PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Protein Structures and Complexes: What they Reveal about the Interactions that Stabilize them

Janet M. Thornton, Malcolm W. MacArthur, Ian K. McDonald, David T. Jones, John B. O. Mitchell, C. Lilian Nandi, Sarah L. Price and Marketa J. J. M. Zvelebil

Phil. Trans. R. Soc. Lond. A 1993 **345**, 113-129
doi: 10.1098/rsta.1993.0123

Protein structures and complexes: what they reveal about the interactions that stabilize them

By Janet M. Thornton[1], Malcolm W. MacArthur[1,2], Ian K. McDonald[1], David T. Jones[1,3], John B. O. Mitchell[1], C. Lilian Nandi[1], Sarah L. Price[4] and Markéta J. J. M. Zvelebil[1]

[1] *Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College London, Gower Street, London WC1E 6BT, U.K.*

[2] *Crystallography Department, Birkbeck College, Malet Street, London WC1E 7HX, U.K.*

[3] *National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, U.K.*

[4] *Department of Chemistry, University College London, 20 Gordon Street, London WC1H 0AJ, U.K.*

The rapid increase in the number of high-quality protein structures provides an expanding knowledge resource about interactions involved in stabilizing protein three-dimensional structures and the complexes they form with other molecules. In this paper we first review the results of some recent analyses of protein structure, including restrictions on local conformation, and a study of the geometry of hydrogen bonds. Then we consider how such empirical data can be used as a test bed for energy calculations, by using the observed spatial distributions of side chain/atom interactions to assess three different methods for modelling atomic interactions in proteins. We have also derived a new empirical solvation potential which aims to reproduce the hydrophobic effect. To conclude we address the problem of molecular recognition and consider what we can deduce about the interactions involved in the binding of peptides to proteins.

1. Introduction

The Brookhaven Protein Structure Databank (PDB) (Bernstein *et al.* 1977) now includes more than 1000 sets of protein coordinates. These data represent a huge knowledge-base about the interactions within proteins and between proteins and their ligands. From these data we can obtain information about stereochemistry, local preferred geometry, the relative strengths of various interactions and details about how proteins recognize each other and their cognate ligands. The small molecule Cambridge Structure Database (CSD) (Allen *et al.* 1983) provides complementary data consisting of accurate bond lengths and angles together with intermolecular interactions (e.g. hydrogen bonds). In this paper we shall present a review of some of the basic observations and principles pertaining to molecular recognition that can be extracted by detailed examination of these data. Intramolecular interactions and theoretical potentials will be discussed, followed by consideration of the recognition of peptides by proteins.

Phil. Trans. R. Soc. Lond. A (1993) **345**, 113–129
Printed in Great Britain

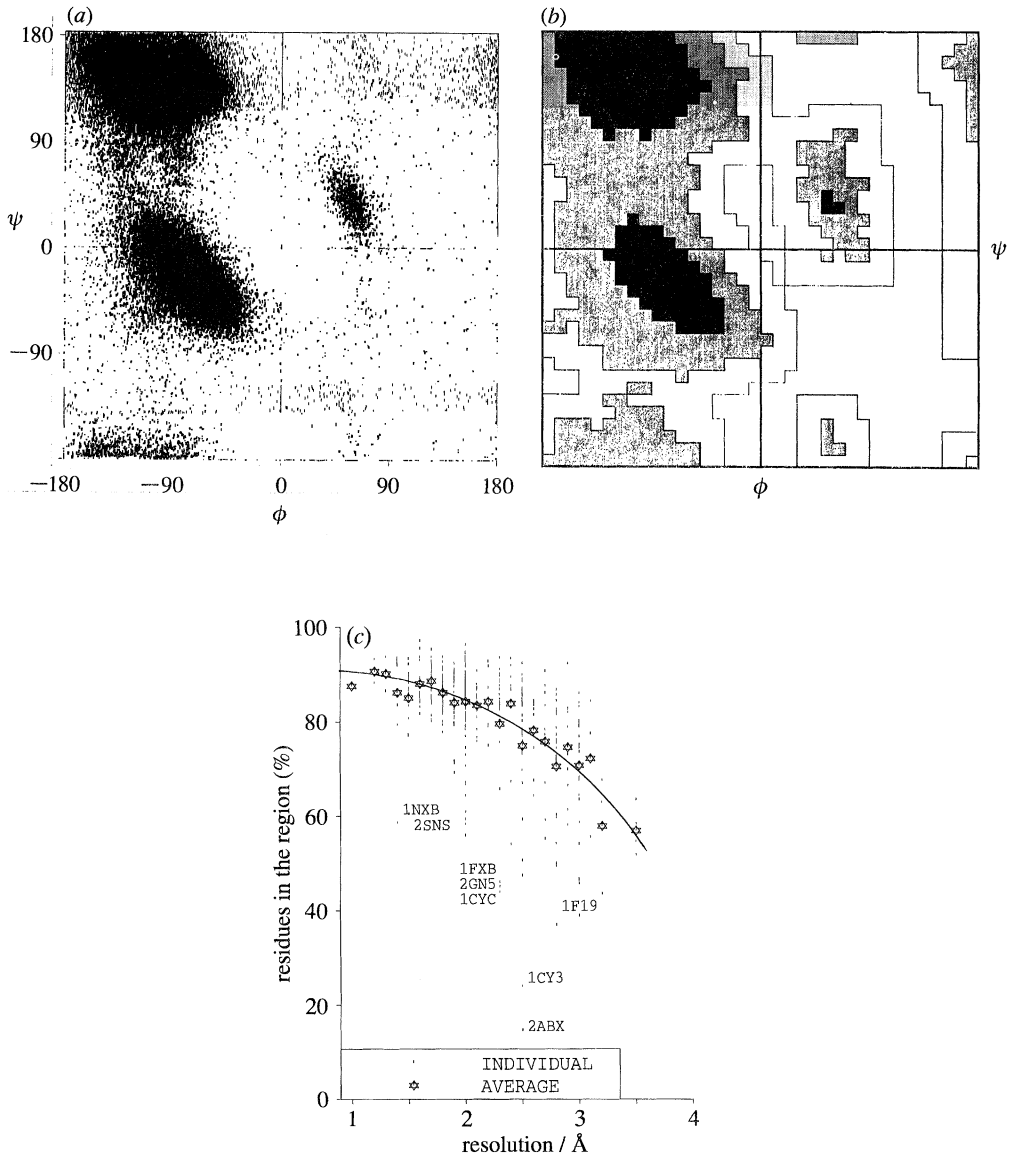


Figure 1. (*a*) Distribution of ϕ, ψ angles for 121 870 residues from 462 protein structures. Proline and glycine were not included. The angles were calculated from Brookhaven coordinates using the program SSTRUC (Smith, personal communication). (*b*) Digitized Ramachandran type plot derived from the actual distribution shown in (*a*). The population within each $10° \times 10°$ pixel was calculated and used to define 4 areas: the black areas are core, the dark grey allowed, and the pale grey, generous, the blank disallowed. (*c*) Plot to show the percentage of residues from 462 structures which fall into the core regions of the digitized distribution plot in (*b*) as a function of resolution. The symbols used are indicated in the figure. There are several outliers relative to other structures at the same resolution, and these are marked on the plot, using the Brookhaven code for identification.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

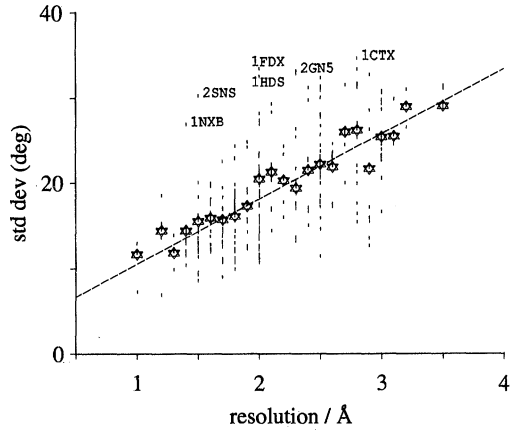


Figure 2. Correlation of the pooled standard deviations of χ_1 angles with resolution. The vertical dashes represent the standard deviations within individual proteins, and the stars show the standard deviations for the sets of values at each resolution. Major outliers are shown using the code from the PDB.

