Table I. $^1J_{C_\alpha-H_\alpha}$, $^3J_{HN-H_\alpha}$, and $^3J_{HN-C_\beta}$ Coupling Constants Measured for Cyclo[Pro[1]-Pro[2]-Ala[3]-Ala[4]-Ala[5]] in DMSO

residue	$^1J_{C_\alpha-H_\alpha}$ (Hz)	ϕ^a (deg)	ψ^a (deg)	$^3J_{HN-H_\alpha}$ (Hz)	ϕ^b (deg)	$^3J_{HN-C_\beta}$ (Hz)	ϕ^c (deg)
Ala[3]	135.8	-120 ± 30	-140 ± 40	8.0	-153	2.0	-90
		60 ± 30	40 ± 40		-87		-30
					44		63
					78		177
Ala[4]	140.4	-150 ± 10	-60 ± 30	6.7	-160	3.0	-60
		-90 ± 10	120 ± 30		-80		73
		90 ± 10			32		167
		90 ± 10			88		
Ala[5]	143.2	-170 ± 10	-60 ± 20	6.6	-161	2.5	-81
		-70 ± 10	120 ± 20		-79		-39
		10 ± 10			31		68
		110 ± 10			89		172

[a] Values and ranges estimated from energy profiles following eqs 1 and 2. Each of the ϕ values can be paired with each of the ψ values. [b] Values of 9.4, -1.1, and 0.4 were used for *A*, *B*, and *C* following the standard Karplus equation.[16] [c] Values of 4.7, -1.5, and -0.2 were used for *A*, *B*, and *C* following the standard Karplus equation.[16]

I) were obtained with an accuracy of 0.3 Hz by HMQC without decoupling. The slices containing the coupling were removed from the 2D data set and processed as 1D spectra, allowing for greater zero filling. The best way to visualize the restraint from the C^α–H^α coupling is by a plot of the energy as determined by eqs 1 and 2 as a function of ϕ and ψ, shown in Figure 1. The ϕ and ψ torsions of the minima obtained from examination of the energy profiles are included in Table I. The figure indicates that, with larger coupling constants, the minima are more clearly defined. The previously determined $^3J_{HN-H^\alpha}$ and $^3J_{HN-C^\beta}$ values,[4,9] both about the ϕ dihedral, are included in Table I for comparison.

The MD simulations were carried out in vacuo and in DMSO[10] using the GROMOS program[11] following a protocol previously described.[12] Starting structures were created by application of dihedral angle restraints to the ϕ,ψ of the three alanines to 180°; the cyclic compound cannot obtain these constraints and therefore is of high energy and is removed from the structures found in solution, which suits our purposes.

Starting from this structure, the *J* restraints were applied following different procedures: (1) application of both the ^1J and ^3J couplings, with equal force constants (separate simulations using force constants of 0.25, 0.5, 1.0, and 5.0 kJ mol^{-1} Hz^{-2}); (2) application of only the ^1J for 20 ps and then slowly increasing the force constant of the ^3J couplings; and (3) similar to procedure 2 but starting with the ^3J couplings. These numerous simulations each in DMSO and in vacuo resulted in only two conformations, each of which are minima in regard to the coupling constants. The first conformation has ϕ and ψ values of (-67°,155°), (-68°,151°), (61°,-104°), (-83°,-15°), and (-81°,-45°) for Pro[1] to Ala[5], respectively, and is close to the conformation observed from NOE-restrained MD [corresponding ϕ and ψ values of (-73°,155°), (-65°,148°), (74°,-83°), (-115°,-20°), and (-63°,-42°)]. The distance restraint violation (using the 18 NOEs measured for this compound) is 19 pm. The second conformation with ϕ and ψ values of (-72°,160°), (-66°,-51°), (-99°,-103°), (-83°,78°), and (-159°,-62°) contains a γ-turn about Ala[4] and is well removed from the NOE structure (distance restraint violation of 50 pm). MD simulations starting with either of these two structures and applying the NOEs quickly produce the conformation previously reported, in agreement with NOEs and couplings (distance restraint violation of 8 pm).

