Application of Direct Phasing Methods for Crystal Structure Analysis Using PCs. 2. Automation of the Symbolic-Addition Method

Kenji Okada

Research and Development Center, Ricoh Company, Ltd., Kouhoku-ku, Yokohama 223, Japan

In the first paper,[1] the automation of heavy-atom methods for crystallographic phasing was described. When organic compounds consist only of light atoms, direct methods are usually used to determine the molecular structure. We describe an improved PC version of direct-methods DIRECTER program. DIRECTER is based on the symbolic-addition method using the tangent formula with the multisolution process to determine the atomic coordinates of all atoms automatically without any chemical assumptions or human intervention. DIRECTER selects three *origin sets* (origin + enantiomorph + Σ_1) satisfying space group conditions, determines the *starting sets* for phase generation, refines the phases with the tangent formula, calculates E-maps, and automatically draws the two-dimensional projection of the structure on a plotter. The details of the improved DIRECTER program are described and the test results obtained are shown. This program is part of the DS*SYSTEM[2,3] and is suitable for use by organic chemists as well as professional crystallographers.

INTRODUCTION

The symbolic-addition method was proposed by Karle and Karle.[4] Germain and Woolfson reported the LASM program[5] which adopted the automated symbolic-addition method. In the last 25 years the following direct methods programs have been described: the DIRDIF series[6] in the case that a partial structure is known, an automatic system MULTAN87,[7] an automated procedure REDUCE[8] which employs estimated values of structure (semiinvariant phase combination), SIM-PEL87[9] which uses optional symbolic phase determination by a dynamic-programming principle, the integrated direct methods MITHRIL[10] which builds around a heavily modified MULTAN80 plus symbolic-addition module for centrosymmetric structures, an integrated direct method package SIR88,[11] SAPI[12] for automatic solution of structures having a pseudo-translational symmetry, and a phase annealing method in SHELEX-90[13] for larger "small-molecular" structures.

We initiated our research of direct methods in 1970 and aimed at developing a fully automatic system which we named the SEARCHER program.[14] We paid attention to the symbolic-addition method and reported our first version of DIRECTER[15] for automatic structure analysis developed on CDC6600 mainframe computer. This program has been written for the Cray-1 version by adding vector facilities and optimization[16] and rewritten in Fortran 77 and implemented on 32-bit personal computers (PCs) with several modifications and corrections of program code.[2,3] The PC 1993 version of DIRECTER has been modified to change the algorithms of peak summation in the Fourier synthesis, to achieve a reduction of calculation time and better conform to user requirements. The structure analysis of more than 20 organic compounds has been determined by the DIRECTER program.

DESCRIPTION

Theoretical Background. If the organic compounds consist of only light atoms, direct methods are usually used to determine the molecular structure. The normalized structure factor magnitude $E(\mathbf{h})$ was defined as follows[4]

$$E(\mathbf{h})^2 = \frac{|F_o(\mathbf{h})|^2}{\epsilon \sum_{j=1}^{N} f_j^2(\mathbf{h})} \quad (1)$$

where $\mathbf{h}$ is the vector index of reflection (h,k,l), $F_o(\mathbf{h})$ is the structure factor of vector index $\mathbf{h}$, N is the number of atoms in a unit cell $(j = 1, 2, ..., N)$, $f_j(\mathbf{h})$ is the atomic scattering factor of the jth atom, and ϵ is the small integer which is space group dependent. To determine the phases in direct methods, we specify the origin defining and enantiomorph reflections (*origin set*). The origin defining reflections have been fixed by space type (P,A,B,C), centro-/noncentrosymmetry, general/restricted phases, invariant/semiinvariant vector/module/phases, linear combinations of phase, independent set of phases, and so on. The enantiomorph reflections should be defined at noncentrosymmetry. The large structure factor magnitude $E_c(\mathbf{h})$ is defined by eq 2 and the phase $\phi(\mathbf{h})$ of $E_c(\mathbf{h})$ is associated with the average value that is taken only over those k_r corresponding to the large $|E(\mathbf{h})|$ values as specified by Σ_2 relationships.

$$E_c(\mathbf{h}) = |E(\mathbf{h})| \exp(i\phi(\mathbf{h})) \quad (2)$$

$$\phi(\mathbf{h}) \simeq \langle \phi(\mathbf{k}) + \phi(\mathbf{h} - \mathbf{k}) \rangle_{k_r} \quad (3)$$

where k_r refers to a set of data restricted to the largest magnitude $|E(\mathbf{h})|$. For centrosymmetric crystals all phases are either 0 or π, and/or equivalently the signs of the structure factors are either + or –, then eq 3 leads to eq 4

$$s(E(\mathbf{h})) \simeq s(\sum_{k_r} E(\mathbf{k}) E(\mathbf{h} - \mathbf{k})) \quad (4)$$

where $s()$ means "sign of". The probability $P_+(E(\mathbf{h}))$ that $E(\mathbf{h})$ is positive and the sign is correct is expressed as follows:

$$P_+(E(\mathbf{h})) = \frac{1}{2} + \frac{1}{2} \tanh(\sigma_3 \sigma_2^{-3/2} |E(\mathbf{h})| \sum_{k_r} E(\mathbf{k}) E(\mathbf{h} - \mathbf{k})) \quad (5)$$