2. Experimental observations derived from many structures

(*a*) *Local conformation*

Within a protein the local conformation of a residue can be summarized by main chain and side chain torsion angles. It has long been known that these are restricted to preferred ranges by steric interactions (Ramachandran *et al.* 1963) and the preference for staggered rotamers (Janin *et al.* 1978). A study of high resolution, well refined structures in the PDB can now reveal the observed extent of these restrictions (Morris *et al.* 1992). From a distribution of ϕ, ψ values in these structures (figure 1*a*) it is possible to define three core regions (figure 1*b*) occupying only 14% of ϕ, ψ space. As expected these regions correspond to the α_R, β and α_L regions of a Ramachandran plot. By calculating the percentage of residues in these core regions and plotting this value against the resolution of the structure (figure 1*c*) we find that, in high resolution structures, 90% of the residues adopt core ϕ, ψ values. This is a very high percentage and indicates that local steric hindrance is a very important factor in determining local conformation, even in the coil or loop regions.

Further evidence for the importance of this effect is seen in the χ_1 values. The substituents at the tetrahedral sp^3 C_β prefer to adopt a staggered conformation relative to the sp^3 bonds of the C_α atom (Janin *et al.* 1978). The standard deviation in χ_1 values from the closest of the three favoured staggered states (60°, 180°, −60°) can be calculated and plotted for the entire protein against resolution (figure 2). In high resolution structures this standard deviation decreases to only ±15°.

It is possible to calculate expected target values for all torsion angles for a protein structure which has been determined to high resolution and refined (see table 1). For example the χ_3 disulphide torsion angle (C_β–S–S–C_β) adopts only two preferred values corresponding to the left-handed ($-85.8 \pm 10.7°$) and right-handed ($96.8 \pm 14.8°$) conformers. Similarly the peptide bond angle ω also exhibits a tight distribution for both the *trans* and *cis* conformers. However, for the ω angles there is no correlation between the standard deviation and resolution, since the standard deviation is determined primarily by the restraints applied during the refinement procedure. The

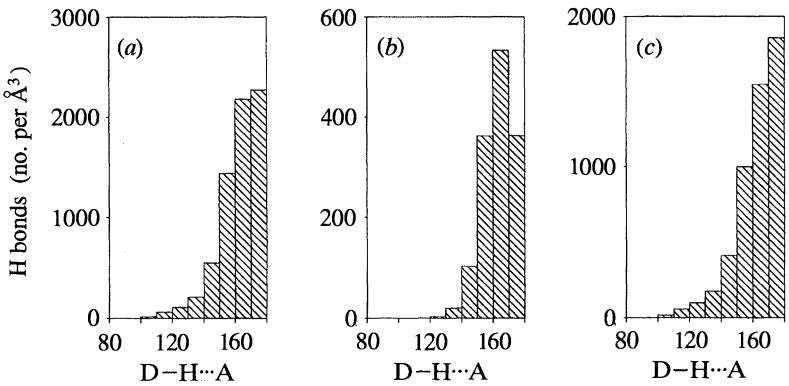


Figure 3. The D–Ĥ...A frequency distributions derived from the high resolution data set and normalized to account for differences in volume. To qualify as a hydrogen bond, the interaction must satisfy all criteria except those for the angle of the hydrogen, which is varied here. Secondary structures are defined by a modified version of the Kabsch & Sander algorithm (Kabsch & Sander 1983). (*a*) All hydrogen bonds in the dataset, including main chain and side chain groups, and protein-solvent hydrogen bonds. (*b*) α-helical main chain $i...i+4$ hydrogen bonds. (*c*) All main chain hydrogen bonds, excluding α-helical $i...i+4$ bonds.

Table 1. *Summary of geometric and conformational parameters in well resolved protein structures*

ϕ, ψ in Ramachandran *core* region	> 90 %
χ_1 (gauche +) angles	$-66.7° \pm 15.0°$
χ_1 (gauche −) angles	$+64.1° \pm 15.7°$
χ_1 (trans) angles	$183.6° \pm 16.8°$
pooled standard deviations of χ_1 angles	15.7°
disulphide χ_3 angles LH	$-85.8° \pm 10.7°$
disulphide χ_3 angles RH	$+96.8° \pm 14.8°$
proline ϕ angles	$-65.4°$
α-helix ϕ angles	$-65.3° \pm 11.9°$
α-helix ψ angles	$-39.4° \pm 11.3°$
hydrogen bond energy (kcal mol^{-1})	-2.03 ± 0.75
peptide bond ω angles (trans)	$179.6° \pm 4.7°$
C_α–N–C′–C_β virtual dihedral angles	$33.9° \pm 3.5°$

tightness of restraints for these angles should be determined either from small molecule studies or from the very high-resolution protein structures which are now being solved and refined without such restraints.

(*b*) *Hydrogen bonds*

In a similar manner the geometry of hydrogen bonds and the hydrogen bonding properties of the individual amino acid can be explored by generating distance and angle distributions from highly resolved protein structures. We have recently constructed an atlas detailing the results from a statistical analysis of the hydrogen bonds formed by the different polar residues and groups in proteins (McDonald & Thorton 1993). Here we consider only two facets of this analysis; firstly the effects of secondary structure formation on the geometry of the main chain hydrogen bonds; then a study of the extent to which buried main chain donors and acceptors satisfy their hydrogen bonding potential.

Our algorithm for locating hydrogen bonds involves two steps. A set of possible

positions for a hydrogen (H) attached to each protein donor is generated, since the hydrogens are not visible in most electron density maps; then a search is made through all atoms to find donor (D) and acceptor (A) pairs which fit the specified geometric criteria. For this analysis we used the following criteria for a hydrogen bond: H...A distance < 2.5 Å; D–Ĥ...A angle $> 90°$; H...Â–AA angle $> 90°$, where AA is the atom attached to the acceptor, usually preceding it along the chain. These criteria were chosen to be in agreement with previous work (Baker & Hubbard 1984) and were found to reflect the observed distributions adequately. For the analysis a dataset of 63 polypeptide chains from 61 non-homologous ($< 30\%$ sequence identity) and non-analogous protein structures was used (Orengo 1993). The structures all have been determined to a resolution 2.0 Å or better and are well refined (R-factor $< 20\%$).

Figure 3 shows several frequency distributions of the D–Ĥ...A angle, calculated from these proteins, which have all been normalized to account for allowed volume. Figure 3*a* shows the distribution for all intramolecular hydrogen bonds; figure 3*b* shows this distribution just for the α-helical main chain hydrogen bonds; figure 3*c* shows the distribution for main chain hydrogen bonds excluding those found in helices. It is well known that a linear arrangement of the donor-hydrogen-acceptor atoms is energetically the most favourable (Taylor *et al.* 1983). Figure 3*b* shows clearly that the hydrogen bonds in the α-helix are somewhat distorted, presumably reflecting the balance made between competing interactions in the well-packed helix.

Similarly figure 4 compares the angle at the acceptor for main chain hydrogen bonds in different secondary structures and those involving side chain acceptors. For an sp^2 hybridized oxygen (e.g. the peptide carbonyl), the preferred acceptor angle H...Â–AA is known to be 120° (Taylor *et al.* 1983), corresponding to the alignment of the hydrogen with the lone pair of the acceptor. The side chain hydrogen bonds clearly show this preference, with a peak in the distribution between 120–130°. In contrast the main chain hydrogen bonds are more distorted with H...AA angles closer to 180°. Ideally each carbonyl oxygen would form two hydrogen bonds, one for each lone pair. However, accessibility to the main chain carboxyl oxygen is often restricted by local secondary structure and 75% of buried carbonyls form only a single hydrogen bond, which tends to be more linear. A similar observation was made for small molecule structures, where linearity is observed in single NH...OC bonds (Taylor *et al.* 1983).

