, KRH, ST, FHWY, and AG. These similarity sets are more restrictive than those used in making the modeling alignment (see text) but reflect more traditional groupings.

[b] n.d., Not determined

Alignment of the α-agglutinin domain III sequence

Despite its identification as an IgV homolog, an accurate sequence alignment was problematic because the sequence of α-agglutinin domain III has very low identity to any of the reference structures (Table 1). A previous alignment (Wojciechowicz et al., 1993) preserved consensus residue identities in β-strands B, C, C′, D, E, and F, but many parts of the alignment were due to single-residue identities or nonidentical hydrophobic residues at analogous positions. Therefore, there were many alignments with scores similar to one another, and small changes in scoring parameters caused large shifts in the alignment. We attempted to build a model based on the published alignment, which was the highest scoring alignment in a sequence-based algorithm (Genetics Computer Group, 1994). However, this model contained a 10-residue insertion in the sequence between the E and F strands (the E–F loop), a region where there are no insertions or gaps in known structures. In addition, the D–E loop in this model was much shorter than in known IgV domains and contained too few residues to connect the two strands. These problems indicated the need for an alternate alignment procedure.

Sequence alignments of homologous proteins have been used to improve the accuracy of the secondary structure predictions by facilitating the choice of a consensus structure for each resi-



due position in the alignment (Cohen et al., 1991; Nierman & Kirschner, 1991; Nishikawa & Noguchi, 1991). Such methods pick a consensus secondary structure element for homologous positions of each sequence. We inverted the procedure, using the shapes of the predicted secondary structure curves of the reference proteins as models for the alignment of the predicted secondary structure of α-agglutinin domain III. Thus, a standard set of secondary structure predictions was carried out on the reference proteins, and the profiles and associated sequences were aligned based on the SCRs in the three-dimensional structures (Chou & Fasman, 1978; Garnier et al., 1978). This procedure generated a profile characteristic of IgV domains. The alignment of the Chou-Fasman β-strand predictor for the reference proteins is shown in the lower part of Figure 1. The α-agglutinin domain III sequence was then added to the alignment, with the consensus Cys residues placed in β-strands B and F. Regions of high β-strand potential were then aligned over the conserved strands in the reference structures, matching the β-strand predictors for the reference structures for all strands that gave high β-strand scores. This match identified sequences predicted to be in the A, B, D, E, F, and G strands.

The assignments from this alignment were confirmed by examining Chou-Fasman predictors of β-turns, which are often found between the strands in IgV domains, and the β-strand hydrophobic moment (Eisenberg et al., 1984), which is expected to be high for β-strands that have one side packed into the interior of the domain and the other exposed to solvent (Fig. 1, top). This procedure identified sequences of α-agglutinin domain III predicted to correspond to each of the strands, loops, and turns of IgV domains, although ambiguity of the exact alignments of the potential C, C′, and C″ strands remained due to an extra peak in the β-strand prediction profile (Fig. 1, top). The turn predictor was strongest for the A strand, and this sequence was modeled on the "kinked" structure common in this strand (Ryu et al., 1990).

Individual residues within the β-strands of the reference proteins were then aligned by a simple similarity algorithm developed to identify distant homologs (Beckman & Bork, 1993). At each sequence position in the SCRs, the aligned residues of the reference structures were classified as uniformly conserved (specific residue), hydrophobic, turn-like, or no consensus. This simplified scheme includes Ala, Glu, Gln, Lys, Pro, Thr, and Trp in both hydrophobic and turn-like categories, because most have long hydrocarbon chains in addition to polar groups or are common at turns despite their hydrophobic characteristics (Bork, 1991, 1992). Each predicted β-strand of α-agglutinin domain III was then aligned with the appropriate consensus sequence with no gaps allowed. Exact matches to uniformly conserved residues were given a score of +2, and matches to h or t were given a score of +1. For most of the potential β-strands of α-agglutinin domain III, the reference protein consensus was matched optimally by a single alignment (Table 2), thereby resolving the ambiguities in strands C, C′, and C″. Neighboring Trp residues in the C strand initially resulted in ambiguity; however, one of two potential alignments (lower alignment, Table 3) had a higher score in this similarity algorithm, and was used to build the model. This procedure also indicated the existence of large insertions in the A–B and F–G loops in comparison to the reference proteins.

The validity of the alignments was checked by alignments of random sequences of identical composition for representative β-strands. In each case, the α-agglutinin domain III sequence aligned at a level better than best alignment of the scrambled sequences (data not shown).

Initial coordinate assignments

Based on the SCR residues in the reference structures, 61 residues from the β-strands of α-agglutinin domain III were initially assigned the coordinates of Ig KOL using the HOMOLOGY

Table 2. *Coordinate assignments for β-strands in α-agglutinin domain III*[a]

β-Strand	Domain III sequence (consensus)	Comments
A	YCPNGYFLG	Aligned by placement of Cys 202 in a sequence position physically close to Cys 300. The strand was subsequently modified to allow formation of the disulfide.
B	NNNVDLDCSS (GttVtLtCt)	Consensus match
C	FNDWWFPQS (Wh.Qt)	High β-strand potential and consensus match
C′	NLWITL (h.hhhh)	High β-strand potential and match with simplified consensus
C″	KLY	C′ and D strand connected directly and minimized
D	EMLWV (tRht)	M substitution for consensus R. Strand was assigned by high β-strand potential and was shortened to maintain connectivity
E	QSLPANVNTI (tthhLhhtth)	Assigned by high β-strand potential. Sequence conforms to simplified consensus.
F	FQYTCLDT (thYhCth)	Consensus match
G	FIVY (hhVh)	High β-strand potential and match with simplified consensus

[a] Amino acids are shown in single letter code. The consensus categories are h (hydrophobic): ACEFGIKLMPQTVWY; and t (turn-like): ADEGHKNPQRSTW.

module of the INSIGHT modeling package (Biosym) running on a Silicon Graphics Indigo. Ig KOL was chosen as the reference protein because it is the best resolved of the reference structures used for the alignments. The A and G strand sequences are not conserved in IgV domains and were not uniformly placed in the reference structures; therefore, these strands of domain III were arbitrarily assigned coordinates of the Ig KOL structure. The C″ strand was treated as an unassigned region due to differences of the α-agglutinin sequence from the reference structures in and around this strand (see below).