where

$$\sigma_n = \sum_{j=1}^{N} Z_j^{\,n}$$

PHASING METHODS FOR CRYSTAL STRUCTURE ANALYSIS

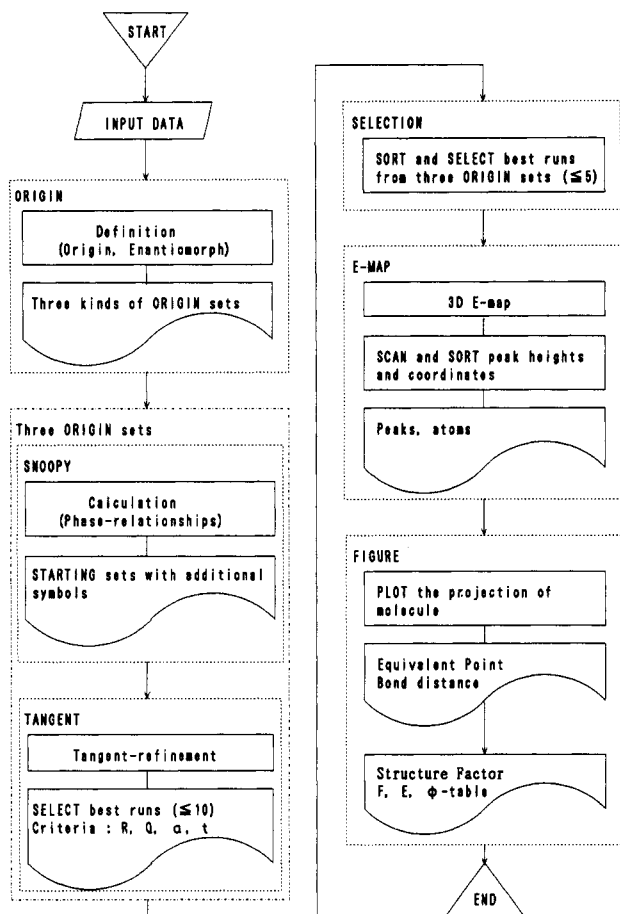*J. Chem. Inf. Comput. Sci., Vol. 34, No. 3, 1994* **595**

Figure 1. A simplified DIRECTER program flow chart. DI-RECTER reads input data, defines three *origin sets* (origin + enantiomrph + Σ_1) with phased symbols, selects additional symbols as *starting set*, refines the phase with tangent formula, selects of the better five *run sets*, calculates E-map, calculates structure factors, and draws two-dimensional projections of the molecules in a unit cell.

Z_j is the atomic number of the jth atom in a unit cell. For noncentrosymmetric crystals the equation (6) of phase relationships is most useful instead of eq 3.

$$\phi(\mathbf{h}) \simeq \frac{\sum_{k_r} |E(\mathbf{k})E(\mathbf{h}-\mathbf{k})|(\phi(\mathbf{k}) + \phi(\mathbf{h}-\mathbf{k}))}{\sum_{k_r} |E(\mathbf{k})E(\mathbf{h}-\mathbf{k})|} \qquad (6)$$

After *origin set* which consists of the origin defining and enantiomorph reflections is specified, we add some phases with additional symbols $p, s, m, a, b, \ldots$ to take the opportunity of fixing an extra unknown phase easily (*starting set*). The phase $\phi(\mathbf{h})$ is refined to 0, π, or $\pm\pi/2$ for two-dimensional reflection and is kept within the range $-\pi < \phi(\mathbf{h}) < +\pi$ for three-dimensional reflection. Tangent formula followed by eq 6 is employed for new phase indications.

$$\tan(\phi(\mathbf{h})) \simeq \frac{B(\mathbf{h})}{A(\mathbf{h})} \qquad (7)$$

where

$$A(\mathbf{h}) = \sum_{k_r} |E(\mathbf{k})E(\mathbf{h}-\mathbf{k})| \cos(\phi(\mathbf{k}) + \phi(\mathbf{h}-\mathbf{k}))$$

$$B(\mathbf{h}) = \sum_{k_r} |E(\mathbf{k})E(\mathbf{h}-\mathbf{k})| \sin(\phi(\mathbf{k}) + \phi(\mathbf{h}-\mathbf{k}))$$

The probability $P_+(E(2\mathbf{h}))$ that the phase $\phi(\mathbf{h})$ of $E(2\mathbf{h})$ is

positive is expressed as follows:

$$P_+(E(2\mathbf{h})) = \frac{1}{2} + \frac{1}{2}\tanh(2\sigma_3\sigma_2^{-3/2}|E(2\mathbf{h})|\sum_1) \qquad (8)$$

where

$$\sum_1 = \frac{1}{4}\sigma_3\sigma_2^{-3/2}\sum_{\mathbf{h}=2\mathbf{h}_\mu}(E(\mathbf{h}\mu)^2 - 1)$$

And the variance $V(\mathbf{h})$ of phase $\phi(\mathbf{h})$ is given as follows[4]

$$V(\mathbf{h}) = \frac{\pi^2}{3} - [I_0(\alpha(\mathbf{h}))]^{-1}\sum_{n=0}^{\infty}\frac{I_{2n}(\alpha(\mathbf{h}))}{n^2} +$$

$$4[I_0(\alpha(\mathbf{h}))]^{-1}\sum_{n=0}^{\infty}\frac{I_{2n+1}(\alpha(\mathbf{h}))}{(2n+1)^2} \qquad (9)$$

where $I_n(x)$ is the n-ordered modified Bessel function of x argument

$$\alpha(\mathbf{h}) = \{[\sum_{k_r} K(\mathbf{h}, \mathbf{k}) \cos(\phi(\mathbf{k}) + \phi(\mathbf{h}-\mathbf{k}))]^2 +$$

$$[\sum_{k_r} K(\mathbf{h}, \mathbf{k}) \sin(\phi(\mathbf{k}) + \phi(\mathbf{h}-\mathbf{k}))]^2\}^{1/2}$$

$$K(\mathbf{h}, \mathbf{k}) = 2\sigma_3\sigma_2^{-3/2}|E(\mathbf{h})E(\mathbf{k})E(\mathbf{h}-\mathbf{k})|$$

When the phase $\phi(\mathbf{h})$ was unknown, Germain and Woolfson have defined[5] the $\alpha(\mathbf{h})$ as the mean value of $\langle\alpha(\mathbf{h})^2\rangle$ for a given set of $K(\mathbf{h}, \mathbf{k})$.

