J|A|C|S
ARTICLES
Published on Web 02/07/2004

SOS-NMR: A Saturation Transfer NMR-Based Method for Determining the Structures of Protein−Ligand Complexes

Philip J. Hajduk,* Jamey C. Mack, Edward T. Olejniczak, Chang Park, Peter J. Dandliker, and Bruce A. Beutel

Contribution from the Pharmaceutical Discovery Division, Abbott Laboratories, Abbott Park, Illinois 60064

Received November 7, 2003; E-mail: philip.hajduk@abbott.com

Abstract: An NMR-based alternative to traditional X-ray crystallography and NMR methods for structure-based drug design is described that enables the structure determination of ligands complexed to virtually any biomolecular target regardless of size, composition, or oligomeric state. The method utilizes saturation transfer difference (STD) NMR spectroscopy performed on a ligand complexed to a series of target samples that have been deuterated everywhere except for specific amino acid types. In this way, the amino acid composition of the ligand-binding site can be defined, and, given the three-dimensional structure of the protein target, the three-dimensional structure of the protein−ligand complex can be determined. Unlike earlier NMR methods for solving the structures of protein−ligand complexes, no protein resonance assignments are necessary. Thus, the approach has broad potential applications − especially in cases where X-ray crystallography and traditional NMR methods have failed to produce structural data. The method is called SOS-NMR for structural information using Overhauser effects and selective labeling and is validated on two protein−ligand complexes: FKBP complexed to 2-(3′-pyridyl)-benzimidazole and MurA complexed to uridine diphosphate *N*-acetylglucosamine.

Introduction

Structure-based drug design is a powerful tool for accelerating drug discovery. On the basis of the structures of protein−ligand complexes, new compounds can be designed that optimize intermolecular interactions and/or improve the physical characteristics of a lead compound without disrupting physical association between the drug and the target. Two of the most successful techniques for obtaining structural information on protein−ligand complexes are X-ray crystallography and NMR spectroscopy. One major disadvantage of X-ray crystallography is the need to obtain suitable crystals of the protein−ligand complex. Despite the advances in the field of crystallography over the past decade, the production of crystals suitable for structure-based design continues to be a major hurdle.[1] High-resolution structure determination using NMR spectroscopy also suffers from many disadvantages.[2,3] The greatest limitation is the need to observe, resolve, and assign the many signals that arise from the protein resonances. In practice, this limits the application of high-resolution NMR structural studies to those targets that have molecular weights less than about 30 kDa. NMR approaches that utilize extensive deuteration can be applied to larger systems,[4,5] but the need to observe and assign the protein resonances still remains a major hurdle.

In the absence of high-resolution structural data on protein−ligand complexes, computational docking strategies can be used to attempt prediction of not only the ligand-binding site,[6−9] but also the structure of the protein−ligand complex for use in structure-based design.[10] In these cases, the structure of the target biomolecule for use in the docking simulations must be obtained either from the experimentally derived structure of the target protein itself (either unliganded or complexed to a substrate or substrate analogue) or from homology modeling. However, computational algorithms for predicting structures of protein−ligand complexes are still maturing and often give rise to multiple solutions that cannot be reliably discriminated on the basis of the predicted binding energies.[11]

Here, we describe an NMR-based strategy for deriving structural information on protein:ligand complexes that provides an alternative to traditional X-ray crystallography and NMR methods for structure-based drug design. The method is called SOS-NMR for structural information using Overhauser effects and selective labeling and relies on performing saturation transfer difference (STD) experiments[12] on a ligand complexed to a series of selectively labeled protein samples. The data from SOS-NMR define the amino acid composition of the ligand-

(1) DeLucas, L. J. *Drug Discovery Today* **2001**, *6*, 734−744.
(2) Opella, S. J. *Methods Enzymol.* **2001**, *339*, 285−313.
(3) Roberts, G. C. K. *Drug Discovery Today* **2000**, *5*, 230−240.
(4) Pellechia, M.; Meininger, D.; Dong, Q.; Chang, E.; Jack, R.; Sem, D. S. *J. Biomol. NMR* **2002**, *22*, 165−173.
(5) Tugarinov, V.; Kay, L. E. *J. Am. Chem. Soc.* **2003**, *125*, 13868−13878.
(6) Liang, J.; Edelsbrunner, H.; Woodward, C. *Protein Sci.* **1998**, *7*, 1884−1897.
(7) Pettit, F. K.; Bowie, J. U. *J. Mol. Biol.* **1999**, *285*, 1377−1382.
(8) Leclerc, F.; Karplus, M. *Theor. Chem. Acc.* **1999**, *101*, 131−137.
(9) Glick, M.; Robinson, D. D.; Grant, G. H.; Richards, W. G. *J. Am. Chem. Soc.* **2002**, *124*, 2337−2344.
(10) Kuntz, I. D.; Meng, E. C.; Shoichet, B. K. *Acc. Chem. Res.* **1994**, *27*, 117−123.
(11) Abagyan, R.; Totrov, M. *Curr. Opin. Chem. Biol.* **2001**, *5*, 375−382.
(12) Mayer, M.; Meyer, B. *Angew. Chem., Int. Ed.* **1999**, *38*, 1784−1788.

10.1021/ja039480v CCC: $27.50

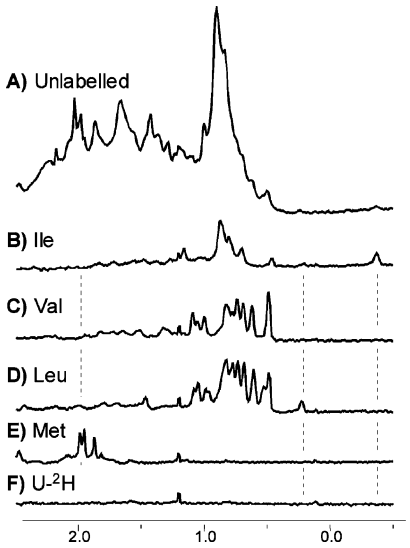

Figure 1. 1D proton NMR spectra showing the aliphatic region for (A) unlabeled FKBP, (B) perdeuterated, Ile-protonated FKBP, (C) perdeuterated, Val-protonated FKBP, (D) perdeuterated, Leu-protonated FKBP, (E) perdeuterated, Met-protonated FKBP, and (F) perdeuterated FKBP. Dashed lines indicate the positions of well-resolved peaks for each amino acid type, indicating little to no biosynthetic scrambling between these amino acid types. Also of note is the complete absence of signal in the spectrum of the perdeuterated protein, indicating essentially complete deuteration.

binding site and, given the three-dimensional structure of the protein target, can be used to determine the structure of the protein−ligand complex. This approach is demonstrated on two protein−ligand complexes: FKBP complexed to 2-(3′-pyridyl)-benzimidazole (**1**) and MurA complexed to uridine diphosphate *N*-acetylglucosamine (UDP-GlcNAc, **2**).