The HOMOLOGY module includes two procedures for assigning coordinates to residues from the loops between the SCRs. A "loop search" program searches the Protein Data Bank for peptide structures of a given length connecting structure elements with geometry that matches the ends of the assigned β-strands (Bajorath et al., 1993); coordinates were assigned to α-agglutinin domain III residues in the A-B, B-C, C-C′, D-E, and E-F loops using reference loops from a library of structures from members of the Ig superfamily (Table 4). A conformation was chosen for the A-B loop that extended the A and B strands to accommodate the extra residues, in conformity to the β-strand prediction (Fig. 1). Because there was no suitable reference for the F-G loop, it was assigned using a "loop generation" algorithm, which generates random structures of a designated length connecting two residues with assigned coordinates. This region was assigned as a poorly packed extended chain to allow room for addition of glycans (see below). Because the C′-C″-D sequence is shorter in α-agglutinin domain III in comparison to the reference structures, residues from the end of the C′ strand to the start of the D strand were manually positioned (Fig. 2). The region following the G strand was also manually positioned close to the extended A and B strands. The completed structure was then annealed and energy minimized in vacuo by molecular mechanics until the RMS energy difference was <0.5 kcal/mol per cycle. In several cases, loop regions were minimized by sequential trials of side-chain conformations (Ponder & Richards, 1987).

Table 3. *Two alternate alignments for the strand C region*[a]

Molecule	Sequence	Score
α-Agglutinin domain III	NDWWFPQS	
	W: ::	5
	DWWFPQSY	
	:W: Q::	8
Consensus	**hθWh.qtt**	
Ig KOL	MYWVAQRP	9
CD8	CSWLFQPR	8
CD4	FHWKNSNQ	7

[a] Residue identities are scored +2 and similarities +1. The consensus categories are h (hydrophobic): ACEFGIKLMPQTVWY; and t (turn-like): ADEGHKNPQRSTW. The symbol ***θ*** denotes a residue that is polar in all three reference proteins and is aromatic in two of them. Similarly, "q" denotes a residue that is Gln in two reference proteins and many other members of the Ig superfamily.

Knowledge-based refinement

The model was refined to accommodate information obtained by peptide mapping (Chen et al., 1995). The disulfide bond in α-agglutinin domain III is between the canonical Cys residue from the F strand (Cys 300) and a noncanonical Cys residue from the A strand (Cys 202) rather than the usual B strand consensus Cys residue. The α-carbons of Cys 202 and Cys 300 were in reasonable proximity in the initial model, and the S atoms were 5 Å apart. Manual adjustment of the A and G strands allowed formation of the disulfide bond. Side chains of neighboring residues were manually torsioned to remove steric interference with disulfide formation.

Domain III of α-agglutinin is heavily *N*- and *O*- glycosylated (Chen et al., 1995). Yeast *N*-glycans are branched high mannose structures with 50–100 residues per chain (Klis, 1994). *N*-glycosylated residues Asn 248 and Asn 306 were substituted with a

Table 4. *Sources of coordinates for nonconserved regions*

Loops	Coordinate source	Comments
A–B	6FAB	Light chain variable domain F–G loop. This loop fit strand packing criteria and peptide mapping data
B–C	3FAB	B–C loop in 3FAB
C–C′	2FB4	F–G loop in reference structure
C′–C″–D	–	Extended conformation
D–E	2RHE	D–E loop in 2RHE
E–F	2FB4	SCR in reference proteins
F–G	–	Coordinate search to find extended conformation
post-G	–	Manual placement to accommodate strand packing and mutagenesis data

truncated, energy-minimized model glycan of $Man_3GlcNAc_2$. Twelve of 23 Ser and Thr residues in domain III are *O*-glycosylated (Chen et al., 1995) (Table 5), with all known *O*-glycosylation sites occurring C-terminal to Ser 282. These residues were modified by addition of Manα1 → 2Manα model oligosaccharides (Klis, 1994) (Table 5) that had been constructed and energy minimized prior to addition to the protein. Some amino acid side chains were torsioned manually to accommodate the added oligosaccharides.

Most of the identified protease cleavage sites (Chen et al., 1995) were exposed on the surface of the protein in the model. Two exceptions (Asp 240 and Arg 317) required refinement. Manual torsioning of the backbone near each of these residues removed these charged groups from the hydrophobic core and exposed them at the domain surface. The alteration of the backbone at Asp 240 also relieved steric strain associated with the position of the Thr 315 oligosaccharide, because side-chain atoms of these two amino acid residues were initially only 3–4 Å apart.

After each of the above modifications, individual regions were minimized to 0.5 kcal/mol, followed by unconstrained minimization of the entire domain. The energy of the glycosylated model was +2,092 kcal/mol, of which +1,842 kcal/mol is due to interactions within the oligosaccharides (see below).

Site-specific mutagenesis has been used to identify a putative binding site for the ligand **a**-agglutinin (de Nobel, Lipke, Kurjan, in prep.). His 292, within the E–F loop, has been identified as essential for binding and accessible to modifying reagents (Cappellaro et al., 1991). The A–B loop and post-G strand region are predicted to be in close proximity to the E–F loop in the model, and mutagenesis identified residues in the A–B and E–F loops and post-G strand that are also important for α-agglutinin activity. The nonglycosylated model was further refined to accommodate the results of several mutations. The identification of Tyr 216 as a critical residue allowed the choice of a model with Tyr 216 and His 292 in close proximity over models that differed in the positioning of the A–B loop (Fig. 2). The lack of conservation of the G strand made its orientation in the model ambiguous. In early versions, Gln 323 was next to His 292, and Tyr 322 was 10 Å distant on the opposite side of the domain. However, several mutations of Tyr 322 resulted in 20-fold decreases in activity, whereas a mutation of Gln 323 reduced activity only twofold. ϕ and ψ angle rotations of residues 322–324 moved the Tyr residue into the putative binding site 7 Å from His 292 (Fig. 3) and moved Gln 323 behind the putative binding site. The accommodation of the mutation results each affected the geometry of the region near His 292 and allowed a choice of a specific configuration for the A–B loop and the post-G strand as part of an