$$\langle\alpha(\mathbf{h})^2\rangle = \sum_{\mathbf{h}} K(\mathbf{h}, \mathbf{k})^2 +$$

$$2\sum_{\substack{\mathbf{k}, \mathbf{h}-\mathbf{k} \\ \mathbf{k}\neq\mathbf{h}-\mathbf{k}}} K(\mathbf{h}, \mathbf{k})K(\mathbf{h}, \mathbf{h}-\mathbf{k})\cdot\frac{I_1(K(\mathbf{h}, \mathbf{k}))}{I_0(K(\mathbf{h}, \mathbf{k}))}\cdot\frac{I_1(K(\mathbf{h}, \mathbf{h}-\mathbf{k}))}{I_0(K(\mathbf{h}, \mathbf{h}-\mathbf{k}))} \qquad (10)$$

where

$$\frac{I_1(k)}{I_0(k)} \simeq 0.5658k - 0.1304k^2 + 0.0106k^3$$

Also Motherwell has defined[17] the $\alpha(\mathbf{h})$ as follows:

$$\alpha(\mathbf{h}) = \begin{cases} K(\mathbf{h},\mathbf{k}) & (k_r = 1) \\ \langle\alpha_\mathbf{h}^2\rangle^{1/2} & (k_r > 1) \end{cases} \qquad (11)$$

Drew, Larson, and Motherwell have defined[18] the magnitude $|E_c(\mathbf{h})|$ that the phase was already determined as follows:

$$|E_c(\mathbf{h})| = \frac{\sigma_3\sigma_2^{-3/2}}{k_r}(A(\mathbf{h})^2 + B(\mathbf{h})^2)^{1/2} \qquad (12)$$

And the residual factor R_{Karle} was expressed as follows

$$R_{\text{Karle}} = \frac{\sum_{k_r}||E_o(\mathbf{h})| - k|E_c(\mathbf{h})||}{\sum_{k_r}|E_o(\mathbf{h})|} \qquad (13)$$

where $E_o(\mathbf{h})$ is the magnitude $E(\mathbf{h})$ defined by eq 1 and k is a constant chosen to give $\Sigma|E_o(\mathbf{h})| = k\Sigma|E_c(\mathbf{h})|$. Also the consistency parameter $t(\mathbf{h})$ and the consistency index Q_{Drew} were expressed as eqs 14 and 15.[18]

$$t(\mathbf{h}) = \frac{(A(\mathbf{h})^2 + B(\mathbf{h})^2)^{1/2}}{\sum_{k_r}|E(\mathbf{k})E(\mathbf{h}-\mathbf{k})|} \qquad (14)$$

Table 1. The Molecular Properties of the Structures Solved Using DIRECTER[a]

	I	II	III	IV	V	VI	VII
A. crystal data							
formula	$C_{15}H_{16}O_4$	$C_9H_{11}O_6N$	$C_{11}H_{14}O$	$C_{19}H_{22}O_3S$	$C_{23}H_{32}O_7$	$C_{15}H_{24}O_2$	$C_{20}H_{30}O_2$
mol wt	260.3	229.2	162.2	330.4	420.5	236.4	302.5
space group	$P2_12_12_1$	$C2$	$P2_12_12_1$	$P\bar{1}$	$P2_1$	$P3_2$	$P2_1$
Z	4	4	4	2	2	3	4
a (Å)	13.633(3)	15.454(5)	9.781(3)	9.127(1)	12.512(2)	13.104(1)	14.058(1)
b (Å)	9.102(2)	6.547(3)	9.923(3)	11.955(1)	7.436(2)	13.104(1)	11.911(1)
c (Å)	10.153(2)	11.425(4)	9.428(3)	8.783(1)	12.252(2)	6.756(1)	11.753(1)
α (deg)	90.0	90.0	90.0	93.08(1)	90.0	90.0	90.0
β (deg)	90.0	121.27(3)	90.0	117.70(1)	102.62(1)	90.0	111.73(2)
γ (deg)	90.0	90.0	90.0	85.49(1)	90.0	120.0	90.0
V (Å³)	1259.8(2)	988.1(8)	915.1(5)	845.7(1)	1112.4(3)	1004.6(1)	1828.1(5)
D_m (g·cm⁻³)	1.367	1.527	1.165	1.282			
D_x (g·cm⁻³)	1.373	1.540	1.177	1.298	1.255	1.172	1.097
B. problem	molecular conformation	molecular conformation	molecular conformation	structure unknown	molecular conformation	molecular conformation	molecular conformation
C. refinement	full-matrix	full-matrix	full-matrix	full-matrix	full-matrix	full-matrix	full-matrix
temp factor	all-aniso	all-aniso	all-aniso	all-aniso	all-aniso	all-aniso	all-aniso
R	0.074	0.050	0.055	0.058	0.071	0.029	0.058
D. experiment							
λ	Mo Kα	Mo Kα	Mo Kα	Mo Kα	Mo Kα	Cu Kα	Cu Kα
scan-range	$\theta \le 27.5°$	$\theta \le 27.5°$	$\theta \le 27.5°$	$\theta \le 27.5°$	$\theta \le 27.5°$	$\theta \le 70°$	$\theta \le 70°$
scan-technique	$\omega/2\theta$	$\omega/2\theta$	$\omega/2\theta$	$\omega/2\theta$	$\omega/2\theta$	$\omega/2\theta$	$\omega/2\theta$
filter	Zr	Zr	Zr	Zr	Zr	Ni	Ni
F(obs) $\ge 3\sigma$	1158	1092	878	2468	2374	1337	2830
E(obs)	134 ($\ge$1.35)	188 ($\ge$1.35)	141 ($\ge$1.50)	299 ($\ge$1.50)	323 ($\ge$1.50)	172 ($\ge$1.35)	450 ($\ge$1.46)
E. reference	X-1	X-12	X-80	X-144	X-164	K-85	K-88

[a] Key: I, linderane; II, showdomycin; III, spirodienone II; IV, 17β-acetoxy-18-methyl-thia-A-norestra-3,9(10),11-trien-2-one; V, 3β,17β-diacetoxy-8,9-seco-5α-androstane-8,9,11-trione; VI, β-coryophyllene derivative; VII, 7,13-abietadien-18-oic acid.

$$Q_{Drew} = \frac{\sum_h (E(\mathbf{h}) - t(\mathbf{h})E(\mathbf{h}))}{\sum_\mathbf{h} |E(\mathbf{h})|} \qquad (15)$$

The E-map $\rho(\mathbf{x})$ expressed by eq 16 is given by the Fourier transformation of $|E(\mathbf{h})|$.

