Stereostructure of luminamicin, an anaerobic antibiotic, via molecular dynamics, NMR spectroscopy, and the modified Mosher method

Hiroaki Gouda*, Toshiaki Sunazuka†, Hideaki Ui*, Masaki Handa*, Yusuke Sakoh*, Yuzuru Iwai†, Shuichi Hirono*‡, and Satoshi Ōmura†‡

*School of Pharmaceutical Sciences, Kitasato University, Shirokane, Minato-ku, Tokyo 108-8641, Japan; and †Kitasato Institute for Life Sciences, Graduate School of Infection Control Sciences, Kitasato University, and The Kitasato Institute, Shirokane, Minato-ku, Tokyo 108-8641, Japan

The absolute stereostructure of luminamicin, an anaerobic antibiotic, has been determined by using conformational analysis via high-temperature molecular dynamics, NMR spectroscopy, and the modified Mosher method. It was found that luminamicin has the S, S, R, R, R, R, S, S, S, R, and S configurations at C2, C4, C7, C9, C10, C11, C12, C13, C16, C28, and C29, respectively. This configuration is the same as that found in nodusmicin, which has a chemical structure quite similar to luminamicin. The structure of luminamicin consists of three different rings, i.e., a decalin ring, a 10-membered macrolactone ring, and a 14-membered macrolactone ring. The resulting three-dimensional structure of luminamicin shows an interesting feature in that the maleic anhydride functionality in conjugation with the enol ether group of the 14-membered macrolactone is nearly perpendicular to the plane of the other two rings.

The emergence of resistance against generally used antibiotics will be a long-lasting serious clinical problem. We need to continue developing new medicines that have unique mechanisms of action. We have found previously uncharacterized antianaerobe antibiotics of actinomycetes origin, thiotetromycin (1), clostomicin (2), luminamicin (3), and lustromycin (4, 5). The structure of luminamicin (Fig. 1) was identical to that of coloradocin, which was isolated by McAlpine (6). Luminamicin showed selective activity against anaerobic and microaerophilic bacteria, including pathogenic species of *Clostridium*, *Neisseria*, and *Haemophilus*, whereas it was not active against most aerobic bacteria (3, 6). Interestingly, structurally related antibiotics, nodusmicin (7) and nargenicin (8), are reported to inhibit some aerobic bacteria. Therefore, the additive macrocyclic structure of luminamicin seems to be important in exerting its selective antianaerobic microbial activity. In conjunction with our continuing program directed toward the structure elucidation and synthesis of important antimicrobial natural products, we describe here the elucidation of the absolute stereochemistry of luminamicin.

The preliminary NMR study of luminamicin could explain only relative configuration with respect to C4, C7, C9, C10, C11, C12, and C13, but could not provide any absolute configuration of chiral carbon atoms of luminamicin (6). Recently, we have determined the stereostructure of choloropeptin I, an unusual chlorinated hexapeptide with selective anti-HIV activities, by inhibiting the binding between HIV gp120 envelope protein and CD4 protein, using a combination method of conformational analysis via high-temperature molecular dynamics (MD) and NMR spectroscopy (9–11). Among six amino acids involved in choloropeptin I, three were established to be all *R* configurations by acidic hydrolysis, but the other three had not been assigned. In that study, we just needed to consider eight different diastereomers. Therefore, NMR spectroscopy has been applied successfully to determine the stereochemistry of choloropeptin I. However, in the case of luminamicin, we have no absolute configuration of chiral carbon atoms and need to consider both

diastereomers and enantiomers. Therefore, only applying NMR spectroscopy is not enough to determine the *absolute* stereochemistry of luminamicin, because it can provide only *relative* structural information. However, the modified Mosher method is well known to give an absolute configuration for a chiral carbon atom with a hydroxyl group (12, 13). Luminamicin has two chiral carbon atoms attached to a hydroxyl group. Therefore, we can expect that the modified Mosher method will be able to assign an absolute configuration for each of these carbons atoms. In this study, we use a combination of conformational analysis via high-temperature MD, NMR spectroscopy, and the modified Mosher method to determine the absolute stereochemistry of luminamicin. Once we obtain at least one absolute configuration by the modified Mosher method, we can assign absolute configurations for other chiral carbon atoms using relative structural information derived from NMR spectroscopy. We will show that this combination method is very powerful, even when it is necessary to consider both diastereomers and enantiomers.

Methods

NMR Experiments. Luminamicin (1) was prepared as described in ref. 3. The sample of 10 mg was dissolved in 0.6 ml of CDCl₃. All NMR spectra were recorded on a Varian INOVA600 spectrometer operating at 600 MHz for the proton frequency at 30°C. A relaxation delay of 2.0 s was used in all experiments. The 1D proton spectrum used to estimate coupling constants was measured with a spectral width of 9,012 Hz and a data block size of 27,000, so the digital resolution was 0.67 Hz per point. An unshifted sinc bell was applied to the free induction decay NMR signal. Chemical shifts were referenced with respect to residual solvent signal (7.26 ppm).

For calculation of distance constraints, the rotating-frame Overhauser effect (ROE) data were collected by the 2D-TROESY (transverse rotating-frame Overhauser effect spectroscopy) pulse sequence of Hwang and Shaka (14, 15) with mixing times of 100, 150, 200, 250, and 300 ms, which is designed to suppress total correlation spectroscopy cross peaks. The spinlock pulse was 180°(*x*) 180°(*x*), with a field strength of 6,200 Hz. 2D-TROESY spectroscopy was performed with a spectral width of 5,273 Hz in the phase-sensitive mode (16). A total of 512 blocks were acquired with data points of 2,048. Before 2D Fourier transformation, the acquired data were multiplied by a Gauss function in t2 and by a shifted sine square function in t1 and were zero-filled once along the t1 direction.
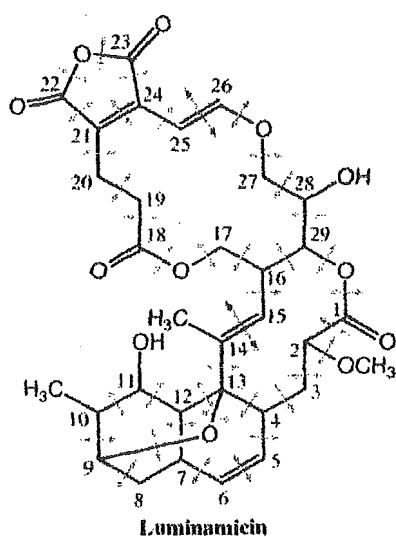
Fig. 1. Chemical structure of luminamicin. The dihedral angles used to cluster similar conformers are indicated by blue arrows. The dihedral angles constrained to be tarns in MD calculations are indicated by red dashed arrows.
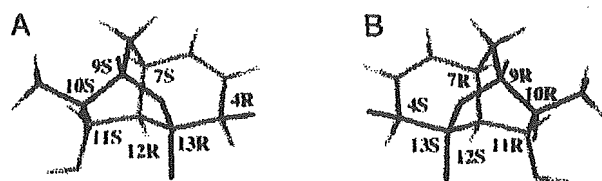

Fig. 2. Two possible stereostructures of the decalin ring of luminamicin. (A) C4, C7, C9, C10, C11, C12, and C13 have the R, S, S, S, S, R, and R configurations, respectively. (B) C4, C7, C9, C10, C11, C12, and C13 have the S, R, R, R, R, S, and S configurations, respectively.

Conformational Analysis of Luminamicin.

The preliminary NMR study of luminamicin (**1**) suggested that there are only two possible absolute configurations on a decalin ring system including C4–C13, i.e., the R, S, S, S, S, R, and R or the S, R, R. R, R, S, and S configurations at C4. C7, C9, C10, C11, C12, and C13, respectively (Fig. 2). Therefore, we first prepared initial structures of the 16 diastereomers differing in the configuration at C2, C16, C28, and C29 for luminamicin with either of these two decalin configurations, using SYBYL 6.91 (Tripos, St. Louis). As a result, a total of 32 initial structures were prepared. Then, the conformational analysis of these initial structures was performed by using the program CAMDAS 2.1 (Conformational Analyzer with Molecular Dynamics and Sampling) (17). This program, developed in our laboratory, generates the energetically stable conformations of a target molecule by performing the high-temperature MD calculation and sampling conformations along the MD trajectory. CAMDAS then clusters similar conformations based on values of dihedral angles defined before calculation. The following calculations were performed for each initial structure by using CAMDAS. Ten MD calculations were simultaneously performed by using different conformers. Each of the MD calculations was carried out for 1,000 ps with an integral time step of 1 fs. The lengths of covalent bonds were fixed with SHAKE algorithm through the MD (18). The temperature of the system was maintained at 1,200 K to enhance the sampling efficiency. The Merck Molecular Force Field was used to evaluate the potential energy surface of the molecule (19). To mimic the shield effects of solvent molecules on electrostatic interactions, the electrostatic potential term was neglected. The 14-methyl group was suggested to be trans to H15 by its ^{13}C chemical shift value of 15.1 ppm (20). The H25 was also indicated to be trans to H26 on the basis of the large coupling constant between these protons of 13.5 Hz. Therefore, values of the dihedral angles of C13-C14-C15-C16 and H25-C25-C26-H26 were constrained to be 180° ± 10° in MD calculations to keep their conformations trans. The constraint energy term was quadratic, and the force constants were 100 kcal·mol^{-1}·rad^{-2}. Conformers were sampled at 100-step intervals, thus producing 10,000 conformations for each MD calculation. A total of 100,000 conformations were preclustered with a dihedral angles deviation threshold of ±30°. A total of 37 dihedral angles were

used to cluster similar conformations, which are indicated by arrows in Fig. 1. Each of the conformers obtained after preclustering was then minimized until the root mean square (RMS) of the potential-energy gradient fell below 0.001 kcal·mol^{-1}·Å^{-1}. The minimized conformers were reclustered with a dihedral angle deviation threshold of ±30°, furnishing a final conformer set. All finally obtained conformers maintained chirality of their initial structures, although chirality restraints were not used in MD calculations. This result indicated that no inversion of chiral centers occurred during MD calculations.

Modified Mosher Method. *28-[(+)-α-Methoxy-α-(trifluoromethyl)phenyl-acetoxy]luminamicin (2).* To a solution of 1 (9.9 mg, 16.1 μmol) in CH$_2$Cl$_2$ (800 μl) at room temperature was added (+)-MTPA (11.3 mg, 48.3 μmol), EDCI [1-ethyl-3-(3-dimethylaminopropyl)carbodiimide] (9.3 mg, 48.3 μmol), and DMAP (4-dimethylaminopyridine) (0.4 mg, 3.22 μmol), and the mixture was stirred for 1 h and quenched with H$_2$O (5 ml). The resultant mixture was extracted with CHCl$_3$ (3 × 5 ml), and the combined extracts were washed with brine (5 ml), dried over Na$_2$SO$_4$, filtered, and concentrated. Flash column chromatography (Benzene/Acetone 20/1) furnished 2 (6.2 mg, 7.47 μmol, 46% yield) as a colorless solid: [α]$_D^{25}$ +34.4 °(c 0.29, CHCl$_3$); TLC R_f 0.61 (Benzene/Acetone 2/1); mp 108–109°C; IR (KBr) 3,442, 1,761, 1,178 cm^{-1}; ^{1}H-NMR (600 MHz, CDCl$_3$) δ 0.98 (3H, d, J = 6.9 Hz), 1.42 (1H, m), 1.45 (1H, m), 1.69 (3H, s), 1.70 (1H, m), 1.80 (1H, m), 1.90 (1H, m), 2.19 (1H, m), 2.21 (1H, m), 2.48 (1H, m), 2.62 (1H, m), 2.72 (1H, m), 2.90 (1H, m), 2.96 (1H, m), 3.27 (3H, s), 3.34 (1H, m), 3.36 (1H, m), 3.64 (1H, m), 3.78 (1H, dd, J = 12.1, 5.0 Hz), 4.22 (1H, m), 4.26 (2H, m), 4.59 (1H, d, J = 12.1 Hz), 5.47 (1H, dd, J = 9.9, 2.5 Hz), 5.62 (1H, dd, J = 9.6, 5.5 Hz), 5.78 (1H, d, J = 13.5 Hz), 5.89 (1H, m), 5.90 (1H, d, J = 6.3 Hz), 6.04 (1H, dd, J = 9.6, 7.4 Hz), 7.91 (1H, d, J = 13.5 Hz); ^{13}C-NMR (150 MHz, CDCl$_3$) δ 15.1, 15.9, 18.7, 27.9, 29.6, 32.9, 33.0, 37.3, 38.1, 38.3, 40.9, 58.0, 64.2, 68.8, 70.1, 70.4, 70.6, 76.4, 77.4, 82.7, 96.3, 122.3, 128.2, 130.3, 133.9, 138.1, 143.0, 156.9, 164.0, 165.7, 171.2, 172.7; HRMS calculated for C$_{42}$H$_{46}$O$_{14}$F$_3$ [M+H]$^+$ 831.2840, found 831.2859.