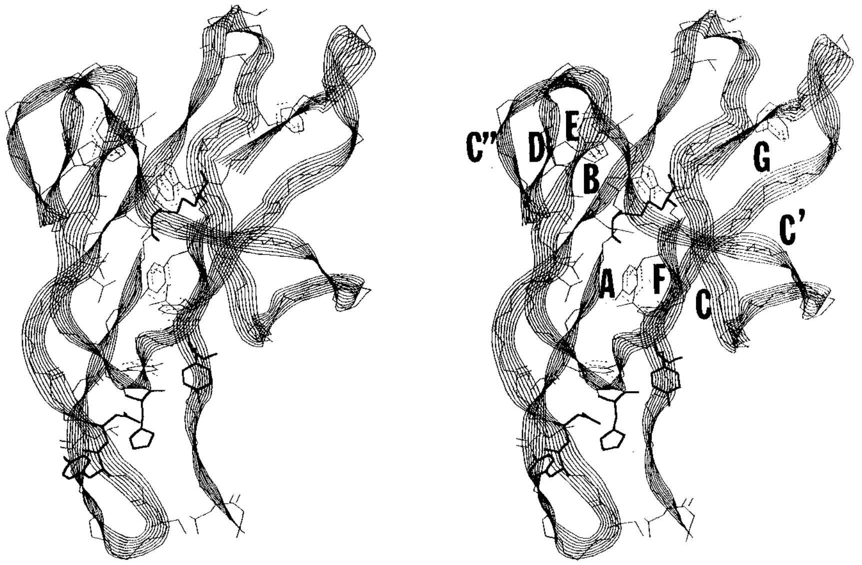


Fig. 2. Stereo ribbon drawing of α-agglutinin domain III. The disulfide bond (Cys 202 and Cys 300) and active-site residues (Tyr 216, Asp 217, His 292, and Tyr 322) are drawn with thick lines. Hydrophobic residues contributing to the core are shown in light lines (see text). The β-strands are marked.

extended structure. After modifications, the model was minimized to −3,612 kcal/mol using CHARMm molecular mechanics (Molecular Simulations, Inc.).

Solvent-accessible surface areas were calculated for each of the three reference proteins and for the agglutinin model by the method of Lee and Richards (1971) using the program ACCESS (Richmond & Richards, 1978), which was kindly supplied by Dr. Richards. Volumes were computed by the Voronoi method (Richards, 1974) with the program VOLUME (Richards, 1985), also supplied by Dr. Richards. Results for each protein were tabulated for the 20 amino acids, those of the three reference proteins being combined into a single set of tables for comparison with the agglutinin model.

Results and discussion

Description and testing of the model

The model of α-agglutinin domain III shows a compact domain, with an extension formed by the extended A and B strands together with a continuation of the G strand (Fig. 2).[5] Although the disulfide bond between the A and F strands is novel among Ig folds, it was easily accommodated in the model. With the exception of Thr 289 (shown just above and to the right of the binding site residues in Fig. 4B), the glycosylation sites form a crescent around one side of the domain (Fig. 4). If the oligosaccharides on the two Asn residues are of typical size (50–100 residues), an entire face of the domain would be covered with carbohydrate.

Two hydrophobic surfaces are exposed in the model. A surface above the putative binding site (Fig. 4B), which is free of carbohydrate, includes residues in the C, C′, and C″ strands (Phe 243, Pro 244, Tyr 247, Leu 261, Trp 262, Ile 263, Leu 265, Leu 269, and Tyr 270). Note that even if some of the β-strands are improperly oriented in the model, there would be a hydrophobic surface, because there are sequential hydrophobic residues in the C and C′ strands, and some of these residues must be surface exposed in an Ig-like structure. The C–C′–C″ β-sheet surface is a common site for interactions between Ig domains (Williams & Barclay, 1988). We propose that this hydrophobic region forms an interface between domain III (Fig. 4B) and domain I or II. There is a smaller hydrophobic surface on the opposite side of the structure composed of residues in the ABED sheet (Met 200, Tyr 201, Pro 203, Val 223, Leu 225, Val 277, and Pro 284) (Fig. 4A). This patch may be covered by the carbohydrates attached to the loop between the F and G strands, which include six *O*-linked saccharides and one *N*-linked saccharide.

The putative binding site identified by mutagenesis (de Nobel et al., in prep.) is on one face of the extension of the A, B, and G strands, shown below the large hydrophobic face in Figures 3 and 4B. This region is not a complete surface and therefore the binding site may include residues contributed by do-

Table 5. *Solvent-accessible residues in α-agglutinin domain III*[a]

Proteolysis sites	*N*-glycosylation sites	*O*-glycosylation sites
Lys 213	Asn 248	Ser 282
Asp 240	Asn 306	Thr 289
Asp 253		Thr 299
Asp 266		Thr 307
Lys 268		Thr 308
Glu 274		Thr 311
Asn 278		Thr 314
Glu 295		Thr 315
Arg 317		Ser 316
Gln 323		
Arg 325		

[a] Data are from Chen et al. (1995).

[5] Coordinates will be deposited in the Protein Data Bank.