Results

FKBP. The FK506 binding protein (FKBP) is a 14 kDa protein that forms a complex with the potent immunosuppressant FK506.[13] We have previously reported the results of NMR-based screening to identify a number of ligands for this target.[14] One of these compounds, 2-(3′-pyridyl)-benzimidazole (**1**, see Figure 2), binds to FKBP with a K_D value of 260 μM and, based on chemical shift changes, binds to the FK506 binding site on FKBP. To demonstrate the applicability of SOS-NMR to define the structure of a protein−ligand complex, six samples of FKBP were prepared. Samples of unlabeled and perdeuterated protein served as positive and negative controls. Next, four additional samples of perdeuterated FKBP were prepared in which the amino acids Ile, Val, Leu, and Met were selectively protonated. Inspection of the ^{1}H NMR spectra of the selectively protonated protein samples indicated that a high level of specific incorporation was achieved for each amino acid (Figure 1). Enhancements of the magnetization transfer between the ligand and any of these four selectively protonated samples (relative to the perdeuterated control) can be interpreted as a direct contact between the ligand and at least one residue of that amino acid type.

The results of the saturation transfer difference experiments on the benzimidazole **1** and the FKBP samples are shown in Figure 2. Large and uniform NOEs are observed between the ligand and unlabeled FKBP (Figure 2A), indicating that all

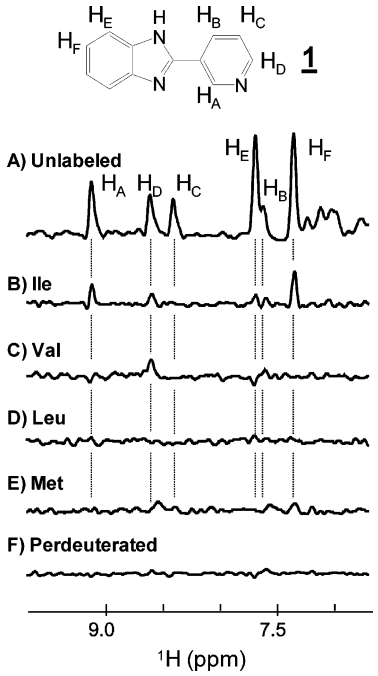

Figure 2. STD-NMR spectra of 2-(3′-pyridyl)-benzimidazole (**1**) in the presence of (A) unlabeled FKBP, (B) perdeuterated, Ile-protonated FKBP, (C) perdeuterated, Val-protonated FKBP, (D) perdeuterated, Leu-protonated FKBP, (E) perdeuterated, Met-protonated FKBP, and (F) perdeuterated FKBP. The resonances corresponding to the ligand are indicated.

portions of the ligand are in contact with the protein. This is in sharp contrast to the selectively protonated samples, where the NOE intensities are nonuniform and indicative of specific contacts between the ligand and the protein. For example, enhancements for the Ile-protonated sample (Figure 2B) are present only for the H_A and H_D positions on the pyridyl group and the H_F position on the benzimidazole, indicating that these protons are in contact with at least one isoleucine residue. In contrast, the Val-protonated sample (Figure 1C) shows a significant contact only to H_D. For the Leu- and Met-protonated samples (Figure 1D, E), no significant enhancements beyond the deuterated control were observed, indicating that no portion of **1** is in contact with any of these residues on the protein. These NOE intensities were quantified into normalized SOS-NMR intensities, I(SOS), by scaling the fraction of the observed NOE that can be attributed to interaction with the selectively protonated amino acid by the weighted STD-NMR intensity, I(STD-NMR), observed for the unlabeled sample (see Methods for details). The results of this analysis are shown in Table 1.

As conventional methodologies for generating recombinant, isotopically labeled proteins will incorporate any given amino acid at all sequence positions coding for that residue type, the resulting NOE information must be defined as an ambiguous restraint in structure calculation protocols. For example, the H_D proton of **1** exhibited NOE enhancement with the Ile-protonated sample and is allowed to make potential contact with the Hα, Hβ, or Hγ of any valine residue in FKBP (Val23, Val24, Val55, Val63, Val68, Val98, or Val101). In contrast, the H_D proton of **1** did not exhibit any enhancement with the Leu- or Met-protonated samples and therefore cannot make contact with any proton of any leucine or methionine residue in FKBP (Leu30, Leu50, Leu74, Leu97, Leu103, Leu104, Leu106, Met29, Met49, and Met66). Ambiguous distance restraints were generated from the magnetization transfer data in a manner analogous to NMR

(13) Liu, J.; Farmer, J. D., Jr.; Lane, W. S.; Friedman, J.; Weissman, I.; Schreiber, S. L. *Cell* **1991**, *66*, 807−815.
(14) Muegge, I.; Martin, Y. C.; Hajduk, P. J.; Fesik, S. W. *J. Med. Chem.* **1999**, *42*, 2498−2503.

Table 1. Spectral Intensities and Derived Distance Restraints for **1** in the Presence of FKBP[a]

proton	1D int. U-^{1}H-FKBP[c]	*I*(STD-NMR)[d]	*I*(SOS)[b] U-^{1}H-FKBP	U-^{2}H-FKBP	U-^{2}H-{^{1}H-I}-FKBP	U-^{2}H-{^{1}H-V}-FKBP	U-^{2}H-{^{1}H-L}-FKBP	U-^{2}H-{^{1}H-M}-FKBP
HA	7.7	94	94 (721)[e]	0 (0)	40 (307) {<**4 Å**}[f]	3 (22) {>5 Å}	8 (65) {−}	6 (47) {−}
HB	5.2	56	56 (292)	0 (73)	0 (76) {>5 Å}	3 (84) {>5 Å}	0 (0) {>5 Å}	0 (0) {>5 Å}
HC	6.0	64	64 (385)	0 (0)	0 (0) {>5 Å}	0 (0) {>5 Å}	9 (52) {−}	0 (0) {>5 Å}
HD	4.6	100	100 (455)	0 (44)	15 (107) {<**5 Å**}	36 (190) {<**4 Å**}	0 (0) {>5 Å}	0 (0) {>5 Å}
HE	12.8	89	89 (1141)	0 (0)	10 (123) {−}[g]	0 (0) {>5 Å}	0 (0) {>5 Å}	5 (63) {>5 Å}
HF	14.7	81	81 (1191)	0 (0)	27 (400) {<**4 Å**}	0 (0) {>5 Å}	5 (75) {>5 Å}	6 (87) {−}

[a] All spectra were recorded with 1.0 mM **1** and 2.5 μM FKBP. The saturation time was 1.4 s, and the CPMG relaxation delay was 10 ms. [b] Normalized SOS-NMR intensities calculated as described in the text. [c] Intensity (arbitrary units) observed in a one-dimensional (1D) T_2-filtered spectrum acquired under conditions identical to those of the STD-NMR experiment. 1D intensities were identical for all samples. [d] Fractional normalized STD-NMR intensity calculated as described in the text. [e] Values in parentheses are the actual intensities (arbitrary units) observed in the STD-NMR spectra. [f] Values in brackets are distance restraints (in Å) derived from the NOE intensity data. [g] Peaks with normalized NOE intensities in the range of 5−15% were not used to assign a distance restraint because this is close to the experimental error.

