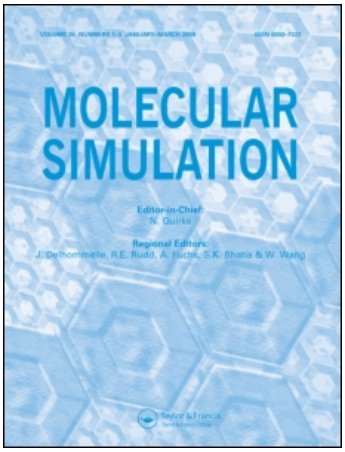

Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Preliminary Analysis of Water Molecule Distributions in Proteins

Julia M. Goodfellow[a]; Narmada Thanki[a]; Janet M. Thornton[a]
[a] Department of Crystallography, Birkbeck College, London

To cite this Article Goodfellow, Julia M. , Thanki, Narmada and Thornton, Janet M.(1989) 'Preliminary Analysis of Water Molecule Distributions in Proteins', Molecular Simulation, 3: 1, 167 — 182
To link to this Article: DOI: 10.1080/08927028908034626
URL: http://dx.doi.org/10.1080/08927028908034626

Molecular Simulation, 1989, Vol. 3, pp. 167–182
Reprints available directly from the publisher
Photocopying permitted by license only

PRELIMINARY ANALYSIS OF WATER MOLECULE DISTRIBUTIONS IN PROTEINS

JULIA M. GOODFELLOW, NARMADA THANKI and JANET M. THORNTON

Department of Crystallography, Birkbeck College, London WC1E 7HX

(*Received January 1988; in final form May 1988*)

In order to check the validity of force fields used in simulation studies it is necessary to have sufficient experimental structural data available with which to compare the simulated results. Accumulation of a large enough quantity of structural data is especially important for interactions for which the stereochemical constraints are expected to be fairly weak such as in the biologically important application of modelling the aqueous environment of proteins. We present a preliminary analysis of such experimental structural data for water molecules around the main chain and some sidechain atoms from 16 high resolution proteins. In general the distribution of water molecules around polar groups are fairly broad. However, where distinct preferential geometries are seen, they agree well with the expected stereochemistry and iso-energy contour maps. The broader distributions are shown to be due to the effects of a water molecule making multicontacts to the protein surface which may be characteristic of a specific secondary structure or sequence.

KEY WORDS: Proteins, aqueous environments, polar graphs

INTRODUCTION

Computer simulation and molecular mechanics studies of the aqueous hydration of macromolecules have increased over the past few years because of the increase in computational power generally available. Such studies are important for the modelling of proteins and their interactions in such applications as drug design and protein engineering. These simulations have included studies on the hydration of amino acids [1,2] oligopeptides [3,4] oligonucleotides [5,6] and whole protein molecules [7,8]. Many force fields are available for macromolecules [9–12] themselves and their interactions with water [13]. Because the reliability of the results from simulations depends on the realistic nature of these atom potential energy functions [14] it is important to be able to compare the simulated results with sufficient experimental data.

Several studies have used crystal hydrate data from x-ray or neutron crystallography in order to compare the structural results from simulations with experimental data [1,5,7,8]. These comparisons are often difficult because of insufficient structural data on water molecule interactions around the protein surface for one specific protein. The lack of data is due to several factors inherent in the crystallography of large molecules and is discussed in detail by Savage [15].

There have been many reviews of protein hydration such as those by Edsall and McKenzie [16,17]. Periodically, the structural data have been surveyed by Finney [18] Edsall and McKenzie [16,17] Baker and Hubbard [19] and Saenger [20]. However, information in these reviews on water interactions with sidechain atoms is limited. The aim of this study is to present a more detailed view of protein hydration based on a large amount of high-resolution crystallographic data for a large number, 16,

proteins. The data are presented graphically as distributions around potential hydrogen bonding main chain and side chain atoms and analysed in terms of spherical polar coordinates in order to look for preferred orientations.

Such distributions based on empirical data will provide models for the pattern of tightly bound water molecules around amino acid side-chains. We are then able to make detailed comparisons of the results of energy minimization studies using different potential energy functions with the distributions for the strongly bound waters. The optimum potential energy functions can then be used in full molecular simulations to establish a complete description (i.e. not only the tightly bound water molecules) of amino acid hydration.

METHOD

The sixteen high-resolution protein structures used in this analysis are indicated in Table 1. The atomic coordinates for both protein and water molecules are taken from the Brookhaven Data Bank [21] and used to generate any symmetry related water molecules within 5.0 Å of the protein surface.

Analysis of the interactions with main chain or sidechain atoms of a given residue involved the following operations:

(i) Generation of a reference group of atoms, with fixed geometry, for each residue type from the data of Momany et al. [9]. The reference groups are given in Table 2.

(ii) Calculation of the matrix needed to rotate each residue in the database of 16 proteins to the coordinate system of the appropriate reference group of atoms and rejection of the experimental x-ray data where the least squares fit to the fixed reference group is bad.

(iii) Application of this rotation matrix to the coordinates of any water molecule associated with that residue.

Table 1

PROTEIN	*FILN*[a]	*REFERENCES*
Actinidin	2ACT	Baker (1980) J. Mol. Biol. 152, 737
Cytochrome (Rice)	1CCR	Ochi et al (1984) J. Mol. Biol. 166, 407
Cytochrome (Tuna)	4CYT	Takano et al (1980)
Carboxypeptidase A	1CPA	Rees et al (1983) J. Mol. Biol. 168, 367
DHFR (L.casei)	3DFR	Bolin et al (1982) J. Biol. Chem. 257, 13650
DHFR (E. coli)	4DFR	Bolin et al (1982) Ibid
Erythrocruorin	1ECD	Steigmann et al (1979) J. Mol. Biol. 127 309
Insulin Dimer	1INS	Baker et al (1985) Cryst.. of Mol. Biol. (Moras)
Human Lysozyme	1LZI	Artymuik et al (1981) J. Mol. Biol. 152, 737
Myoglobin (deoxy)	1MBD	Phillips (1980) J. Mol. Biol. 142 531
Plastocyanin (Cu++)	1PCY	Guss et al (1983) J. Mol. Biol. 169 521
BPTI	5PTI	Wlodawer et al (1984) J. Mol. Biol. 180, 301
Ribonuclease A	1RN3	Borkakoti et al (1982) Acta Cryst. B38 2210
Thermolysin	3TLN	Holmes et al (1982) J. Mol. Biol. 160 623
Beta-Trypsin	1TPP	Walter et al (1983) Acta Cryst B38 1462
Rubredoxin	4RXN	Watenpaugh et al (1979) J. Mol. Biol. 138, 615

a Filename in Brookhaven Protein Data Bank.