$$\rho(\mathbf{x}) = \frac{1}{V}\sum_\mathbf{h} |E(\mathbf{h})| \exp(i\phi(\mathbf{h})) \exp(-2\pi i \mathbf{h}\cdot\mathbf{x}) \qquad (16)$$

where $\phi(\mathbf{h})$ is the phase angle obtained by eq 7. Based on possible atoms from the E-map the calculated structure factor $F_c(\mathbf{h})$ is expressed by eq 17.

$$F_c(\mathbf{h}) = \sum_{j=1}^{N} f_j(\mathbf{h}) \exp(-B_j \sin^2 \theta/\lambda^2) \exp(-2\pi i \mathbf{h}\cdot\mathbf{r}_j) \qquad (17)$$

where B_j is the atomic temperature factor of atom j, $\mathbf{r}_j$ is the position vector of atom j (x_j, y_j, z_j), and $\sin \theta$ is the Bragg angle. And the reliability factor R (R-factor) expressed by eq 18 was obtained from structure factors.

$$R = \frac{\sum_\mathbf{h} ||kF_o(\mathbf{h})| - |F_c(\mathbf{h})||}{\sum_\mathbf{h} |F_c(\mathbf{h})|} \qquad (18)$$

where k is the scale factor to $|F_o(\mathbf{h})|$, $k\sum|F_o(\mathbf{h})| = \sum|F_c(\mathbf{h})|$.

DIRECTER. DIRECTER is based on the symbolic-addition method using the tangent formula with a multisolution process. The algorithm of PC 1993 version is illustrated in Figure 1 as a flow chart. After the loading of the data, the following six logical steps are carried out in sequence of calculations: (i) ORIGIN, define three kinds of *origin set* with phased symbols; (ii) SNOOPY, select additional symbols to generate up to 95% of the potential reflections (*starting sets*); (iii) TANGENT, refine the phase and select of the better 10 *run sets* from each *starting set*; (iv) SELECTION,

select out the better 5 *run sets*; (v) E-map, calculate Fourier synthesis by $E_c(\mathbf{h})$, look for the peaks, store the coordinates of each peak and calculate the structure factors; (vi) FIGURE, draw two-dimensional projections of the molecules in a unit cell. DIRECTER has been reconstructed from the several programs as follows: MULTAN71[19] for the definition of *origin set*; ASIG[20] for Σ_2 calculation; SNOOPY[17] for *starting set* selection; TANTWO[18] and TANGFORM[20] for tangent formula; SEARCHER[14,21] for the E-map calculation and two-dimensional projection figure. We improved DIRECTER on the input and output file assignments by adopting interactive prompt facilities, on the peak selection in Fourier synthesis, and on the optimization of DO-loop sequences in the computation at Σ_2 relationships, *starting set* selection, tangent formula, Fourier summation, and structure factor calculation.

DIRECTER reads the data in three sets. The first contains the cell constants, the name of space group, the chemical symbols of atoms, and numbers of atoms in a unit cell. The second set of data includes the control data for the calculations of *origin set*, *starting set*, tangent formula, E-map, and two-dimensional projection. The third set contains the reflection data $[h, k, l, |F_o(\mathbf{h})|, |E_o(\mathbf{h})|, \sigma(I_o(\mathbf{h}))]$. These $|E_o(\mathbf{h})|$ are presorted in decreasing magnitude. We usually use $|E_o(\mathbf{h})| \ge$ EMIN (input), EMIN = 1.50, or the number of reflection data equal nearly to 10 times of number of non-hydrogen atoms of one molecule. In the case of the cell constants are unbalanced, for example $a > 2b$, it is desirable sometimes to eliminate high ordered reflection that Σ_2 relationships should not be biased.

Origin Set (*ORIGIN* + *Enantiomorph* + Σ_1). DIRECTER specifies three kinds of *origin set* with symbolic phases. The first *origin set* is selected from the convergence map by $|E_o(\mathbf{h})| \ge$ EMINO (input, usually EMIN) from EMIN reflections. If the magnitude of the projected vector is Σ_1 and the probability $P_+(E(\mathbf{2h}))$ greater than 0.95, this reflection is added to the *origin set* as a "known phase". After the first *origin set* is defined, the second *origin set* must satisfy $|E_o(\mathbf{h})| \ge$

Table 2. Computational Results of the Structures Solved by DIRECTER

compound	space group	semiinvariant vector module[a]	EMIN[b]	set	ORIGIN EMINO[c]	SNOOPY EMINS[d]	AMINS	E3MINS	TANGENT AMINT	run sets	SELECTION R_{Karle}	FIGURE E-map R	atoms
X-1	$P2_1 2_1 2_1$	(h,k,l)	1.35	1	1.50	1.50	2.5	6.0	2.5	64	0.194–0.209	5th	18/19
		(2,2,2)	134		134, 5	7, 77%						0.212	
		3	$\|h,k,l\| \le 10$	2	1.60	1.50	2.5	6.0	2.5	64			
					71, 4	7, 77%							
				3	1.70	1.50	2.5	6.0	2.5	64			
					38, 4	7, 77%							
X-12	C2	(k,l)	1.35	1	1.35	1.50	3.0	4.0	2.5	128	0.177–0.183	4th	16/16
		(0,2)	188		188, 4	5, 99%						0.172	
		2		2	1.45	1.50	3.5	4.0	2.5	1024			
					140, 4	7, 99%							
				3	1.55	1.50	4.0	4.0	2.5	2048			
					110, 5	7, 98%							
X-80	$P2_1 2_1 2_1$	(h,k,l)	1.50	1	1.50	1.50	5.0	10.0	2.5	16	0.178–0.188	1st–5th	12/12
		(2,2,2)	141		141, 6	7, 82%						0.195–0.216	
		3		2	1.60	1.50	5.0	10.0	2.5	16			
					106, 6	7, 82%							
				3	1.70	1.50	1.50	10.0	2.5	16			
					96, 6	7, 82%							
X-144	$P\bar{1}$	(h,k,l)	1.50	1	1.50	1.50	5.0	10.0	2.5	8	0.199–0.247	5th	19/23
		(2,2,2)	229		229, 7	8, 83%						0.450	
		3	$\|h,k,l\| \le 9$	2	1.70	1.50	5.0	10.0	2.5	8			
					195, 7	8, 83%							
				3	1.90	1.50	5.0	10.0	2.5	8			
					146, 6	7, 83%							
X-164	$P2_1$	(h,k,l)	1.50	1	1.50	1.50	4.5	8.0	2.5	8	0.198–0.212	1st	25/30
		(2,0,2)	323		323, 5	6, 75%						0.274	
		3		2	1.70	1.50	4.5	8.0	2.5	16			
					193, 5	6, 75%						5th	
				3	1.80	1.50	4.5	8.0	2.5	8		0.291	
					149, 5	6, 75%							
K-85	$P2_3$	(h−k,l)	1.35	1	1.50	1.50	3.5	4.0	2.5	2048	0.156–0.167	5th	17/17
		(3,0)	172		172, 4	7, 100%						0.170	
		2	$\|h,k,l\|, 10$	2	1.60	1.50	4.0	4.0	2.5	1024			
					76, 6	9, 100%							
				3	1.70	1.50	3.5	4.0	2.5	256			
					56, 8	10, 100%							
K-88	$P2_1$	(h,k,l)	1.46	1	1.50	1.50	4.5	8.0	2.5	128	0.215–0.217	3rd	34/44
		(2,0,2)	450		450, 6	10, 75%						0.378	
		3		2	1.60	1.50	4.5	8.0	2.5	64			
					327, 6	9, 75%							
				3	1.70	1.50	4.5	8.0	2.5	64			
					238, 5	7, 75%							