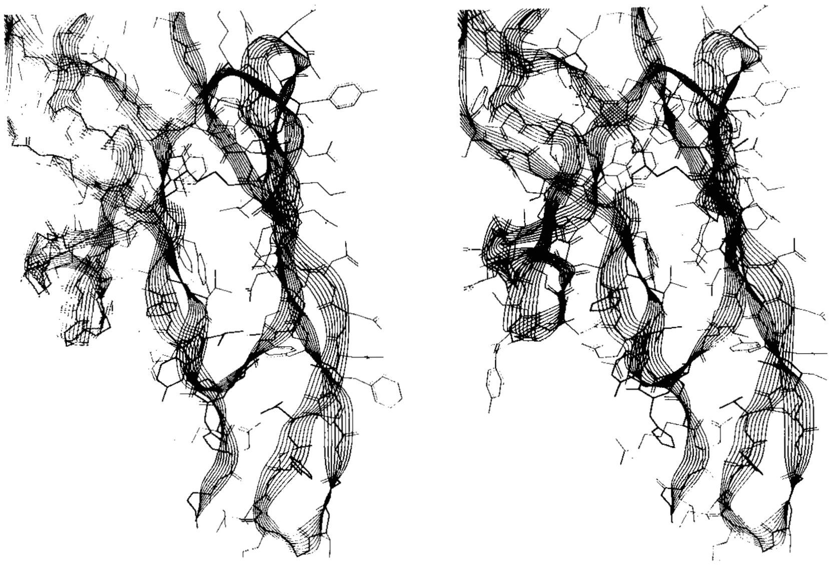


Fig. 3. Stereo drawing of the extended region containing the active-site residues. The backbone and proposed binding site residues are highlighted in thick black lines. Mutations in these residues reduce binding activity threefold or more: Tyr 216, Asp 217, Asp 291, His 292, Leu 294, Phe 296, and Tyr 322. The binding data are from de Nobel et al., in prep.

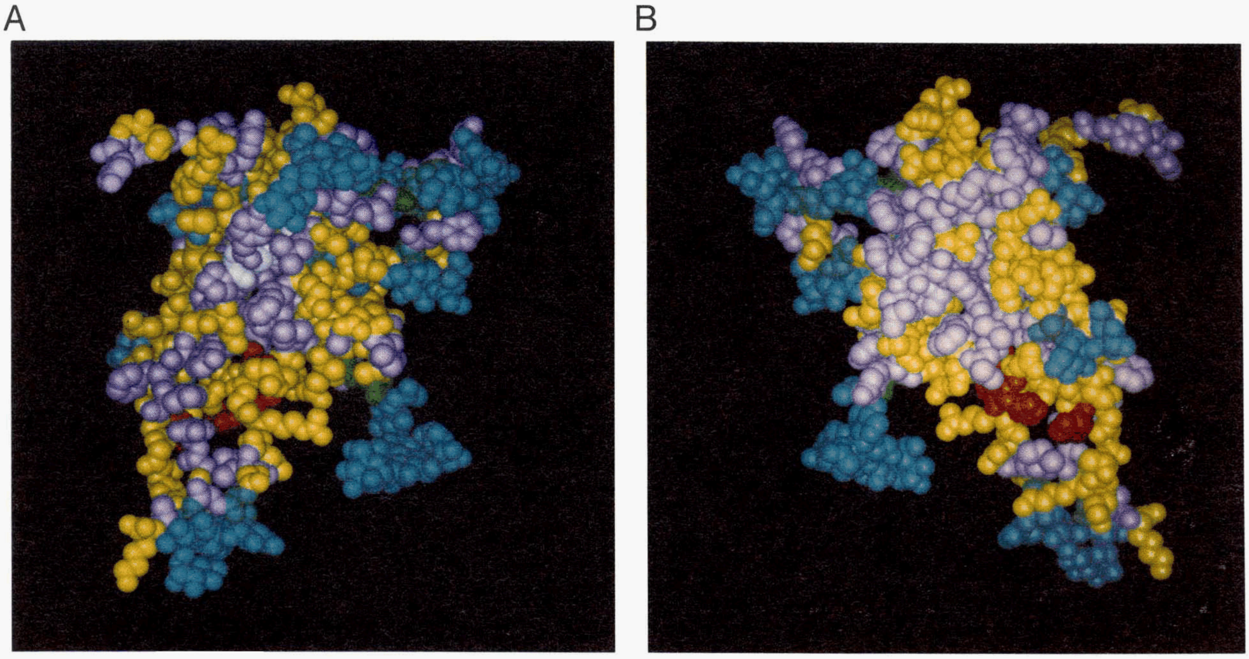


Fig. 4. Surface features of the glycosylated model of domain III of α-agglutinin. **A** and **B** are rotated 180° about the Y-axis, relative to each other. Carbohydrate residues are shown in blue, and glycosylated Asn residues are green. Hydrophobic side chains are lavender. Active site residues Tyr 216, His 292, and Tyr 322 are red. Residues N-terminal to domain III are shown extending to the top left in A and top right in B.

mains I and/or II, which probably interact with this region (Chen et al., 1995).

Validation of the model

We have used molecular genetic, biophysical, and biochemical data to test the model. It is consistent with general considerations of domain structure. There is a hydrophobic core composed of Leu 225, Val 230, Val 232, Phe 238, Trp 242, Trp 262, Trp 276, Leu 283, Met 284, Ala 285, Val 287, Tyr 298, and Val 321 (Fig. 2). The disulfide bond is buried in the interior of the structure and incorporates a rotamer of cystine that is extended in length and characteristic of immunoglobulin folds (Fig. 2) (Richardson, 1981). The Cα–Cα distance is 7 Å. Furthermore, the vast majority of Asn residues are in loop regions or at the edges of β-sheets, which is characteristic of known protein structures (Richardson, 1981).

Truncations of *AGα1* have indicated that the C-terminal boundary of the active protein is close to residue 324 (de Nobel et al., in prep.). That is, a truncation at residue 315 inactivated the protein, whereas truncations at residue 325 or beyond had no effect on activity. A frameshift mutation at residue 324 resulted in partial activity. In the model, the G strand of the domain III is composed of residues 319–322. Therefore, the C-terminal boundary of the active protein correlates with the C-terminus of domain III.