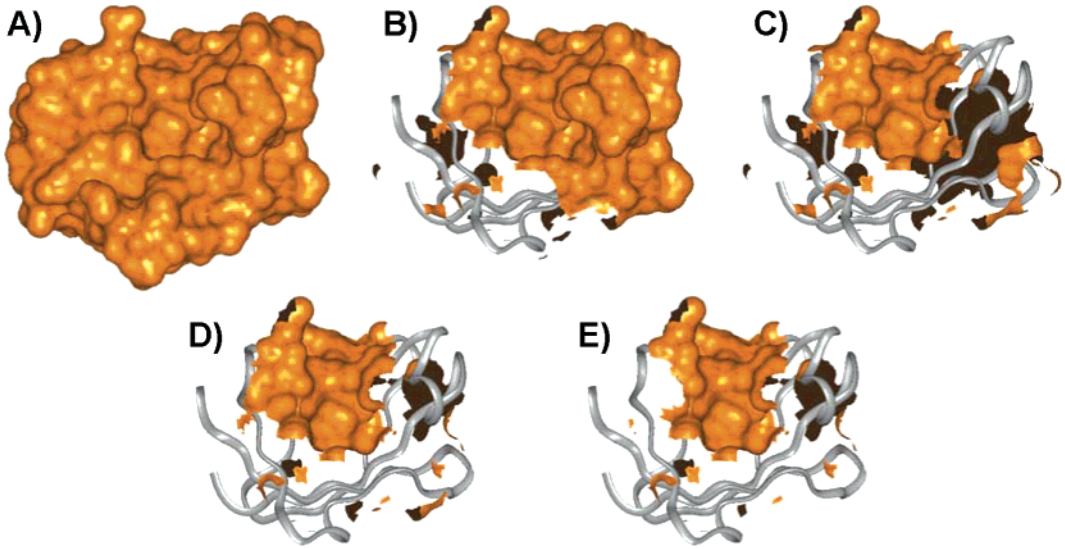

Figure 3. Binding site determination for **1** on FKBP using the SOS-NMR data as described in the text. (A) Complete protein surface. (B) Surface area within 10 Å of any isoleucine residue. (C) Surface area from (B) that is also within 10 Å of any valine residue. (D) Surface area from (C) that is not within 5 Å of any leucine residue. (E) Surface area from (D) that is not within 5 Å of any methionine residue.

structure calculations in the case of spectral degeneracy or missing assignments.[15] Distance restraints of 3, 4, 5, and >5 Å were assigned to protons with *I*(SOS) values of >50%, 25−49%, 15−25%, and <5%, respectively. Resonances with *I*(SOS) values between 5% and 15% were considered in the range of the experimental error and not used in the simulations. The results of this process for **1** complexed to FKBP are shown in Table 1.

To calculate the structure of the complex using the SOS-NMR data, the program DOCK 4.0[10] was implemented using the protein structure determined in complex with FK506 (PDB code 1FKJ). As DOCK requires a known binding site, the binding site for **1** on FKBP was determined using the SOS-NMR data. Starting from the total solvent accessible surface area of the protein (Figure 3A), regions of the surface were cumulatively rejected when not in proximity to Ile (Figure 3B), and Ile and Val (Figure 3C). Regions of the surface were also rejected that were in proximity to Leu (Figure 3D) and Met (Figure 3E). After this analysis, only a single contiguous surface remains (Figure 3E) that corresponds to the FK506 binding site on FKBP (see Methods for details). The benzimidazole **1** was then docked to this binding site, and the resulting 34 low energy conformations were assigned an NOE penalty on the basis of the ambiguous restraint list. Figure 4 shows the low energy

(15) Nilges, M.; Odonoghue, S. I. *Prog. Nucl. Magn. Reson. Spectrosc.* **1998**, *32*, 107−139.

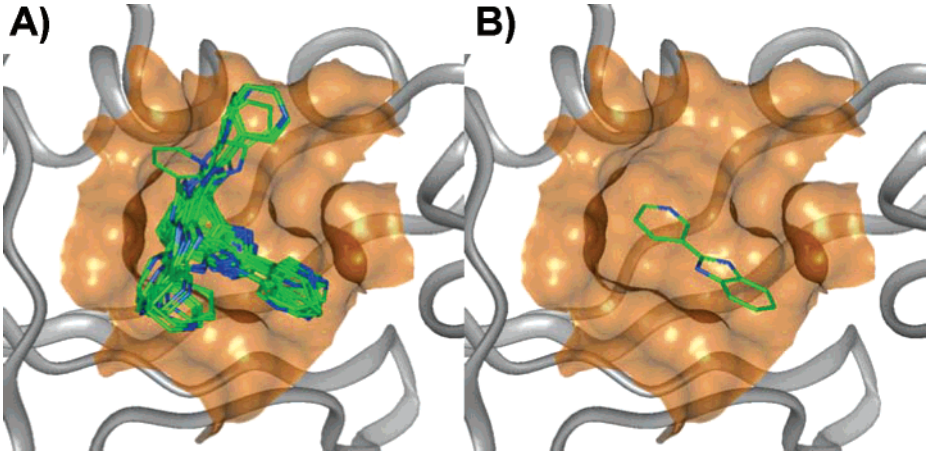

Figure 4. Conformations of **1** (green carbon atoms) complexed to the FK506 binding site of FKBP (gray ribbons, orange surface) generated using the program DOCK. Low energy conformations are shown (A) before and (B) after filtering with the SOS-NMR data.

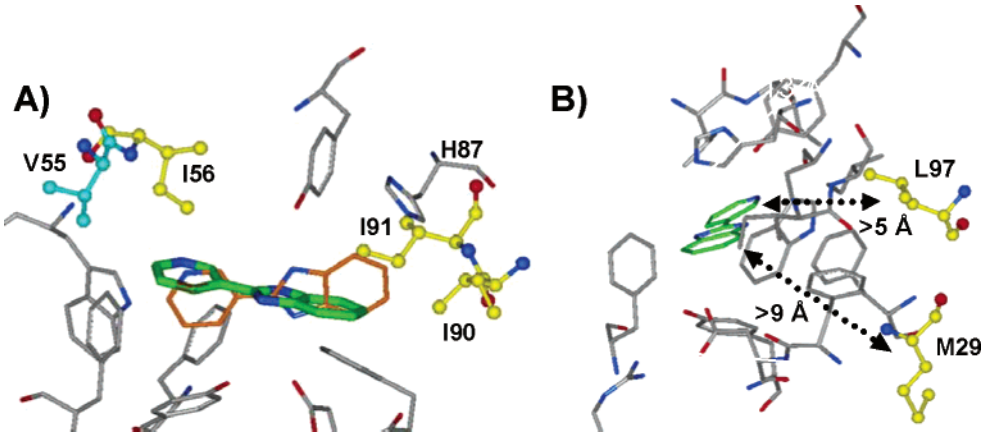

Figure 5. Structure of FKBP complexed to **1** (green carbon atoms) as determined using the SOS-NMR constraints. (A) Positive SOS-NMR contacts to Ile and Val residues are fulfilled by V55 (cyan carbon atoms) and I56, I91, and I90 (yellow carbon atoms). The location of **1** as observed by X-ray crystallography is shown with orange carbons. (B) Negative SOS-NMR contacts to Leu and Met residues are also fulfilled, as the nearest residues (L97 and M29, yellow carbon atoms) are more than 5 Å from the ligand.

conformations identified before (Figure 4A) and after (Figure 4B) application of the ambiguous restraint list. Only a single low energy conformation remains.