Table 2

RES	*NRES*[a]	*FITTED ATOMS*	*ATOM*[b]	*NW*[c]
SER	199	CA,CB,OG	OG	113
THR	161	CB,OG1,CG2	OG1	114
TYR	130	CG,CD1,CD2,CE1,CE2,CZ,OH	OH	84
Main	2608	CA,C',O	O	1197
Main	2587	C',O,N,CA	N	562

a Total Number of Residues
b Atom of Interest for Analysis
c Number of Water molecules within 3.5 Å

An interaction is defined by use of a geometric criterion i.e. the distance between non-hydrogen atoms is less than or equal to 3.5 Å. The data is accumulated for each residue type. It is analysed graphically on an Evans and Sutherland PS330 graphics system using Frodo [22] and numerically using spherical polar coordinates centered on the atom of interest as depicted in Figure 1. The energy minimization has been carried out using programs (CARTE and SITE) which have been written by Dr. F. Vovelle (CNRS, Orleans). These programs use the method of Powell [23] to minimize either three rotational degrees of freedom for a water molecule on a grid point (CARTE) or a full six (i.e. 3 translational and 3 rotational) degrees of freedom. The potentials are taken from the OPLS parameterization of Jorgensen [12] together with the TIPS4P [25] model [24] for water itself.

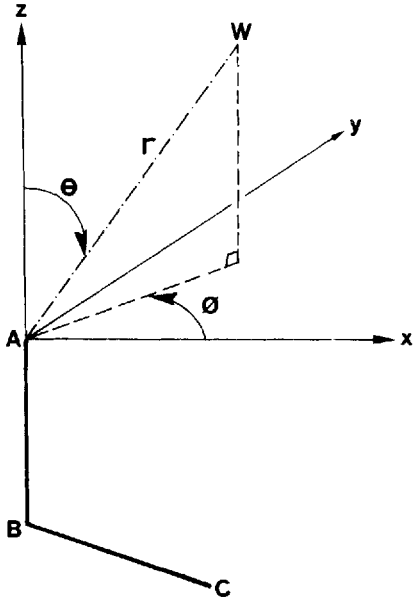

Figure 1 Diagram showing the use of spherical polar coordinates (r,θ,ϕ) to represent the water distributions around a group of atoms (A,B,C) centered on atom A.

RESULTS

Overall Hydration

In order to obtain an overall view of hydration, the number of interactions with water molecules located within 3.5 Å of each main chain or side chain non-hydrogen atom was calculated. These data are presented in Table 3 for the twenty residue types. It is immediately apparent that the percentage of interacting main chain atoms show a much narrower range of values than the similar quantity associated with the side chains. When these data are plotted as in Figure 2, this result is more clearly seen. Moreover, it becomes apparent that the residues below the plot of percentage of interacting main chain groups are mainly hydrophobic in character whereas those above this line are predominantly hydrophilic. If we take residue types cysteine (C) and glutamic acid (E) as examples of hydrophobic and hydrophilic residues respectively, we see that whereas the percentage of side chain interactions is consistent with this property (i.e. 12% and 77% for C and E respectively), the percentage of main chain interactions is similar at 47% and 51% for C and E respectively.

A more detailed view of the hydration of residue types is presented in Figure 3 in which a histogram of the number of residues of a given type is plotted as a function of the number of water molecules interacting with each residue. The polar and acidic residues on the left of Figure 3 quite frequently have contact with two or more water molecules. In contrast, the apolar residues on the right of Figure 3 have relatively few contacts even to one water molecule within 3.5Å and rarely contact more than one water molecule.

Table 3

RES[a]	*NRES*[b]	*MAIN CHAIN*		*SIDE CHAIN*	
		%R[c]	*NW*[d]	*%R*[c]	*NW*[d]
GLY	232	67	268	0	0
ALA	214	60	181	20	51
VAL	177	37	91	11	24
LEU	167	38	84	14	30
ILE	148	41	86	14	23
PRO	102	55	74	32	45
SER	199	47	132	51	143
THR	161	51	107	62	149
PHE	105	42	58	16	20
TYR	130	45	77	66	135
TRP	46	46	26	63	36
LYS	154	59	120	61	194
ARG	88	57	80	82	165
HIS	68	56	53	74	114
ASP	154	62	144	84	305
GLU	124	51	99	77	216
ASN	131	52	101	61	142
GLN	98	51	73	72	146
CYS	68	47	47	12	9
MET	42	31	23	29	18

a Residue name
b Total Number of Residues
c Percentage of residues within 3.5 Å of a water molecule
d Number of Water Molecules within 3.5 Å of residue

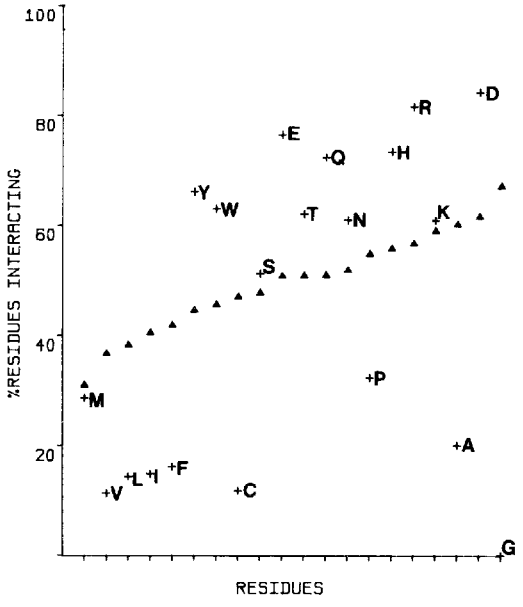

Figure 2 Plot of the percentage of residues interacting with water molecules (3.5 Å) for main chain atoms (▲) and side chain atoms (+). The data are plotted in ascending order of main chain interactions.