The utilization of one-bond $^1J_{C^\alpha-H^\alpha}$ couplings as conformational restraints in MD simulations has been illustrated for a model cyclic pentapeptide. The results indicate that coupling constants, especially when more than one coupling about a torsion is available, are a valuable source of conformational restraints. Dynamics, either free rotation of a side chain or multiple backbone conformations,[13] has been purposely avoided in this simple example since the constant restraints may not be appropriate. The approach of time-dependent restraints,[14] as has been utilized for NOEs,[15] seems to be a viable alternative for cases involving dynamics.

Acknowledgment. The Deutsche Forschungsgemeinschaft and Fonds der Chemischen Industrie are gratefully acknowledged for financial support. S.G.G. acknowledges support from the Ministry of Science and Technology of Slovenia.

Registry No. Cyclo[Pro-Pro-Ala-Ala-Ala], 135866-42-1.

(9) Kurz, M. Ph.D. Thesis, Technical University of Munich, Munich, Germany, 1991.
(10) Mierke, D. F.; Kessler, H. *J. Am. Chem. Soc.* **1991**, *113*, 7466-7470.
(11) van Gunsteren, W. F.; Berendsen, H. J. C. Groningen molecular simulation library manual (GROMOS), Biomos B. V., Groningen, 1987.
(12) Kurz, M.; Mierke, D. F.; Kessler, H. *Angew. Chem., Int. Ed. Engl.* **1992**, *31*, 210-212.
(13) Kessler, H.; Griesinger, C.; Müller, A.; Lautz, J.; van Gunsteren, W. F.; Berendsen, H. J. C. *J. Am. Chem. Soc.* **1988**, *110*, 3393-3398.
(14) Torda, A. E.; Brunne, R. M.; Huber, T.; Kessler, H.; van Gunsteren, W. F. *J. Biomol. NMR*, submitted for publication.
(15) Torda, A. E.; Scheek, R. M.; van Gunsteren, W. F. *Chem. Phys. Lett.* **1989**, *157*, 289-294.
(16) (a) Karplus, M. *J. Chem. Phys.* **1959**, *30*, 11-15. (b) Karplus, M. *J. Am. Chem. Soc.* **1963**, *85*, 2870-2871.

Prediction of Water Binding Sites on Proteins by Neural Networks

Rebecca C. Wade,*,† Henrik Bohr,‡ and Peter G. Wolynes‡

European Molecular Biology Laboratory
Meyerhofstrasse 1, 6900 Heidelberg, Germany
Noyes Laboratory, School of Chemical Sciences
University of Illinois, 505 South Mathews Avenue
Urbana, Illinois 61801
Received May 18, 1992

The ability to predict ligand binding sites on biological macromolecules is an important goal in biotechnology. Because water plays a crucial role in the binding of ligands to proteins, we focus here on the prediction of water binding sites on proteins. We describe neural networks trained using crystallographic data to predict water sites on the basis of amino acid sequence and secondary structure. These networks make predictions on the atomic scale and surprisingly produce results comparable to those from other known methods of predicting water sites, even though the latter use tertiary structure information. The networks may be used to analyze relationships between the positions of water sites and protein sequence and secondary structure.

Feed-forward networks with one hidden layer were employed. These are known to have the ability to generalize molecular biology data.[1-6] Two different networks were used to determine (1)

* Author to whom correspondence should be addressed.
† European Molecular Biology Laboratory.
‡ University of Illinois.

(1) Qian, N.; Sejnowski, T. J. *J. Mol. Biol.* **1988**, *202*, 865-884.
(2) Bohr, H.; Bohr, J.; Brunak, S.; Cotterill, R. M. J.; Lautrup, B.; Nørskov, L.; Olsen, O. H.; Petersen, S. B. *FEBS Lett.* **1988**, *241*, 223-228.
(3) Holley, L. H.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 152-156.
(4) Bohr, H.; Bohr, J.; Brunak, S.; Cotterill, R. M. J.; Fredholm, H.; Lautrup, B.; Petersen, S. B. *FEBS Lett.* **1990**, *261*, 43-46.