His 292, which has been shown to be essential for activity, is accessible to modifying reagents (Cappellaro et al., 1991) and is exposed to solvent in the model. The A–B loop and post-G strand region are predicted to be in close proximity to His 292, which is in the E–F loop (Fig. 3). Mutations were made in 16 residues predicted to be within 10 Å of His 292 in the model (de Nobel et al., 1995). Mutations of five residues resulted in 20–200-fold decreases in activity and mutations of two additional residues resulted in 4–10-fold decreases in activity. Mutations in nine other residues had little effect on activity. The use of the model to identify a putative binding site composed of residues from three discontinuous patches confirmed the proximity of the A–B and E–F loops and the post-G strand. As controls, six residues in the C–C′ and C′–C″ loops, which are predicted to be distant from His 292, were mutated. None of the control mutations had significant effects on activity, again providing evidence consistent with the model.

The quality of the model was also assessed by examining the surface exposure of 11 glycosylation sites, 11 proteolytic sites, and the 4 Cys residues. Most of the identified trypsin and staphylococcal protease V8 cleavage sites (Chen et al., 1995) were exposed on loops of the protein in initial models. After refinement, the two exceptions (Asp 240 and Arg 317) were also exposed at the surface (Fig. 5). Conversely, the four Cys residues, which include two cysteines with free sulfhydryls in addition to the disulfide-bonded Cys residues, are inaccessible to reducing agents and alkylating agents (Chen et al., 1995). These residues are buried in the interior of the modeled domain.

N- and *O*-glycosylation sites must be exposed to solvent and be able to accommodate model oligosaccharides. All of the determined glycosylation sites in the C-terminal region of domain III, starting with the *O*-glycosylated Ser 282 residue at the beginning of the E strand and extending past domain III into the proposed stalk formed by the C-terminal half of α-agglutinin are exposed on the surface (Fig. 4) (Chen et al., 1995).

Native globular proteins have been shown to be densely packed, the packing density of the protein interior approaching that of organic solids (Richards, 1974). Volume packing thus

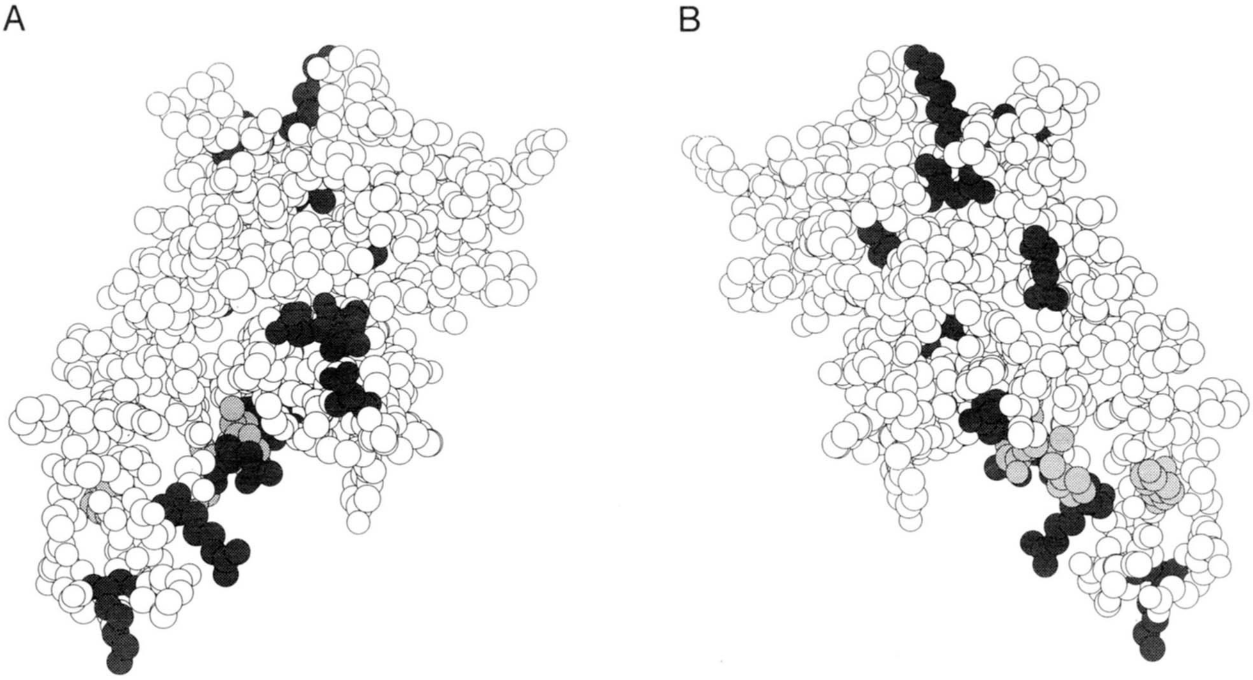


Fig. 5. Position of staphylococcal V8 and tryptic sites in domain III of α-agglutinin. Cleaved residues are shaded darkly, and binding site residues are lightly shaded. The two views correspond to those in Figure 4.

provides a criterion by which models can be compared with the structures determined by crystallography or NMR. Solvent-accessible surface area (Lee & Richards, 1971; Shrake & Rupley, 1973) is another criterion. We therefore compared the model with the three crystallographically determined structures with which α-agglutinin shows similarity in sequence.

The average accessible surface areas and average volumes of the reference proteins and for the model are shown in Figure 6. It is seen that, for 17 of the 20 amino acids, the accessibility of the model exceeds that of the reference proteins, and for 18 of the 20, the agglutinin volumes exceed those of the reference set. The differences in volumes, it should be noted, are an underestimate, although a small one, as the method of calculating the Voronoi volumes (Richards, 1974) yields slightly smaller values for surface amino acids than for the same amino acids when they are buried (data not shown). This would have a larger effect on the volumes of the more accessible residues in agglutinin than on the reference proteins, thus reducing the volume differences between it and the reference set.

The greater access of solvent to the model's surface and the greater volumes occupied by its amino acids show that the model is not as well packed as a crystallographically determined native structure would be. Comparison of Figure 6A with Figure 6B shows that the accessibility differences are generally greatest for the less polar amino acids, which one would expect to be least accessible. The differences, particularly in volume, are generally not large. Large differences in surface exposure are seen in rare amino acids (Met, Trp, Arg, His) and aromatics. For the rare amino acids, poor packing of a single residue greatly affects the mean values. In the case of His, Tyr, and Arg, residues were deliberately exposed during model building to help visualize the binding site or a proteolytic site. However, for each of the 20 amino acid types the distributions of accessible surface and of volume for the model overlap heavily those of the reference set (data not shown). Thus, many regions of the model do meet the packing criteria of native globular proteins.