A detailed view of the interactions between **1** and FKBP is shown in Figure 5. The pyridyl moiety of **1** is near Ile56 and Val55, while the benzimidazole group is near Ile90 and Ile91 (Figure 5A). No portion of the ligand is near a leucine or methionine, as the nearest of these residues is more than 5 Å from the ligand (Figure 5B). This is in excellent agreement with the SOS-NMR data shown in Figure 2 and is also in agreement with the structure of the FKBP/**1** complex determined by X-ray crystallography (Figure 5). The pairwise rmsd for **1** as determined by the SOS-NMR data and the X-ray crystal structure is 1.2 Å for all heavy atoms. Significantly, this difference is due primarily to a 1.1 Å shift in the position of His87 (labeled in Figure 5) in the structure of FKBP when complexed to FK506 versus **1**. Not surprisingly, SOS-NMR docking of **1** to the protein structure determined in complex with **1** resulted in rmsd values <1 Å (data not shown). However, even with the observed conformational changes, the structure derived from the SOS-NMR approach is in good correspondence with the information derived from traditional high-resolution methods.

MurA. An additional example of the use of SOS-NMR to obtain structural information on protein−ligand complexes is shown on the complex of MurA with uridine diphosphate *N*-acetylglucosamine (UDP-GlcNAc) (**2**, see Figure 6).[16] MurA is a 48 kDa protein involved in bacterial cell wall synthesis and would not typically be amenable to structural analysis using NMR. For the SOS-NMR studies, five samples of MurA were prepared. Samples of unlabeled and perdeuterated protein were prepared to serve as positive and negative controls. Next, three additional samples of perdeuterated MurA were prepared in which the amino acids Trp, Phe, and His were selectively protonated.

The results of the saturation transfer difference experiments on the UDP-GlcNAc and MurA samples are shown in Figure 6, and the spectral intensities and derived distance restraints are given in Table 2. As with FKBP, large and uniform NOEs are observed between the ligand and unlabeled MurA, while the NOE intensities for the selectively protonated samples are nonuniform and indicative of specific contacts between the ligand and the protein. For example, enhancements for the His-protonated sample are present only for the uridine ring (Figure 6B). Likewise, the Trp-protonated sample only exhibits NOEs to the methyl protons on the *N*-acetyl group (Figure 6C), while the Phe-protonated sample shows enhancements for the ribose and pyranose ring protons (Figure 6D). Construction of the binding surface and docking to MurA again yielded only a single

(16) Skarzynski, T.; Mistry, A.; Wonacott, A.; Hutchinson, S. E.; Kelly, V. A.; Duncan, K. *Structure* **1996**, *4*, 1465−1474.

Table 2. 1D and STD-NMR NOE Intensities for **2** in the Presence of MurA[a]

proton	1D int. U-^{1}H-MurA[c]	I(STD-NMR)[d]	I(SOS)[b] U-^{1}H-MurA	U-^{2}H-MurA	U-^{2}H-{^{1}H-H}-MurA	U-^{2}H-{^{1}H-W}-MurA	U-^{2}H-{^{1}H-F}-MurA
HA	4.1	38	38 (20)[e]	0 (0)	38 (20) {<**4 Å**}[f]	0 (0) {>5 Å}	0 (0) {>5 Å}
HB	4.8	100	100 (60)	0 (0)	0 (0) {>5 Å}	0 (0) {>5 Å}	0 (0) {>5 Å}
HC	4.0	95	95 (48)	0 (0)	0 (0) {>5 Å}	6 (3) {−}[g]	0 (0) {>5 Å}
HD	6.0	51	51 (38)	0 (0)	0 (0) {>5 Å}	0 (0) {>5 Å}	26 (19) {<**4 Å**}
HE	6.0	51	51 (38)	0 (0)	0 (0) {>5 Å}	0 (0) {>5 Å}	26 (19) {<**4 Å**}
HF	15.1	54	54 (104)	0 (0)	0 (0) {>5 Å}	21 (40) {<**5 Å**}	0 (0) {>5 Å}

[a] All spectra were recorded with 0.5 mM **2** and 20 μM MurA. The saturation time was 1.0 s. [b] Normalized NOE intensities calculated as described in the text. [c] Intensity (arbitrary units) observed in a one-dimensional (1D) T_2-filtered spectrum acquired under conditions identical to those of the STD-NMR experiment. 1D intensities were identical for all samples. [d] Fractional normalized STD-NMR intensity calculated as described in the text. [e] Values in parentheses are the actual intensities (arbitrary units) observed in the STD-NMR spectra. [f] Values in brackets are distance restraints (in Å) derived from the NOE intensity data. [g] Peaks with normalized NOE intensities in the range of 5−15% were not used to assign a distance restraint because this is close to the experimental error.

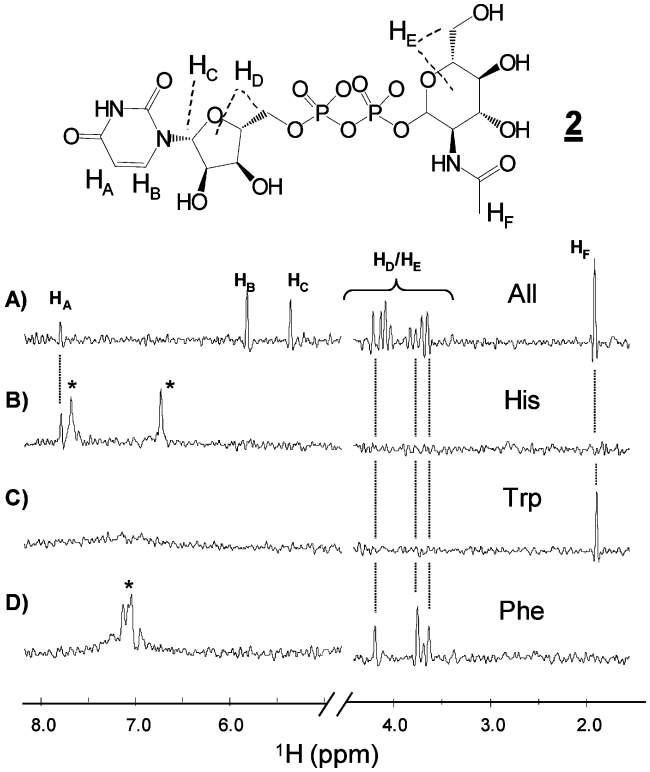

Figure 6. STD-NMR spectra of UDP-GlcNAc (**2**) in the presence of (A) unlabeled MurA, (B) perdeuterated, His-protonated MurA, (C) perdeuterated, Trp-protonated MurA, and (D) perdeuterated, Phe-protonated MurA. The resonances corresponding to UDP-GlcNAc are indicated. Peaks marked with an asterisk are protein peaks (histidine resonances in (B) and phenylalanine resonances in (D)).