0002-7863/92/1514-8284$03.00/0 © 1992 American Chemical Society

Table I. Performance of Neural Networks Predicting Water Sites in Proteins and Comparison to Other Methods

neural network[a]	proteins[b]	neural network performance[c]			other methods				
						$Q1$ within given distance (Å)[d]			
		Q	$Q1$	C	ref	0.5–0.6	1.0	1.4–1.5	1.8
1	training set	0.89	0.93	0.76					
	test set	0.64	0.88	0.14					
2	training set	0.90	0.83	0.77					
	test set	0.59	0.44	0.09					
2	1bp2	0.56	0.38	0.03	12	0.24	0.42	0.60	0.81
2	1hsy	0.55	0.35	0.01	13, 14		0.48	0.64	
2	2cpp	0.59	0.42	0.06	12	0.16	0.41	0.59	0.72
2	2prk	0.57	0.44	0.07	12	0.16	0.40	0.59	0.71
2	2rhe	0.59	0.54	0.16	12	0.19	0.40	0.56	0.71
2	3grs	0.59	0.45	0.13	11	0.08	0.33	0.67	
2	3tln	0.58	0.43	0.07	11	0.15	0.44	0.72	

[a] Network 1 predicted whether each residue was hydrated. Network 2 predicted whether each atom in each residue (except C and H atoms) had water and protein ligands. Results are only given for prediction of water ligands. [b] The following proteins from the Brookhaven Protein Databank[15] were used. Forty training proteins (6913 residues): 1alc, 1cho, 1ctf, 1gcr, 1mba, 1paz, 1rdg, 1sn3, 1snc, 1srn, 1ubq, 1utg, 1ypi, 256b, 2act, 2aza, 2cdv, 2cga, 2fbj, 2gbp, 2hhb, 21tn, 2ovo, 2sec, 2wrp, 3app, 3c2c, 3dfr, 3rnt, 4dfr, 4ins, 4pep, 4pti, 5cpa, 5cpv, 5cyt, 5pcy, 5rxn, 8dfr, 9pap. Nine testing proteins (2224 residues): 1bp2, 1ccr, 1hsy (coordinates supplied by Artymiuk and Blake), 2cpp, 2prk, 2rhe, 3grs, 3sgb, 3tln. [c] Q, proportion of neurons correctly predicted; $Q1$, proportion of neurons with actual output 1 correctly predicted; C, Matthews's correlation coefficient.[16,17] [d] To determine proportion $Q1$, an experimentally observed water site was considered to be correctly predicted if a predicted water site lay within the given distance of it.

whether each residue in a test protein has a water ligand and (2) whether each atom in the protein makes a close contact with a water molecule.

The input for both networks was a "window" of 17 residues centered on the residue for which predictions were being made. Each residue was represented by input neurons specifying its type and secondary structure. Its type was given by seven neurons corresponding to its physical properties (size, hydrophobicity, polarity, charge, aliphaticity, aromaticity, and whether it is proline). Its secondary structure class was given by the Kabsch and Sander assignment[7] and specified by five binary neurons. Secondary structure was input because it may have an important influence on protein hydration. Thanki et al.[8] found that solvent networks extend the hydrogen-bonding structure of secondary structures. Water molecules are often seen at particular locations along α-helices (e.g., bridging between i and $i-3$ or $i-4$ residues), β-sheets, and turns. The hydration of Ser and Thr side chains, but not Tyr side chains, also appears to be dependent on secondary structure.[9] In principle, secondary structure may be deduced from the amino acid sequence alone using a neural network although, at present, accuracy is limited to about 65% on average.[1-3,6] Therefore, input of secondary structure to the networks should not require the tertiary structure of a test protein to be known.

The output from the two networks was as follows: (1) one binary output neuron for each residue indicating whether the residue had ≥ 1 water ligand or not, and (2) two binary output neurons for each atom in each residue, except C and H atoms, the first indicating whether a water was close to the particular atom and the second whether a protein atom which was not in the same or an adjacent residue was close. Ligands close to each protein atom were found by searching within a radius of 3.5 Å. This distance was chosen to detect all ligands making hydrogen bonds and close van der Waals contacts with the protein atoms.