The H...Â–AA distributions for random coil and β-strand residues also have a large peak due to hydrogen-bond-like interactions between 80–90° (which fall outside the above definition of the hydrogen bond). These constitute intra-residue bonds, most often between the amide and carbonyl of the same residue, when it adopts a β conformation, but also including a few side chain/main chain bonds in serine, threonine and cysteine. Although these hydrogen bonds fall within the accepted criteria, and they clearly represent a favourable electrostatic interaction, they are very distorted and can be considered as secondary interactions which occur frequently because they form part of a larger stable structure.

A study has also been made to consider the extent to which the hydrogen-bond potential of protein main chain groups is satisfied, either by solvent molecules or intramolecular interactions. It has long been noticed that it is relatively uncommon to find unsatisfied hydrogen-bond donors or acceptors buried in the core of the protein. Clearly an unsatisfied group is energetically unfavourable, but in some circumstances this lack of interaction may be compensated for by other interactions.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

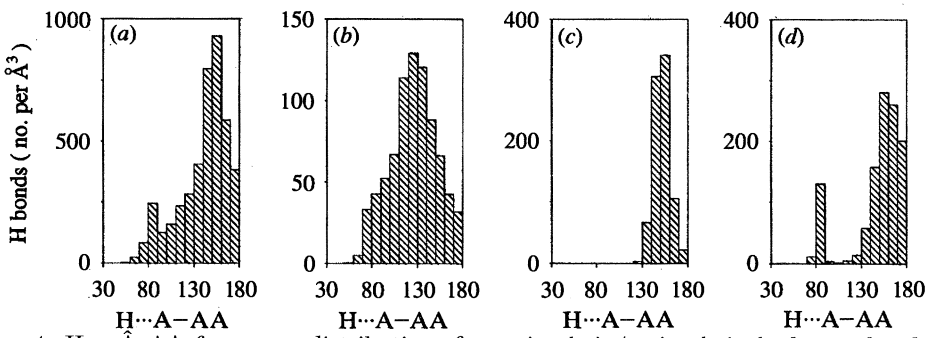


Figure 4. H...Â–AA frequency distributions for main chain/main chain hydrogen bonds. To qualify as a hydrogen bond the interaction must satisfy all criteria except those for both angles at the acceptor, which are varied here. (*a*) All hydrogen bonds. (*b*) Hydrogen bonds accepted by side chains. (*c*) α-Helical main chain $i...i+4$ hydrogen bonds only. (*d*) β-sheet main chain hydrogen bonds only.

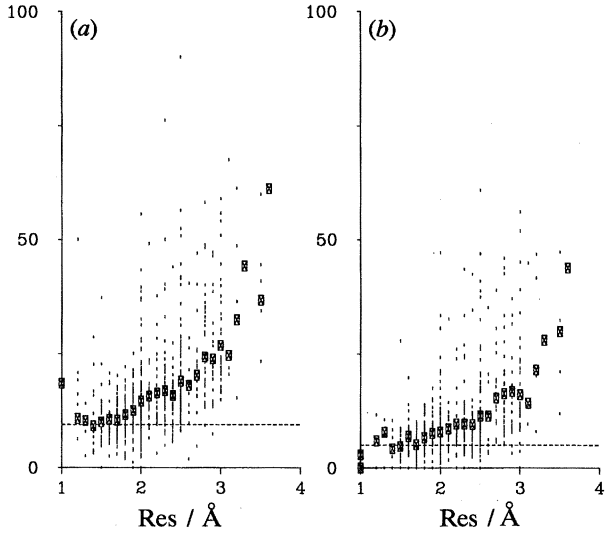


Figure 5. The percentage of buried donor and acceptor groups which do not form hydrogen bonds. Dashes represent individual proteins; boxes indicate *a*, the mean percentage of unsatisfied groups for that resolution; the horizontal dashed line represents the average value for the high resolution dataset: (*a*) main-chain NH groups; (*b*) main-chain CO groups.

Therefore a study of the high-resolution protein dataset was made. Only buried donors or acceptors were considered, since any surface group will presumably have its hydrogen-bonding potential satisfied by solvent molecules. A buried atom is defined in this study as one having zero solvent accessibility, according to the method of Lee & Richards (1971), calculated the program ACCESS, written by Dr Simon Hubbard. Accessibility was calculated for each individual polypeptide chain, excluding all water molecules and any other heteroatoms.

Using the criteria given above, for proteins in the high resolution dataset, we find that 9.4% of buried main chain NH groups and 5.0% of buried main chain CO groups are unsatisfied. If it is assumed that any exposed groups are solvated then we find that only 5.7% of all main chain amides and 2.1% of all main chain carbonyls are unsatisfied. Figure 5 plots the observed percentages for each protein of known structure against the resolution of the structure. It is immediately apparent that as

the resolution of the protein improves, the percentage of unsatisfied donors and acceptors decreases. Indeed the figures quoted above are likely to be an overestimate, since the percentages continue to decrease even beyond a resolution of 2 Å and an R-factor of 20%.

We have briefly examined these unsatisfied groups to ask if there is any common cause. We do not find a single dominant effect, although several examples involve prolines. If a proline interrupts a helix or strand, then the partner carbonyl which would have hydrogen-bonded to the proline NH, is often buried but cannot form a bond. Many of the unsatisfied groups are quite close to suitable partners, but fall just outside the hydrogen bond criteria. Often the angle of approach is not ideal, or the separation just too long. However, even if these near-misses are taken into account, there still remain some donors and acceptors which do not form hydrogen bonds. They are rare and this suggests that it is relatively expensive not to satisfy hydrogen-bonding potential. This is a major constraint on protein folding which limits the repertoire of stable protein folds. It provides a rationale for the high frequency of the secondary structures, in which all the main chain donors and acceptors are satisfied, allowing the main chain peptide groups to be buried in the hydrophobic core.

3. Modelling interactions in proteins

(*a*) *The hydrophobic effect and recognizing protein folds*

The database of protein structures can be used to address the problem of the hydrophobic effect and its contribution to protein folding. One of the earliest observations about protein structures was the presence of a tightly packed hydrophobic core, from which water is generally excluded. More recent comparisons of related protein sequences show that the buried core residues are well conserved, suggesting that they are crucial for the intact native structure. Many different hydrophobicity scales for the 20 amino acids have been derived, using a variety of experimental systems (Fauchene & Pliska 1983; Eisenberg 1984).

Our interest in an energetic representation of the hydrophobic effect arose when we were considering how to identify protein topology from amino acid sequence. Experimental observations as described above suggested that it would be crucial to take this term into effect. Furthermore, Novotny *et al.* (1984) elegantly demonstrated the inability of conventional potentials to distinguish between the correct and incorrect folds. Their test problem was very simple. The sequences of myohemerythrin and an immunoglobulin domain of identical length were exchanged. Both the two native structures and the two 'misfolded' proteins were then subjected to energy minimization using the CHARMm (Brooks *et al.* 1983) force-field. It proved to be impossible to distinguish between the native and misfolded structures on the basis of the calculated energy sums. Subsequently Novotny *et al.* showed that the potential could be modified to take into account the effect of solvent. Indeed Eisenberg & McLachlan (1986) were able to distinguish correct models from misfolded models by using a simple solvation energy model alone. Later work (Holm & Sander 1992; Baumann *et al.* 1989) has emphasized the dominant effect of solvation in determining the fold of a protein.