28-[(−)-α-Methoxy-α-(trifluoromethyl)phenylacetoxy]luminamicin (3). Previous procedure was performed with 1 (8.9 mg, 14.4 μmol), (−)-MTPA (10.1 mg, 43.2 μmol), EDCI (8.3 mg, 43.2 μmol), and DMAP (0.4 mg, 2.88 μmol) in CH$_2$Cl$_2$ (720 μl) to afford 3 (4.1 mg, 4.94 μmol, 34% yield) as a colorless solid: [α]$_D^{26}$ +12.9 ° (c 0.185, CHCl$_3$); TLC R_f 0.63 (Benzene/Acetone 2/1); mp 115–116°C; IR (KBr) 3.521, 1.761, 1,178 cm^{-1}; ^{1}H-NMR (600 MHz, CDCl$_3$) δ 0.98 (3H, d, J = 6.9 Hz), 1.39 (1H, m), 1.43 (1H, m), 1.60 (3H, m), 1.70 (1H, dd, J = 13.2, 11.0 Hz), 1.77 (1H, m), 1.88 (1H, m), 2.17 (1H, m), 2.20 (1H, m), 2.46 (1H, m), 2.62 (1H, ddd, J = 15.1, 12.4, 2.7 Hz), 2.71 (1H, ddd, J = 14.9, 7.2, 2.7 Hz), 2.76 (1H, m), 2.90 (1H, ddd, J = 15.1, 7.2, 2.5 Hz), 3.27 (3H, s), 3.33 (1H, m), 3.35 (1H, m), 3.64 (1H, dd, J = 5.0, 1.9 Hz), 3.76 (1H, dd, J = 12.1, 4.7 Hz), 4.18 (1H, dd, J = 12.1, 2.2 Hz), 4.30 (2H, d, J = 3.9 Hz), 4.55 (1H, dd, J = 12.1, 1.7 Hz), 5.46 (1H, dd, J = 9.6, 2.5 Hz), 5.49 (1H, dd, J = 10.2, 5.5 Hz), 5.78 (1H, d, J = 13.5 Hz), 5.89 (1H, m), 5.89 (1H, m), 6.03 (1H, ddd, J = 10.2, 4.1, 3.9

Table 1. Vicinal coupling constants and torsinal constraints

Torsional angle	$^3J_{HH}$, Hz	Torsional constraints
H2-C2-C3-H3a (2.49 ppm)	12.4	180° ± 40°
H3a (2.49 ppm)-C3-C4-H4	1.9	-90° ± 40° or 90° ± 40°
H16-C16-C17-H17b (4.05 ppm)	<1.0	-90° ± 40° or 90° ± 40°
H16-C16-C17-H17a (4.87 ppm)	<1.0	-90° ± 40° or 90° ± 40°
H19b (2.46 ppm)-C19-C20-H20a (2.67 ppm)	2.2	-90° ± 40° or 90° ± 40°
H19b (2.46 ppm)-C19-C20-H20b (3.29 ppm)	12.6	180° ± 40°
H19a (2.86 ppm)-C19-C20-H20b (3.29 ppm)	2.0	-90° ± 40° or 90° ± 40°
H28-C28-C29-H29	10.7	180° ± 40°

Hz), 7.94 (1H, d, J = 13.5 Hz); ^{13}C-NMR (150 MHz, CDCl₃) δ 14.9, 16.0, 18.7, 27.9, 29.6, 32.9, 33.0, 37.0, 38.1, 38.3, 41.1, 58.0, 64.3, 68.9, 69.8, 70.5, 70.6, 76.4, 77.4, 82.8, 96.5, 122.5, 128.2, 130.3, 134.1, 138.0, 142.8, 156.8, 164.0, 165.7, 171.1, 172.8; HRMS calculated for C₄₂H₄₅O₁₄F₃ [M]⁺ 830.2734, found 830.2761.

Results and Discussion

Distance and Dihedral Constraints Derived from NMR Experiments. Torsional constraints were obtained by applying the Karplus equation to vicinal proton–proton scalar coupling constants obtained by the 1D proton spectrum with high resolution (21). A proton–proton coupling constant >10 Hz was treated as indicating an anti H–H orientation and a dihedral-angle estimate of 180° ± 40°. A coupling constant <3 Hz was considered indicative of a *gauche* orientation, i.e., -90° ± 40° or 90° ± 40°. The resulting torsional constraints are given in Table 1.

To obtain distance constraints, 2D-TROESY spectra were measured with mixing times of 100, 150, 200, 250, and 300 ms. A plot of the volume of the cross-peak versus mixing time showed linearity up to 200 ms. Therefore, the proton–proton distance constraints were based on the integrated cross-peaks from the 200-ms spectrum. Fig. 7, which is published as supporting information on the PNAS web site, shows a part of the 2D-TROESY spectrum measured with a mixing time of 200 ms. Volumes of the five nonoverlapping geminal proton cross-peaks

were averaged and used for calibrating measured ROE volumes. A distance of 1.8 Å was used as a distance reference for geminal proton cross-peaks. This calibration yielded the theoretically expected value, 2.4 Å, for the distance between H5 and H6. Distance constraints were classified into three categories corresponding to 1.8–2.7, 1.8–3.5, and 1.8–5.0 Å, corresponding to strong, medium, and weak ROEs, respectively. For the distance constraints related to methylene protons at C27 (H27a,b), of which chemical shifts are overlapping, or methyl protons of 2-methoxy, 10-methyl, or 14-methyl groups, carbon atoms attached to these protons were used to estimate target distances. In such a case, 1.0 Å was added to the upper boundary of the constraints (22). For example, the strong ROE observed between H4 and the 14-methyl group is converted to a distance constraint between H4 and carbon atom of 14-methyl group (14MeC), the range of which is 1.8–3.7 Å. A total of 43 distance constraints were obtained as shown in Table 2.

Determination of Absolute Stereochemistry of Luminamicin. Table 3 lists the number of energetically possible conformers obtained from CAMDAS calculation for 32 different configurations. About 2,500~3500 distinct conformers were obtained for each configuration. Then, we calculated RMS deviations for all conformers generated by CAMDAS to determine which configurations can adopt conformers accommodating the distance and dihedral constraints. All distance constraints were treated with equal

Table 2. Distance constraints obtained by the 2D-TROESY experiments

Atom A	Atom B	The upper bound, Å	Atom A	Atom B	The upper bound, Å
H2	2MeC	3.7	H9	H15	5.0
H2	H3b (1.41 ppm)	2.7	H10	H11	3.5
H2	H4	2.7	H10	H15	3.5
H2	H5	3.5	10MeC	H11	3.7
2MeC	H25	4.5	H12	14MeC	3.7
2MeC	C27	5.5	14MeC	H16	3.7
H3b (1.41 ppm)	H4	3.5	14MeC	H29	3.7
H3b (1.41 ppm)	H5	2.7	H15	H16	3.5
H3a (2.49 ppm)	H4	3.5	H15	H17b (4.05 ppm)	2.7
H3a (2.49 ppm)	H15	2.7	H16	H17b (4.05 ppm)	2.7
H4	H5	2.7	H16	H17a (4.87 ppm)	2.7
H4	H12	3.5	H16	H28	5.0
H4	14MeC	3.7	H16	H29	2.7
H6	H7	2.7	H17a (4.87 ppm)	H28	2.7
H6	H8a (1.43 ppm)	3.5	H19b (2.46 ppm)	H20a (2.67 ppm)	2.7
H7	H8b (1.69 ppm)	2.7	H19a (2.86 ppm)	H20b (3.29 ppm)	2.7
H8a (1.43 ppm)	H9	2.7	H20b (3.29 ppm)	H25	2.7
H8b (1.69 ppm)	H9	2.7	H25	C27	3.7
H8b (1.69 ppm)	10MeC	3.7	H25	H28	2.7
H8b (1.69 ppm)	H11	3.5	C27	H29	3.7
H9	H10	2.7	H28	H29	3.5
H9	10MeC	3.7			

Table 3. RMS deviations of CAMDAS-generated conformers of luminamicin from experimentally determined values

Configuration with respect to the decalin ring							Configuration of C2, C16, C28, and C29				Smallest RMS deviation from exptl distance constraints, Å	RMS deviation from exptl dihedral constraints, °*	Number of conformers obtained by CAMDAS
C4	C7	C9	C10	C11	C12	C13	C2	C16	C28	C29			
R	S	S	S	S	R	R	R	R	R	R	0.34	0.0	3,185
R	S	S	S	S	R	R	R	R	R	S	0.32	0.9	2,369
R	S	S	S	S	R	R	R	R	S	R	0.00	0.0	3,287
R	S	S	S	S	R	R	R	R	S	S	0.17	18.6	2,749
R	S	S	S	S	R	R	R	S	R	R	0.36	18.5	3,022
R	S	S	S	S	R	R	R	S	R	S	0.27	0.0	2,769
R	S	S	S	S	R	R	R	S	S	R	0.33	19.0	3,016
R	S	S	S	S	R	R	R	S	S	S	0.33	9.5	2,931
R	S	S	S	S	R	R	S	R	R	R	0.39	37.2	3,294
R	S	S	S	S	R	R	S	R	R	S	0.51	0.0	3,064
R	S	S	S	S	R	R	S	R	S	R	0.36	8.1	3,375
R	S	S	S	S	R	R	S	R	S	S	0.38	18.7	3,235
R	S	S	S	S	R	R	S	S	R	R	0.41	19.2	2,954
R	S	S	S	S	R	R	S	S	R	S	0.19	0.0	3,392
R	S	S	S	S	R	R	S	S	S	R	0.43	20.0	2,895
R	S	S	S	S	R	R	S	S	S	S	0.17	0.2	3,298
S	R	R	R	R	S	S	R	R	R	R	0.52	11.4	3,297
S	R	R	R	R	S	S	R	R	R	S	0.54	14.3	3,160
S	R	R	R	R	S	S	R	R	S	R	0.43	0.0	3,477
S	R	R	R	R	S	S	R	R	S	S	0.27	36.5	3,105
S	R	R	R	R	S	S	R	S	R	R	0.38	0.0	3,109
S	R	R	R	R	S	S	R	S	R	S	0.36	8.1	3,321
S	R	R	R	R	S	S	R	S	S	R	0.49	13.8	3,067
S	R	R	R	R	S	S	R	S	S	S	0.53	8.9	3,252
S	R	R	R	R	S	S	S	R	R	R	0.39	0.0	2,806
S	R	R	R	R	S	S	S	R	R	S	0.33	19.0	2,955
S	R	R	R	R	S	S	S	R	S	R	0.35	0.0	2,909
S	R	R	R	R	S	S	S	R	S	S	0.36	18.5	2,886
S	R	R	R	R	S	S	S	S	R	R	0.17	18.7	2,569
S	R	R	R	R	S	S	S	S	R	S	0.00	0.0	3,204
S	R	R	R	R	S	S	S	S	S	R	0.50	18.9	2,521
S	R	R	R	R	S	S	S	S	S	S	0.37	0.0	3,329

*Calculated for the conformer having the smallest RMS deviations from the distance constraints.

weight, although they were classified as strong, medium, and weak. For constraints including nonoverlapping geminal protons (3-, 8-, 17-, or 19-, or 20-methylene protons), careful analysis was done when estimating target dihedral angles and proton–proton distances. For example, H3$_a$ (2.49 ppm) is included in a total of four constraints, i.e., two dihedral constraints and two distance constraints as shown in Tables 1 and 2. The proton corresponding to H3$_a$ should simultaneously satisfy all of these four constraints. Because we cannot establish the stereospecific assignments on 3-methylene protons (H3$_a$ and H3$_b$) in advance, we first need to consider all combinations when estimating target dihedral angles or target proton–proton distances. For example, we first estimate a total of four dihedral angles, i.e., two dihedral angles are combination between H2 and two 3-methylene protons, and other two are combination between H4 and two 3-methylene protons. Then, we determine which of two 3-methylene protons can simultaneously satisfy two dihedral constraints included in Table 1. The identified proton can be considered to correspond to H3$_a$ (2.49 ppm). Next, for the identified 3-methylene proton, we estimate two target distances related to H3$_a$ (2.49 ppm) in Table 2. The similar procedure was also performed for the constraints on 8-, 17-, 19-, or 20-methylene protons. Table 3 includes the smallest RMS distance deviations obtained for each configuration. The conformers satisfying all experimental constraints were found in only two configurations. For luminamicin with the R, S, S, S. S, R, and R configurations at C4, C7, C9, C10, C11, C12, and C13, respectively, C2, C16, C28, and C29 should

have the R, R, S, and R configurations, respectively. On the other hand, for luminamicin with the S, R, R, R. R, S, and S configurations at C4, C7, C9, C10, C11, C12, and C13, respectively, the results embodied the S, S, R, and S configurations at C2, C16, C28, and C29, respectively. Obviously, the former configuration is an enantiomer of the latter one. This means that NMR constraints could not distinguish between enantiomers. This was an expected result, because distance and dihedral constraints can provide only relative structural orientation. The final determi-
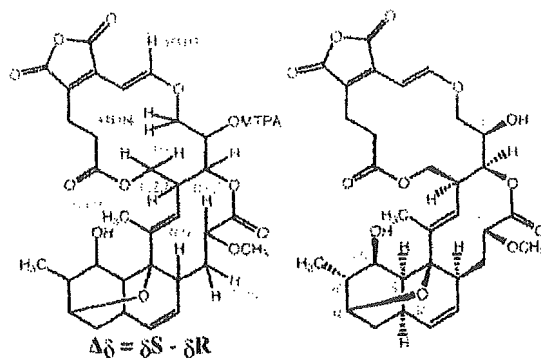

$$\Delta\delta = \delta S - \delta R$$

Fig. 3. Δδ Values of MTPA esters from luminamicin (1).

Luminamicin   Nodusmicin
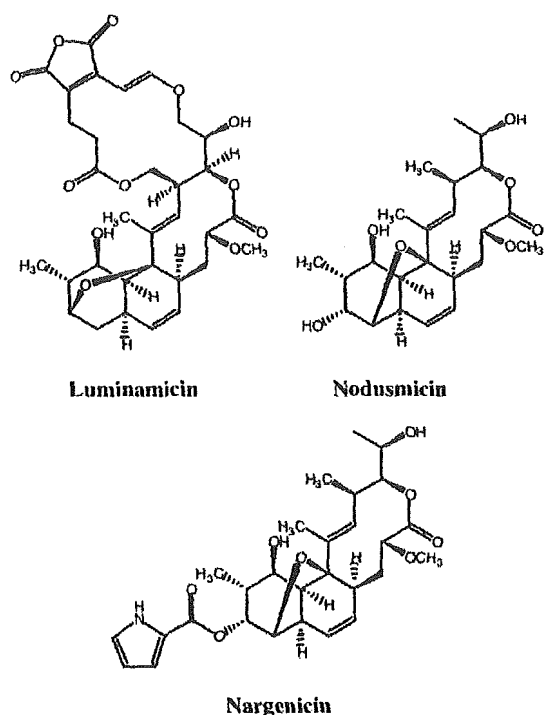
Nargenicin

Fig. 4. Structures of luminamicin, nodusmicin, and nargenicin.