[a] Number of origin defining reflections. [b] Number of reflections used. [c] Number of used reflections and number of origin defining reflections with enantiomorphs and Σ_1 reflections. [d] Number of starting reflections and percentage of generated reflections.

(EMINO + 0.1) and be different from the first set, and the third should be $|E_o(\mathbf{h})| \ge$ (EMINO + 0.2). DIRECTER treats the phases of origin as "known-fixed" and the phase of enantiomorph as "known-free" which the combination number of phase reduces to half of general. The value of phase has combined with a symbol in millicycles ($2\pi = 1000$) and a flag (fixed = 2, free = 0). Usually the symbol means 0 or $\pi/2$ for two-dimensional reflection and $(\pi/4, 3\pi/4)$ for general three-dimensional reflection.

Starting Set (SNOOPY). The first NSNPY (input, usually 50) reflections which have large $|E_o(\mathbf{h})|$ are postulated as a potential member of the *starting set*. To choose several potential members, DIRECTER calculates new Σ_2 with the following conditions: $\alpha_\mathbf{h} \ge$ AMINS (input), $|E_o(\mathbf{h})| \ge$ EMINS (input) for $k_r = 1$ and $|E(\mathbf{h})E(\mathbf{k})E(\mathbf{h}-\mathbf{k})| \ge$ E3MINS (input). If a chosen member can generate more than 95% of all data, this potential reflection is added to the *starting set* as the new member that has the symbol and flag "known". The phases of new symbols are restricted by their dimensions, $(0, \pm\pi)$ or $(\pm\pi/2)$ for two-dimensional reflection and $(\pm\pi/4, \pm3\pi/4)$ for general three-dimensional reflection. If the number of generated phases is less than 75%, DIRECTER reduces EMINS, AMINS, and E3MINS automatically to loose the conditions. Usually we use AMINS = 5.0–2.5 ($\sigma_\alpha = 27.3$–$43.0°$), EMINS = 1.50–1.35, and E3MINS = 10.0–6.0.

Tangent Formula (TANGENT). The phases of *starting set* which are given as symbols are replaced by numerical values in a systematic manner. These phases are used to generate new phases by the tangent formula subject to the following certain acceptance criterions: $|E(\mathbf{h})E(\mathbf{k})E(\mathbf{h}-\mathbf{k})| \ge$ E3MINT (input), $\alpha(\mathbf{h}) \ge$ AMINT (input), $t(\mathbf{h}) \ge$ TMINT (input). The process is performed to cover all combinations (*run set*) of phases and is continued by stepping iteration. The step ranges of reflection are usually 50 of the largest $|E(\mathbf{h})|$ for the first 7 cycles, 100 for the next 8 cycles, and all data for 10 cycles (total 25 cycles). If the phase changes greatly from cycle to cycle, the phase is adjusted to 0, $\pm\pi$, or $\pm\pi/2$ for two-dimensional reflection or is rejected by setting a maximum allowed shift $\Delta\phi(\mathbf{h}) = \pi/4$. A value of flag = 2 for a symbolic phase means the phase value is kept fixed until the last five cycles, flag = 0 means the phase is allowed to refine from the first cycle. The final values of symbolic phases are usually different from the starting values, but the correct solution often shows the least movement of these values. In order to choose the "best 10 *run sets*" of phases with which to compute five E-maps, we compute the figures R_{Karle}, Q_{Drew}, $\langle\alpha\rangle$, and $\langle t\rangle$.[18] Usually we use the following values: E3MINT = 10.0–8.0, AMINT = 2.5, TMINT = 0.25. As the number of atoms increases, the computed values of t usually decrease and R_{Karle} increases.

598 *J. Chem. Inf. Comput. Sci., Vol. 34, No. 3, 1994*

OKADA

Table 3. Comparison of Computer Speeds Using DIRECTER

compound	formula space group	CDC6600[a]	80386 (20 MHz)[a]	i486 + ODP (25 MHz)[a]
X-1	$C_{15}H_{16}O_4$	1 h 33 min 18 s	1 h 20 min 34 s	19 min 12 s
	$P2_12_12_1$	4.859	4.196	1.000
X-12	$C_9H_{11}O_6N$	1 h 19 min 27 s	10 h 15 min 08 s	1 h 41 min 58 s
	$C2$	0.779	6.023	1.000
X-80	$C_{11}H_{14}O$	14 min 49 s	53 min 47 s	10 min 19 s
	$P2_12_12_1$	1.436	5.213	1.000
X-144	$C_{19}H_{22}O_3S$	29 min 12 s	41 min 04 s	9 min 54 s
	$P\bar{1}$	2.949	4.148	1.000
X-164	$C_{23}H_{32}O_7$	33 min 56 s	2 h 13 min 10 s	22 min 47 s
	$P2_1$	1.489	5.845	1.000
K-85	$C_{15}H_{24}O_2$		25 h 04 min 24 s	6 h 39 min 56 s
	$P3_2$		3.762	1.000
K-88	$C_{20}H_{30}O_2$		5 h 06 min 01 s	1 h 01 min 00 s
	$P2_1$		5.017	1.000
ratio		2.310	4.886	1.000

[a] Upper, actual CPU time; lower, ratio.