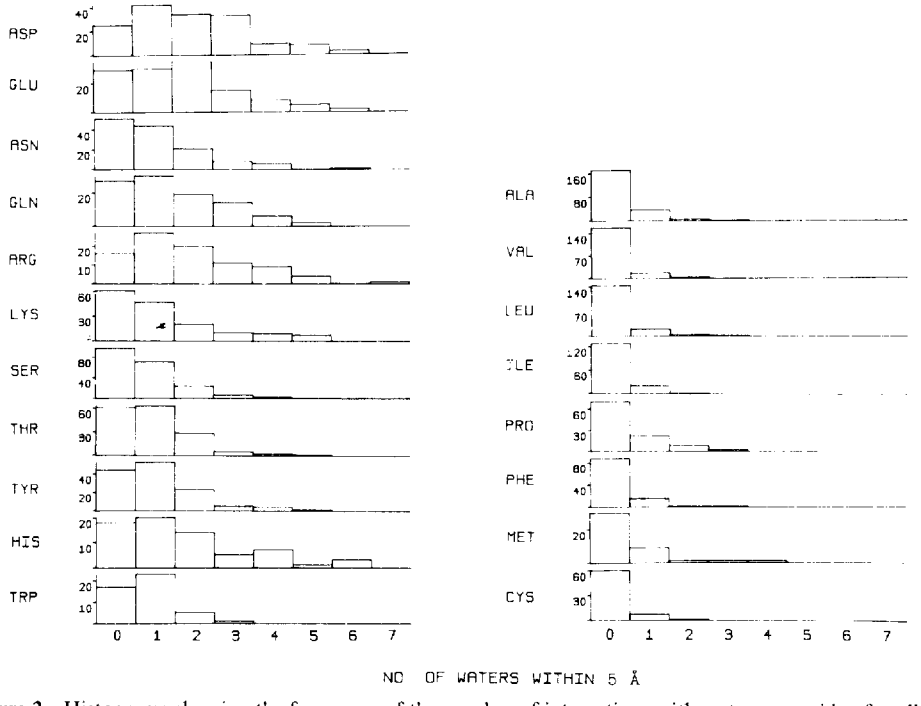

Figure 3 Histograms showing the frequency of the number of interactions with water per residue for all side chains based on interactions with side chain atoms.

Main Chain Hydration

In Figure 4a and 4b, we present the distribution of water molecules around the main chain carbonyl and amino groups respectively. Although these distributions are broad, clustering around both the CO and NH groups can be seen clearly. The analysis in terms of spherical polar coordinates shows that (a) the peak in the distance plot is at a lower value of r for CO compared with NH interactions, and (b) the θ and ϕ plots show good agreement with the expected stereochemistry of these potential hydrogen bonding groups. Although there is a distinct grouping of water molecules into clusters, which correspond to expected hydrogen bond geometries, the spread in the distributions is large.

Side Chain Hydration

We have started our analysis of side-chain hydration by looking at the hydroxyl groups of tyrosine, serine and threonine. These residues were chosen because there are a relatively large number interacting with water molecules within our sample of 16 proteins and because they are usually well-defined in electron density maps. The

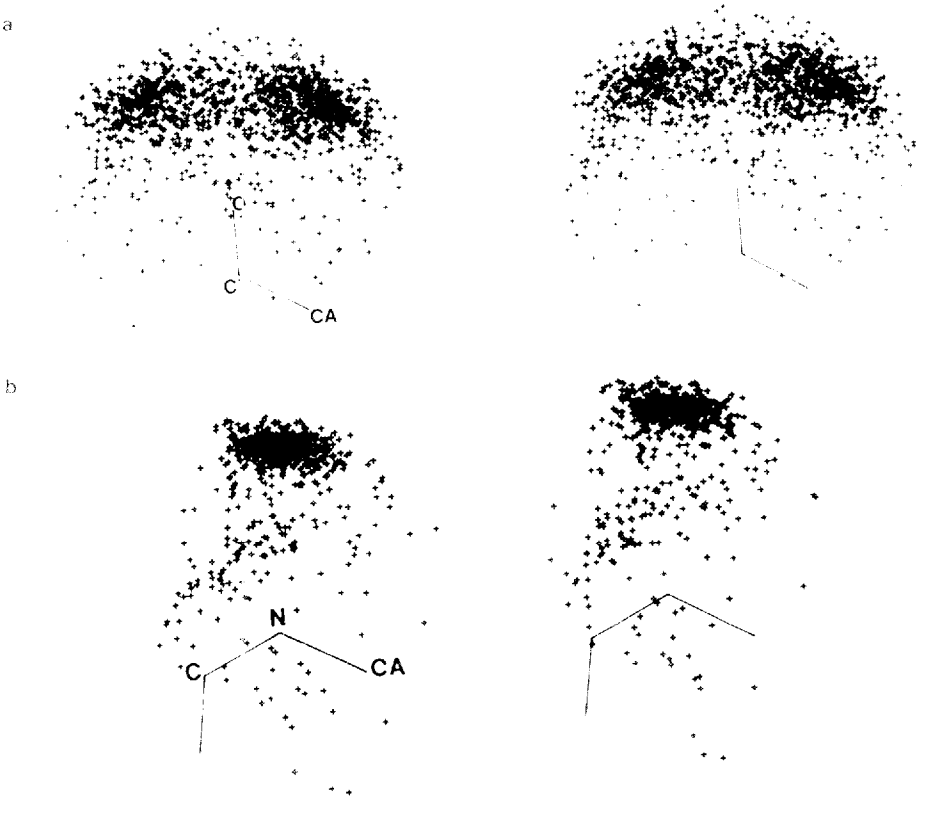

Figure 4 Stereo plots showing distribution of water molecules interacting with (a) main chain carbonyl group and (b) main chain amino groups.