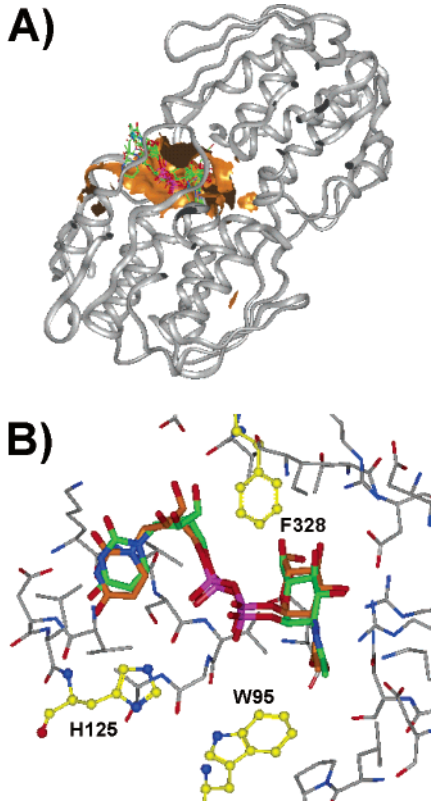

Figure 7. (A) Allowed binding surface (shown in orange) for **2** on MurA using the SOS-NMR data as described in the text. Also shown in (A) are conformations of **2** (green carbon atoms) generated using the program DOCK. (B) Structure of MurA complexed to **2** (green carbon atoms) as determined using the SOS-NMR constraints. Only a single conformation of **2** is allowed after filtering with the SOS data. Positive SOS-NMR contacts to Trp, His, and Phe residues are fulfilled by W95, H125, and F328 (yellow carbon atoms). The location of **2** as observed in the X-ray crystal structure of MurA complexed to UDP-GlcNAc (PDB 1UAE) is shown with orange carbons.

conformation consistent with the SOS-NMR data (Figure 7A). The results are in excellent agreement with the X-ray crystal structure of MurA complexed to UDP-GlcNAc (pairwise rmsd

Table 3. Analysis of Amino Acid Composition of Ligand-Binding Sites

amino acid	total on surface[a]	total in binding site[b]	surface (%)[c]	binding site (%)[d]	side-chain enrichment in ligand-binding site[e]	enrichment in protein–protein interfaces[f]	biosynthetic scrambling of side chain[g]
Ala	3691	190	6.2	5.3	0.85		Val, Leu, others
Arg	4076	207	6.9	5.8	0.84	2.47	
Asn	3164	173	5.4	4.8	0.90	0.93	Asp
Asp	4602	252	7.8	7.0	0.90	1.67	Asn, Met, Thr, Lys
Cys	516	58	0.9	1.6	1.85	0.00	
Gln	2987	133	5.1	3.7	0.73	0.58	Glu, His
Glu	5613	205	9.5	5.7	0.60	0.68	Gln, Pro, Arg
Gly	0	0	0	0		0.45	
His	1835	180	3.1	5.0	1.62	1.49	
Ile	2522	218	4.3	6.1	1.43	1.79	
Leu	4490	339	7.6	9.5	1.25	0.01	
Lys	5090	197	8.6	5.5	0.64	1.17	
Met	1215	131	2.1	3.7	1.78	0.54	
Phe	2158	225	3.7	6.3	1.72	0.56	Tyr
Pro	3205	95	5.4	2.7	0.49	1.25	
Ser	3806	214	6.4	6.0	0.93	0.21	Cys, Gly
Thr	3650	209	6.2	5.8	0.94	0.28	Ile
Trp	945	131	1.6	3.7	2.29	3.91	
Tyr	2367	192	4.0	5.4	1.34	2.29	
Val	3165	233	5.4	6.5	1.21	0.00	
total	59 097	3582					

[a] The total number of times at least one side-chain heavy atom of a particular amino acid was on the surface of the protein in the set of 272 crystal structures analyzed. [b] The total number of times at least one side-chain heavy atom of a particular amino acid was within 5 Å of at least one heavy atom of the ligand in the set of 272 crystal structures analyzed. [c] Of all surface amino acids identified in this analysis (total = 59 097), percentage that the side chain of a particular amino acid occurs on the surface of the protein. [d] Of all amino acids identified in the ligand-binding sites (total = 3582), percentage that the side chain of a particular amino acid occurs in the ligand-binding site. [e] Ratio of binding site to surface frequency. [f] Taken from ref 23. [g] Taken from ref 28.

of 0.6 Å for all heavy atoms).[16] In this structure, Trp95 is located in the binding site and is near the *N*-acetyl group, while His125 is located near the uridine moiety. The only phenylalanine in the substrate-binding pocket, Phe328, is located near the diphosphate moiety and is within NOE distance of both the ribose and the pyranose rings of UDP-GlcNAc (see Figure 7B), consistent with the SOS-NMR data.

Binding Pocket Analysis. To assess the generality of SOS-NMR to identify ligand-binding sites, an analysis was performed on crystal structures of 272 unique proteins in complex with either their natural ligands or synthetic inhibitors. In the first step of this analysis, potential ligand-binding sites (referred to as centroids) on each protein were simulated by the systematic placement of solvent molecules along the entire protein surface (see Methods). Next, the minimum distance of each centroid to each of the 20 amino acid types was calculated. These data were then used to ascertain whether the centroid corresponding to the crystallographically observed binding site could be distinguished from all other sites as the number of amino acids used in the analysis was increased. The order of addition of the amino acids was the following: (1) Val, (2) Leu, (3) Ile, (4) Met, (5) Lys, (6) Arg, (7) Pro, (8) Thr, (9) Ala, (10) Cys, (11) Trp, (12) Phe, (13) His, (14) Tyr, (15) Glu, (16) Gln, (17) Asp, (18) Asn, (19) Gly, (20) Ser. Shown in Figure 8 is the percentage of proteins in the dataset for which the actual ligand-binding site could be identified as a function of the number of amino acids used in the simulated SOS-NMR analysis. As can be observed from this figure, an average of six amino acids is necessary for SOS-NMR to unambiguously identify ligand-binding sites, comprising the set of {Val, Leu, Ile, Met, Lys, and Arg}. In addition to this analysis, a statistical survey of the occurrence of individual amino acid side chains in the ligand-binding site was performed, and the results are shown in Table 3.