Selection (SELECTION). At this step, DIRECTER sorts the "best 10 *run sets*" obtained from each *starting set* in increasing order of R_{Karle}, eliminates equivalent *runs* that have the same values of R_{Karle}, Q_{Drew}, $\langle\alpha\rangle$, and $\langle t\rangle$, and identifies the five most probable *run sets* for computing *E*-maps.

E-map. The three-dimensional Fourier synthesis of *E*-maps is computed for the asymmetric unit at the predetermined grid intervals ($\Delta x, \Delta y, \Delta z$) to select a resolution in the range 0.1–0.25 Å. The centers of Fourier peaks are calculated using a suitable second-order polynomial by Lagrangian interpolation from 27 values in three neighboring sections. When the distance between peaks in neighboring Fourier sections is less than SAMAX (0.5 Å), the peak-pair is considered to be the same atom. The limitation in peak selection was improved from a predetermination to an automatic adjustment. Compared with peak height and RHOMIN (input), the number of peaks is adjusted by revise the initial RHOMIN. The total number of atoms (NSF) selected from *E*-map is set to 1.1 times the number of non-hydrogen atoms for the structure, to allow for included solvent and ghost atoms in selection process. After NSF atoms are obtained, DIRECTER calculates structure factors and the *R*-factor.

2D-Projection (FIGURE). The outputs of DIRECTER are two-dimensional projection diagrams of the molecules with the scale of 1 Å = 1.0–2.5 cm, the atomic coordinates, and the bond distances in a unit cell on A4 paper.

APPLICATIONS AND RESULTS

We summarize the results of molecular properties of seven organic compounds solved by DIRECTER in Table 1. Table 2 shows a summary of the computational results in the semivariant vector/module, *origin set*, *starting set*, tangent formula, R_{Karle} of peak selections, *E*-maps, and the number of obtained atoms. A comparison of computational times for a large CDC6600 computer with a number smaller recent PCs is summarized in Table 3.

X-80: $C_{11}H_{14}O$, $P2_12_12_1$, $Z = 4$. The number of origin defining reflections is three and the seminvariant vector/module is $(h,k,l) = (2,2,2)$. The input data are shown in Figure 2. We used 141 reflections with EMIN = 1.50. At ORIGIN, the *origin set 1* is defined by six reflections assigned with the symbols *A*, *B*, an enantiomorph (1,3,7) and (2,0,0) added by Σ_1 with EMINO = 1.50. The *origin sets 2* and *3* were defined by six reflections with the symbols *A*, *B* and an enantiomorphs (8,2,4) from 106 [EMINO = 1.60] and (1,3,1) from 96 reflections [EMINO = 1.70], respectively. At SNOOPY, seven reflections with the symbol *C* were selected

• Example Input data of *DIRECTER*

```
----*----1----*----2----*----3----*----4----*----5----*----6----*----7----*----8
* X-80   C11.H14.O  Z=4  P212121 , ORIGIN TO E-MAP *
9.7813    9.9232    9.4286    0.0       0.0       0.0
P212121                      300010
    4    8    44    6   56    1
1.0
                      30                           C-
    2  100    0    50    3  100    0  100    1  100    0   50
    2.0    3  1.5
   71    1(3I5,F10.4,F10.6,I5)
ORIGIN
END
----*----1----*----2----*----3----*----4----*----5----*----6----*----7----*----8
```

• Example Reflection data : h, k, l, $|F_o(h)|$, $|E_o(h)|$, $\sigma(I_o(h))$

```
----*----1----*----2----*----3----*----4----*----5----*----6----*----7----*----8
    1    2    0   327.78    4.2226    3
    0    7   10    18.05    3.1236    3
    1    3    7    54.27    2.9940    3
    8    2    4    38.07    2.9012    3
    4    2    0   128.85    2.8513    3
  . . .
  . . .
  . . .
    8    6    7     9.08    1.5121    3
    4    0    0   108.85    1.5119    3
    1    1    9    18.77    1.5082    3
    2    7    9    10.40    1.5033    3
    5    6    3    24.47    1.5023    3
999999999999999999999999999999999999999999
----*----1----*----2----*----3----*----4----*----5----*----6----*----7----*----8
```

Figure 2. Input data of DIRECTER program. DIRECTER read the title, the cell parameters, the name of space group, the chemical composition (element type and number of atoms in the unit cell), the control data for the calculation of *E*-maps, reflection data, and key words for calculation. The reflection data containing $|F_o(h)|$ should be previously sorted in decreasing order of $|E_o(h)|$.

as the *starting set 1* that could generate 115 reflections (82%) with EMINS = 1.50, AMINS = 5.0, and E3MINS = 10.0. It calculated in 16 *run sets* with AMINT = 2.5 at TANGENT. The *starting sets 2* and *3* indicating seven reflections with the symbol *C* were selected and 16 *run sets* were calculated, respectively. The *starting sets* are shown in Table 4 and the list of the more probable correct phase sets at SELECTION is shown in Table 5. R_{Karle} of the lower five values had from 0.178 to 0.188 which indicated all 12 atoms of one molecule in all *E*-maps. After FIGURE, these *run sets* have indicated *R*-factors from 0.214 to 0.215. Figure 3 shows the two-dimensional projection of the b^*-axis of top set (set = 2, run = 2).