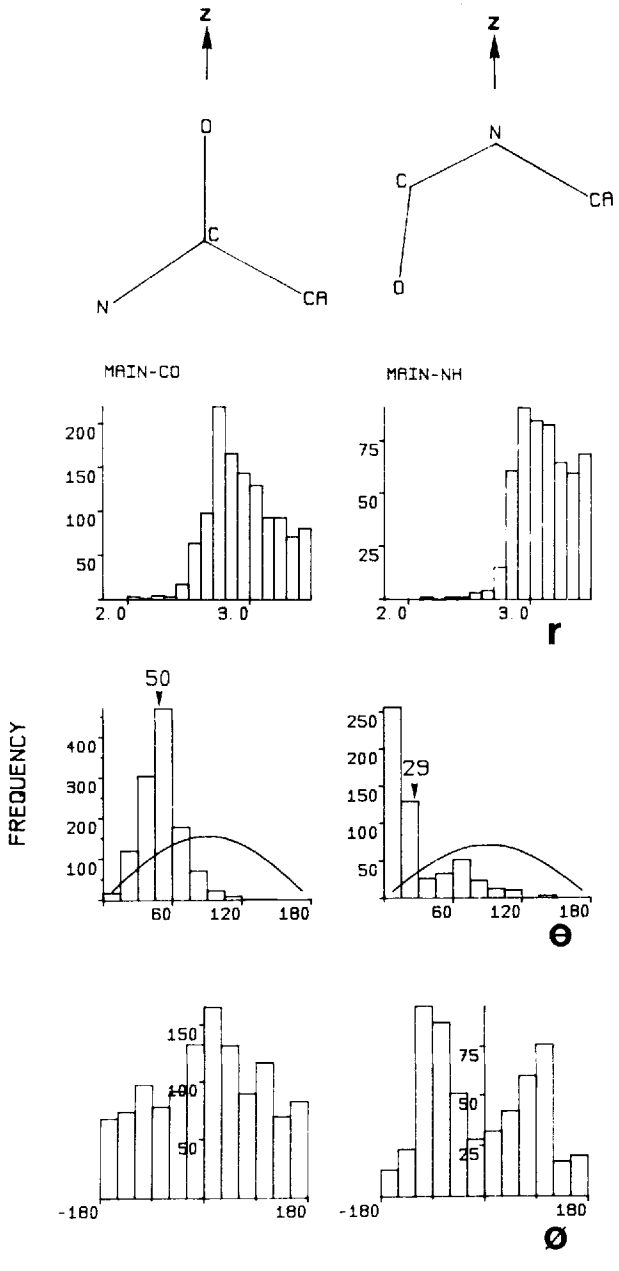

Figure 5 Histograms showing the analysis of the water distributions for main chain atoms as a function of the spherical polar coordinates r, theta and phi

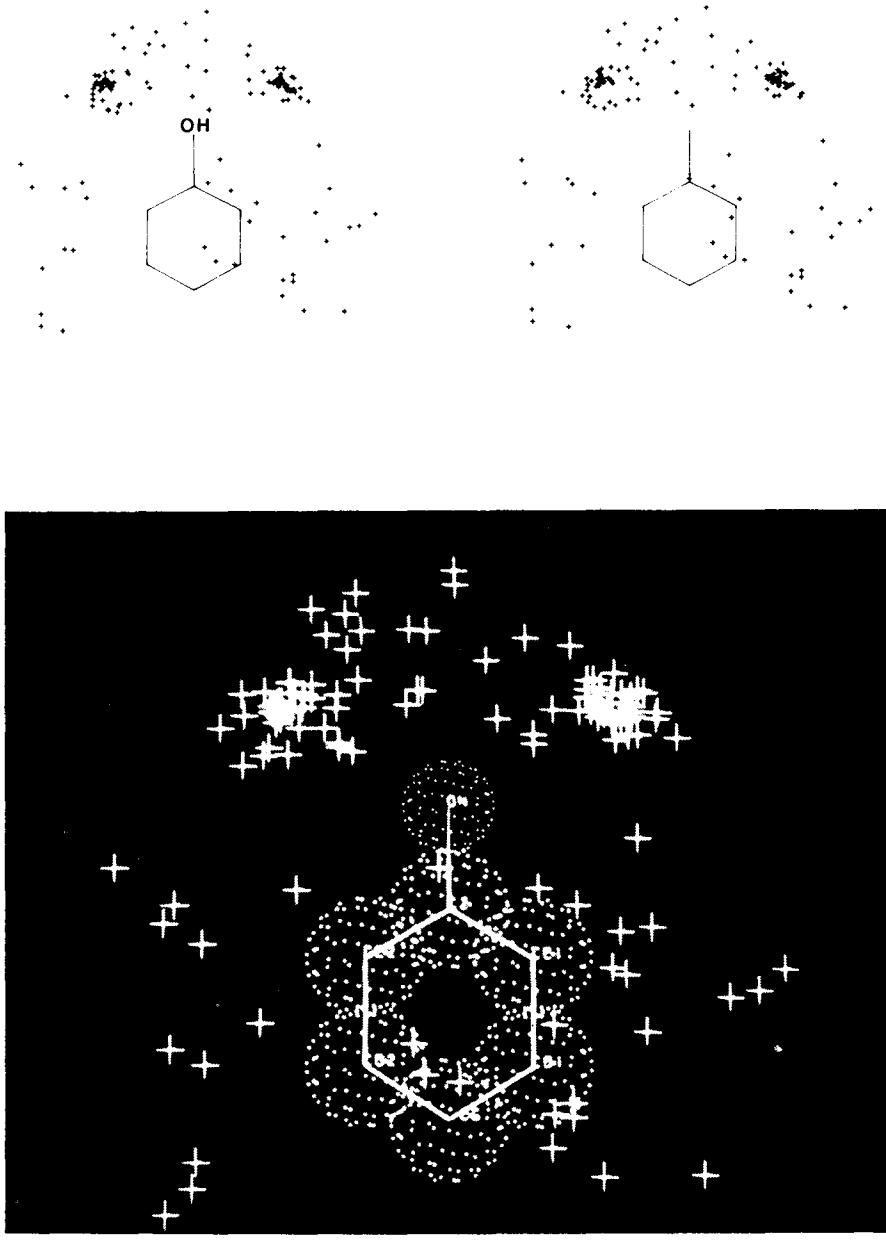

Figure 6 Stereo plots of the water molecule distributions around the hydroxyl group of tyrosine. (See colour plate V.)

distributions are presented graphically in Figures 6,7 and 8 for tyrosine, serine and threonine respectively. Wide distributions of water molecules are seen around these three hydroxyl groups but there is distinct clustering for tyrosine (Figure 6). The spherical polar coordinate plots (Figure 9) highlight the difference between on one hand tyrosine and on the other serine and threonine. The former residue shows a clear distance peak at approximately 2.75 Å. The θ plot peaks between 60° and 70° and the ϕ plot at 0° and 180° consistent with the expected sp^2 hybridization caused by delocalisation [27] of the lone pair into the aromatic ring. In contrast we expect the serine and threonine plots to be consistent with sp^3 hybridisation and staggered conformations. However, they are too broad to give clear information on orientational preferences.