The networks were trained and tested with proteins whose crystal structures were solved at high (<2.0 Å) resolution (see footnote b, Table I). The proteins were nonhomologous and included a wide range of tertiary structure types. While the networks appeared to train well (about 90% correct), their predictive performance was not as good (about 60% correct; see Table I). Neural networks are capable of generalizing even when a proportion of the input data is incorrect, creating "noise". In this case, however, the amount of noise in the data sets may be such that it prevents the networks from developing better predictive abilities. For instance, noise arises because not all water molecules hydrating proteins are detected by X-ray crystallography, (see, for example, ref 10), different criteria may be used to assign water sites in different structures, and crystal contacts influence the location of bound water molecules. The predictive performance of the networks can, however, be expected to improve as more accurate protein structures becomes available and if these structures are processed before training in order to reduce their noise, e.g., by omitting water sites whose positions are physically unrealistic or dependent on crystal symmetry.

The performance of the second network is compared for seven proteins with that of other prediction methods in Table I. These require the three-dimensional coordinates of the protein for which predictions are made. Vedani and Huhta[11] used the directionality of hydrogen bonds to determine solvation sites. Pitt and Goodfellow[12] employed a knowledge-based approach to determine solvent positions around polar groups. Wade and Goodfellow[13,14] used an empirical energy function to predict water sites. Although the predictions of the networks and the other methods cannot be compared directly because they employ different definitions of water sites, a comparison of $Q1$, the proportion of experimentally observed water sites predicted correctly, shows that the network performs similarly to the other methods if they are required to have an accuracy of about 1.0 Å (roughly equivalent to a water site being 2.8 ± 1.0 Å from a protein atom). Thus, despite being based on amino acid sequence alone, the neural networks may provide useful tools for predicting binding sites with reference to particular atoms as well as residues.

Acknowledgment. We thank Drs. R. Goldstein, J. Bryngelson, and Z. Schulten for helpful discussions. R.C.W. thanks Professor J. A. McCammon for his help and Drs. P. Artymiuk and C. Blake for providing the coordinates of human lysozyme. This work was supported by NIH Grant PHS GM 44557-01.

Registry No. Water, 7732-18-5.

(5) Brunak, S.; Engelbrecht, J.; Knudsen, S. *Nature* **1990**, *343*, 123.
(6) Knellner, D. G.; Cohen, F. E.; Landridge, R. *J. Mol. Biol.* **1990**, *214*, 171–182.
(7) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577–2637.
(8) Thanki, N.; Umrania, Y.; Thornton, J. M.; Goodfellow, J. M. *J. Mol. Biol.* **1991**, *221*, 669–691.
(9) Thanki, N.; Thornton, J. M.; Goodfellow, J. M. *Protein Eng.* **1990**, *3*, 495–508.
(10) Finer-Moore, J. S.; Kossiakoff, A. A.; Hurley, J. H.; Earnest, T.; Stroud, R. M. *Proteins* **1992**, *12*, 203–222.
(11) Vedani, A.; Huhta, D. W. *J. Am. Chem. Soc.* **1991**, *113*, 5860–5862.
(12) Pitts, W. R.; Goodfellow, J. M. *Protein Eng.* **1991**, *4*, 531–537.
(13) Wade, R. C. D. Phil. Thesis, University of Oxford, Oxford, U.K., 1988.
(14) Wade, R. C.; Goodford, P. J. Submitted for publication.
(15) Bernstein, F. C.; Koetzlke, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennedy, O.; Shimanouchi, T.; Tasumi, M. *J. Mol. Biol.* **1977**, *112*, 535–542.
(16) Matthews, B. W. *Biochem. Biophys. Acta* **1975**, *405*, 442–451.
(17) Stolorz, P.; Lapedes, A.; Xia, Y. *J. Mol. Biol.* **1992**, *225*, 363–377.