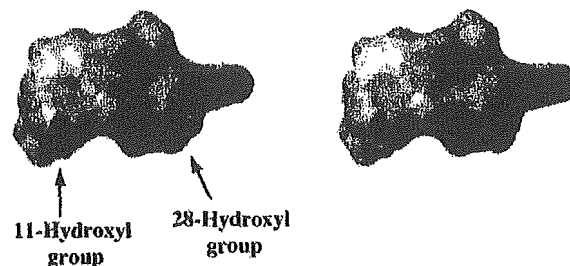**Fig. 6.** Stereoview of surface representation of luminamicin. This view direction is the same as that of Fig. 5B.

11-Hydroxyl   28-Hydroxyl
group   group

nation of absolute stereostructure of luminamicin was done by using the result of the following modified Mosher method.

1 was treated with (*R*)-(+)- and (*S*)-(−)-2-methoxy-2-trifluoromethyl-2-phenylacetic acid (MTPA) in the presence of EDCI and DMAP to afford the (*R*)-(+)- and (*S*)-(−)-MTPA esters (2 and 3) (see Scheme 1, which is published as supporting information on the PNAS web site). Fig. 3 shows the $\Delta\delta$ values ($\delta_S - \delta_R$) obtained from the ¹H-NMR data of 2 and 3, which are included in Table 4, which is published as supporting information on the PNAS web site. The $\Delta\delta$ values for H₃-14-Me, H-16, and H-29 were negative, whereas positive $\Delta\delta$ values were obtained for H-26 and H₂-27, thus indicating an *R* configuration. The absolute configurations therefore at C2, C4, C7, C9, C10, C11, C12, C13, C16, C28, and C29 of 1 were assigned as *S, S, R, R, R, R, S, S, S, R,* and *S,* respectively.

This configuration is the same as that found in nodusmicin (7) and nargenicin (8), which have a chemical structure quite similar with luminamicin (Fig. 4).

Description of the 3D Structure of Luminamicin. The process described above led to a set of two energy-minimized structures for luminamicin, which have Merck Molecular Force Field energies of 135.6 and 142.1 kcal/mol, respectively. Fig. 5 shows stereopairs of the best-fit superposition of the heavy atoms for these two structures. The structures only differ with respect to the orientation of the 2-methoxy group. As shown is Fig. 1, luminamicin consists of three different rings, i.e., a decalin, a
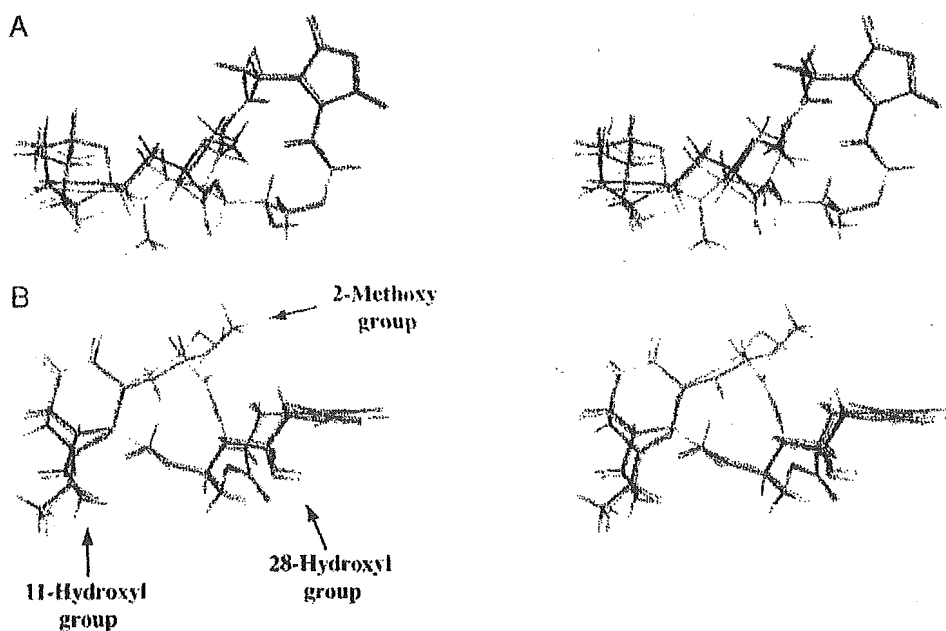
A

B

2-Methoxy
group

28-Hydroxyl
group

11-Hydroxyl
group

Fig. 5. Stereopairs of the superposition of the resulting two 3D structures of luminamicin. Black and green structures have Merck Molecular Force Field energies of 135.6 and 142.1 kcal/mol, respectively. These are the results of the best fit of the heavy atoms. *B* is rotated 90° in relation to *A*.

10-membered macrolactone, and a 14-membered macrolactone. The earlier NMR study described only the stereochemistry of the decalin ring (Fig. 2). Both C3 and C14 of the 10-membered macrolactone ring occupy equatorial positions relative to the decalin ring, as indicated in Fig. 2. The two strong ROEs between H4 and 14-methyl group and H12 and 14-methyl group conclusively established the spatial orientation of the 14-methyl group. It must be on the same side with H4 and H12 of the decalin ring. The configuration of C2 was well defined by the ROEs between H2 and H4, 2-methoxy group and H27$_{a,b}$, and 2-methoxy group and H25. These ROEs result in the 2-methoxy group oriented to the 14-membered macrolactone. The 10- and 14-membered macrolactones were clearly found to be cis-fused in this study, because critical strong ROEs were observed between H16 and H29, H16 and 14-methyl group, and H29 and 14-methyl group, indicating that these protons are on one side of the molecule. Both of H17$_a$ and H17$_b$ have gauche orientation relative to H16, because both correlations of 16H to H17$_a$ and H17$_b$ are strong in the TROESY spectrum. The small values of $J_{16,17a}$ (<1.0 Hz) and $J_{16,17b}$ (<1.0 Hz) are also consistent with these gauche orientations. The orientation of olefinic group (C25 and C26) was established by the strong correlations of H25 to H20$_b$ and H27 in the TROESY spectrum. H25 is oriented inside the molecule and H26 is turned outside. The stereochemistry of C28 was determined by the strong ROE between H25 and H28, resulting in the 28-hydroxyl group oriented outside the molecule. The resulting 3D structure of luminamicin shows an interesting feature that the maleic anhydride functionality in conjugation

with olefinic group of the 14-membered macrolactone is nearly perpendicular to the plane of the decalin and 14-membered macrolactone rings, as shown in Fig. 5.

Luminamicin has two hydroxyl groups at C11 and C28, which are expected to react with (R)-MTPA or (S)-MTPA. However, it appears that these reagents can react with only the 28-hydroxyl group. With respect to this phenomenon, we have obtained some relevant information from the 3D structure of luminamicin. As shown in Fig. 6, two hydroxyl groups at C11 and C28 are placed on one side of luminamicin. In addition, the 28-hydroxyl group is more exposed to the solvent, because its solvent-accessible surface area (36.9 Å) is larger than that of the hydroxyl group at C11 (29.7 Å). Therefore, we can consider that the reagents appear to more easily react at C28, and the (R)-MTPA or (S)-MTPA ester group attached at C28 most likely hinder the reaction occurring at C11.

In conclusion, the absolute stereostructure of luminamicin, an anaerobic antibiotic, has been determined by using the conformational analysis via high-temperature MD, NMR spectroscopy, and the modified Mosher method. Luminamicin (1) could be a new lead for medicines to compare with vancomycin, which is used clinically in pseudomembranous colitis therapy.

1. Ōmura, S., Iwai, Y., Naagawa, A., Iwata, R., Takahashi. Y., Shimizu, H. & Tanaka, H. (1983) J. Antibiot. 36, 109–114.
2. Ōmura, S., Imamura, N., Oiwa, R., Kuga, H., Iwata, R., Masuma, R. & Iwai, Y. (1986) J. Antibiot. 39, 1407–1412.
3. Ōmura, S., Iwata, R., Iwai, Y., Taga, S., Tanaka, Y. & Tomoda, H. (1985) J. Antibiot. 38, 1322–1326.
4. Tomoda, T., Iwata, R., Takahashi. Y., Iwai, Y., Oiwa, R. & Ōmura, S. (1986) J. Antibiot. 39, 1205–1210.
5. Handa, M., Ui, H., Yamamoto, D., Monma, S., Iwai, Y., Sunazuka, T. & Ōmura, S. (2003) Heterocycles 59, 497–500.
6. Rasmussen, R. R., Scherr, M. H., Whittern, D. N., Buko, A. M. & McAlpine, J. B. (1987) J. Antibiot. 40, 1383–1393.
7. Whaley, H. A., Chidester, C. G., Mizsak, S. A. & Wnuk, R. J. (1980) Tetrahedron Lett. 21, 3659–3662.
8. Celmer, W. D., Chmurny, G. N., Moppett, C. E., Ware, R. S., Watts, P. C. & Whipple, E. B. (1980) J. Am. Chem. Soc. 102, 4203–4209.
9. Matsuzaki, K., Ikeda, H., Ogino, T., Matsumoto, A., Woodruff, H. B., Tanaka, H. & Ōmura, S. (1994) J. Antibiot. 47, 1173–1174.
10. Matsuzaki, K., Ogino, T., Sunazuka, T., Tanaka, H. & Ōmura, S. (1997) J. Antibiot. 50, 66–69.

11. Gouda, H., Matsuzaki, K., Tanaka, H., Hirono, S., Ōmura, S., McCauley, J. A., Sprengeler, P. A., Furst, G. T. & Smith, A. B. (1996) J. Am. Chem. Soc. 118, 13087–13088.
12. Ohtani, I., Kusumi, T., Kashman, Y. & Kakisawa, H. (1991) J. Org. Chem. 56, 1296–1298.
13. Dale, J. A., Dull, D. L. & Mosher, H. S. (1969) J. Org. Chem. 34, 2543–2549.
14. Hwang, T.-L. & Shaka, A. J. (1992) J. Am. Chem. Soc. 114, 3157–3159.
15. Hwang, T.-L. & Shaka, A. J. (1992) J. Magn. Reson. B 102, 155–165.
16. States, D. J., Haberkorn, R. A. & Ruben, D. J. (1982) J. Magn. Reson. 48, 286–292.
17. Tsujishita, H. & Hirono, S. (1997) J. Comput. Aided Mol. Des. 11, 305–315.
18. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. C. (1977) J. Comput. Chem. 23, 327–341.
19. Halgren, H. (1990) J. Am. Chem. Soc. 112, 4710–4723.
20. de Haan, J. W. & van de Ven, L. J. M. (1973) Org. Magn. Reson. 5, 147–153.
21. Karplus, M. (1963) J. Am. Chem. Soc. 85, 2870–2871.
22. Steinmetz, W. E., Sadowsky, J. D., Rice, J. S., Roberts, J. J. & Bui, Y. K. (2001) Magn. Reson. Chem. 39, 163–172.

Comparison of Consensus Scoring Strategies for Evaluating Computational Models of Protein—Ligand Complexes

Akifumi Oda,*,† Keiichi Tsuchida,† Tadakazu Takakura,† Noriyuki Yamaotsu,‡ and Shuichi Hirono‡

Discovery Laboratories, Toyama Chemical Co., Ltd., 2-4-1 Shimookui, Toyama 930-8508, Japan, and School of Pharmaceutical Sciences, Kitasato University, 5-9-1 Shirokane, Minato-ku, Tokyo 108-8641, Japan

Here, the comparisons of performance of nine consensus scoring strategies, in which multiple scoring functions were used simultaneously to evaluate candidate structures for a protein—ligand complex, in combination with nine scoring functions (FlexX score, GOLD score, PMF score, DOCK score, ChemScore, DrugScore, PLP, ScreenScore, and X-Score), were carried out. The systematic naming of consensus scoring strategies was also proposed. Our results demonstrate that choosing the most appropriate type of consensus score is essential for model selection in computational docking; although the vote-by-number strategy was an effective selection method, the number-by-number and rank-by-number strategies were more appropriate when computational tractability was taken into account. By incorporating these consensus scores into the FlexX program, reasonable complex models can be obtained more efficiently than those selected by independent FlexX scores. These strategies might also improve the scoring of other docking programs, and more-effective structure-based drug design should result from these improvements.

1. INTRODUCTION

Drug design that is based on the three-dimensional (3D) structures of biopolymers, which are candidate drug targets, is commonly referred to as structure-based drug design (SBDD). Predictions of the 3D structures of protein—ligand complexes play an important role in SBDD.[1-3] Over the past 15 years, a variety of computational docking programs that predict protein—ligand complex structures have been developed.[4-7] FlexX is one of the most useful docking programs,[4] in which ligand molecules are divided into small fragments and reconstructed in the active sites of target proteins guided by physicochemical interactions. It is widely used for computational docking trials.

Because docking programs generally produce numerous model candidates for one system, it is essential to evaluate the predicted models and determine which are the most suitable. Although free energy calculations of these systems are required for this purpose, accurate calculations of free energies using molecular simulations are very time-consuming. Therefore, various scoring functions that approximately estimate the binding free energies of protein—ligand systems using simple functions without molecular simulations have been developed.[4-13] Scoring functions can be classified into three groups: empirical, knowledge-based, and force-field-based scoring functions.[3,14] Empirical scoring functions are fit to reproduce experimental data, such as experimentally obtained binding energies and conformations, as a sum of several parametrized functions and are the most widely employed. Knowledge-based scoring functions are derived from experimental structures and are represented by relatively

simple atomic interaction-pair potentials. Force-field-based scoring functions are derived from molecular mechanics force-fields and are represented by physicochemical-interaction terms, such as van der Waals potentials and Coulombic interactions. These scoring functions rank the complex structure candidates, and the most highly ranked models are adopted.

As well as their role in model selection, scoring functions have three essential functions in the computational docking process.[1,15] First, during the steps of model construction, the scores of the docking models that are under construction are calculated using scoring functions. The obtained values are then utilized in the next construction step. Thus, scoring functions are required not only in the selection of constructed models but also in the model-construction steps themselves. Second, one or a few predicted models are selected from a large number of model candidates using scoring functions (as mentioned above). Third, in virtual screening trials, in which numerous ligands are docked into one target protein, scoring functions can identify those that potentially represent favorable drug candidates. These three roles are all important for SBDD.