Several reasons for the lack of clustering around serine and threonine are possible including the variation in χ_1 torsion angles, and the effects of multicontacts with main chain or local side chain atoms. The expected contacts will vary with sequence and secondary structure. These problems are not independent of each other as different types of secondary structure may show preference for different χ_1 angles [25] or more easily lead to multicontacts. In order to investigate these possibilities, we have plotted the distributions for residues in alpha helix, beta sheet and turn conformations as defined by Kabsch and Sander [26]. Comparison of these distributions for tyrosine

Figure 7 Stereo plots of the water molecule distributions around the hydroxyl group of serine.

Figure 8 Stereo plots of the water molecule distributions around the hydroxyl group of threonine.

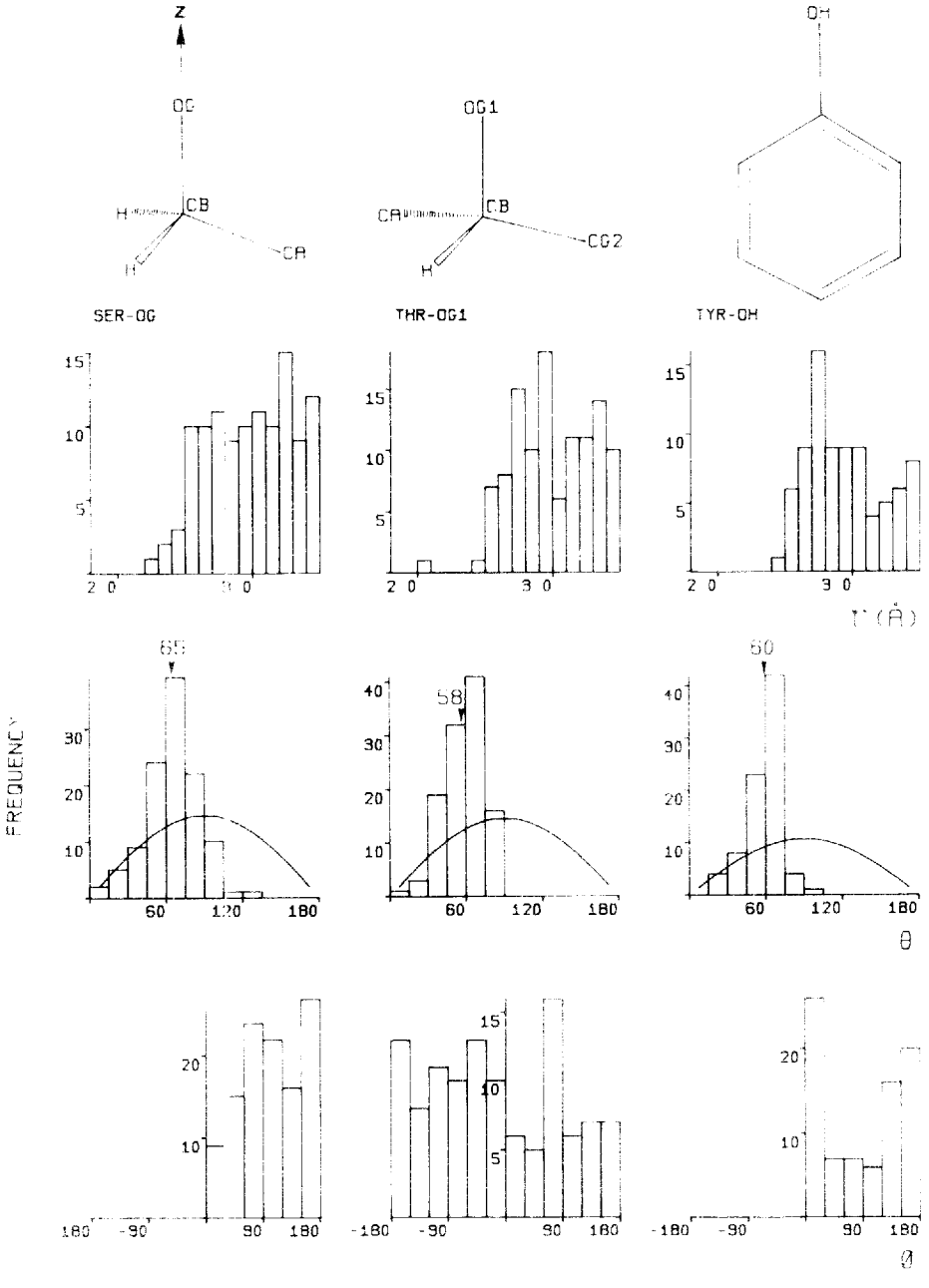

Figure 9 Histograms of the spherical polar coordinate analysis for serine, threonine and tyrosine distributions.