Most residues in the model have volume and surface accessibility with values typical of the reference structures. The exceptions point to regions of poor physical structure in the model. Figure 7 shows that these residues are clustered in sequence position: 13 residues with abnormally large surface exposure, and 28 (including 5 residues that also have excessive surface exposure) that occupy larger than normal volumes. Three of these 36 flagged residues were positioned to facilitate visualization of the binding site (Tyr 216, His 292, and Tyr 322). Because the criterion for plotting was a 2-SD difference, one would expect four to six residues would exceed the threshold, even in a well-packed structure of this size.

The location of poorly packed regions is illustrative of the strengths and weaknesses of the model. Six residues in the A strand and the A–B loop have unusually large volumes. This is partly due to exposure of hydrophobic residues on the domain surface next to the probable surface of domain II, which is contiguous with the A strand. The B–C loop is also poorly packed, as expected for a region with no constraints built in. As mentioned above, there is substantial difference in sequence and sequence length in the C–C′–C″–D region, and the model is not well packed in this region. Note that, although this region is poorly compacted, there is little abnormal surface exposure. Finally, there is a high degree of surface exposure in the F strand and F–G loop. This region is highly glycosylated, and therefore, the areas that were calculated in the absence of carbohydrate residues are substantial overestimates.

As expected, the three-dimensional model deviates significantly from the structure of Ig KOL (Table 1). As shown by the residue packing analyses, this difference is greatest in the C″–

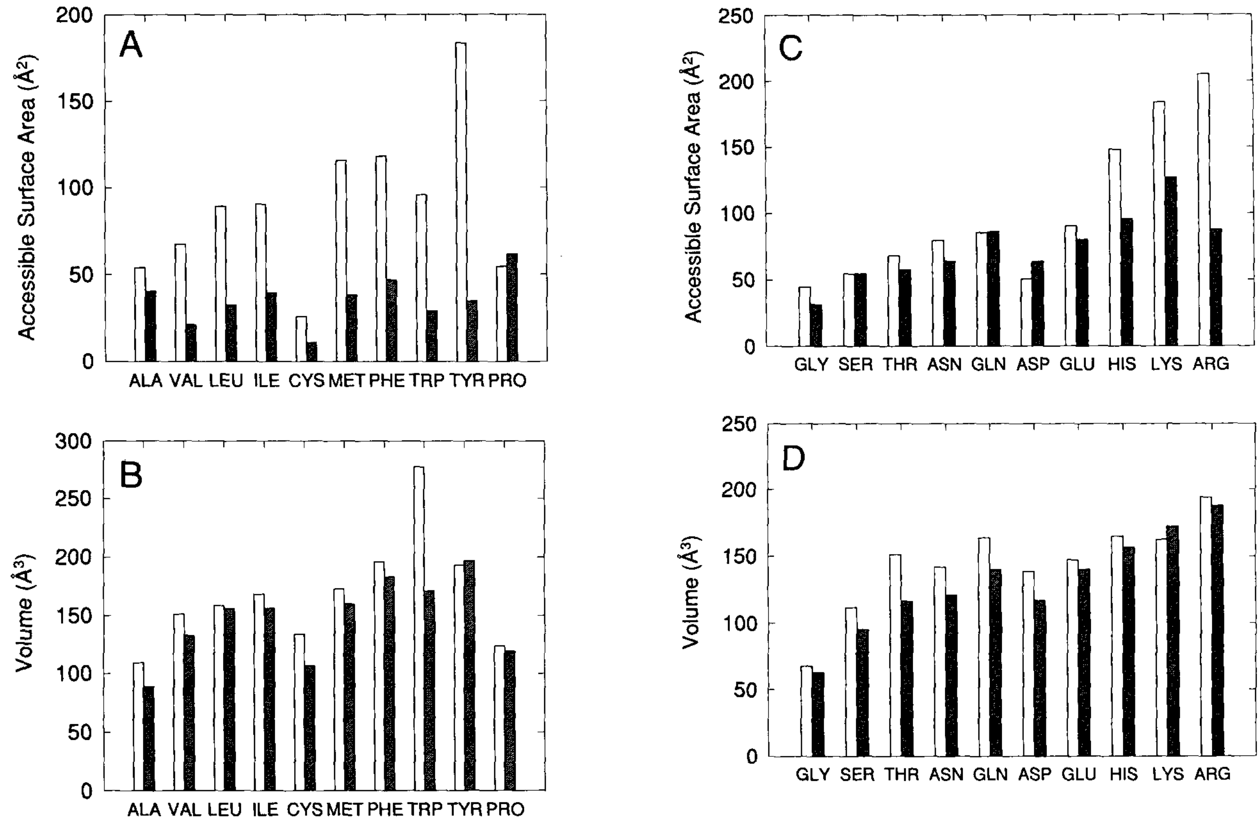


Fig. 6. Mean accessible surface area and mean volume for each type of residue. **A:** Surface area for hydrophobic residues. **B:** Volume for hydrophobic residues. **C:** Surface area for polar and charged residues. **D:** Volume for polar and charged residues. Residues in α-agglutinin are represented in the white bars and residues of the pooled reference structures are shown in the shaded bars.