Scoring functions play significant roles in SBDD, and various functions have been proposed. However, building a scoring function that can make use of every protein—ligand system remains the "final frontier" of computational docking studies.[16] Every existing scoring function has specific advantages and disadvantages, and there is no de facto standard. Although this might suggest that different scoring functions are required for different protein—ligand systems, discussions about which scoring function is most suitable for a particular system are often difficult when the complex structures are not experimentally observed. Recently, the concept of a consensus score was proposed for model selection in computational docking and virtual screening

* Corresponding author phone: +81 76 431 8218; fax: +81 76 431 8208; e-mail: AKIFUMI_ODA@toyama-chemical.co.jp.
† Toyama Chemical Co., Ltd.
‡ Kitasato University.

COMPARISON OF CONSENSUS SCORING STRATEGIES

J. Chem. Inf. Model., Vol. 46, No. 1, 2006 **383**

Table 1. Consensus Scores

threshold	consensus scoring method according to ref 19[a]	
	rank-by-number	rank-by-rank
number-by	number-by-number	number-by-rank
rank-by	rank-by-number	rank-by-rank
percent-by	percent-by-number	percent-by-rank

[a] Note that the names of the consensus scoring strategies in the current study differ from those given in ref 19.

for model selections with three types of selection criteria: in the first, models with consensus score values that are less than or equal to $x_{threshold}$ are selected; in the second, the models are ranked according to the consensus scores and the top $y_{threshold}$ models are selected; in the third, the top $z_{threshold}$ percent candidates are selected. No systematic studies of these three criteria have been reported previously. Hence, in the present study, we compared these criteria for both rank-by-number and rank-by-rank consensus scores. A total of six types of average-based consensus score were compared in terms of their effectiveness for model selection. The first criterion, under which the consensus score values themselves are used for model selection, was denoted using the prefix "number-by-" in this study, whereas the second and third criteria were indicated by the prefixes "rank-by-" and "percent-by-", respectively. Because names such as number-by-rank-by-number (referring to the rank-by-number consensus score with the number-by criterion) were considered to be too long, we chose to describe this type of consensus scoring strategy as number-by-number. In the same way, the terms number-by-rank, rank-by-number, rank-by-rank, percent-by-number, and percent-by-rank were used. Note that the meanings of rank-by-number and rank-by-rank in this study differ from those employed in ref 19. The AASS corresponded to the number-by-number, rank-by-number, and percent-by-number approaches.[30] The six types of average-based consensus score are shown in Table 1. In this study, autoscaled scores are used for the by-number strategies. Optimizations of the thresholds, $x_{threshold}$, $y_{threshold}$, and $z_{threshold}$, were also carried out.

By contrast, the rank-by-vote approach uses a majority-vote-based consensus score. According to this strategy, if the score value of a model meets the standard that is set for a vote, the model is awarded one vote. This procedure is repeated for all of the scoring functions that are included in the consensus score, and the models that have many votes are eventually selected. Some previous studies[17-20] employed the rule that "models whose scores are within the top $z_{threshold}$ percent win one vote", and the rule that "models whose scores are less than, or equal to, $x_{threshold}$ win one vote" was used in CScore.[21] In addition to these criteria, the idea that

the "top $y_{threshold}$ models obtain one vote" could also be applied. No previous comparative studies have examined these three standards, and they have all been referred to using the same term—rank-by-vote. In the current study, we proposed different names for these three approaches: the majority-vote-based consensus scoring strategy that awarded votes to "models whose scores are within the top $z_{threshold}$ percent" was referred to as vote-by-percent, the strategy that awarded votes to "models whose scores are less than, or equal to, $x_{threshold}$" was referred to as vote-by-number, and the consensus scoring strategy in which the "top $y_{threshold}$ models" were awarded votes was referred to as vote-by-rank. Thus, the rank-by-vote strategy described in ref 19 was referred to as the vote-by-percent strategy in the current study, and the CScore strategy was referred to as the vote-by-number strategy. In the present study, these three types of majority-vote-based consensus scores were investigated together with six types of average-based consensus score. In contrast to the average-based consensus scores, in which only one threshold is used, the vote-by strategies require two types of threshold: one for the voting standard and another for the number of votes. For example, in the vote-by-percent strategy, the model candidate with a score within the top $z_{threshold}$ percent obtains one vote for one scoring function (the $z_{threshold}$ is the threshold for the voting standard), and a candidate with a total number of votes for all of the scoring functions that is greater than, or equal to, the $w_{threshold}$ will eventually be selected (the $w_{threshold}$ is the threshold for the number of votes). Figure 3 illustrates the vote-by strategy. Both thresholds of the vote-by strategies were optimized in our study.

2.5. Evaluations of Consensus Scores. To evaluate the consensus scores, we initially determined the thresholds that enabled reasonable solutions to be obtained for all of the systems without exception. Using these thresholds, the number of model candidates could be reduced, and we investigated how many remained in each candidate group. For example, in the number-by strategies, the $x_{threshold}$ values were optimized in order to reduce the number of models as much as possible while at least one reasonable model remained in the filtered sets of candidates, the consensus score values of which were less than, or equal to, the $x_{threshold}$ for all 220 protein−ligand systems without exceptions. Similarly, the $y_{threshold}$ and $z_{threshold}$ values were optimized for the rank-by and percent-by strategies, respectively. These thresholds are illustrated in Figure 4. As shown in the figure, the model with the best score from all of the reasonable models was identified, and its score value, rank, and percentage were investigated. These values were obtained for all 220 protein−ligand systems, and the highest values

Model	A score	B score	C score		Number of votes		Elected or excluded
1	0.525	0.564	0.671		0		excluded
2	0.434	0.000	0.911		2		elected
3	0.000	0.232	0.000	vote $\xrightarrow{}$ $x_{threshold} = 0.5$	3	selection $\xrightarrow{}$ $w_{threshold} = 2$	elected
4	1.000	1.000	1.000		0		excluded
5	0.566	0.777	0.620		0		excluded
6	0.303	0.412	0.405		3		elected

Figure 3. Vote-by strategies.

Model	RMSD	Score
A	3.442	0.100
B	5.260	0.200
C	1.952	0.300
D	0.085	0.400
E	6.435	0.500

$x_{threshold} = 0.300$

$y_{threshold} = 3$

$z_{threshold} = 60\%$

Figure 4. Thresholds. The solution with a RMSD $\le$ 2.0 Å was selected as a focus, and its score value ($x_{threshold}$), rank order ($y_{threshold}$), and top % ($z_{threshold}$) were adopted as the thresholds.

among the 220 scores, ranks, and percentages were set as the $x_{threshold}$, $y_{threshold}$, and $z_{threshold}$ values, respectively. Using these thresholds, at least one reasonable model could be obtained for all 220 systems. In contrast to average-based consensus scores, not only $x_{threshold}$, $y_{threshold}$, $z_{threshold}$ but also $w_{threshold}$ need to be optimized for majority-vote-based strategies. When the thresholds for voting standards ($x_{threshold}$, $y_{threshold}$, or $z_{threshold}$) were defined, each model obtained a certain number of votes by vote-by strategies. For example, the situation in which model 1 wins three votes, model 2 wins one vote, and model 3 wins five votes and only models 1 and 2 are "reasonable models", is considered. For this situation, the appropriate $w_{threshold}$ is three, because when the $w_{threshold}$ is greater than three, no reasonable models are selected. When $w_{threshold}$ is less than three, it is possible to select reasonable models, but a higher threshold is more appropriate. In this way, an appropriate $w_{threshold}$ was investigated for each vote-by strategy and each $x_{threshold}$, $y_{threshold}$, and $z_{threshold}$. The ratio of the remaining models to all of the candidates was referred to as the compression ratio and was calculated using the following formula:

$$p_{compress} = \sum_{220\,systems} n_{remain} \Big/ \sum_{220\,systems} n_{all} \qquad (3)$$

Here, $p_{compress}$ is the compression ratio, n_{remain} is the number of remaining models using the threshold, and n_{all} is the total number of models. The compression ratio was used as an indicator of the ability of each consensus score strategy.

Using conditions under which reasonable models could be obtained for all 220 systems without exception, the compression ratio was generally relatively large (that is, not well-compressed). Therefore, the numbers of protein–ligand systems for which reasonable models could be selected using several predefined threshold values were also elucidated, to investigate the scenario in which reasonable models could be obtained, not for all 220 systems, but for the majority of the systems, with few exceptions. Using these predefined thresholds, a good compression ratio was expected at the expense of the accuracy of the modeling of a few of the protein–ligand systems. For this investigation, not only the compression ratio but also the ratio of accurate modeling (that is, the ratio of accurately modeled systems to the total number of systems) was evaluated using the following formula:

$$p_{accurate} = n_{accurate} / n_{all} \qquad (4)$$

Here, $p_{accurate}$ is the ratio of accurate modeling, $n_{accurate}$ is the number of protein–ligand systems in which reasonable models can be obtained using the predefined threshold, and n_{all} is the total number of systems (in this study, $n_{all} = 220$

for the complete test set, $n_{all} = 57$ for those complexes with high-affinity ligands only, and $n_{all} = 122$ for groups including complexes with $n \ge 250$). The tradeoff between accuracy and efficiency was investigated using the $p_{compress}$ and $p_{accurate}$ values for each consensus score.

In the current study, the combinations of scoring functions with the best $p_{accurate}$ and $p_{compress}$ values were investigated for each strategy and threshold. However, the computational cost of evaluating all 511 combinations would have been extremely high. Therefore, from a practical standpoint, the results produced using all nine functions could be discussed for the simple consensus scoring methods without exploring the combinations; thus, the computational cost that was involved in searching for the best combination was reduced. To further simplify the procedure, we also investigated the consensus scores using only five of the scoring functions: FlexX score, GOLD score, PMF score, DOCK score, and ChemScore. These functions were included in the CScore module of the SYBYL 6.9 program. Because all of these scores could be calculated simultaneously in a single CScore trial, using the consensus scores with these five functions was the simplest method for our study. Not only exhaustive investigations of 511 combinations but also tests of these simplified consensus scores were carried out in this study.

Consensus scores for the experimental structures of protein–ligand complexes were also calculated in order to consider the wider applications of the consensus scoring strategies. As the experimental structures were regarded as the "correct" answers, the consensus scores for these structures were expected to reflect the ability of the scoring strategies in applications with real complexes. The number-by-number strategy was used for this purpose, because it was one of the most useful consensus scoring strategies and could easily be compared among different protein–ligand complex systems. The models that were calculated by FlexX were used as parent populations for the consensus scoring. The experimental structure was added to n models obtained by FlexX, so the parent population for the consensus scoring included $n + 1$ models.

The ranks and percentages of models might be biased by differences of the numbers of model candidates between test complexes. For example, although we considered that the "top three models of six candidates" means the same as the "top three models of 500 candidates" in rank-by strategies, the latter three models appeared to be more highly selected than the former models. Because the flexibilities of ligands and the sizes of active sites were very different from each other for 220 test systems, it is not a practical setting that FlexX generates completely the same number of candidates for all systems. Thus, we extracted the 122 complex systems that have more than or equal to 250 candidates, and we compared the abilities of consensus scoring strategies for these 122 systems. By using this test set, the bias caused by differences of the numbers of candidates was reduced, and the dependencies of abilities of consensus scores on n were investigated.

3. RESULTS AND DISCUSSION

3.1. Selection of Reasonable Models for the 220 Protein–Ligand Systems. To determine the thresholds under which at least one reasonable model could be selected

COMPARISON OF CONSENSUS SCORING STRATEGIES

J. Chem. Inf. Model., Vol. 46, No. 1, 2006 **385**

Table 2. Comparison of the Nine Types of Consensus Score

(a) Number-by, Rank-by, and Percent-by Strategies

	threshold	$p_{compress}$ of best combination
number-by-number	$x_{threshold} = 0.452$	0.399
number-by-rank	$x_{threshold} = 151$	0.519
rank-by-number	$y_{threshold} = 151$	0.519
rank-by-rank	$y_{threshold} = 151$	0.519
percent-by-number	$z_{threshold} = 62.0\%$	0.617
percent-by-rank	$z_{threshold} = 62.4\%$	0.623

(b) Vote-by Strategies

	threshold for vote or not	threshold for number of votes ($w_{threshold}$)/number of functions included in best combination	$p_{compress}$ of best combination
vote-by-number	$x_{threshold} = 0.5$	5 votes/7 voters	0.371
vote-by-rank	$y_{threshold} = 150$	2 votes/3 voters	0.510
vote-by-percent	$z_{threshold} = 70\%$	3 votes/3 voters	0.502

for all 220 protein–ligand systems without exception, nine consensus scoring strategies and 511 combinations of scoring functions (that is, a total of 4599 types of consensus scores) were examined. The compression ratios that were produced using these thresholds were also investigated, that is, the compression ratios under the condition $p_{accurate} = 1.0$. Although all 511 combinations for each strategy were systematically evaluated, the combination that gave the best (smallest) compression ratio is focused on here (this was defined as the "best combination"). The best combinations are summarized in Table 2, in which both the thresholds and the compression ratios are described. For example, in the case of the number-by-number strategy, the best combination included the FlexX score, GOLD score, PMF score, DOCK score, PLP, and ScreenScore (as shown in Table S2 of the Supporting Information, and see below for discussion). After the autoscaling of the six scores, the average value—that is, the number-by-number consensus score—was calculated for each model. The $x_{threshold}$ of this combination for the number-by-number strategy was 0.452, which meant that at least one reasonable model was produced for each of the 220 protein–ligand systems by selecting the models with number-by-number consensus scores that were less than, or equal to, 0.452 (Table 2a). Using this strategy and threshold value, the number of model candidates to be explored was expected to be reduced to about 40% of the current size, because the compression ratio was equal to 0.399 (Table 2a). The vote-by-number strategy gave the best compression ratio for selecting suitable models for all 220 systems (Table 2a). In the best combination of vote-by-number, when the auto scaled value of one score was less than, or equal to, 0.5 for one model, it obtained one vote because $x_{threshold} = 0.5$ (Table 2b). When one model received more than, or equal to, five votes from the seven voters, it was selected because $w_{threshold} = 5$. Although the vote-by-percent strategy has been discussed for use in virtual screening trials in some previous reports,[17–20] our result suggests that a vote-by-number strategy, such as CScore, is a more appropriate majority-vote-based consensus score system for model selections. The second-best was the number-by-number strategy, which gave the best compression ratio of all average-based consensus scores.