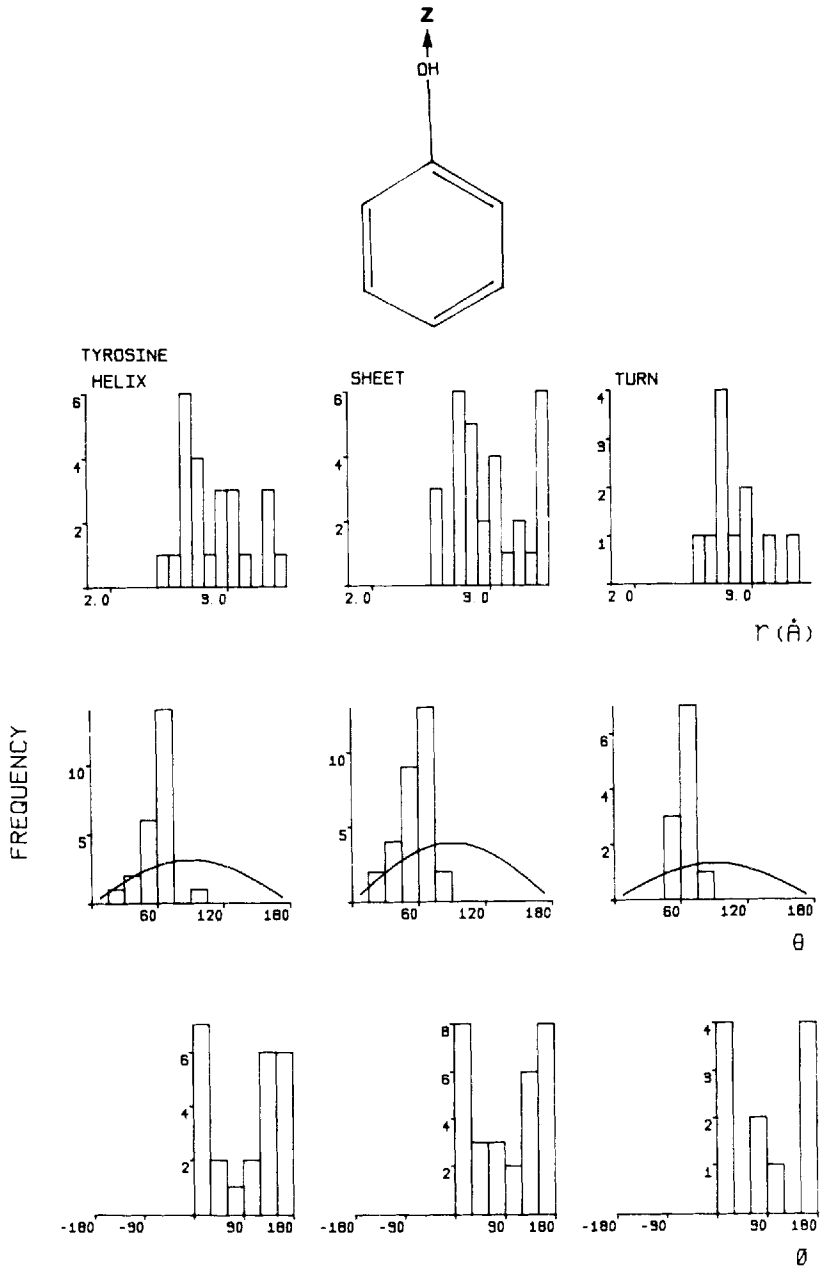

Figure 10 Histograms of the spherical polar coordinate analysis for tyrosine for residues in alpha helix, beta sheet and turn conformations.

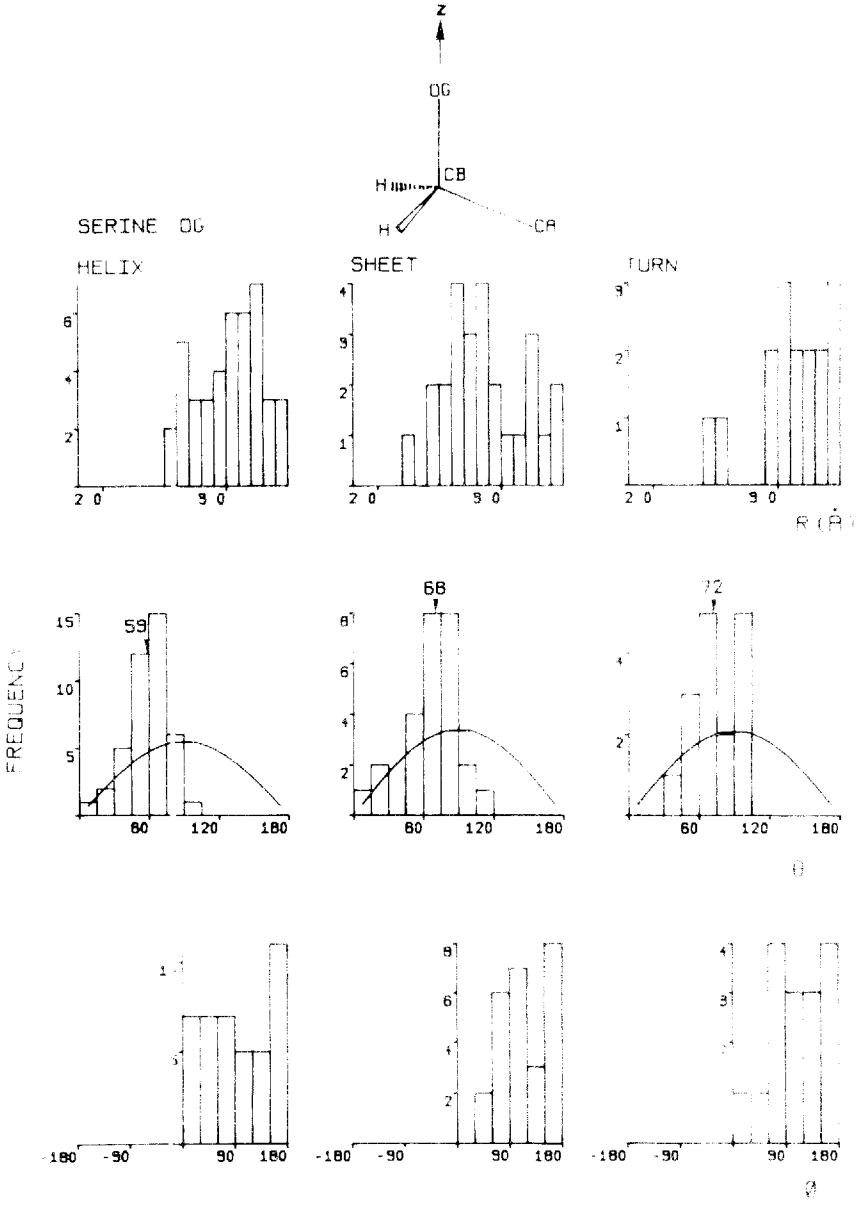

Figure 11 Histograms of the spherical polar coordinate analysis for serine for residues in alpha helix, beta sheet and turn conformations.