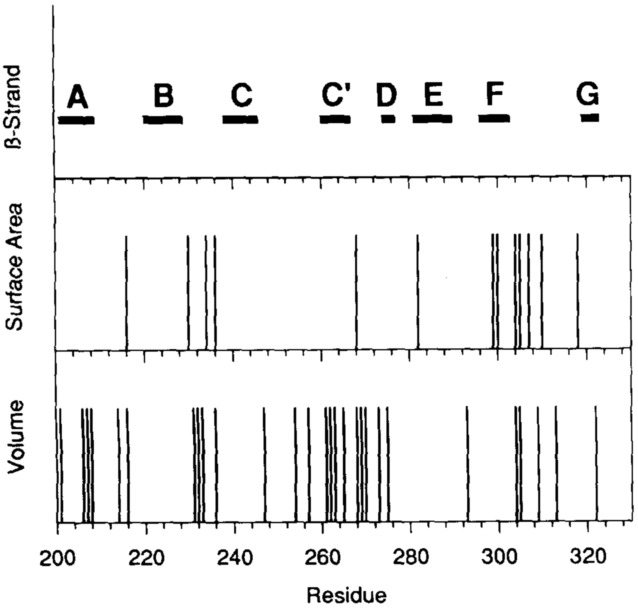


Fig. 7. Position of residues with atypical values for volume and accessible surface area. Sequence position is shown for those residues whose surface area or volume values exceed the mean for the reference structures by >2 SDs. The sequence positions of the β-strands in the model are shown at the top.

C′-C-F-G β-sheet. This difference is caused by shortened sequence between strands C′ and D and by the disulfide bond.

Secondary structure-based alignments

There is commonality of structure among proteins that share little or no sequence similarity; the degree of sequence similarity between many members of the Ig superfamily is low, and several structures are known that have the Ig fold, but no demonstrable sequence similarity (DeVos et al., 1992; Holmgren et al., 1992; Juy et al., 1992; Overduin et al., 1995; Shapiro et al., 1995). Standard sequence-based alignment procedures were inadequate for alignment of α-agglutinin domain III to reference IgV structures due to the low level of sequence identity between these highly diverged proteins. Therefore, an unorthodox approach to alignment was developed and used to construct this model. This alignment technique worked well for α-agglutinin domain III, which had a maximum of 15% sequence identity to the reference proteins (Table 1). The similarity scores were 33–43%, however, reflecting the conservation of the hydrophobic core (Chen et al., 1995). The modeling technique should therefore accommodate structures of this degree of similarity.

As an additional test of the alignment procedure, we aligned a similar region of the sequence derived from the *ALS1* gene from *Candida albicans*. The *ALS1*-derived protein sequence is

27% identical and 39% similar to α-agglutinin in the N-terminal 350 residues but has low similarity to Ig superfamily members (Hoyer et al., 1995). Secondary-structure-based alignment of the two yeast sequences with the reference protein consensus resulted in an alignment in the domain III region with 26% identity and 52% similarity between the yeast sequences. In addition, the *ALS1* and *AGα1* sequences conformed to the consensus from the model proteins similarly: they each conformed to consensus at 50 of 69 consensus positions, including all Cys residues. Neither sequence matched the consensus at 14 positions. At only five positions was the consensus match discrepant for the two sequences (data not shown). Therefore, the alignment procedure appears to be generally useful and applicable for distantly related sequences.

The alignment procedure has also been carried out for the first two domains of α-agglutinin, which are more distantly related to the Ig superfamily (Wojciechowicz et al., 1993; Chen et al., 1995). Preliminary models based on these alignments appear reasonable upon visual inspection but have not yet been tested.

The quality of the basic structure of the model was indicated by the minor nature of required alterations after accommodation of the disulfide bond in the first refinement; glycosylated residues and all but two protease-sensitive residues were exposed adequately. The two remaining proteolysis sites were exposed after a local modification. The use of the model to identify a ligand binding site by mutagenesis (de Nobel et al., in prep.) provided indication of the utility of the model. Because the modeling techniques and biochemical tests give only approximate spatial information, the model is not equivalent to an X-ray or NMR structure in its biophysical properties, as indicated by volume packing and surface accessibility analyses, as well as visual inspection. However, the success of the α-agglutinin domain III model implies that homology modeling using a secondary structure-based alignment procedure will be more generally applicable to evolutionarily distant proteins in systems not yet amenable to more definitive structural techniques. It is also clear that models need not meet all the requirements of packing characteristics of native proteins to be useful.

Acknowledgments

This work was supported by a MARC Faculty Fellowship grant to P.N.L., as well as grants from NIGMS (P.N.L. and J.K.) and the Research Centers in Minority Institutions program of the NIH. Critical support was also provided by the New Jersey Agriculture Experiment Station and the Center for Advanced Food Technology. (CAFT) of Rutgers University (paper D-01405-3-95 and D-10535-5-95) (P.C.K.). CAFT is a New Jersey Commission on Science and Technology Center.

References

Bajorath J, Stenkamp R, Aruffo A. 1993. Knowledge-based model building of proteins: Concepts and examples. *Protein Sci 2*:1798–1810.

Bates PA, Luo J, Sternberg MJ. 1992. A predicted three-dimensional model for the carcinembryonic antigen (CEA). *FEBS Lett 301*:207–214.

Bates PA, McGregor MJ, Islam SA, Sattenau QJ, Sternberg MJ. 1989. A predicted three-dimensional structure for the human immunodeficiency virus binding domains of CD4 antigen. *Protein Eng 3*:13–21.

Beckman G, Bork P. 1993. An adhesive domain detected in functionally diverse receptors. *Trends Biochem Sci 18*:40–41.

Berendt AR, McDowell A, Craig AG, Bates PA, Sternberg MJE, Marsh K, Newbold CI, Hogg N. 1992. The binding site on ICAM-1 for *Plasmodium falciparum*-infected erythrocytes overlaps, but is distinct from, the LFA-1-binding site. *Cell 68*:71–81.

Bork P. 1992. The modular architecture of vertebrate collagens. *FEBS Lett 307*:49–54.

Cappellaro C, Baldermann C, Rachel R, Tanner W. 1994. Mating type-specific cell–cell recognition of *Saccharomyces cerevisiae*: Cell wall attachment and active sites of **a**- and α-agglutinin. *EMBO J 13*:4737–4744.

Cappellaro C, Hauser K, Mrsa V, Watzele M, Watzele G, Gruber C, Tanner W. 1991. *Saccharomyces cerevisiae* **a**- and α-agglutinin: Characterization of their molecular interaction. *EMBO J 10*:4081–4088.