Our approach to calculating an empirical solvation potential is based on the observed solvent accessibility distributions for the 20 amino acids derived from a dataset of high resolution structures. The potential simply measures the frequency

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

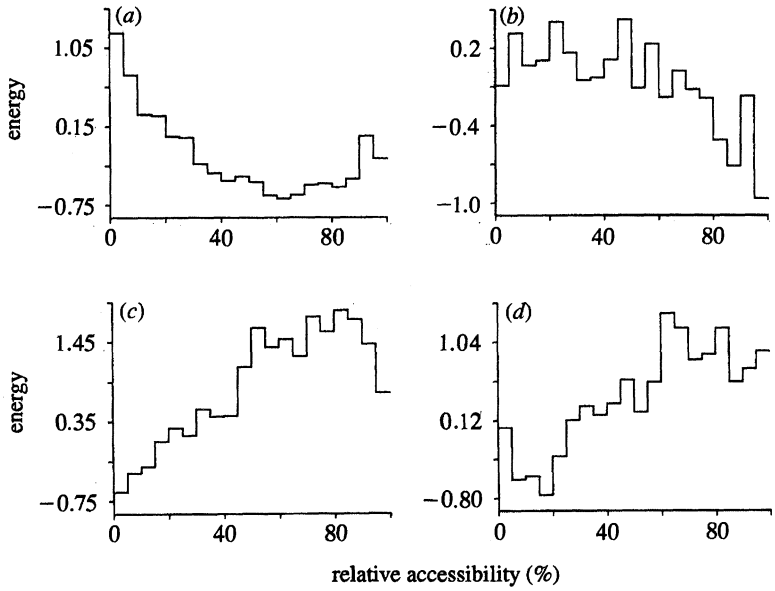


Figure 6. Empirical solvation potentials. (*a*) Glutamic acid. (*b*) Glycine. (*c*) Leucine. (*d*) Tyrosine.

with which each amino acid species is found with a certain degree of solvation. We define the solvation potential for a given amino acid residue a as

$$\Delta E^{a}_{\mathrm{solv}}(r) = -kT \ln\left[\frac{f^{a}(r)}{f(r)}\right],$$

where r is the percentage residue accessibility (relative to residue accessibility in GGXGG extended pentapeptide). Residue accessibilities here were calculated using the DSSP program of Kabsch & Sander (1983). The solvation potentials were generated with a histogram sampling interval of 5%. The potentials, which are illustrated in figure 6, clearly show the hydrophobic nature of the amino acids. For example, the energy of a completely buried aspartic or glutamic acid is high, and gradually decreases as the accessibility to solvent increases. Leucine and isoleucine show the opposite effect, with low energy when buried, increasing to high energy when exposed.

This potential can be evaluated using the method previously described by Hendlich *et al.* (1990) in which each sequence in a library of folds is fitted to all other sequential structural segments in the library and the solvation energy term summed in each case. (The library of folds consists of a set of non-homologous structures taken from the PDB (Jones *et al.* 1993).) For example a sequence of 120 residues will be threaded onto all 120 residue structural segments in the library and the solvation energy calculated for each structure. This is done by calculating the accessibility for each position in a structure and then calculating the energy involved in locating the particular amino acid in the sequence at that site. These energies are summed over the whole sequence and are then compared. If the potential is a good discriminator, the lowest energy should correspond to the native fold. In our tests, the fold library

comprised 102 polypeptide chains, of which the longest must be excluded as there are no other sequential structural segments available for testing. After jack-knifing to exclude any related structure in the derivation of the potentials, the solvation potential recognized 84 out of 101 (83%) possible folds.

This is a surprisingly good result and illustrates that this relatively simple potential is a rather powerful discriminator between folds and also underlines the importance of the hydrophobic effect in determining the structure of a protein.

(*b*) *A comparison of three theoretical approaches to modelling atomic interactions in proteins*

The large number of high quality protein structures provide experimental three dimensional distributions of side chain/side chain, side chain/atom and atom/atom interactions. These data can be used to benchmark theoretical studies by comparing the calculated energy minima with the observed spatial distributions. We have calculated experimental distributions for many different types of atom/side chain interactions using the program suite SIRIUS (Singh & Thornton 1993). Here we present only the comparison of some theoretical calculations on phenylalanine/carboxylate interactions with the experimental data; a fuller version will appear elsewhere (Mitchell *et al.* 1993). A previous study had suggested that the oxygens tend to lie in the plane of the ring (Thomas *et al.* 1982). Three theoretical methods were tested: first, the drug design program GRID (Goodford 1985), which is suited to calculating side chain/atom distributions; secondly, the empirical potential CHARMm (Brooks *et al.* 1983); and thirdly, the distributed multipole analysis (DMA) electrostatic approach (Stone 1981).

The experimental interaction geometries were extracted from a set of 62 high resolution structures (Singh & Thornton 1993), taken from the PDB (Bernstein *et al.* 1977). All carboxylate/aromatic interactions were extracted using the criterion that the oxygen/carbon separation should be less than the sum of the van der Waals radii plus 1 Å.

All the interacting pairs were transformed to the frame of an idealized phenylalanine to produce the distributions. The coordinate system was defined so that the origin lies at the centre of the aromatic ring, the xy plane in the plane of the aromatic ring with the x-axis passing through the C_γ atom, and the z-axis perpendicular to the plane of the ring. The position of an atom relative to the ring can be defined in spherical polar coordinates where ϕ is the angle in the xy plane and θ the angle between the position vector of the atom and its projection onto the xy plane. $\phi = 0°$ lies along the x-axis; $\theta = 0°$ lies in the xy plane and $\theta = 90°$ lies above the ring. A random filling of space (ignoring volume acclusion) would give rise to an evenly spread ϕ distribution, while the variation of the available volume with θ means that the expected θ distribution is cosinusoidal, with the largest number of observations expected at $\theta = 0°$ and the smallest number at $\theta = 90°$.

The program GRID (Goodford 1985) uses an empirical energy function to calculate the interaction energy between a probe and its target molecule. The intermolecular potential detailed by Goodford includes a standard Lennard–Jones term and a hydrogen bonding potential. Its electrostatic model is based on atomic charges, with the dielectric effect incorporated by a macroscopic dielectric constant for the protein interior DPRO, and the dielectric constant of the environment DWAT. Both terms are involved in evaluating the electrostatic energy using a 'method of images' approach. Since we are interested in the interaction of the probe with the side chain

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

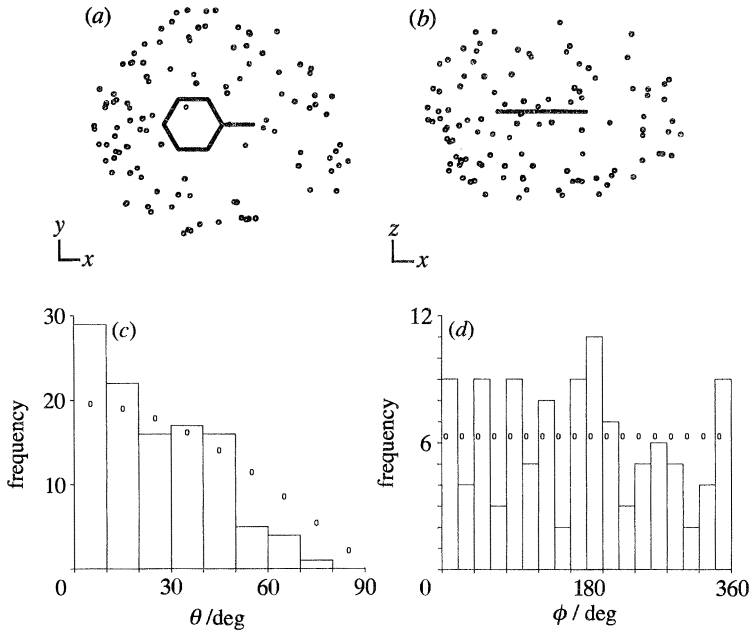


Figure 7. (*a*), (*b*) Scatter plots viewed from two perpendicular directions showing the distribution of 110 carboxylate oxygens around phenylalanine in 62 high resolution protein structures. (*c*) The histogram shows the distribution of the out-of-plane angle θ for these phenylalanine/carboxylate oxygen interactions; θ is defined such that $\theta = 0°$ corresponds to an interaction in the molecular *xy* plane, and $\theta = 90°$ to one directly above the origin, which is the ring centre. The statistically expected values are shown by the small oval symbols. (*d*) The corresponding histogram for the in-plane angle ϕ. This angle is defined such that $\phi = 0°$ corresponds to the positive *x*-axis, which passes through C_γ of phenylalanine.

rather than the main chain, the phenylalanine is modelled by toluene (C_β upwards). GRID evaluates the interaction energy of a single atom (in this case a carboxylate oxygen) at each point on a three dimensional grid. A cubic box of side 20 Å, is centred on the aromatic ring, with grid points at 0.5 Å intervals. The output is a set of energies, one corresponding to each grid point, which is then used to create energy contour maps.