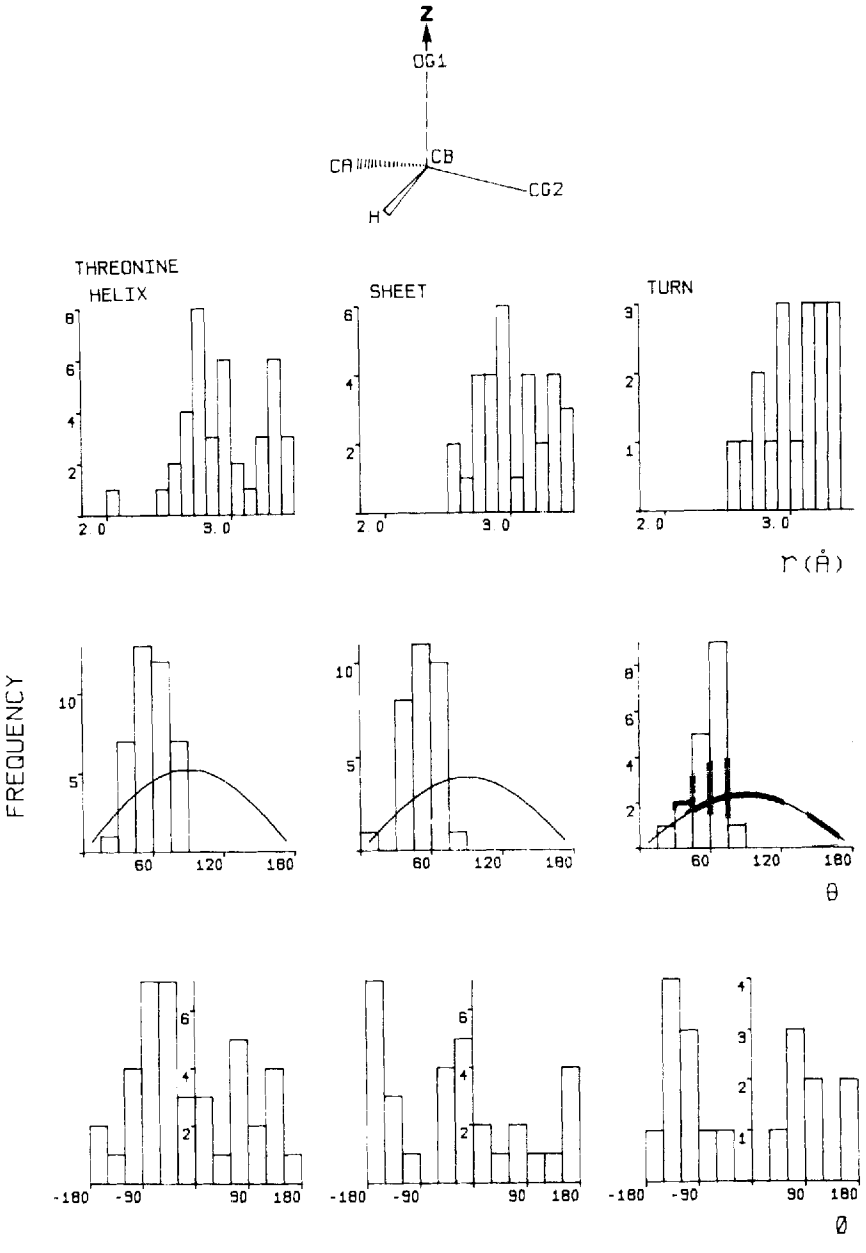

Figure 12 Histograms of the spherical polar coordinate analysis for threonine for residues in alpha helix, beta sheet and turn conformations.

(Figure 10) show that they are similar whatever the secondary structure. In contrast the distributions for serine (Figure 11) are very different depending on the secondary structure environment of the residue. Those for threonine (Figure 12) show some differences especially in phi plots.

Initial analysis of the multicontacts made by water molecules interacting with serine hydroxyl atoms shows that patterns occur repeatedly. For example, in alpha helical regions the serine hydroxyl groups interact with main chain carbonyls of residues i–3 or i–4. A detailed analysis is underway but complicated by the large number of possible multicontact interactions.

The isoenergy contour plots for tyrosine have been calculated and displayed graphically on a PS330 (Figure 13). These calculations have been carried out with the hydroxyl hydrogen on one side of the tyrosine ring whereas the experimental data shows that hydrogen can occur on either side. Comparisons of this calculated energy map with the experimental distribution show that there is good agreement between the energy minima and the preferential positions for water molecule binding.

DISCUSSION

We have demonstrated that by accumulating information from 16 well-refined protein structures, there are sufficient data to show the distribution of water molecule sites around some side-chains as well as main chain polar atoms. This is in spite of the problems associated with determining water molecule positions from electron density maps. We must emphasize that crystallographic data contain information only on the ordered water molecule sites. There must be other water molecules present in these crystal hydrates but their positions are disordered and thus there will be no clear electron density in which to place them.

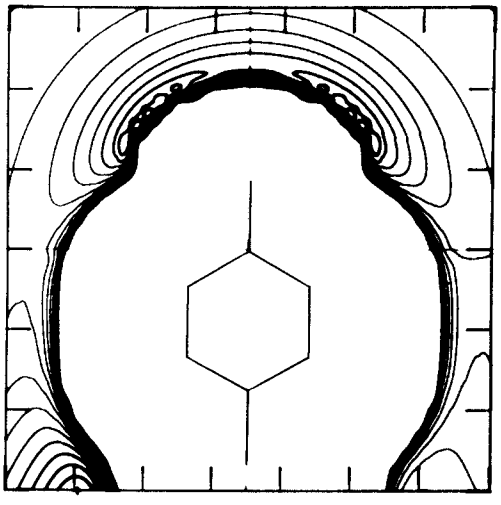

Figure 13 Iso-energy contour map for the interaction of a water molecule with tyrosine residue.

Our analysis of the number of residues interacting with solvent molecules shows that side-chain and main-chain groups of the same residue interact to a different degree. A similar percentage of main chain polar atoms are in contact with ordered water molecules whether the residue is polar or apolar.

It is only the percentage of side-chain atoms in contact with ordered water molecules which seems to correlate with the hydrophobicity of the residue. The implication is that main chain polar atoms of hydrophobic residues are as likely to be near the surface of the protein and thus able to interact with solvent as main chain polar atoms from residues with polar side-chains.

The distribution of water molecules around polar atoms shows that the preferred geometry is consistent with the expected stereochemistry of the protein atoms. However, this preferred geometry is superimposed on a rather broad distribution indicating the relatively weak stereochemical constraints for hydrogen bonding of water molecules. This broad distribution of sites could also be a consequence of errors in the crystallographic data. However, our methodology involved the removal of residues with a low occupancy or whose geometry was such that the least squares fit to the fixed reference group of atoms was poor. This procedure should remove the most unreliable of the data points.

The distribution of water molecule sites around the hydroxyl side-chains of serine, threonine and tyrosine were again quite broad with only the latter (tyrosine) distribution showing a clear preferential binding geometry in the plane of the ring. Again this conforms to the expected stereochemistry of the tyrosine oxygen being sp^2 hybridisation whereas the serine and threonine oxygen atoms are expected to have sp^3 hybridisation.