X-144: $C_{19}H_{22}O_3S$, $P\bar{1}$, $Z = 2$. We used 229 reduced reflections with $|h,k,l| \leq 9$ and EMIN = 1.50. A total of 19 of 23 atoms of the fifth *E*-map are determined by DIRECTER. The other four atoms were obtained from SEARCHER.[1-3] The structure of X-144 was reported.[22]

K-88: $C_{20}H_{30}O_2$, $P2_1$, $Z = 4$. We used 450 reflections with EMIN = 1.46 for ORIGIN. And these reflections are reduced automatically to 371 reflections for the limitation of Σ_2 of DIRECTER for SNOOPY and TANGENT. The third *E*-map indicated $R = 0.378$ with 34/44 atoms of one molecule. The other 10 atoms were obtained from SEARCHER. The structure of K-88 was reported.[23]

CONCLUSION

We have Fortran coded a PC 1993 version of the direct-methods DIRECTER program that was based on the symbolic-addition method. DIRECTER was able to get equal atoms or molecular fragments surely compared with the MULTAN series[7,24,25] which is used most widely in the world. That is a reason why DIRECTER makes all combinations of symbols to calculate the tangent formula. DIRECTER can be run automatically from start to finish or can be restarted from any step. And it is very useful that the structural results can

Table 4. *Starting Sets* of X-80

			starting set 1								*starting set 2*								*starting set 3*			
h	k	l	$\|E(\mathbf{h})\|$	$\phi(\mathbf{h})$	symbol^d	values	h	k	l	$\|E(\mathbf{h})\|$	$\phi(\mathbf{h})$	symbol^d	values	h	k	l	$\|E(\mathbf{h})\|$	$\phi(\mathbf{h})$	symbol^d	values
2	0	0^a	2.12	0	2		2	0	0^a	2.11	0	2		2	0	0^a	2.11	0	2	
1	2	0	4.22	250	2		1	2	0	4.22	250	2		1	2	0	4.22	250	2	
0	1	7	2.36	250	2		8	0	5	2.31	250	2		8	0	5	2.31	250	2	
8	0	5	2.31	250	2		1	5	0	2.24	250	2		0	5	1	2.01	250	2	
1	3	7^b	2.99	0	A0	125, 375	1	3	7	2.99	0	A0	125, 375, 625, 875	8	2	4	2.90	0	A0	125, 375, 625, 875
8	2	4	2.90	0	B0	125, 375, 625, 875	8	2	4^b	2.90	0	B0	125, 375	1	3	1^b	2.55	0	B0	125, 375
2	3	0^c	2.19	0	C0	1000, 1500	2	3	0^c	2.19	0	C0	1000, 1500	2	3	0^c	2.19	0	C0	1000, 1500

^{a} Selected from Σ_1. ^{b} Enantiomorph, the number of phase reduces to the half of general. ^{c} Selected from SNOOPY. ^{d} A flag (2 = fixed, 0 = free) and/or combined with a symbol.

Table 5. List of the More Correct Phase Sets of X-80^a

no.	R_{Karle}	Q_{Drew}	$\langle t \rangle$	$\langle \alpha \rangle$	Σ_2	NDET	set	run	A	B	C
1	0.1776	0.3586	0.6300	70.81	140	6820	2	2	375	125	1000
							1	2	same as above		
2	0.1776	0.3586	0.6300	70.81	140	6820	3	1	125	125	1000
3	0.1818	0.3772	0.6128	68.48	139	6820	1	9	125	125	1500
4	0.1875	0.3685	0.6204	69.32	138	6761	3	13	125	375	1500
5	0.1883	0.3752	0.6140	68.85	138	6761	3	15	625	375	1500
6	0.1958	0.3888	0.6004	59.20	127	5822	1	13	125	625	1500
7	0.2023	0.3801	0.6088	63.74	135	6299	2	14	375	375	1500
8	0.2099	0.4196	0.5675	57.95	127	5935	2	12	875	125	1500
9	0.2213	0.4192	0.5741	58.94	132	6124	1	3	125	375	1000
10	0.2222	0.4175	0.5740	62.32	131	6382	3	11	625	125	1500

^{a} DIRECTER normalizes the $\phi(\mathbf{h})$ in the range of $0 < \phi(\mathbf{h}) < 4\pi$. $C = 1000$ means $\phi(\mathbf{h}) = 0$ and 1500 means π.

X-80 C11.H14.O Z=4 P212121 , ORIGIN TO E-MAP ×
14 ATOMS , SET= 2, RUN= 2, R(KARLE) = .1776, R = .2140
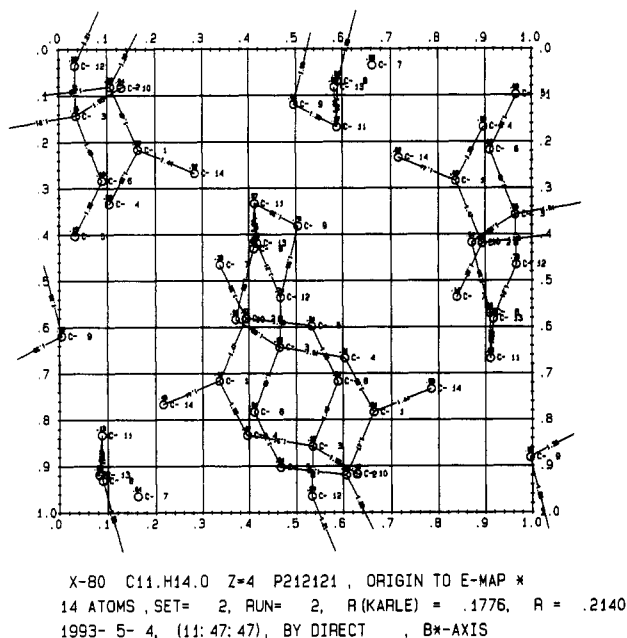1993- 5- 4, (11: 47: 47), BY DIRECT , B×-AXIS

Figure 3. Output of DIRECTER program. DIRECTER draws the molecules in the unit cell, and lists atomic coordinates and bond distances.

be output for inspection as a two-dimensional projection in a unit cell. If some fragments are obtained by DIRECTER, SEARCHER can be used as a powerful tool to find out remaining atoms. Nonexpert crystallographers should have no problem determining the crystal structures of their own organic compounds using DIRECTER.

DIRECTER needs from 10 min to 7 h running time on the latest 32-bit PC (i486+overdrive processor, 25MHz). The computational time was reduced dramatically by program modifications over the last two years. The performance of the new PC version is twice as fast as the older CDC6600 version of the past. Now a days PCs are widely distributed throughtout the world and easy to use, and the PC version of

DIRECTER is easy to operate and is very cost effective for these kinds of computations.