The CHARMm potential is the basis of one of the most commonly used molecular modelling packages. The intermolecular energy consists of a charge/charge electrostatic term and a Lennard–Jones repulsion–dispersion term. Partial charges are placed on each atom, the values used here being based on those supplied with version 21 of CHARMm.

In contrast, the DMA method calculates the electrostatic term by placing a charge, dipole, quadrupole, etc., on each atomic centre. The multipoles are derived directly from *ab initio* quantum mechanical calculations of the molecular wavefunction, and they represent the anisotropy of the local charge distribution. This rather sophisticated electrostatic term is combined with a 'hard spheres' model to ensure reasonable interaction distances and prevent the two molecules from collapsing in one another (Buckingham & Fowler 1985). This approach is reasonable since the orientational variation of the energy is usually dominated by the electrostatic component. The DMAS are derived from 6–31G* wavefunctions; the *ab initio* calculations were carried out using CADPAC (Amos & Rice 1987) and the Cambridge

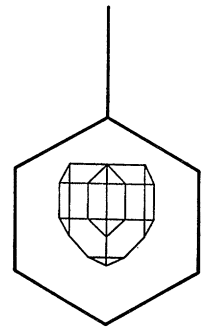


Figure 8. GRID contours for the interaction of a carboxylate oxygen probe with a phenylalanine side-chain, which is modelled by toluene (with DWAT = 80.0). The contours show the region with an interaction energy lower than -6.7 kJ mol^{-1}.

Direct SCF program (C. W. Murray, J. S. Andrews & R. D. Amos), with the basis sets being taken from the CADPAC library. A 2.0 Å radius is used for the combined –CH atoms and 1.4 Å for oxygen. All the minimizations and energy calculations are carried out using A. J. Stone's program ORIENT.

For both the CHARMm and DMA approaches, the phenylalanine/carboxylate interaction is modelled by toluene and acetate. No attempt is made to include the effects of solvation in either calculation ($\epsilon = 1$ in CHARMm). To locate energy minima the molecules are positioned in 50 different starting orientations, and the energy minimized using the algorithms provided. The energy of a fully stacked structure is also evaluated.

The results of this study are shown in figures 7–10. The experimental distribution of carboxylate oxygens around phenylalanine (figure 7) shows a clear preference for the oxygen to lie close to the plane of the aromatic ring. Over 90% of the carboxylates lie at $\theta < 50°$, with no oxygens lying directly over the aromatic ring. A comparison of the observed and expected θ distributions (figure 7) shows significant non-random behaviour.

What do the three different energy calculations predict? According to GRID the most favourable position for a carboxylate is directly above the ring centre, as shown by the energy contour plot (figure 8). The energy involved is small (-8.6 kJ mol^{-1}) and consists mainly of the Lennard–Jones contribution, which is minimized directly over the ring. The electrostatic contribution is negligible ($+0.1$ kJ mol^{-1}) because the program assigns very small negative charges ($-0.032e$) to the aromatic carbons. This model does not take into account the π-electron density above and below the ring which would be expected to interact more strongly and repulsively with the oxygen.

In contrast, the CHARMm results (figure 9) indicate that structures in which the oxygens lie close to or in the plane of the aromatic ring are favoured. Any structure in which the oxygen is stacked above the plane of the ring is energetically unfavourable. The minimum energy is -18 kJ mol^{-1}, of which -5.4 kJ mol^{-1} is electrostatic. The main difference compared to the GRID calculation is the size of the charges on the ring atoms ($-0.1e$ on ring carbons and $0.1e$ on the attached hydrogens). Even these significantly underestimate the overall quadrupole of a benzene ring (Price 1985). The lowest energy structure in the CHARMm calculations has the plane of the carboxylate almost perpendicular to the plane of the aromatic ring.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

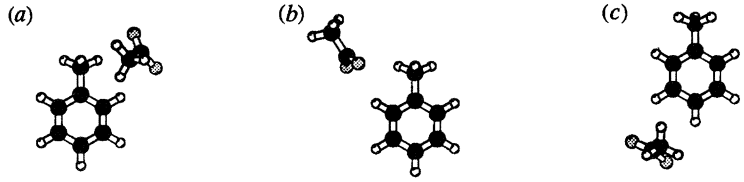


Figure 9. The three lowest energy toluene/acetate CHARMm minima. All have their carboxylate oxygens approximately co-planar with the aromatic ring, but the molecular planes are almost perpendicular. Electrostatic energy: −5.4 kJ mol^{-1} (*a*); −6.6 kJ mol^{-1} (*b*); −7.4 kJ mol^{-1} (*c*). Total energy: −18 kJ mol^{-1} (*a*); −17 kJ mol^{-1} (*b*), (*c*).

The DMA calculations produce three minima (figure 10), and in all these structures the two carboxylate oxygens lie close to or in the plane of the aromatic ring with the two molecules almost co-planar. The electrostatic energies are between −13 and −17 kJ mol^{-1}. This compares to an equivalent energy of −439 kJ mol^{-1} for a charge/charge interaction between methylguanidinium (modelling arginine) and acetate.

The observed avoidance of large θ values can be understood in terms of an electrostatic repulsion between the negatively charged carboxylate atoms and the π-electron cloud of the aromatic ring. The DMA model gives a correct prediction, with the lowest energy structures having the oxygens in the plane of the ring. Stacked structures are repulsive. The energy difference between stacked and co-planar structures is more than 30 kJ mol^{-1}, suggesting that electrostatics give ample reason for the former to be avoided. CHARMm gives very similar oxygen positions, although its non-inclusion of the 'lone-pairs' (which favour co-planar interactions) leads to a highly non-planar interaction geometry. In contrast GRID is not suited to the phenylalanine/carboxylate system since it fails to treat the quadrupole of the aromatic ring.

This study illustrates the power of using the experimental distributions to determine the approximations which can be used in modelling proteins. Even though the aromatic rings are often considered to be non-polar and their electrostatic contribution neglected, this term is clearly strong enough to influence their interactions and packing in protein structures. In modelling these side chains it is therefore essential that this term be taken into account.