The very uniform distribution of water molecules around the hydroxyl group of serine was not expected. The further analysis shows that binding to the $-OH$ group is influenced by main-chain polar groups, whose disposition depends on the local secondary structure. The effect is that different distributions are observed for serines in helices and strands.

The importance of these distributions is their future use for modelling proteins and their interactions with small ligands. This modelling could involve either the direct use of these experimentally determined distributions for a specific protein or an energy minimization approach (such as that developed by Goodford [27]). Perhaps the optimum strategy is a combination of both approaches in which (a) the force fields used in the minimization procedure are checked by comparison of the simulation results of water molecule binding to the experimental distributions and (b) the experimental distributions are used to delineate 'probable' zones for binding and then full minimization calculations can be performed from several starting points within each zone. Our initial energy calculations for tyrosine look promising in that the energy minima appear to correspond well with the preferred binding sites found experimentally.

Acknowledgements

We would like to acknowledge support from the Science and Engineering Research Council (project grant number GR/D/38729). We would like to thank Dr Francoise Vovelle and Dr Mansoor Saqi for assistance in generating the iso-energy contour maps.

References

[1] J.M. Goodfellow, J.L. Finney & P. Barnes, "Monte Carlo Computer Simulation of Water-Amino Acid Interactions", *Proc. Roy. Soc B214*, 213 (1982)
[2] J.M. Goodfellow, "Computer Simulation of Hydration Networks around Amino Acids", *Int. J. Biol. Macro.* **9**, 273 (1987)
[3] A.T. Hagler, J. Moult & D.J. Osguthorpe, "Monte Carlo Simulation of the Solvent Structure in Crystals of a Hydrated Cyclic Peptide", *Biopolymers* **19**, 395 (1980)
[4] P. Rossky & M. Karplus, "Solvation. A Molecular Dynamics Study of a Dipeptide in Water", *J. Am. Chem. Soc.* **101**, 1913 (1979)
[5] R.J. Elliott & J.M. Goodfellow, "Monte Carlo Simulation of Nucleotide Crystal Hydrates and their counter-ions", *J. Theor. Biol.* **127**, 403 (1987)
[6] D.L. Beveridge, P.V. Maye, B. Jayaram, G. Ravishanker & Mezei, "Aqueous Hydration of Nucleic Acid Constituents: Monte Carlo Simulation Studies", *J. Biomol. Struc. & Dyn.* **2**, 261 (1984)
[7] A.T. Hagler & J. Moult, "Computer Simulation of the Solvent Structure around Biological Macromolecules", *Nature*, **272**, 222 (1979)
[8] W.F. van Gunsteren, H.J.C. Berendsen, J. Hermans, W.G.J. Hol & J.P.M. Postma, "Computer Simulation of the Dynamics of Hydrated Protein Crystals and its Comparison with x-ray data", *Proc. Natl. Acad. Sci* **80**, 4315 (1983)
[9] F.A. Momany, R.F. McGuire, A.W. Burgess & H.A. Scheraga, "Energy Parameters in Polypeptides VII", *J. phys. Chem.* **79**, 2361 (1975)
[10] S. Lifson, A.T. Hagler & P. Dauber "Consistent Force Field Studies of Intermolecular Forces in Hydrogen Bonded Crystals", *J. Am. Chem. Soc.* **53**, 4544 (1979)
[11] S.J. Weiner, P.A. Kollman, D.T. Nguyen & D.A. Case, "An All Atom Force Field for Simulations of Proteins and Nucleic Acids", *J. Comp. Chem.* **7**, 230 (1986)
[12] W.L. Jorgensen & C.J. Swenson, "Optimized Intermolecular Potential Functions for Amides and Peptides. Hydration of Amides", *J. Am. Chem. Soc.* **107**, 1489 (1985)
[13] J.L. Finney, J.E. Quinn & J.O. Baum, "The Water Dimer Potential Surface", *Water Sci. Rev.* **1**, 33 (1986)
[14] J.L. Finney, Goodfellow, J.M. Howell, P.L. & Vovelle, F., "Computer Simulation of Aqueous Biomolecular Systems", *J. Biomol. Struct. & Dyn.* **3**, 599 (1985)
[15] H. Savage, "Water Structure in Vitamin B_{12} Coenzyme Crystals", *Biophys. J.* **50**, 947 (1986)
[16] J.T. Edsall & H.A. McKenzie, "Water and Proteins I", *Adv. Biophys.* **10**, 137 (1978)
[17] J.T. Edsall & H.A. McKenzie, "Water and Proteins II", *Adv. Biophys.* **16**, 53 (1983)
[18] J.L. Finney "Organization and Function of Water in Protein Crystals" in 'Water. A Comprehensive Treatise', (Ed. F. Franks), vol 6, 47–122, Plenum (1979)
[19] E.N. Baker and R.E. Hubbard "Hydrogen Bonding in Globular Proteins", *Prog. Biophys. Mol. Biol* **44**, 97 (1984)
[20] W. Saenger, "Structure and Dynamics of Water Surrounding Biomolecules", *Ann. Rev. Biophy.* **93**, 114 (1987).
[21] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi & M. Tasumi, "The Protein Data Bank: A computer-based Archival File for Macromolecular Structure", *J. Mol. Biol.* **112**, 535 (1977)
[22] T.A. Jones, "A Graphics Model Building and Refinement System for Macromolecules", *J. Appl. Cryst.* **11**, 268 (1978)
[23] M.J.D. Powell "An efficient method for finding the minimum of a function of several variables without calculating derivatives", *Computer Journal*, **7**, 155 (1964)
[24] W.L. Jorgensen, J. Chandrashekhar, J.D. Madura, R. Impey & M.L. Klein, "Comparison of Simple Functions for Simulating Liquid Water", *J. Chem. Phys.* **79**, 926 (1983)
[25] M.J. McGregor, S.A. Islam & M.J.E. Sternberg, "Analysis of the Relationship between Side-Chain Conformation and Secondary Structure in Globular Proteins", *J. Mol. Biol.* **198**, 295 (1987)
[26] W. Kabsch & C. Sander, "Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features", *Biopolymers* **22**, 2577 (1983)
[27] P.J. Goodford, "A Computational Procedure for Determining Energetically Favourable Binding Sites on Biologically Important Macromolecules", *J. Med. Chem.* **28**, 849 (1985)