The combinations that gave the top 10 compression ratios for all nine consensus scoring strategies are shown in Table S2 of the Supporting Information. As shown in Table S2a

and g, for number-by-number and vote-by-number strategies, which had compression ratios that were superior to those of the other approaches, FlexX score, PLP, and ScreenScore were included in many of the top 10 consensus scores. It has been reported previously that these scoring functions work well when they are used independently.[20,22] Our results suggest that they are appropriate not only for independent scoring but also for consensus scoring. However, although it was reported that the GOLD score and DOCK score implemented in the CScore module were less successful when used alone, they appear high up in the lists of successful consensus scoring methods shown in Table S2a and g (Supporting Information). These force-field-based scoring functions are based on a different concept from that used to develop the empirical and knowledge-based scoring functions, and they seem to compensate for the shortcomings of the empirical scoring functions (such as FlexX score, PLP, and ScreenScore). By contrast, ChemScore and X-Score appeared in relatively few of the top 10 combinations. These are both empirical scoring functions, which are similar in form to FlexX score and ScreenScore, so mutual complementarity between them might not be expected. Therefore, ChemScore and X-Score do not seem to be appropriate for use together with FlexX score or ScreenScore, although they might perform well independently.

3.2. Efficient Selection of Reasonable Models with Some Exceptions. The investigations of the abilities of consensus scores discussed in the previous section were carried out under conditions in which reasonable models were selected for all protein–ligand test sets without exception (that is, $p_{accurate} = 1.0$). In this scenario, the values of the compression ratios tend to become overly large, because the thresholds are set to a high value if there are only a few systems for which reasonable models are difficult to search; this impairs the effectiveness of the consensus scores, even if they work well for most protein–ligand systems. In fact, the compression ratio was around 40%, even for the vote-by-number strategy, which was the best approach for this purpose. Therefore, the tradeoff between the $p_{accurate}$ and $p_{compress}$ values was investigated using several threshold values in order to make effective selections of reasonable models for as many protein–ligand systems as possible. In the current section, we discuss number-by, rank-by, and percent-by strategies, in which only one threshold is used. In the following section, consensus scores that are easy to use without exploring the combinations of scoring functions are described, and the tradeoff between the $p_{accurate}$ and $p_{compress}$ values is also discussed for vote-by strategies, in which two types of thresholds (that is, thresholds for the voting standard and for the number of votes) are required.

Figure S1 of the Supporting Information shows the ratios of the accurate modeling ($p_{accurate}$) for the number-by, rank-by, and percent-by strategies using several threshold values. In this section, the combination with the best $p_{accurate}$ among all of the 511 combinations is defined as the "best combination". The best combination for each threshold is summarized in Figure S1. Note that the meaning of best combination in this section is different from that in the previous section (where the term referred to the best compression ratio, rather than the best ratio of accurate modeling). The best combination of scoring functions for each threshold is shown in Table S3 of the Supporting Information together with the $p_{accurate}$
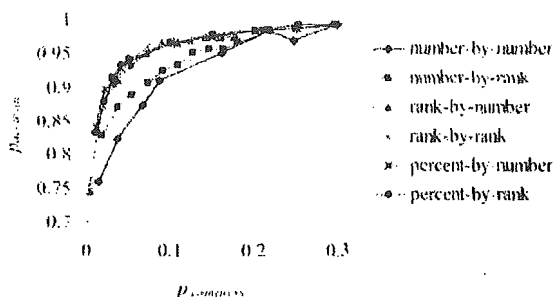
Figure 5. Compression ratios versus the ratios of accurate modeling when several threshold values were used.

Table 3. Combinations of Scores That Gave the Top 10 Ratios of Accurate Modeling When $x_{threshold}$ and $y_{threshold}$ Were Small

(a) Number-by-Number ($x_{threshold}$ = 0.2)

			scoring functions							
F	G	PM	DO	C	Dr	PL	S	X	$p_{accurate}$	$p_{compress}$
1					√	√	√			0.9091	0.0882
2						√	√			0.9091	0.0957
3							√			0.9091	0.1179
4						√				0.9045	0.1183
5					√					0.9045	0.1186
6					√	√				0.9000	0.1050
7					√		√			0.8955	0.0909
8						√	√	√		0.8818	0.0706
9	√					√	√			0.8818	0.0856
10									√	0.8818	0.1099

(b) Rank-by-Number ($y_{threshold}$ = 10)

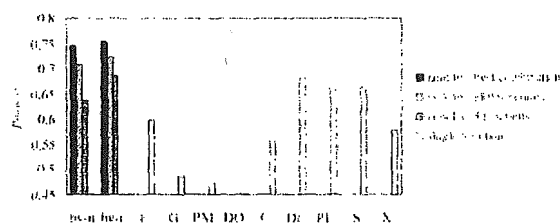|			scoring functions							
F	G	PM	DO	C	Dr	PL	S	X	$p_{accurate}$	$p_{compress}$
1				√		√				0.9091	0.0372
2	√	√		√	√	√				0.9045	0.0371
2	√	√		√				√		0.9045	0.0371
2				√	√	√	√			0.9045	0.0371
2	√	√		√	√	√	√			0.9045	0.0371
2	√	√			√	√	√			0.9045	0.0371
2	√	√		√		√	√			0.9045	0.0371
2		√			√	√		√		0.9045	0.0371
2		√			√	√	√	√	√	0.9045	0.0371
10					√	√	√	√		0.9045	0.0372

Figure 6. Ratios of accurate modeling for rank-by-number, rank-by-rank, and nine independent scoring functions when $y_{threshold}$ = 1 was used. The "by-n" and "by-r" mean rank-by-number and rank-by-rank, respectively.

and $p_{compress}$ values. In the rank-by and percent-by strategies, the behaviors of the by-number and by-rank consensus scores depending on the thresholds were similar to one another (Figures S1c and d). This indicates that, for these strategies, the results do not depend on whether by-number or by-rank approaches are used when the same thresholds are adopted.

Although Figure S1 (Supporting Information) illustrates the dependency of the $p_{accurate}$ values on thresholds, the tradeoff between the $p_{accurate}$ and $p_{compress}$ values cannot be discussed on the basis of this figure. Thus, the relationships between these parameters are explored in Figure 5, on the basis of the $p_{accurate}$ and $p_{compress}$ values presented in Table S3 (Supporting Information). Figure 5 shows that the $p_{compress}$ values of both of the number-by strategies were much worse than those of the rank-by and percent-by strategies when they were compared at the same values of $p_{accurate}$. This suggests that the rank-by and percent-by strategies are more appropriate for the effective selection of model candidates than the number-by strategies, which differs from the result we obtained for model selection with $p_{accurate}$ = 1.0 (in which the number-by-number strategy was more appropriate). Therefore, different strategies should be adopted for different purposes, for example, "$p_{accurate}$ = 1.0 is indispensable" or "a good balance between $p_{accurate}$ and $p_{compress}$ is desired".

For the number-by-number and rank-by-number strategies, the top 10 combinations of scoring functions in terms of $p_{accurate}$ are shown in Table 3. The $x_{threshold}$ and $y_{threshold}$ values were as small as possible while maintaining $p_{accurate}$ ≥ 0.9 for the best combinations (the top 10 combinations for the other strategies are shown in Table S4 in the Supporting Information). Table 3 demonstrates that while all of the top 10 combinations (excluding the number 1 combination) included more than, or equal to, four scoring functions in the rank-by-number strategy, all of the top 10 combinations in the number-by-number strategy included less than, or equal to, three functions; in particular, four of the top 10 combinations included only one scoring function and, thus, were not consensus scoring approaches. This suggests that the concept of consensus scoring does not work well for the number-by-number strategy, in contrast to the rank-by-number strategy, for the purpose of highly effective model selection at the expense of the accurate modeling of a few exceptional systems.

For the fast selection of model candidates, only the top model is frequently investigated. This is the situation with model selection. In Figure 6, the $p_{accurate}$ values of rank-by-rank and rank-by-number with $y_{threshold}$ = 1 are illustrated. The results of independent scoring by nine

scores are also shown in the figure. For rank-by-number and rank-by-rank, not only the results of the best combination but also those of combinations including all nine functions and five CScore functions, that is, FlexX score, GOLD score, PMF score, DOCK score, and ChemScore, are illustrated. As shown in this figure, $p_{accurate}$ is around 0.75 for two types of rank-by strategies, and they were superior to all nine independent scoring functions. The $p_{accurate}$ values of the best combinations of two rank-by strategies were similar to one another, and this result was consistent with those of $y_{threshold}$ > 1. However, for the combinations including all nine functions and five CScore functions, the results of rank-by-rank were better than those of rank-by-number. In fact, for rank-by-number, although the best combination and the combination including all nine functions were superior to all independent scoring functions, the result of the combination including five CScore functions was worse than the results of the independent scoring of DrugScore, PLP, and ScreenScore. On the other hand, for rank-by-rank, $p_{accurate}$ values obtained by not only best combination and the combination including all nine functions but also the combination including five CScore functions were better than
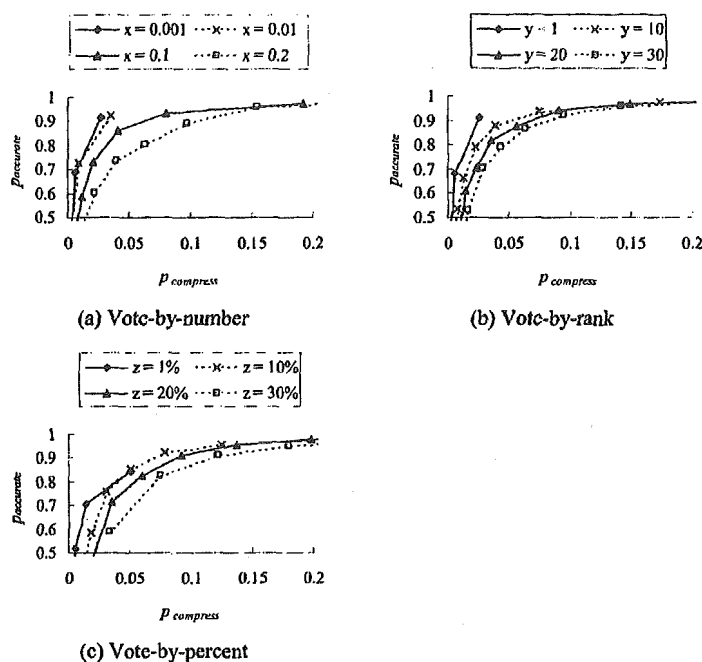
COMPARISON OF CONSENSUS SCORING STRATEGIES

J. Chem. Inf. Model., Vol. 46, No. 1, 2006 **387**

(a) Vote-by-number

(b) Vote-by-rank

(c) Vote-by-percent

Figure 7. Results obtained using all nine functions depending on the thresholds.

those of all independent scoring functions. These results suggest that rank-by-rank is more robust in terms of combination of scoring functions than rank-by-number when $y_{threshold} = 1$.

3.3. Easy-to-Use Consensus Scores without Exploring the Combinations of Scoring Functions. For practical applications of consensus scores, exploring the best combination of scoring functions is highly expensive in terms of computational cost. Although thorough investigations of the combinations are desirable to allow detailed discussions, the computational cost is frequently as important as accuracy in drug design trials. Thus, in this section, consensus scores that are accurate and easy to use without the need to exhaustively explore the best combinations are discussed. One of these systems includes all nine functions and another includes the five CScore functions. In particular, the latter can be calculated using only a single CScore trial and is, therefore, the most simple consensus score considered in this study.

A comparison of the $p_{compress}$ values of the nine consensus scoring strategies under the condition $p_{accurate} = 1.0$ was carried out. The results are shown in Figure S2 of the Supporting Information, and they were similar to those of the best combinations mentioned in Section 3.1. These results support the finding that the two strategies, that is, vote-by-number and number-by-number, are appropriate for model selection of all systems without exception. Furthermore, for these two strategies, the $p_{compress}$ values calculated by both the combinations including all nine functions and five CScore functions were not much worse than those of the best combination. It suggests that these strategies also have advantages in terms of robustness of the combination of scoring functions.

The $p_{accurate}$ values obtained by the number-by, rank-by, and percent-by strategies depending on the thresholds are

shown in Figure S3 of the Supporting Information (this is similar to the investigations mentioned in Section 3.2). The $p_{accurate}$ values of the rank-by and percent-by strategies using all nine functions or the five CScore functions were not much worse than those of the best combinations, in contrast to the number-by strategies. This means that, for the rank-by and percent-by strategies, the combinations using all nine functions or the five CScore functions, without exploring the best combinations, were effective in saving computational costs. Although the $p_{accurate}$ values of the rank-by-number and rank-by-rank strategies were similar to each other by using the best combinations, the rank-by-number strategy gave better $p_{accurate}$ values than the rank-by-rank strategy using all nine functions and the five CScore functions when $y_{threshold}$ was between 5 and 20. These results were different from those of $y_{threshold} = 1$, mentioned in Section 3.2, in which rank-by-rank was more robust than rank-by-number in terms of combinations of scoring functions.

In addition to the number-by, rank-by, and percent-by consensus scoring strategies, the dependencies of the $p_{accurate}$ and $p_{compress}$ values on the $w_{threshold}$ (the thresholds for the number of votes) were discussed for the vote-by strategies with the combinations including all nine functions. These discussions are useful for model selection by vote-by strategies with good balances between the $p_{accurate}$ and $p_{compress}$ values.