Chen MH, Shen ZM, Bobin S, Kahn PC, Lipke PN. 1995. Evidence for multiple immunoglobulin-like domains with atypical disulfides in *Saccharomyces cerevisiae* α-agglutinin. *J Biol Chem 270*:in press.

Chou PY, Fasman GD. 1978. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol Relat Areas Mol Biol 47*:45–148.

Cohen BI, Presnell SR, Cohen FE. 1991. Pattern-based approaches to protein structure prediction. *Methods Enzymol 202*:252–268.

DeVos AM, Ultsch M, Kossiakoff AA. 1992. Human growth hormone and extracellular domain of its receptor. *Science 255*:358–363.

Eisenberg D, Weiss RM, Terwilliger TC. 1984 The hydrophobic moment predicts periodicity in protein hydrophobicity. *Proc Natl Acad Sci USA 81*:140–144.

Garnier J, Osguthorpe DJ, Robson B. 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol 120*:97–120.

Genetics Computer Group. 1994. *Program manual for the Wisconsin package, version 8*. Madison, Wisconsin, USA.

Harpaz Y, Chothia C. 1994. Many of the immunoglobulin family domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains. *J Mol Biol 238*:528–539.

Holmgren A, Kuehn MJ, Branden CI, Hultgren SJ. 1992. Conserved immunoglobulin-like features in a family of periplasmic pilus chaperones in bacteria. *EMBO J 11*:1617–1722.

Hoyer LL, Scherer S, Schatzman AR, Livi GP. 1995. *Candida albicans ALS1*: Domains related to a *Saccharomyces cerevisiae* sexual agglutinin separated by a repeated motif. *Mol Microbiol 15*:39–54.

Huber R, Diesenhofer J, Colman PM, Matsushima M, Palm W. 1976. Crystallographic structure studies on an IgG molecule and an Fc fragment. *Nature 264*:415–420.

Jentoft N. 1990. Why are proteins *O*-glycosylated? *Trends Biochem Sci 15*:291–295.

Juy M, Amit AG, Alzari PM, Poljak RM, Claeyssens M, Beguin P, Aubert JP. 1992. Three-dimensional structure of a thermostable bacterial cellulase. *Nature 357*:89–91.

Klis FM. 1994. Review: Cell wall assembly in yeast. *Yeast 10*:851–869.

Leahy DJ, Axel R, Hendrickson WA. 1992. Crystal structure of a soluble form of the human T cell coreceptor CD8 at 2 Å resolution. *Cell 68*:1145–1162.

Lee B, Richards FM. 1971. The interpretation of protein structure: Estimation of static accessibility. *J Mol Biol 55*:379–400.

Lipke PN, Kurjan J.1992. Sexual agglutinins in budding yeasts: Structure, function and regulation of yeast cell adhesion proteins. *Microbiol Rev 56*:180–194.

Lipke PN, Wojciechowicz D, Kurjan J. 1989. *AGα1* Is the Structural gene for the *Saccharomyces cerevisiae* α-agglutinin, a cell surface glycoprotein involved in cell–cell interactions during mating. *Mol Cell Biol 9*:3155–3165.

Nierman T, Kirschner K. 1991. Use of homologous sequences to improve protein secondary structure prediction. *Methods Enzymol 202*:45–59.

Nishikawa K, Noguchi T. 1991. Predicting protein secondary structure based on amino acid sequence. *Methods Enzymol 202*:31–44.

Overduin M, Harvey TS, Bagby S, Tong KI, Yau P, Takeichi M, Ikura M. 1995. Solution structure of the epithelial cadherin domain responsible for selective cell adhesion. *Science 267*:386–389.

Ponder JW, Richards FM. 1987. Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol 193*:775–791.

Richards FM. 1974. The interpretation of protein structures: Total volume group volume distributions and packing density. *J Mol Biol 82*:1–14.

Richards FM. 1985. Calculation of molecular volumes and areas for structures of known geometry. *Methods Enzymol 115*:440–469.

Richardson JS. 1981. The anatomy and taxonomy of protein structure. *Adv Protein Chem Biochem 35*:167–339.

Richmond TJ, Richards FM. 1978. Packing of α-helices: Geometrical constraints and contact areas. *J Mol Biol 119*:537–555.

Ryu SE, Kwong PD, Truneh A, Porter TG, Arthos J, Rosenberg M, Dai X, Zuong N, Axel R, Sweet RW, Hendrickson WA. 1990. Crystal structure of an HIV-binding fragment of human CD4. *Nature 348*:419–426.

Shapiro L, Fannon AM, Kwong PD, Thompson A, Lehmann MS, Grubel

G, Legrand JF, Als-Nielsen J, Colman DR, Hendrickson WA. 1995. Structural basis of cell-cell adhesion by cadherins. *Nature 374*:327-337.

Shrake A, Rupley JA. 1973. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J Mol Biol 79*:351-371.

Sudarsanam S, March CJ, Srinivasan S. 1994. Homology modeling of divergent proteins. *J Mol Biol 241*:143-149.

Wang J, Yan Y, Garrett TPJ, Liu J, Rodgers DW, Garlick RL, Tarr GE, Husain Y, Reinherz EL, Harrison SC. 1990. Atomic structure of a fragment of human CD4 containing two immunoglobulin-like domains. *Nature 348*:411-418.

Wells CA, Saavedra RA, Inouye H, Kirschner DA. 1993. Myelin Po glycoprotein: Predicted structure and interaction of extracellular domains. *J Neurochem 61*:1987-1995.

Williams AF, Barclay AN. 1988. The immunoglobulin superfamily–domains for cell surface recognition. *Annu Rev Immunol 6*:381-405.

Wojciechowicz D, Lu CF, Kurjan J, Lipke PN. 1993. Cell surface anchorage and ligand-binding domains of the *Saccharomyces cerevisiae* cell adhesion protein α-agglutinin, a member of the immunoglobulin superfamily. *Mol Cell Biol 13*:2554-2563.