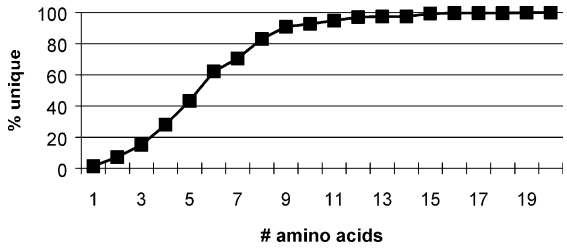

Figure 8. Simulated SOS-NMR analysis on X-ray crystal structures of 272 unique proteins complexed to either natural ligands or synthetic inhibitors. Plotted is the percentage of proteins in the test set for which the ligand-binding site was uniquely defined as a function of amino acids types used in the analysis.

Discussion

STD-NMR is a powerful tool for monitoring ligand binding and for mapping the epitopes of ligands that make contact with the target.[17] Because the technique monitors modulation of the ligand resonances upon selective saturation of the protein, it has the significant advantage of being applicable to virtually any biomolecular target regardless of size, composition (e.g., glycoproteins, protein−RNA complexes, etc.), or oligomeric state. In addition, only a small amount of protein target is required to perform the experiments (each example presented here required less than 1 mg of protein per amino acid label). However, STD-NMR on unlabeled proteins does not yield information about the specific region of the protein that is in close spatial proximity to the ligand or about the three-dimensional structure of the protein−ligand complex. This is due to the fact that, although only a small number of protein protons are saturated directly by the selective pulse, all protons on the protein are subsequently saturated due to efficient spin-diffusion processes that occur in large biomolecules. The use

(17) Mayer, M.; Meyer, B. *J. Am. Chem. Soc.* **2001**, *123*, 6108−6117.

of saturation transfer NMR with highly deuterated proteins has been shown to improve ligand screening and reduce the effects of nonspecific binding, provided that the binding site on the protein and the amino acids that make contact with the ligand are known.[4] SOS-NMR further exploits the benefits of extensive deuteration to generate ambiguous restraints for a series of selectively protonated protein samples so that both the binding site and the three-dimensional structure of the protein−ligand complex can potentially be defined. Significantly, even in the absence of the target structure, the SOS-NMR data can still be used to construct the amino acid composition of the ligand-binding site, yielding the location of polar and nonpolar amino acids that are in contact with the ligand. This type of information can be extremely valuable in medicinal chemistry efforts to optimize lead compounds.

Selective deuteration has long been used to improve the quality of protein spectra,[18] and even to probe the interfaces between protein−protein[19] and protein−ligand complexes.[4,20] However, all of these techniques require the observation and assignment of the protein resonances, limiting their application to highly soluble, low molecular weight systems. SOS-NMR overcomes these limitations and allows for structure determination of protein−ligand complexes without direct observation of the biomolecular target. Future technological improvements promise to extend the applicability of SOS-NMR and resolve some of the ambiguities that result from labeling all amino acids of a particular type. For example, the routine and cost-effective production of highly deuterated proteins from yeast and baculovirus systems promises to extend the range of accessible targets beyond that which can be obtained from bacterial systems. Cell-free expression systems can also enable the study of additional targets, but holds the unique potential of labeling a single amino acid in the primary sequence through the use of amber codons and modified transfer RNAs.[21,22] This approach eliminates the possibility of observing NOE intensity from multiple binding sites, which at present must be confirmed through competition experiments with alternative ligands.

For FKBP and MurA, only five and three amino acids, respectively, were necessary for unambiguous identification of the ligand-binding site. The generality of the SOS-NMR approach is highlighted in Figure 8, where it is observed that an average of six amino acids are required for the SOS-NMR approach to unambiguously discriminate the ligand-binding surface from the total surface area of the protein. In fact, the range of three to nine amino acids is sufficient to identify the ligand-binding sites of more than 90% of the proteins in the test set. It is significant to note that, because a fixed order of amino acids was used in the analysis, the actual number of amino acids required to identify the ligand-binding site for many proteins will be much lower than that calculated using this analysis. This is due to the fact that each protein active site will in fact have a unique minimal set of amino acids that differentiate it from all other sites and that is not represented by the order of amino acids used here. For example, using SOS-NMR, we have shown that the active site of MurA can be differentiated from all other sites using only three amino acids (Trp, Phe, and His). However, the simulation required eight amino acids (comprising the set of {Val, Leu, Ile, Met, Lys, Arg, Pro, Thr}). The order used in this analysis reflects pragmatic concerns such as ease of incorporation and minimization of side-chain scrambling, and in most applications the judicious choice of amino acids for SOS-NMR study should reduce the number of amino acids that need to be considered. The amino acid preferences at ligand-binding sites are different than those observed for protein−protein interactions,[23,24] especially with regards to the hydrophobic amino acids Leu, Met, and Val and the charged amino acids Arg, Asp, and Lys. Thus, the distribution of amino acids that comprise ligand-binding sites is not reflective of the relative abundance of amino acids either in the primary sequence or at protein−protein interfaces, suggesting unique and conserved preferences for particular amino acids in ligand binding. Significantly, the amino acids that are found at statistically higher frequencies in ligand-binding sites (e.g., Cys, His, Ile, Leu, Met, Phe, Trp, and Tyr) are those that can be selectively labeled in *E. Coli* without significant biosynthetic scrambling.

Conventional STD-NMR relies on efficient spin-diffusion of the saturation throughout the biomolecule. In highly deuterated systems, this may not be the case, and care must be taken to ensure that all proton spin systems on the target are efficiently saturated. For example, methyl group chemical shifts can vary by more than 2 ppm. If a continuous network of methyl−methyl contacts that involves all protonated methyl groups in the protein does not exist, then irradiation at a single frequency may not be sufficient to saturate all methyl resonances. This can be tested experimentally by inspection of the STD-NMR spectrum in the absence of a T_2-filter to ensure that a difference signal is observed for every peak that is observed in the 1D spectrum. In instances where this is not the case, the frequency of irradiation or the length of the selective pulse can be changed, or saturation at multiple frequencies may be necessary. For example, simultaneous irradiation at 0.5 and 2.0 ppm was sufficient for complete saturation of all protein resonances for each selectively protonated FKBP sample based on inspection of the STD-NMR spectra (data not shown). In contrast, a unique saturation frequency was utilized for each MurA sample based on the label (see Methods). Irradiation strategies will have to be tailored on the basis of the labeling pattern and ligand employed to ensure efficient saturation of the biomolecule and avoid direct saturation of the ligand resonances with the selective pulses.