Program availability. DIRECTER is written is standard Fortran77 and fully tested under MS-DOS 3.x. Only standard options were applied. The limitation of DIRECTER is unified by using DS*SYSTEM common libraries: the number of symbols ≤ 8, the Σ_2 relationships ≤ 10000, and the number of grids in one section $\leq 200 \times 200$. This program needs 520 kbyte of memory and 2 Mbyte of disk space. A laser printer with HP-GL emulation is needed for drawing the two-dimensional projection of these structures. DIRECTER consists of a main segment and DS*SYSTEM libraries totaling about 29 000 source code lines. Documentation and the latest executable file are available from the author.

ACKNOWLEDGMENT

REFERENCES AND NOTES

(1) Okada, K. Application of Automated Structure Analysis to Some Organic Compounds Using PCs: 1. Automated Patterson Techniques. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 871–878.

(2) Okada, K.; Okada, S.; Koyama, H. 32-bit Personal Computer Version of DIRECT-SEARCHER System. *Acta Crystallogr., Sect. A* **1990**, *A47*, C70.

(3) Okada, K.; Koyama, H. 32 bit Personal Computer Version of the DIRECT-SEARCHER System. *J. Appl. Crystallogr.* **1991**, *24*, 1067–1070.

(4) Karle, J.; Karle, I. L. The Symbolic Addition Procedure for Phase Determination for Centrosymmetric and Noncentrosymmetric Crystals. *Acta Crystallogr.* **1966**, *21*, 849–859.

(5) Germain, G.; Woolfson, M. M. On the Application of Phase Relationships to Complex Structures. *Acta Crystallogr., Sect B* **1968**, *B24*, 91–96.

(6) Prick, P. A. J.; Beurskens, P. T. The Application of Direct Methods to Difference Structures. Non-Centrosymmetric Structures Containing Heavy Atoms. III. Enantiomorph and Origin Fixation", *Acta Crystallogr., Sect. A* **1983**, *A39*, 570–576.

(7) Debaerdemaeker, T.; Tate, C.; Woolfson, M. M. On the Application of Phase Relationships to Complex Structures. XXXVI. Developments of The Sayre-Equation Tangent Formula. *Acta Crystallogr., Sect. A* **1988**, *A44*, 353–357.

(8) Pontenagel, W. M. G. F.; Krabbendam, H.; Kroon, J. REDUCE: An Automated Procedure for the Determination of Structure-Factor Phases from the Estimated Values of Structures (Sem)invariant Phase Combinations. *Acta Crystallogr., Sect A* **1987**, *A43*, 60–64.

(9) Peschar, R.; Shenk, H. Optimal Symbolic Phase Determination. *Acta Crystallogr., Sect. A* **1987**, *A43*, 751–763.

(10) Gilmore, C. J.; Brown, S. R. Developments in the MITHRIL Direct Methods Program. *J. Appl. Crystallogr.* **1988**, *21*, 571–572.

(11) Burla, M. C.; Camalli, M.; Cascarano, G.; Giacovazzo, C.; Polidori, G.; Viterbo, D. SIR88: A Direct-Methods Program for the Automatic Solution of Crystal Structures. *J. Appl. Crystallogr.* **1989**, *22*, 389–393.

(12) Hai-Fu, F.; Jin-Zi, Q.; Chao-De, Z.; Yuan-Xin, G.; Heng-Ming, K.; Sheng-Hua, H. The Program Sapi and its Applications. III. Determination of Structures having Pseudo-Translational Symmetry. *Acta Crystallogr., Sect. A* **1990**, *A46*, 99–103.

(13) Sheldrick, G. M. Phase Annealing in SHELEX-90: Direct Methods for Larger Structures. *Acta Crystallogr., Sect. A* **1990**, *A46*, 467–473.

(14) Koyama, H.; Okada, K. Automatic Heavy-Atom Analysis of Some Organic Compounds. *Acta Crystallogr., Sect. B* **1970**, *B26*, 444–447.

(15) Koyama, H.; Okada, K. Automation of the Crystal Structure Analysis. The DIRECT-SEARCHER Program System. *Acta Crystallogr., Sect. A* **1975**, *A31*, S18.

(16) Koyama, H.; Okada, K. Automatic Determination of Crystal Structure, I. Heavy-atom Method, II. Direct-Method. *Acta Crystallogr., Sect. A* **1978**, *A34*, S40.

(17) Motherwell, W. D. S. SNOOPY: A Selection of a Starting Set for Tangent Formula Calculations. Private communication, 1971.

(18) Drew, M.; Larson, A.; Motherwell, S. TANTWO: Generation and Refinement of Phases by the Tangent Formula. University of Cambridge, England, 1971.

(19) Main, P.; Woolfson, M. M.; Germain, G. MULTAN71: A Computer Program for the Automatic Solution of Crystal Structures. University of York, England, and Louvain, Belgium, 1971.

(20) Brenner, S. A.; Gum, P. H. TANGFORM, ASIG: The Tangent Formula Program for the X-ray Analysis of Non-Centrosymmetric Crystals. NRL Report 6697, Naval Research Laboratory: Washington, DC, 1968.

(21) Koyama, H.; Okada, K. SEARCHER: Automatic Heavy-Atom Analysis for the Determination of the Crystal Structure; Osaka, Japan, 1970.

(22) Okada, K.; Koyama, H. Structure of 17β-Acetoxy-18-methyl-1-thia-A-norestra-3,9(10),11-trien-2-one. *Acta Crystallogr., Sect. C* **1993**, *C49*, 1123–1125.

(23) Okada, K.; Koyama, H. Crystal Structure and Conformational Analysis of 7,13-abietadien-18-oic acid. *Bull. Chem. Soc. Jpn.* **1994**, *67*, 807–815.

(24) Main, P.; Hull, S. E.; Lessinger, L.; Germain, G.; Declercq, J-P.; Woolfson, M. M. MULTAN78: A System of Computer Programms for the Automatic Solution of Crystal Structures from X-ray Diffraction Data. University of York, York, England, 1978.

(25) Symersky, J.; Kriz, V. A. Test of MULTAN87. *J. Appl. Crystallogr.* **1989**, *22*, 435–438.