4. Protein-peptide complexes

There are now 11 protein-peptide complexes in the PDB. We have analysed these complexes to learn about the peptide conformation when bound and the interactions which stabilize the complex. The binding proteins include antibodies (Stanfield *et al.* 1990; Rini *et al.* 1992), an MHC class I molecule (Madden *et al.* 1992; Fromont *et al.* 1992), cAMP-dependent kinase (Knighton *et al.* 1991), proteinases (Bailey *et al.* 1993), SH2 domains (Waksman *et al.* 1992, 1993), calmodulin (Ikura *et al.* 1992) and thrombin (Banner & Hadvary 1991). The biological functions and the reasons why the above proteins bind peptide are very different. For example, the interaction between an antibody and a peptide is entirely fortuitous, in that the conformation of the antibody combining site just happens to be able to bind one conformation of the peptide. In contrast the MHC molecule has evolved to bind peptides of a specific length with some sequence restrictions. The peptides which inhibit enzymes have been designed and synthesized to bind into the active site and often mimic cognate

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

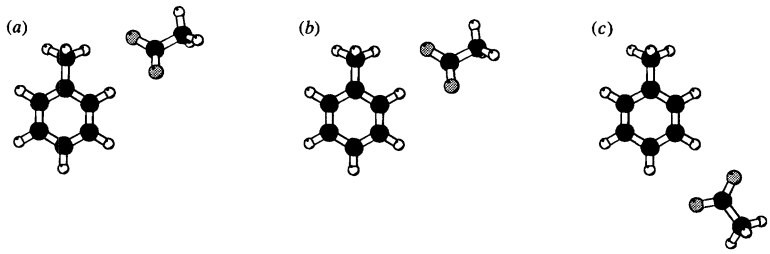


Figure 10. The DMA global (*a*) and local (*b*), (*c*) minima for toluene/acetate. Structures (*a*) and (*b*) have virtually identical electrostatic energies (-17 kJ mol^{-1}), with a difference of only 0.1 kJ mol^{-1}. In all three structures, the two molecules are approximately co-planar. The electrostatic energy of (*c*) is -13 kJ mol^{-1}.

sequences. The role of the SH2 domains is to recognize a specific sequence which usually occurs as part of another protein, but only if it contains a phosphorylated tyrosine. Inspection of the conformations of the complexes show that it is these functional and evolutionary requirements which dictate the binding and interactions of a peptide with a protein.

(*a*) *Conformation*

The peptides bind to the proteins in many different conformations. The majority adopt an extended β-conformation, some include a classic β turn, and the peptide which binds to calmodulin forms a classic α-helix. In general the main chain ϕ, ψ angles fall within the allowed regions of the Ramachandran plot (see figure 11) (85% lie within the core region), showing that the peptides generally bind in a relatively low energy conformation.

(*b*) *Loss of accessible surface area*

The loss of accessible area (ΔASA) on complex formation was calculated for each complex, using the ACCESS program (see above). These ΔASA range from 450–1500 Å^2 in total, but most lie between 600–800 Å^2 (Allen *et al.* 1983). For comparison, the surface area lost in forming a protein–protein complexes ranges between 650–1000 Å^2 (Allen *et al.* 1983) for each protein (Janin and Chothia 1990). This loss of surface area is often taken as an approximate measure of the hydrophobic solvent exclusion effect.

(*c*) *Hydrogen bonds*

The number of intermolecular hydrogen bonds formed between the protein and the peptide ranges between 1–16, with some complexes dominated by these polar interactions and others having very few.

(*d*) *Main chain/side chain specificity*

The relative contribution of main chain and side chain interactions to the binding energy was evaluated using the ΔASA. The main chain can contribute a maximum of 81 Å^2 in glycine, whereas a tryptophan residue was 248 Å^2 (Allen *et al.* 1983) maximum standard accessibility. The main chain contribution ranges from 4–31% of the total ΔASA. Some of the peptides seem to optimize the interaction of the peptide backbone with the protein. This is especially true for the enzyme inhibitors. In contrast other complexes have few interactions between the protein and the main chain. These differences probably reflect the different biological functions of these complexes.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

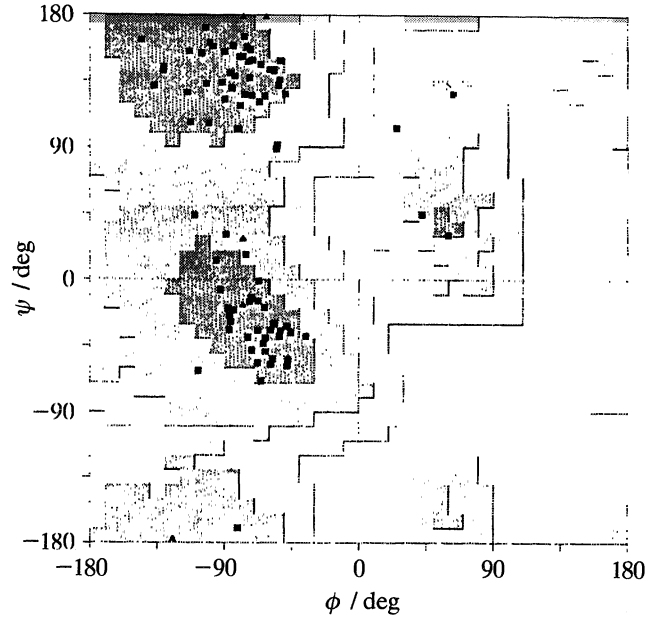


Figure 11. The Ramachandran plot of dihedral angles from the 11 peptides. Most angles (85%) fall within the core region, a few in the allowed and only one in the disallowed area.

The enzymes have evolved to be non-specific proteinases which can cleave a great variety of polypeptide sequences. For these promiscuous proteinases it is logical that the recognition process should depend heavily on the main chain groups, which will be common to all sequences. In contrast the more specific proteinases, which recognize only a single specific sequence have developed several subsite specificities. In a similar manner the MHC class I molecules have a limited specificity, but always recognize a short peptide, usually eight or nine residues long. Here there is specificity at some subsites, much as in the proteinases, but it is essential that some of the residues are exposed on the surface to allow presentation to and recognition by the T-cell receptors. The MHC class I molecule enables binding to occur by having specific recognition sites for the N and C-terminal groups, which are common to all the peptides, combined with buried pockets for only two or three of the side chains. The other side chains protrude from the peptide cleft where they can be easily 'seen' by the receptor.

Recognition of peptides by the SH2 protein domains has rather different biological requirements. Here the peptide must only be recognized when its tyrosine becomes phosphorylated. As would be expected therefore, the recognition site is dominated by the pocket which binds the phosphate attached to the tyrosine. The highly charged phosphate is completely buried in the complex with five hydrogen bonds to the protein. More extended sequence specificity is achieved by the interaction of the peptide side chains with a hydrophobic pocket. Three charged groups on the protein are involved which will certainly increase the binding energy considerably. If the tyrosine is not phosphorylated then these groups would still be buried in the complex but unable to form hydrogen bonds. As shown above (§2*b*) this is a very unlikely and expensive energetically, to the extent that presumably the unphosphorylated peptide can no longer be bound by the SH2 domains. This is exactly what is required for the biological function. Similarly the recognition of the peptide helix by

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

calmodulin acts as a signal. Here the backbone of the peptide is involved in the intramolecular helical hydrogen bonds and plays little part in the interaction. Most of the binding energy must come from side chain interactions, predominantly van der Waals contacts, which will, however, be dependent upon the peptide adopting the correct helical conformation. Binding involves a major conformational rearrangement of the calmodulin, and it is this which generates the signal and changes the affinity for calcium.

In contrast to all the above interactions, the binding of a peptide by an antibody is fortuitous, rather than the result of evolution under biological constraints. The only condition is that the antibody should bind the peptide reasonably tightly. In principle any conformation of the peptide will suffice, as long as it occurs in solution. In practice, the known examples suggest that the bound conformations are the classic low energy states observed in proteins. It is irrelevant whether the interactions are made with the main chain or side chain atoms, or whether the interaction is predominantly polar or hydrophobic. The classic concavity of the antibody binding site may well favour peptide conformations involving turns, and each peptide observed in a complex does indeed incorporate a turn in the binding site.

In summary, peptides can be recognized by proteins in many different conformations, dominated by different energetic terms and involving interactions with different components of the peptide. The dominating factor is the biological function of the complex and its evolutionary history.

References

Allen, F. H., Kennard, O. & Taylor, R. 1983 Systematic analysis of structural data as a research technique in organic chemistry. *Acc. chem. Res.* **16**, 146–153.

Amos, R. D. & Rice, J. E. 1987 CADPAC: the Cambridge analytical derivatives package, Issue 4.0. Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, U.K.

Bailey, D., Cooper, J. B., Veerapandian, B., Blundell, T. L., Atrash, B., Jones, D. M. & Szelke, M. 1993 X-ray crystallographic studies of complexes of pepstatin-A and a statine-containing human renin inhibitor with endothiapepsin. *Biochem. J.* **289**, 363–371.