Conformational changes in the target upon ligand binding are common and must be considered in any strategy that involves docking a ligand to a static target conformation. Significantly, the SOS-NMR data provide experimental contacts to residues that reflect the bioactive conformation. Thus, these experimental contacts can be used to exclude target conformations that do not present the required amino acid types in the context of a pocket that can accommodate the ligand. More subtle conformational changes, such as side-chain motions, will certainly affect the docked orientation of the ligand. This was in fact the case for FKBP, where a 1.1 Å shift in the position

(18) LeMaster, D. M. *Q. Rev. Biophys.* **1990**, *23*, 133−174.
(19) Takahashi, H.; Nakanishi, T.; Kami, K.; Arata, Y.; Shimada, I. *Nat. Struct. Biol.* **2000**, *7*, 220−223.
(20) Ramesh, V.; Syed, S. E. H.; Frederick, R. O.; Sutcliffe, M. J.; Barnes, M.; Roberts, G. C. K. *Eur. J. Biochem.* **1996**, *235*, 804−813.
(21) Ellman, J. A.; Volkman, B. F.; Mendel, G.; Schultz, P. G. *J. Am. Chem. Soc.* **1992**, *114*, 7959−7961.
(22) Weigelt, J.; Wikstrom, M.; Schultz, J.; van Dongen, M. *Comb. Chem. High Throughput Screening* **2002**, *5*, 623−630.
(23) Bogan, A. A.; Thorn, K. S. *J. Mol. Biol.* **1998**, *280*, 1−9.
(24) Hu, Z.; Ma, B.; Wolfson, H.; Nussinov, R. *Proteins: Struct., Funct., Genet.* **2000**, *39*, 331−342.

of His87 (see Figure 5) significantly contributed to a net 1.2 Å shift of the ligand as compared to the crystallographic result. Such changes can potentially be addressed by incorporating protein side-chain flexibility and/or superior scoring functions into the docking simulations.[11] However, as was the case for FKBP, it is expected that such subtle conformational changes will result only in modest changes in ligand orientation and that the resulting structures will still be extremely useful for lead characterization and optimization.

In summary, an NMR-based strategy is described for deriving structural information on ligands complexed to virtually any biomolecule that can be isotopically labeled, including protein targets of high molecular weight and heterogeneous, oligomeric complexes. What is required is a moderately soluble ligand that is in fast exchange on the NMR time scale (typically $K_D > 1$ μM), which is often the case for "leadlike" compounds which tend to be small and only moderately potent.[25,26] Thus, SOS-NMR is expected to have its greatest impact at the early stages of the drug discovery process, where initial structural information can greatly facilitate the optimization of a lead series into more potent compounds.

Methods

Sample Preparation. FKBP[27] was prepared as previously described. The coding sequence of MurA was amplified by PCR with primers encoding 5′- and 3′- restriction sites. The PCR product was digested ligated into the *Nco*I and *Xho*I sites of the pET21d(+) plasmid (Novagen, Madison). The MurA protein used for structural studies was expressed in *E. coli* BL21(DE3) grown on M9 media and was purified using a NiNTA affinity chromatography. Selectively ^{1}H labeled samples were grown in ^2H_2O, on minimal medium containing [U-^{2}H]glucose (CIL) with the protonated amino acid being added to a concentration of 100 mg/L 1 h prior to induction with 1 mM IPTG. No adaptation of the cells to deuterated media was necessary, with inoculation being performed by the addition of 1 mL of LB culture (grown to OD_{600} values between 1 and 2) added directly to 1 L of deuterated medium. Based on ^{1}H NMR spectra (see Figure 1), essentially complete perdeuteration was achieved, and no scrambling between the amino acid types was observed. The final conditions for FKBP were 1.0 mM **1** and 2.5 μM FKBP in a 100% D_2O buffer comprised of 10 mM Na_2PO_4, 140 mM NaCl, 5% DMSO-d_6, pH 7.4. Final conditions for MurA were 500 μM **2** and 20 μM MurA in a 100% D_2O buffer comprised of 50 mM Na_2PO_4, 5 mM ^{2}H-dithiothreitol (CIL), pH 7.0.

NMR Spectroscopy. The pulse sequence for the 1D STD-NMR spectra was as described by Mayer and Meyer[12] except that a CPMG T_2-filter was used instead of a $T_{1\rho}$-filter to remove background protein resonances. Subtraction of the on- and off-resonance spectra was performed internally, and a WATERGATE sequence was used for suppression of the residual solvent signal. For FKBP, all NMR experiments were performed at 300 K on a Bruker Avance DRX500 system equipped with a CryoProbe. On-resonance irradiation of the protein was performed using a train of Gauss-shaped pulses of 2.5 ms length separated by a 2.5 ms delay, and alternating on-resonance irradiation was applied. On-resonance radiation was applied at 0.5 and 2.0 ppm for all samples, with off-resonance irradiation applied at −10.0 ppm. A total of 280 shaped pulses were used in the pulse train, leading to a total saturation time of 1.4 s. All spectra were acquired with 1024 complex points, 1024 scans, a ^{1}H sweep width of 8333 Hz, a T_2-filter delays of 10 ms, and a total recycle time of 2.0 s. For MurA, all NMR experiments were performed at 300 K on a Bruker DRX800 NMR spectrometer. On-resonance irradiation of the protein was performed by using a train of Gauss-shaped pulses of 28 ms length separated by a 7 ms delay. On-resonance irradiation was applied at 6.81, 7.02, and 7.16 ppm for the His-, Trp-, and Phe-protonated samples, respectively. Off-resonance irradiation was applied at −2.8 ppm. A total of 30 shaped pulses were used in the pulse train, leading to a total saturation time of 1.05 s. All spectra were acquired with 8192 complex points, 1536 scans, a ^{1}H sweep width of 10 000 Hz, a total recycle time of 2.0 s, and without a T_2-filter. All NMR spectra were processed and analyzed on Silicon Graphics computers using in-house-written software.

Calculation of Distance Restraints. Distance restraints were calculated according to the following procedure. First, normalized STD-NMR intensities, *I*(STD-NMR), were obtained by dividing the intensity of each peak observed in the STD-NMR spectrum of **1** in the presence of unlabeled protein by the intensity observed in a one-dimensional T_2-filtered spectrum acquired under identical conditions. The fractional intensities were then normalized by scaling the maximum value to 100. This procedure is identical to that employed in group epitope mapping,[17] and the fractional intensities give a relative indication of the proximity of each proton on the ligand to the protein. Next, normalized NOE intensities, *I*(SOS), for each selectively protonated sample were calculated as the fraction of the observed NOE that can be attributed to interaction with the selectively protonated amino acid, scaled by the normalized STD-NMR intensity, *I*(STD-NMR), using the following equation:

$$I(\mathrm{SOS}) = \frac{I(\mathrm{SL}) - I(^2\mathrm{H})}{I(^1\mathrm{H}) - I(^2\mathrm{H})} I(\mathrm{STD-NMR})$$

where *I*(SL), *I*(^{2}H), and *I*(^{1}H) are the intensities observed in the STD-NMR spectra of the selectively labeled, perdeuterated, and unlabeled samples, respectively.