Figure 7 compares the balances between the $p_{accurate}$ and $p_{compress}$ values depending on the thresholds for the vote-by consensus scores including all nine functions, to determine which of the threshold values for the voting standard ($x_{threshold}$, $y_{threshold}$, and $z_{threshold}$) are desirable. For each of the thresholds for the voting standard, the threshold values for the number of votes ($w_{threshold}$) were set from 1 to 9. Figure 7 illustrates the results depending on both of the two thresholds (the complete data are shown in Table S5 in the Supporting

388 *J. Chem. Inf. Model.*, Vol. 46, No. 1, 2006

ODA ET AL.

Information). As shown in the figure, the tradeoff between the $p_{accurate}$ and $p_{compress}$ values obtained by the vote-by-percent strategy was the worst of all the vote-by approaches. For the vote-by-number and vote-by-rank strategies, smaller $x_{threshold}$ and $y_{threshold}$ values gave smaller (better) compression ratios when similar $p_{accurate}$ values were obtained. For example, if $p_{accurate} \geq 0.9$ was required, the best $p_{compress}$ for the vote-by-rank was around 2.6% and was obtained using $y_{threshold} = 1$ and $w_{threshold} = 1$. These thresholds mean that each scoring function votes for only the top model, and the models that win one or more votes are selected. The same results can be obtained using the vote-by-number strategy when the $x_{threshold}$ is small enough, because the autoscaled score value of the top model is always 0. The compression ratio produced by this threshold was the best of all the consensus scores, including not only the vote-by but also the number-by, rank-by, and percent-by strategies for the condition of $p_{accurate} \geq 0.9$. By contrast, it is difficult to adjust the parameters for vote-by strategies because of the requirement for two thresholds; thus, other strategies, such as the rank-by-number approach, might be more suitable for easy-to-use consensus scoring.

3.4. Consensus Scores for Complexes with High Binding Affinities. Sections 3.1−3.3 discussed our investigations of all 220 protein−ligand complex systems. The current section describes the analysis of the 57 complexes with high affinities that are marked in bold in Table S1 of the Supporting Information.

Figure S4 in the Supporting Information shows the compression ratios produced by the nine consensus scoring strategies under the condition that gave reasonable models for all 57 systems without exception (that is, $p_{accurate} = 1.0$). The results that were obtained using the best combinations, all nine functions, and five CScore functions are illustrated. The number-by-number and vote-by-number strategies were appropriate for the 57 complexes, which was consistent with the results obtained for all 220 test complexes. These findings suggest that the results for the 220 complexes will also be useful for designing high-affinity ligands, which could play important roles in drug design trials.

In addition to the calculations under the condition $p_{accurate} = 1.0$, the tradeoffs between the $p_{accurate}$ and $p_{compress}$ values were investigated for the 57 complexes with high-affinity ligands using several threshold values. The results are shown in Figure 8. The balances between the $p_{accurate}$ and $p_{compress}$ values obtained using the best combinations for number-by, rank-by, and percent-by strategies are shown in Figure 8a (similar to those presented in Figure 5). Figure 8b, which corresponds to Figure 7b, illustrates the tradeoffs between the $p_{accurate}$ and $p_{compress}$ values for the vote-by-rank strategy using all nine functions. According to these calculations, the results of the rank-by and percent-by strategies were better than those of the number-by strategies. For the vote-by-rank, $y_{threshold} = 1$ was the best voting standard threshold. These findings were consistent with the earlier discussions of all 220 complexes.

3.5. Consensus Scores for Experimental Complex Structures. In this section, we discuss the abilities of consensus scores for use with experimentally observed complex structures. Calculating the consensus score values for various experimental structures (that is, the "correct answer") revealed how small scores could be obtained for
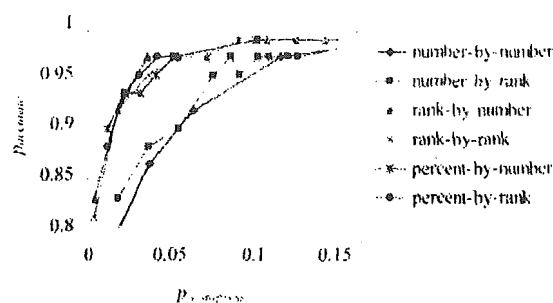
(a) Number-by-, rank-by- and percent-by- strategies.
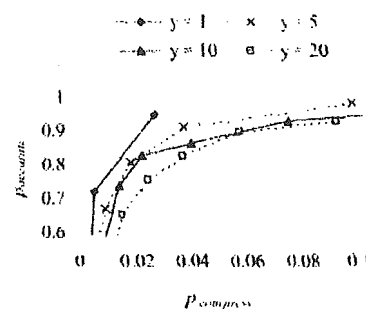
(b) Vote-by-rank strategies (using all nine functions).

Figure 8. Compression ratios versus the ratios of accurate modeling for the test set including complexes with high-affinity ligands.

Table 4. Thresholds of Number-by-Number Consensus Scores for the Experimental Complex Structures

	$x_{threshold}$	
	all complexes	complexes with high-affinity ligands
best combination	0.386	0.271
using all nine functions	0.689	0.513
using the five CScore functions	0.785	0.676

compounds that experimentally docked into target proteins. In this study, the number-by-number strategy was used for this purpose.

Table 4 and Figure 9 present the computational results for the experimental structures produced using the number-by-number strategy. The thresholds under which reasonable models can be selected (that is, the worst value for the number-by-number strategy) are described for all 220 experimental structures and for the 57 complexes with high-affinity ligands. The score value for the test set including complexes with high-affinity ligands was better than that for the whole test set. This suggests that, when the ligand with a better score is selected, a higher binding affinity can be expected using the number-by-number strategy. In addition, the frequency distributions of the consensus score values of the experimental structures are illustrated in Figure 9. For practical reasons, we used the combinations including all nine scoring functions and the five CScore functions. For the high-affinity ligands, the number-by-number values were less than 0.3 for 90% of the 57 experimental structures produced using all nine functions.

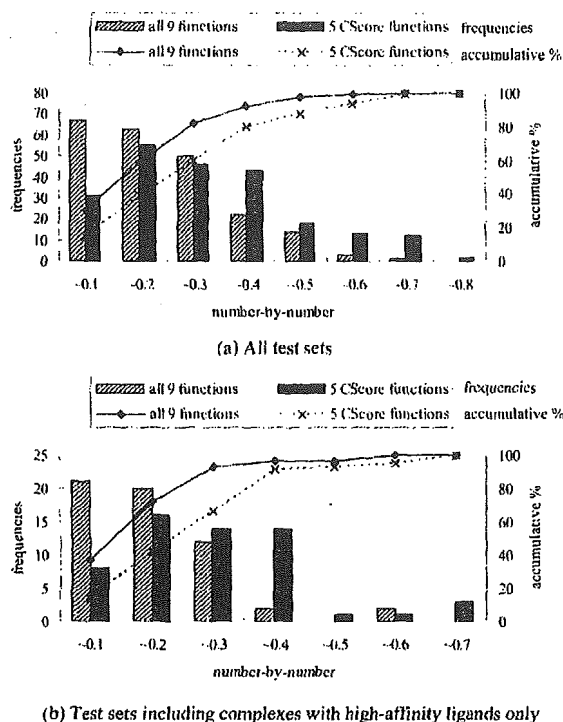**3.6. Consensus Scores for Complexes with More Than or Equal to 250 Candidates Generated by Computational**

Figure 9. Frequency distributions and accumulative percentages of number-by-number values for the experimental complex structures.

Docking. In computational docking, the numbers of generated candidates depend on flexibilities of ligands and sizes of active sites of target proteins. As shown in Table S1 of the Supporting Information, in this study, the number of candidates was between 6 and 500. To find the consensus scoring strategies which can be widely used, various types of test complexes, for which various numbers of candidates were generated, need to be investigated as mentioned in Sections 3.1–3.5. However, the large differences of the numbers of candidates possibly cause bias for the ranks and percentages. In this section, the comparisons of consensus scores for only 122 test complexes which have more than or equal to 250 model candidates generated by FlexX were carried out in order to reduce the bias.

Figure 10 shows a comparison of the $p_{compress}$ values of the nine consensus scoring strategies under the condition $p_{accurate} = 1.0$. As shown in this figure, the results of the best combinations of percent-by strategies were as good as or

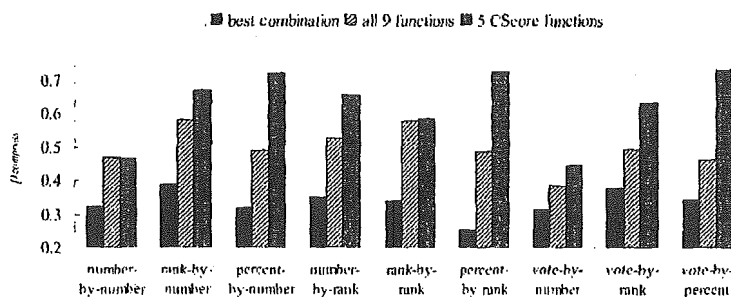better than those of vote-by-number and number-by-number for complexes with $n \geq 250$ although both percent-by strategies gave much worse $p_{compress}$ values than vote-by-number and number-by-number in a test for all 220 complexes, as mentioned in Section 3.1. On the other hand, for the combinations including all nine functions and five CScore functions, results of the percent-by strategies were worse than those of vote-by-number and number-by-number. It indicates that vote-by-number and number-by-number are more appropriate in terms of the robustness of combinations of scoring functions not only for complete test sets but also for complexes with $n \geq 250$. For rank-by-number and vote-by-rank, although $p_{compress}$ values were highly improved in comparison with the results shown in Table 2 in Section 3.1, they remain worse than those of vote-by-number and number-by-number. These results suggest that although ranks and percentages are affected by the number of model candidates (n), vote-by-number and number-by-number are still appropriate under the condition $p_{accurate} = 1.0$ because they workwell regardless of whether n is large or not and they are robust in terms of combinations of scoring functions.

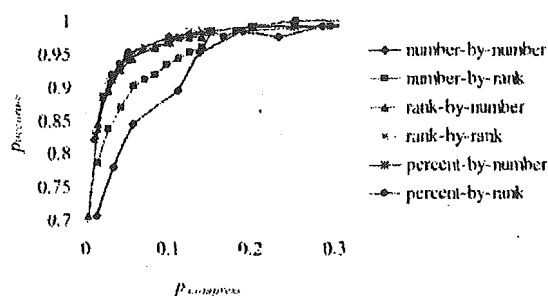In Figure 11, the tradeoffs between $p_{accurate}$ and $p_{compress}$

Figure 11. Compression ratios versus the ratios of accurate modeling for the test set including complexes with $n \geq 250$.

for complexes with $n \geq 250$ are illustrated. As shown in this figure, the results were similar to those in Figure 5 in Section 3.2. Thus, to achieve a good balance between $p_{accurate}$ and $p_{compress}$, rank-by and percent-by strategies were appropriate regardless of the number of candidates. Both the results of studies with conditions under $p_{accurate} = 1.0$ and those of tradeoff studies for complexes with $n \geq 250$ were consistent with those for all 220 complexes, and they indicate that the results of this study are useful for various systems regardless of the numbers of generated candidates within the limitation of "Num.Answers" = 500.

Figure 10. Compression ratios for the focused test set including complexes with $n \geq 250$.

390　*J. Chem. Inf. Model.,* Vol. 46, No. 1, 2006

ODA ET AL.

4. CONCLUSIONS

In this study, we investigated the abilities of consensus scores to evaluate docking models constructed using FlexX. We systematically named the nine types of consensus scoring strategies that have been independently proposed (and previously confused with one another) and compared their performance. All 511 types of combinations including the nine scoring functions were investigated for each of the strategies. Consequently, we found that the number-by-number and vote-by-number strategies were appropriate for use in model selection in all of the systems without exception, and the rank-by-number and percent-by-number strategies were useful for model selection with a good tradeoff between accuracy and efficiency. Considering the scoring functions that were utilized, PLP and DrugScore, which were effective for model selection in single scoring systems, were also appropriate for use in consensus scores. In addition, GOLD score and DOCK score, which were not effective in single scoring systems, were also useful in consensus scoring approaches, as they seemed to compensate for the shortcomings of the other scoring functions. Optimizing the combinations of scoring functions is expensive in terms of computational costs, so consensus scores including all nine functions or the five CScore functions without the need for prior optimizations are particularly useful in practice. Although the vote-by strategies were effective for model selection, they require two types of threshold to be defined, and it is difficult to control the numbers of finally selected models. Thus, we recommend the number-by-number strategy (for all systems without exception) or the rank-by-number strategy (for a good balance between accuracy and efficiency), both of which have abilities similar to those of the vote-by strategies.

In previous papers,[17–20,31,32] the vote-by-percent strategy has been used as a representative of vote-by approaches (denoted as rank-by-vote in refs 17 and 20, and the "intersection approach" in ref 31) for compound selection in virtual screening trials. Some of these studies reported that vote-by strategies did not work well in comparison to other approaches or single scoring.[17,19,31] However, as shown in the current study, the vote-by-percent approach is the least appropriate of all the vote-by strategies, at least for model selection, and other techniques should be used for discussions of the abilities of vote-by strategies. Although the scoring scheme of compound selection in virtual screening is different from that of model selection in computational docking, the vote-by-percent strategy might not be appropriate for virtual screening, as is the case in model selections. In our study, the abilities of consensus scores were discussed only for use in model selection in computational docking, and we intend to systematically investigate the performance of consensus scores in virtual screening in future trials.

In the current study, model selection for all systems without exception, and for most systems with few exceptions, were investigated. This is the first comparison of the use of consensus scores in these two situations, and we found that different types of strategies are required. If vote-by strategies are inadequate because of practical problems, number-by-number and rank-by-number approaches should be suitable for the former and latter situations, respectively. We previously described the by-number-type consensus score AASS[30] and argued that both the number-by-AASS and rank-by-

AASS approaches should be used according to the demands of the specific situation.

Although this study focused on the selection of docking models produced by FlexX, we expect our results to be relevant to other computational docking programs. We intend to investigate these aspects further in a future study.