Baker, E. N. & Hubbard, R. E. 1984 Hydrogen bonding in globular proteins. *Prog. biophys. molec. Biol.* **44**, 97.

Banner, D. W. & Hadvàry, P. 1991 Crystallographic analysis at 3.0 Å resolution of the binding to human thrombin of 4 active site-directed inhibitors. *J. biol. Chem.* **266**, 20085–20093.

Baumann, G., Frommel, C. & Sander, C. 1989 Polarity as a criterion in protein design. *Protein Engng* **2**, 329–334.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. 1977 The protein data bank: a computer based archival file for macromolecular structures. *J. mol. Biol.* **122**, 535–542.

Brooks, B. R., Bruccoleri, B. D., Olafson, D. J., States, D. J., Swaminathan, S. & Karplus, M. 1983 CHARMm: a program for macromolecular energy, minimisation and dynamics calculations. *J. Comput. Chem.* **4**, 187.

Buckingham, A. D. & Fowler, P. W. 1985 A model for the geometries of van der Waal complexes. *Can. J. Chem.* **63**, 2018.

Eisenberg, D. 1984 The hydrophobic moment detects periodicity in protein hydrophobicity. *A. Rev. Biochem.* **53**, 595.

Eisenberg, D. & McLachlan, A. D. 1986 Solvation energy in protein folding and binding. *Nature, Lond.* **319**, 199–203.

Fauchere, J. L. & Pliska, V. 1983 Hydrophobic parameters-pi of amino acid side-chains from the partitioning of N-acetyl amino acid amides. *Eur. J. med. Chem.* **10**, 369.

Fremont, D. H., Matsumura, M., Stura, E. A., Peterson, P. A. & Wilson, I. A. 1992 Crystal

structures of 2 viral peptides in complex with murine MHC class-I H-2K(B). *Science, Lond.* **257**, 919–927.

Goodford, P. J. 1985 A computational procedure for determining energetically favourable binding sites on biologically important macromolecules. *J. med. Chem.* **28**, 849.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. & Sippl, M. J. 1990 Identification of native protein-folds amongst a large number of incorrect models – the calculation of low-energy conformations from potentials of mean-force. *J. mol. Biol.* **216**, 167–180.

Holm, L. & Sander, C. 1992 Evaluation of protein models by atomic solvation preference. *J. mol. Biol.* **225**, 93–105.

Ikura, M., Clare, G. M., Gronenbornm, A. M., Guiang, Z., Kee, C. B. & Bax, A. 1992 Solution structure of a calmodulin-target peptide complex by multidimensional NMR. *Science, Wash.* **256**, 632–637.

Janin, J., Wodak, S., Levitt, M. & Maigret, B. 1978 Conformation of amino acid side-chains in proteins. *J. mol. Biol.* **125**, 357–386.

Janin, J. & Chothia, C. 1990 The structure of protein–protein recognition sites. *J. biol. Chem.* **256**, 1602.

Jones, D. T., Taylor, W. R. & Thornton, J. M. 1992 A new approach to protein fold recognition. *Nature, Lond.* **358**, 86–89.

Kabsch, W. & Sander, C. 1983 Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637.

Knighton, D. R., Zheng, J., Ten Eyck, L. F., Xuang, N., Taylor, S. S. & Sowadski, J. M. 1991 Structure of a peptide inhibitor bound to the catalytic subunit of cyclic adenosine-monophosphate dependent protein-kinase. *Science, Wash.* **253**, 414–420.

Lee, B. & Richards, F. M. 1971 The interpretation of protein structures: estimation of static accessibility. *J. mol. Biol.* **55**, 379.

Madden, D. R., Gorga, J. C., Strominger, J. L. & Wiley, D. C. 1992 The 3-dimensional structure of HLA-B27 at 2.1 Ångstrom resolution suggests a general mechanism for tight peptide binding to MHC. *Cell* **70**, 1035–1048.

Mitchell, J. B. O., Nandi, C. L., Price, S. L., Singh, J., Snarey, M. & Thornton, J. M. 1993 A comparison of three theoretical approaches to the study of side-chain interactions in proteins. *J. chem. Soc. Faraday Trans.* (In the press.)

Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. 1992 Stereochemical quality of protein structure co-ordinates. *Proteins* **12**, 345.

McDonald, I. K. & Thornton, J. M. 1993 In preparation.

McGregor, M. J., Islam, S. A. & Sternberg, M. J. E. 1987 Analysis of the relationship between side-chain conformation and secondary structure for globular proteins. *J. mol. Biol.* **198**, 295–310.

Novotny, J., Bruccoleri, R. E. & Karplus, M. 1984 An analysis of incorrectly folded protein models – implications for structure predictions. *J. mol. Biol.* **177**, 787–818.

Novotny, J., Rashin, A. A. & Bruccoleri, R. E. 1988 Criteria that discriminate between native proteins and incorrectly folded models. *Proteins* **4**, 19–30.

Orengo, C. A., Flores, T. P., Taylor, W. R. & Thornton, J. M. 1993 Identification and classification of protein fold families. *Protein Engng* (In the press.)

Ponder, J. W. & Richards, F. M. 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.

Price, S. L. 1985 A distributed multipole analysis of the charge-densities of some aromatic hydrocarbons. *Chem. Phys. Lett.* **114**, 359.

Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. 1963 Stereochemistry of polypeptide chain configurations. *J. mol. Biol.* **7**, 95–99.

Singh, J. & Thornton, J. M. 1990 SIRIUS, an automated method for the analysis for the preferred packing arrangements between protein groups. *J. mol. Biol.* **211**, 595.

Stanfield, R. L., Fieser, T. M., Lerner, R. A. & Wilson, I. A. 1990 Crystal structures of an antibody to a peptide and its complex with protein antigen at 2.8 Å. *Science, Wash.* **248**, 712–719.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Stone, A. J. 1981 Distributed multipole analysis, or how to describe a molecular charge distribution. *Chem. Phys. Lett.* **83**, 233.

Taylor, R., Kennard, O. & Versichel, W. 1983 Geometry of the N–H . . . O=C hydrogen bond. 1. Lone pair directionality. *J. Am. chem. Soc.* **105**, 5761.

Thomas, K. A., Smith, G. M., Thomas, T. B. & Feldmann, R. J. 1982 Electronic distributions within protein phenylalanine aromatic rings are reflected by the three-dimensional oxygen atom environments. *Proc. natn. Acad. Sci. U.S.A.* **79**, 4843.

Rini, J. M., Schulze-Gahmen, U. & Wilson, I. A. 1992 Structural evidence for induced fit as a mechanism for antibody-antigen recognition. *Science, Wash.* **255**, 959–965.

Waksman, G., Kominos, D., Robertson, S. R. *et al.* 1992 Crystal-structure of the phosphotyrosine recognition domain SH2 of V-SRC complexed with tyrosine-phosphorylated peptides. *Nature, Lond.* **358**, 646–653.