These NOE intensities were then converted into distance restraints for use in structural calculations in the following manner. For any given amino acid type and ligand resonance, normalized NOE intensities less than 5% were treated as having no interaction. Therefore, distances between that ligand atom and all protons on all amino acids of that specific type were set at greater than 5 Å. For NOE intensities greater than 15%, ambiguous distance restraints were assigned. Normalized NOE intensities between 15% and 25% were treated as weak interactions, and at least one distance between that ligand atom and at least one proton from an amino acid of that specific type should be less than 5 Å. Normalized NOE intensities between 25% and 45% were treated as moderate interactions, and at least one distance between that ligand atom and at least one proton from an amino acid of that specific type should be less than 4 Å. Normalized NOE intensities greater than 45% were treated as strong interactions, and at least one distance between that ligand atom and at least one proton from an amino acid of that specific type should be less than 3 Å. Peaks with normalized NOE intensities in the range of 5−15% were not used to assign a distance restraint because this was considered within the experimental error. Note that incomplete incorporation of the protonated amino acid will decrease the observed NOE intensities and potentially result in weaker distance constraints.

Generation of Binding Surfaces. The allowed ligand-binding surfaces were determined by the cumulative removal of the Connolly surface area for those atoms that violated the SOS-NMR data. Thus, for those residue types for which no proton on the ligand was in contact, the surfaces for all atoms within 5 Å of any proton on these residues were removed. For example, as no contacts were observed between **1** and any leucine residues of FKBP, the surfaces for all atoms within 5 Å of any leucine residue were removed. In a similar manner, for those residue types for which at least one proton on the ligand was in contact, the surfaces for all atoms not within a scaled distance of any proton

(25) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308−1315.
(26) Carr, R.; Jhoti, H. *Drug Discovery Today* **2002**, *7*, 522−527.
(27) Egan, D. A.; Logan, T. M.; Liang, H.; Matayoshi, E.; Fesik, S. W.; Holzman, T. F. *Biochemistry* **1993**, *32*, 1920−1927.
(28) Stryer, L. *Biochemistry*, 3rd ed.; W. H. Freeman and Co.: New York, 1988.

on any residue of this type were removed. For example, as positive contacts were observed between **1** and at least one isoleucine residue of FKBP, the surfaces for all atoms not within 10 Å of any atom of an isoleucine residue were removed. The larger distance used for the positive contacts reflects the length of the ligand, as the binding pocket is allowed to accommodate the entire molecule. The end-to-end lengths of **1** and **2** are approximately 8.7 and 13.2 Å, respectively, and positive contact distances of 10 and 14 Å were used in the surface generation.

Structure Calculations. DOCK 4.0[10] was used with default parameters. After docking, all low energy conformations were assigned NOE energies calculated in the usual manner using the ambiguous restraint list.

Binding Site Analysis. For the simulated SOS-NMR analysis, putative ligand-binding sites on a set of 272 protein surfaces (see Supporting Information) were determined in the following manner. First, each protein (in the absence of ligand) was solvated within InsightII (Accelerys). Next, the number of protein atoms within 5.6 Å of each water molecule was calculated. The solvent molecule with the highest number of contacts (i.e., most deeply buried in the protein surface) was chosen as a "centroid" to represent a putative ligand-binding site, and all solvent molecules with 5.6 Å of this centroid were removed. The solvent molecule with the next highest number of contacts was chosen as the next centroid, and this process was repeated until all solvent molecules were analyzed. Solvent molecules with fewer than five contacts to the protein were discarded from the analysis as they did not represent invaginations into the protein surface and were therefore not potential ligand-binding sites. For the set of 272 protein−ligand complexes, an average of 18 centroids were found, and the actual ligand-binding site was represented by a centroid in all cases. A figure depicting this process is given in the Supporting Information.

After identification of the centroids that represented potential ligand-binding sites, the ability of SOS-NMR to discriminate the actual ligand-binding site versus all other putative sites was accomplished in the following manner. First, the centroid nearest the geometric average of the ligand was identified as the centroid representing the actual ligand-binding site. Next, a single amino acid was chosen, and each centroid was assessed as either making contact or not making contact to that amino acid type on the basis of whether the minimum distance to any carbon atom of that amino acid type (excluding carbonyl carbons) was $\leq$5.6 Å. It was then ascertained whether the centroid representing the actual ligand-binding site could be distinguished from all others on the basis of this contact. Another amino acid was then added, contacts were measured, and it was again ascertained whether the centroid representing the actual ligand-binding site could be distinguished from all others on the basis of these two contacts. This was repeated an additional 18 times until the full set of 20 amino acids were included in the list. The order of addition of the amino acids was the following: (1) Val, (2) Leu, (3) Ile, (4) Met, (5) Lys, (6) Arg, (7) Pro, (8) Thr, (9) Ala, (10) Cys, (11) Trp, (12) Phe, (13) His, (14) Tyr, (15) Glu, (16) Gln, (17) Asp, (18) Asn, (19) Gly, (20) Ser.

To assess enrichment of the amino acids within the ligand-binding sites, the surfaces of all proteins were generated within InsightII (Accelerys), and the surface atoms were stored. Next, residues that had at least one side-chain heavy atom in the list of surface atoms were tabulated by residue type. This same list of surface atoms was then examined for contact distances ($<$5.0 Å) to the ligand and again tabulated by residue type. These results are given in Table 1. The frequency with which a particular amino acid occurred either on the surface or in the binding site was taken as the ratio of occurrence for that particular amino acid over all occurrences for all amino acids. For example, at least one side-chain heavy atom of alanine occurred on the surface 3691 times over all 272 proteins in the set, while alanine occurred in the ligand-binding site 190 times. Given total surface and binding-site amino acid occurrences of 59 097 and 3582, respectively, this results in a surface frequency of 6.2% and a binding-site frequency of 5.3% for alanine, yielding an enrichment factor of 0.85.

X-ray Crystallography. Crystals of FKBP complexed to **1** were obtained by cocrystallization using a solution mixture of 0.5 mM FKBP and 10 mM **1** in 5 mM Tris buffer, pH 7.4. The well solution of the cocrystallization plates was 55% saturation ammonium sulfate in 1 M Tris buffer, pH 8.4. The hanging drop was a 1:1 mixture of the well solution and the FKBP-**1** mixture. Diffraction data were collected at 100 K with mineral oil as cryoprotectant at the Advanced Photon Source (Argonne National Laboratories). The resolution was 1.7 Å, and R_{sym} was 4.6%. The space group was $P1$ with $a = 33.877$ Å, $b = 31.845$ Å, $c = 31.845$ Å, and $\alpha = 104.33°$, $\beta = 96.88°$, $\gamma = 113.83°$. There were two FKBP molecules in the unit cell. A total of 292 water molecules were included in the refinement. The final R and R_{free} values were 25.7% and 34.2%, respectively.

Acknowledgment. The authors would like to thank Dr. Robert Meadows for valuable discussions regarding the use of ambiguous distance restraints in binding site identification and structure determination.

Supporting Information Available: A figure depicting the process of centroid generation as described in the Methods section and a table listing all of the protein−ligand complexes used in the binding site analysis (PDF). This material is available free of charge via the Internet at http://pubs.acs.org.

JA039480V