REFERENCES AND NOTES

(1) Kroemer, R. T. Molecular Modelling Probes: Docking and Scoring. *Biochem. Soc. Trans.* 2003, *31*, 980–984.

(2) Leach, A. R. In *Molecular Modelling*, 2nd ed.; Pearson Education Limited: Essex, 2001; Chapter 12, pp 640–726.

(3) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications, *Nat. Rev. Drug Discovery* 2004, *3*, 935–949.

(4) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* 1996, *261*, 470–489.

(5) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* 1997, *267*, 727–748.

(6) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule–Ligand Interactions. *J. Mol. Biol.* 1982, *161*, 269–288.

(7) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* 2004, *47*, 1739–1749.

(8) Muegge, I.; Martin, Y. C. A General and Fast Scoring Function for Protein–Ligand Interactions: A Simplified Potential Approach. *J. Med. Chem.* 1999, *42*, 791–804.

(9) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical Scoring Functions: 1. The Development of a Fast Empirical Scoring Function to Estimate the Binding Affinity of Ligand in Receptor Complexes. *J. Comput.-Aided Mol. Des.* 1997, *11*, 425–445.

(10) Murray, C. W.; Auton, T. R.; Eldridge, M. D. Empirical Scoring Functions. II. The Testing of an Empirical Scoring Function for the Prediction of Ligand–Receptor Binding Affinities and the Use of Bayesian Regression to Improve the Quality of the Model. *J. Comput.-Aided Mol. Des.* 1998, *12*, 503–519.

(11) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based Scoring Function to Predict Protein–Ligand Interactions. *J. Mol. Biol.* 2000, *295*, 337–356.

(12) Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T. Molecular Recognition of the Inhibitor AG-1343 by HIV-1 Protease: Conformationally Flexible Docking by Evolutionary Programming. *Chem. Biol.* 1995, *2*, 317–324.

(13) Wang, R.; Lai, L.; Wang, S. Further Development and Validation of Empirical Scoring Functions for Structure-Based Binding Affinity Prediction. *J. Comput.-Aided Mol. Des.* 2002, *16*, 11–26.

(14) Stahl, M.; Schulz-Gasch, T. Practical Database Screening with Docking Tools. *Ernst Schering Res. Found. Workshop* 2003, *42*, 127–151.

(15) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative Evaluation of Eight Docking Tools for Docking and Virtual Screening Accuracy. *Proteins: Struct., Funct., Bioinf.* 2004, *57*, 225–242.

(16) Willis, R. C. 2001: A Dock Odyssey. *Mod. Drug Discovery* 2001, *4*, 26–28.

(17) Verdonk, M. L.; Berdini, V.; Harshorn, M. J.; Mooij, W. T. M.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual Screening Using Protein–Ligand Docking: Avoiding Artificial Enrichment. *J. Chem. Inf. Comput. Sci.* 2004, *44*, 793–806.

(18) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem.* 1999, *42*, 5100–5109.

(19) Wang, R.; Wang, S. How Does Consensus Scoring Work for Virtual Library Screening? An Idealized Computer Experiment. *J. Chem. Inf. Comput. Sci.* 2001, *41*, 1422–1426.

(20) Stahl, M.; Rarey, M. Detailed Analysis of Scoring Functions for Virtual Screening. *J. Med. Chem.* **2001**, *44*, 1035–1042.

(21) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus Scoring for Ligand/Protein Interactions. *J. Mol. Graphics Modell.* **2002**, *20*, 281–295.

(22) Wang, R.; Lu, Y.; Wang, S. Comparative Evaluation of 11 Scoring Functions for Molecular Docking. *J. Med. Chem.* **2003**, *46*, 2287–2303.

(23) Marsden, P. M.; Puvanendrampillai, D.; Mitchell, J. B. O.; Glen, R. C. Predicting Protein–Ligand Binding Affinities: A Low Scoring Game? *Org. Biomol. Chem.* **2004**, *2*, 3267–3273.

(24) Rarey, M.; Kramer, B.; Lengauer, T. Docking of Hydrophobic Ligand with Interaction-Based Matching Algorithms. *Bioinformatics* **1999**, *15*, 243–250.

(25) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FlexX Incremental Construction Algorithm for Protein–Ligand Docking. *Proteins: Struct., Funct., Genet.* **1999**, *37*, 228–241.

(26) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(27) *SYBYL 6.9*; Tripos Inc.: St. Louis, MO, 2002.

(28) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parametrization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490–519.

(29) *X-Score 1.1*; Department of Internal Medicine, University of Michigan Medical School: Ann Arbor, MI, 2003.

(30) Katsuki, M.; Chuang, V. T. G.; Nishi, K.; Kawahara, K.; Nakayama, H.; Yamaotsu, N.; Hirono, S.; Otagiri, M. Use of Photoaffinity Labeling and Site Directed Mutagenesis for Identification of Key Residue Responsible for Extraordinarily High Affinity Binding of UCN-01 in Human Alpha 1-Acid Glycoprotein. *J. Biol. Chem.* **2005**, *280*, 1384–1391.

(31) Bissantz, C.; Folkers, G.; Rognan, D. Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *J. Med. Chem.* **2000**, *43*, 4759–4767.

(32) Xiang, L.; Hodgkin, E.; Liu, Q.; Sedlock, D. Evaluation and Application of Multiple Scoring Functions for a Virtual Screening Experiment. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 333–344.

Discovery of Nonpeptidic Small-Molecule AP-1 Inhibitors: Lead Hopping Based on a Three-Dimensional Pharmacophore Model

Keiichi Tsuchida,*,† Hisaaki Chaki,‡ Tadakazu Takakura,† Hironori Kotsubo,‡ Tadashi Tanaka,† Yukihiko Aikawa,‡ Shunichi Shiozawa,#,⊥ and Shuichi Hirono§

Discovery Laboratories and Research Laboratories, Toyama Chemical Co., Ltd., 4-1 Shimookui 2-chome, Toyama 930-8508, Japan, Department of Rheumatology, Faculty of Health Science, Kobe University School of Medicine, 7-10-2 Tomogaoka, Suma-ku, Kobe 654-0142, Japan, Rheumatic Disease Division, Kobe University Hospital, 7-5-2 Kusunoki-cho, Chuo-ku, Kobe 650-0017, Japan, and School of Pharmaceutical Sciences, Kitasato University, 5-9-1 Shirokane, Minato-ku, Tokyo 108-8641, Japan

We designed and synthesized small-molecule activator protein-1 (AP-1) inhibitors based on a three-dimensional (3D) pharmacophore model that we had previously derived from a cyclic decapeptide exhibiting AP-1 inhibitory activity. New AP-1 inhibitors with a 1-thia-4-azaspiro[4.5]decane or a benzophenone scaffold, which inhibit the DNA-binding and transactivation activities of AP-1, were discovered using a "lead hopping" procedure. An additional investigation of the benzophenone analogues confirmed the reliability of the pharmacophore model, its utility to discover AP-1 inhibitors, and the potency of the benzophenone derivatives as a lead series.

Introduction

Activator protein-1 (AP-1) is a transcription factor that has a crucial role in cellular signal transduction, as it is responsible for the induction of a number of genes that are involved in cell proliferation, differentiation, and immune and inflammatory responses.[1,2] It has been implicated in various diseases, such as rheumatoid arthritis.[3,4]

AP-1 contains members of the Fos and Jun families, which form either Jun-Jun homodimers or Fos-Jun heterodimers and bind to the consensus DNA sequence 5'-TGAGTCA-3', which is known as the AP-1 binding site.[1] An investigation of the X-ray crystal structure of the basic region-leucine zipper (bZIP) domains of c-Fos and c-Jun bound to a DNA fragment containing the AP-1 binding site revealed that both domains form continuous α-helices, and the heterodimer grips the major groove of the DNA, similar to a pair of forceps.[5]

Natural products such as curcumin,[6] dihydroguaiaretic acid,[7] and an anthraquinone derivative[8] were reported to inhibit the binding of AP-1 to the AP-1 binding site. However, three-dimensional (3D) structural information about the AP-1 binding of these inhibitors that is necessary for structure-based drug design is not well-known. In a previous report,[9] we discovered a new cyclic disulfide decapeptide Ac-cyclo[Cys-Gly-Gln-Leu-Asp-Leu-Ala-Asp-Gly-Cys]-NH2 (peptide 1) that exhibits AP-1 inhibitory activity, using a de novo approach that exploited molecular modeling methods, such as molecular dynamics (MD) simulations, and docking studies based on the 3D structure of the bZIP domains derived from the X-ray structure.[5] Furthermore, we built a 3D pharmacophore model based on the chemical and structural features of peptide 1. These data were obtained from an alanine scan and structural studies involving a combination of MD simulation of the bZIP–peptide 1 complex
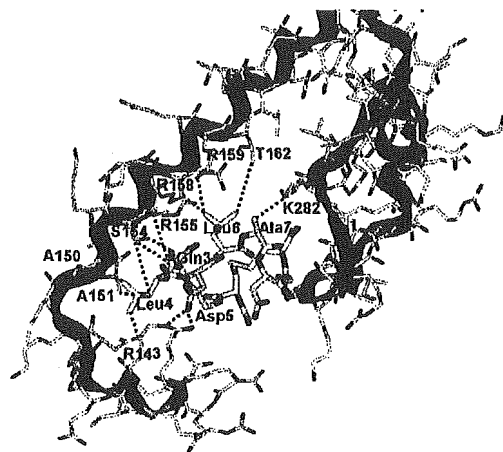

Figure 1. Binding model of peptide 1 (yellow) resulting from MD simulation. Ribbon representation of the basic domains (c-Fos, cyan; c-Jun, magenta). The residues of c-Fos and c-Jun that are involved in interactions are labeled with one-letter codes, and the residues of peptide 1 that are involved in interactions are labeled with three-letter codes. Red broken lines indicate putative hydrogen bonds and green broken lines indicate putative hydrophobic interactions. Hydrogen atoms are not shown for clarity.

with explicit water molecules (Figure 1) and NMR measurements of the peptide in water.[9]

Peptides generally have unfavorable properties for therapeutic drugs, such as poor bioavailability.[10] Several approaches have been reported to convert bioactive peptides into nonpeptidic drug candidates known as peptidomimetics.[10–12] In many cases, these peptidomimetics are peptide-like molecules, such as peptoids, and so further modifications are essential.

To avoid these problems, we designed nonpeptidic small molecules using a molecular modeling method based on a 3D pharmacophore model. This procedure could be described as "lead hopping" and has been employed as part of several recent in silico approaches.[13] There are two general computational approaches, de novo design and 3D database searching, to identify new lead candidates with desired biological activity using a pharmacophore model.[14] Although several successful

* To whom correspondence should be addressed. Phone: +81-76-431-8218. Fax: +81-76-431-8208. E-mail: keiichi_tuchida@toyama-chemical.co.jp.
† Discovery Laboratories, Toyama Chemical Co., Ltd.
‡ Research Laboratories, Toyama Chemical Co., Ltd.
# Kobe University School of Medicine.
⊥ Kobe University Hospital.
§ Kitasato University.

a   b

Figure 3. Fitting images of two scaffolds, 1-thia-4-azaspiro[4.5]decane (a) and benzophenone (b) to the trapezoid shown in Figure 2.
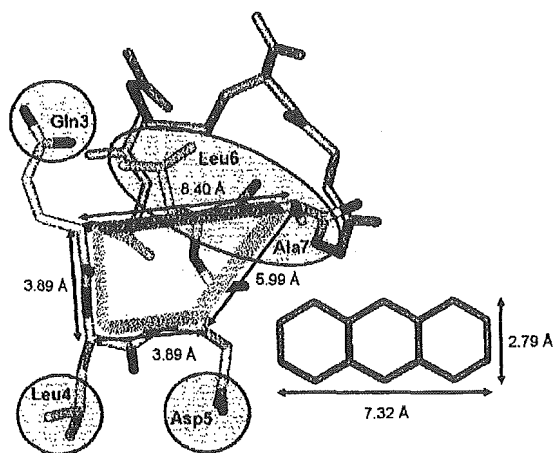

Figure 2. Pharmacophore model derived from peptide **1**. The pharmacophoric residues and the supplementary parts are colored yellow and gray, respectively. The pharmacophoric elements on peptide **1** are shown as circles and an ellipse in magenta. The main chain forms a trapezoid shape, the dimensions of which are shown by the arrows. An anthracene molecule of the same scale is depicted in green for a comparison of the sizes. Hydrogen atoms are not shown for clarity.

Figure 4. Illustration of the lead hopping of benzophenone. The superposition of the pharmacophoric residues of peptide **1** and the benzophenone scaffold is shown. The pharmacophoric elements on peptide **1** are shown as magenta-colored circles and an ellipse. The connections of the substituents from benzophenone are denoted by curved arrows.

examples of these have been reported, the problems inherent in them have also been pointed out.[15,16] In de novo design, output structures are particularly apt to be difficult to synthesize.[15,16] In 3D database searching, novel structures cannot be obtained.[15] Thus, we chose to search the compound library in our company to identify synthetically accessible scaffolds before general 3D database searching. We succeeded in finding several good scaffolds in the library. These scaffolds are synthetically accessible and easy to modify by our organic chemists. Through design efforts based on these scaffolds, we discovered new AP-1 inhibitors: 1-thia-4-azaspiro[4.5]decane and benzophenone derivatives.

In the current paper, we describe the design by a lead-hopping strategy and synthesis of nonpeptidic AP-1 inhibitors and their inhibitory activities as evaluated by AP-1 binding and cell-based reporter gene assays.

The Design of Nonpeptidic AP-1 Inhibitors. Lead-Hopping Strategy. Our pharmacophore model of AP-1 binding compounds is illustrated in Figure 2, which shows the 3D arrangement of the side-chain functional groups of the Gln-Leu-Asp-Leu-Ala residues in peptide **1**.[9] A visual inspection revealed that the main chain of the five residues of peptide **1** has a β-turn-like conformation and a trapezoid shape (Figure 2). Its dimensions, which were measured as the distances between the α-carbons of Gln3-Leu4, Leu4-Asp5, Asp5-Ala7, and Gln3-Ala7, are 3.89, 3.89, 5.99, and 8.40 Å, respectively. Thus, we decided to design candidate small molecules for synthesis and evaluation by combining a suitably sized scaffold placed onto the trapezoid structure with the corresponding pharmacophoric elements, which constitute pivotal functional groups within the side chains of the five pharmacophoric residues of peptide **1**, to retain the 3D arrangement in the pharmacophore model.