Waksman, G., Shoelson, S. E., Pant, N., Cowburn, D. & Kuriyan, J. 1993 Binding of a high-affinity phosphotyrosyl peptide to the SRC SH2 domain – crystal-structures of the complexed and peptide-free forms. *Cell* **72**, 779–790.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

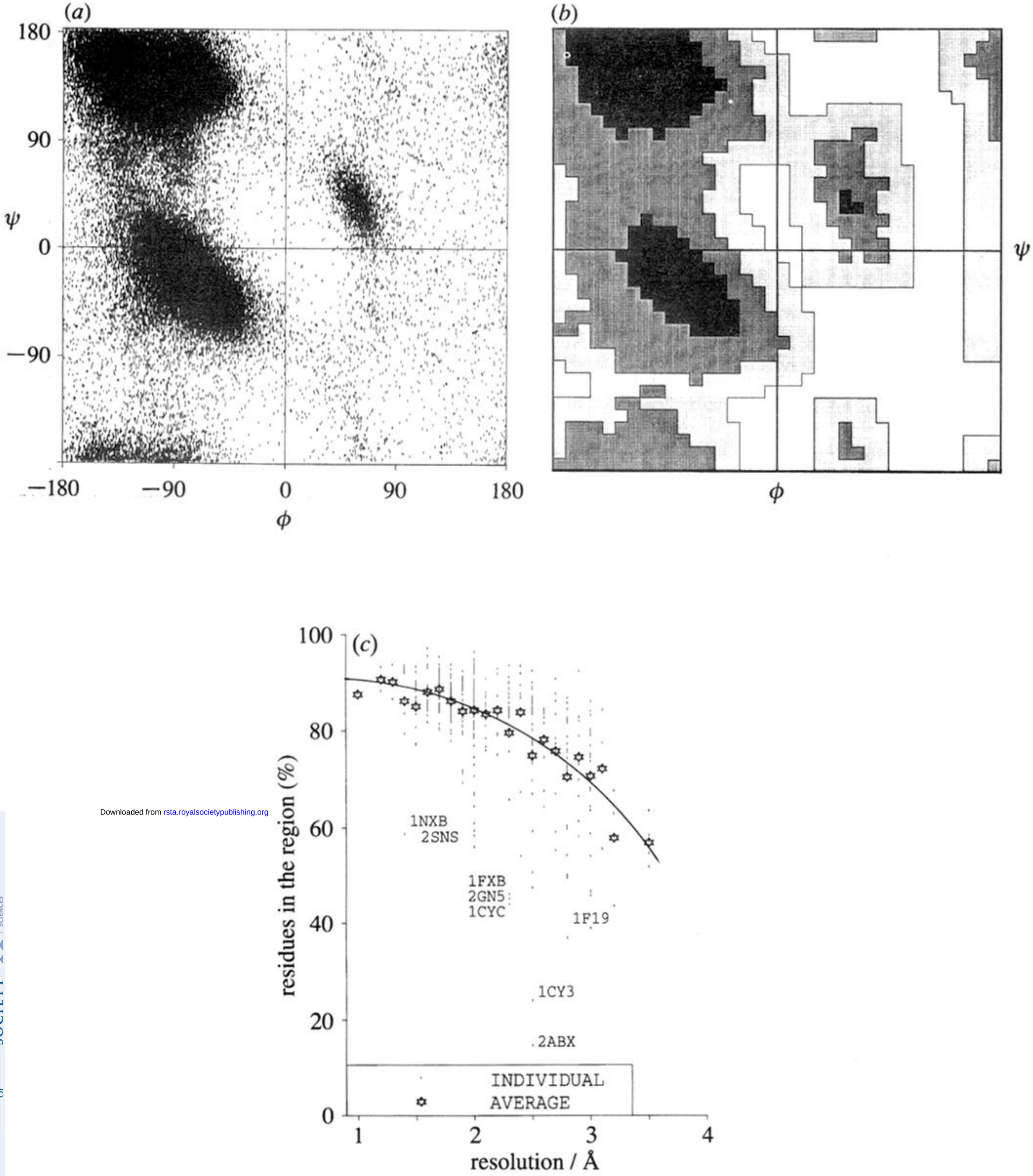
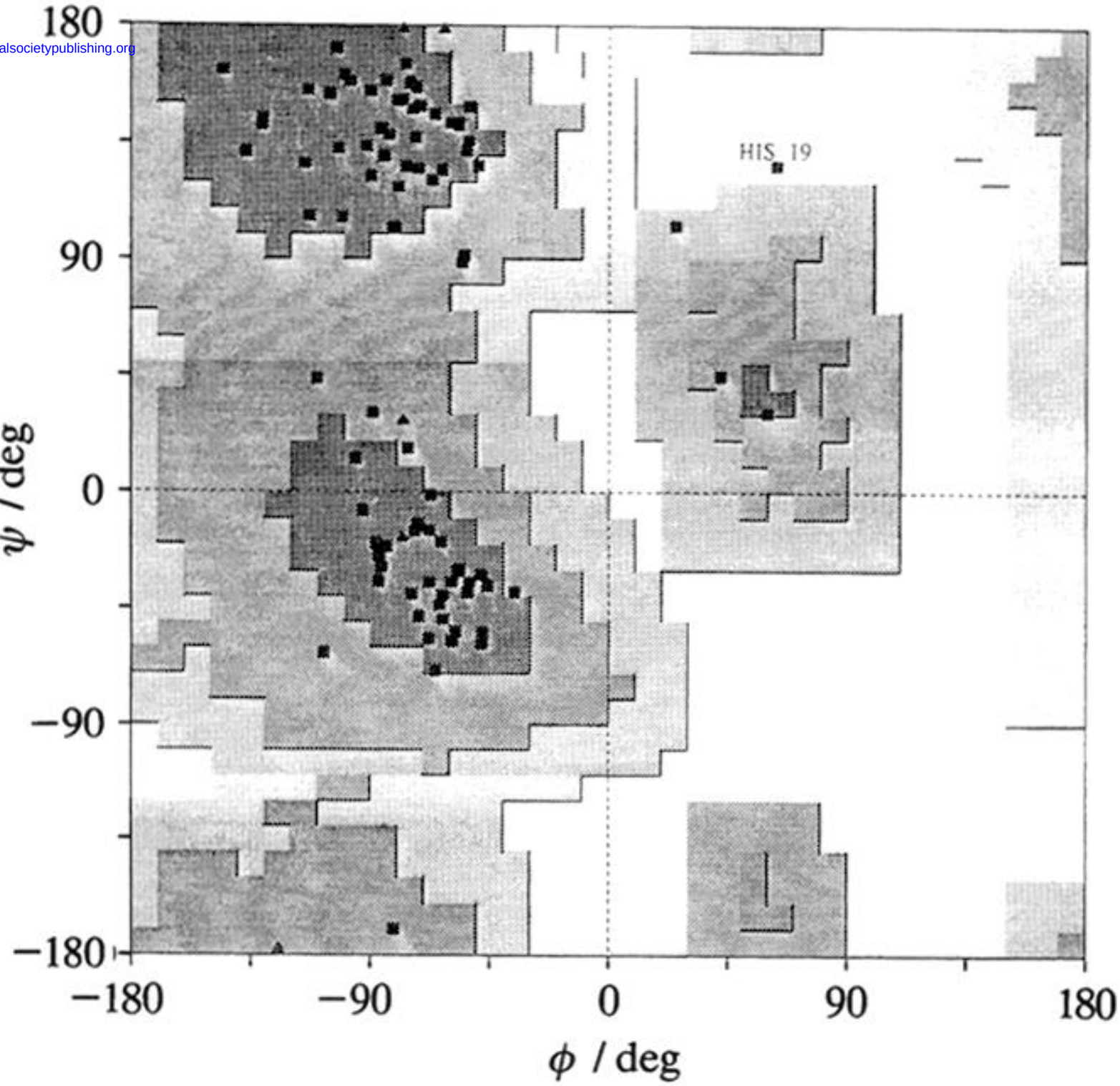


Figure 1. (*a*) Distribution of ϕ, ψ angles for 121 870 residues from 462 protein structures. Proline and glycine were not included. The angles were calculated from Brookhaven coordinates using the program SSTRUC (Smith, personal communication). (*b*) Digitized Ramachandran type plot derived from the actual distribution shown in (*a*). The population within each $10° \times 10°$ pixel was calculated and used to define 4 areas: the black areas are core, the dark grey allowed, and the pale grey, generous, the blank disallowed. (*c*) Plot to show the percentage of residues from 462 structures which fall into the core regions of the digitized distribution plot in (*b*) as a function of resolution. The symbols used are indicated in the figure. There are several outliers relative to other structures at the same resolution, and these are marked on the plot, using the Brookhaven code for identification.



igure 11. The Ramachandran plot of dihedral angles from the 11 peptides. Most angles (85%) fall within the core region, a few in the allowed and only one in the disallowed area.