Scaffold Design. In general, it is assumed that a scaffold should have a certain level of rigidity to maintain the correct spatial arrangement of each of the pharmacophoric elements. In addition, the substitution position to which a pharmacophoric element might be connected via a suitable linker should be flexible to ensure synthetic feasibility. In a different perspective, at least three substituents for each pharmacophoric element could be introduced that would allow sufficient pharmacophoric interactions. As depicted in Figure 2, the trapezoid was as large
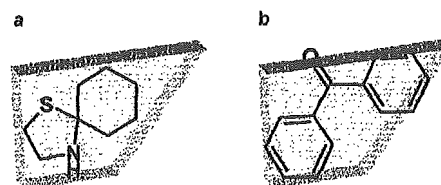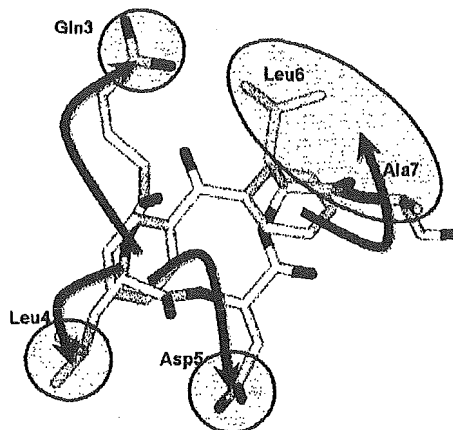
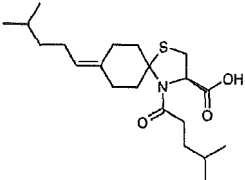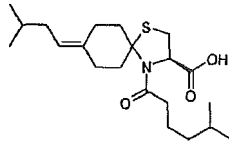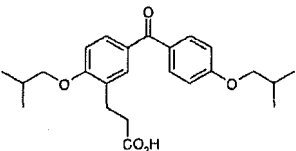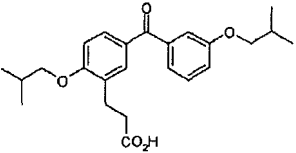as an anthracene molecule, which was therefore assumed to be the maximum size for a scaffold.

Although the main chain of the five pharmacophoric residues of peptide **1** had a planar trapezoid shape, the pharmacophoric elements were not located on the same plane. Thus, to maintain the pharmacophoric arrangement, it was desirable that the scaffold was not planar, but rather was twisted and compact, within limits.

On the basis of these considerations, by searching our compound library, we selected the two scaffolds that are depicted in Figure 3: 1-thia-4-azaspiro[4.5]decane (a), and benzophenone (b) scaffolds.

Definition of Pharmacophoric Elements. Four pharmacophoric elements were defined from the five residues of the pharmacophore model: the carboxamide group of Gln3, the isobutyl group of Leu4, the carboxyl group of Asp5, and an isobutyl group that was adopted as a proxy for Leu6 and Ala7. The reason for using a proxy was that the two residues in question were adjacent to one another and could be regarded as one large hydrophobic element. Among these, the carboxyl group of Asp5 was regarded as a key pharmacophoric element, because substitution of the Asp5 had a significant effect in a previous alanine scan experiment;[9] we therefore assumed that an acidic group at this position was essential for interactions with the basic regions of the bZIP domains. In addition, this carboxyl group appeared to be the only charged group and might have conferred desirable features in terms of drug design. We therefore produced various combinations of three or four pharmacophoric elements, while ensuring that a carboxyl group was incorporated into the newly designed molecules.

Scaffold Placement and Substitution Points. Only those atoms on a scaffold for which the introduction of a suitable substituent was synthetically feasible were viewed as potential

Table 1. Chemical Structure of Compounds 2—5 and Their Inhibitory Activities against AP-1

compd	structure	IC_{50} (μM)	
		binding assay[a]	luciferase assay[b]
1	Ac-c [Cys-Gly-Gln-Leu-Asp-Leu-Ala-Asp-Gly-Cys]-NH₂	64	NT[c]
2		650	11.8
3		460	13.3
4		610	5.0
5		600	5.8

[a] Inhibition of the binding of the AP-1 bZIP peptide to synthetic oligonucleotides containing the AP-1 binding site. [b] Inhibition of the expression of AP-1-luciferase by TPA-stimulated NIH3T3 cells. [c] NT: not tested.

substitution points. Each scaffold was manually placed onto the trapezoid depicted in Figure 2 and interactively checked using a visual display. The position and orientation of the scaffolds were designed so that as many substitution points as possible were pointed in a favorable direction toward the corresponding pharmacophoric elements.
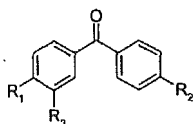
Combination of Scaffold and Pharmacophore Elements. The candidate molecules for the synthesis were constructed by connecting the relevant pharmacophoric elements to suitable substitution points on each scaffold via linkers.

An example of substituent introduction featuring benzophenone is shown in Figure 4. The connections between the pharmacophoric elements at suitable substitution points are indicated by curved arrows. In this case, the pharmacophoric elements corresponding to Gln3, Leu4, and Asp5 could be connected by any atoms on the left-hand ring of the scaffold via linkers. Likewise, the centroid of Leu6 and Ala7 could be connected from any atom on the right-hand ring. The lengths and types of linkers used were carefully selected, taking synthetic feasibility into consideration.

Computational Evaluation of Designed Molecules. We examined the various possible combinations of substitution points, and the lengths and types of linkers, so that each pharmacophoric element was well suited to the pharmacophore model. Each of the designed molecules was computationally evaluated to verify the fact that the 3D pharmacophore requirements were met by a low-energy conformer. Conformational analyses were carried out to obtain low-energy conformers. All unique conformers that were within 6 kcal mol⁻¹ of the lowest energy were collected for each of the designed molecules. The selected low-energy conformers were then systematically superimposed onto peptide 1 using the least-squares fit[17] of three or four of the pharmacophoric atoms: the δ-carbon atom of Gln3, the γ-carbon atoms of Leu4 and Asp5, and the centroid of the γ- and δ-carbon atoms of Leu6 and the β-carbon atom of Ala7. In addition, the overlapping volume between each conformer and the corresponding pharmacophoric residues in peptide 1, and its ratio to the volume of the corresponding residues in peptide 1, were calculated. Those compounds in which at least one conformer had a root-mean-square deviation

Table 2. Benzophenone Derivatives **4, 6−13** and Their Inhibitory Activities in Binding Assay

compd	R_1	R_2	R_3	% inhibition at 500 μM	% inhibition at 1 mM	IC_{50} (μM)[a]
4	O^{i}Bu	O^{i}Bu	CH$_2$CH$_2$CHO$_2$H	28	91	610
6	H	O^{i}Bu	CH$_2$CH$_2$CHO$_2$H	10	14	>2000
7	O^{i}Bu	H	CH$_2$CH$_2$CHO$_2$H	7	9	>2000
8	O^{i}Bu	O^{i}Bu	H	3	−[b]	ND[c]
9	O^{n}Pr	O^{i}Bu	CH$_2$CH$_2$CHO$_2$H	26	55	910
10	O^{i}Bu	O^{n}Pr	CH$_2$CH$_2$CHO$_2$H	25	42	930
11	OCH$_2$Ph	O^{i}Bu	CH$_2$CH$_2$CHO$_2$H	51	97	420
12	O^{i}Bu	OCH$_2$Ph	CH$_2$CH$_2$CHO$_2$H	45	−[b]	ND[c]
13	O^{i}Bu	O^{i}Bu	CH$_2$CH$_2$CONH$_2$	5	−[b]	ND[c]

[a] 50% inhibition concentration was determined from the concentration−response curve (125−2000 μM). [b] % inhibition was not determined at this concentration due to precipitation. [c] ND: not determined.

Scheme 1[a]

16 $m = 2$
18 $m = 1$

2 $m = 2, n = 1$
3 $m = 1, n = 2$
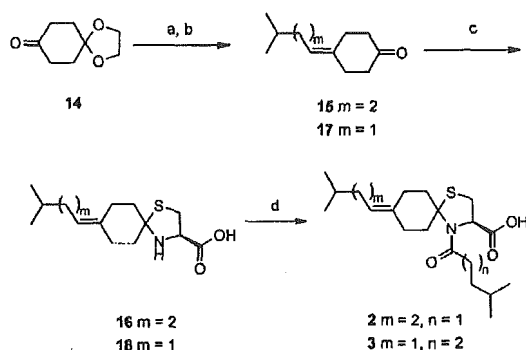
[a] Reagents: (a) Me$_2$CHCH$_2$CH$_2$CH$_2$PPh$_3$Br (for **15**) or Me$_2$CHCH$_2$CH$_2$-PPh$_3$I (for **17**), n-butyllithium, THF; (b) 6 M HCl, 1,4-dioxane; (c) L-cysteine, aq EtOH; (d) 4-methylpentanoic acid (for **2**) or 5-methylhexanoic acid (for **3**), oxalyl chloride, Et$_3$N, CH$_2$Cl$_2$.

Scheme 2[a]

19	20	6

[a] Reagents: (a) (EtO)$_2$P(O)CH$_2$CO$_2$Et, NaH, DMF; (b) H$_2$, Pd−C, EtOH; (c) oxalyl chloride, DMF, CH$_2$Cl$_2$; (d) isobutoxybenzene, AlCl$_3$, CH$_2$Cl$_2$; (e) NaOH−H$_2$O, EtOH.
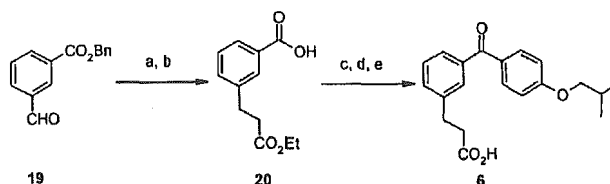
(RMSD) value ≤ 1.0 Å under a three- or four-point superposition with the corresponding residues of peptide **1**, and an overlapping volume ratio ≥ 0.5, were selected as small-molecule inhibitor candidates for synthesis and biological evaluation.

Biological Evaluation of the Compounds Synthesized. The AP-1 inhibitory activities of the compounds synthesized were evaluated using enzyme-linked immunosorbent assay (ELISA)-based AP-1 DNA-binding and cell-based reporter assays (Table 1; compounds **2−5**).

Furthermore, the synthesis of benzophenone derivatives and the evaluation of their inhibitory activities were performed to verify our pharmacophore model (Table 2; compounds **6−13**).

Chemistry

Synthesis of the (R)-4-(4-methylpentanoyl)-1-thia-4-azaspiro-[4.5]decane compound **2** was carried out as outlined in Scheme 1. The starting material, 1,4-cyclohexanedione monoethylene ketal (**14**), was subjected to the Wittig reaction with 4-methylpentyltriphenylphosphonium bromide and n-butyllithium to afford 4-(4-methylpentylidene)cyclohexanone (**15**) after an acid-catalyzed deprotection procedure. Treatment of **15** with L-cysteine in aqueous EtOH produced the spiro[cyclohexane-1,2′-thiazolidine] compound **16**, which was treated with 4-methylpentanoyl chloride in the presence of Et$_3$N to provide the target molecule **2**.[18] The structurally analogous spiro compound **3**, 1-thia-4-azaspiro[4.5]decane-3-carboxylic acid, 8-(3-methylbutylidene)-4-(5-methylhexanoyl), was prepared by the same procedure.

Syntheses of mono- and di-isobutoxy derivatives of 3-(3-benzoylphenyl)propionic acid (**4−7**) were carried out as illustrated in Schemes 2−4. 3-[3-(4-Isobutoxybenzoyl)phenyl]-propionic acid (**6**) was prepared via the Horner−Emmons reaction of benzyl 3-formylbenzoate (**19**) with ethyl diethylphosphonoacetate and NaH, which was followed by Pd−C-catalyzed hydrogenation to yield 3-(ethoxycarbonylethyl)benzoic acid (**20**). The corresponding acid chloride generated by treatment with oxalyl chloride was allowed to react with isobutoxybenzene[19] in the presence of AlCl$_3$ in CH$_2$Cl$_2$ to give an ethyl ester, which on alkaline ester hydrolysis provided compound **6** (Scheme 2). Compound **7** (Scheme 3), which is a regioisomer of the isobutoxy group of **6**, was prepared via isobutyl 3-(2-isobutyloxyphenyl)propionate (**22**), which was subjected to a regioselective Friedel−Crafts benzoylation. Compound **22** was also employed for the synthesis of 3-[2-isobutoxy-5-(4-isobutoxybenzoyl)phenyl]propionic acid (**4**). The regioselective formylation reaction[20] of **22** followed by Na-ClO$_2$−H$_2$O$_2$ oxidation[21] afforded the monocarboxylic acid **23**, and coupling its acyl chloride with isobutoxybenzene afforded compound **4**. The regioisomer **5**, 3-[2-isobutoxy-5-(3-isobutoxybenzoyl)phenyl]propionic acid, was prepared via coupling of **22** with the acyl chloride of 3-isobutoxybenzoic acid (**25**) in the presence of AlCl$_3$ (Scheme 4).

Exchange of either one of the two isobutyloxy groups in **4** with a propyloxy or benzyloxy ether moiety was also considered as outlined in Scheme 5. The methyl ester of **4** (compound **26**) was subjected to AlCl$_3$-catalyzed lactonization accompanied by cleavage of the isobutoxy linkage to produce 6-(4-hydroxybenzoyl)-3,4-dihydrocoumarin (**27**). The liberated phenolic hydroxyl group was then subjected to the Mitsunobu reaction with R^2-OH (2-methylpropanol, propanol, or benzyl alcohol) in the presence of diisopropyl azodicarboxylate (DIAD) and triphenylphosphine in tetrahydrofuran (THF) followed by MeONa-