Towards automated protein structure determination: *BnP*, the *SnB*-PHASES interface

C. M. Weeks*,[I], R. H. Blessing[I], R. Miller[I,II], R. Mungee[I], S. A. Potter[I], J. Rappleye[II], G. D. Smith[I,III], H. Xu[I], and W. Furey[IV]

[I] Hauptman-Woodward Medical Research Institute & Dept. of Structural Biology, SUNY at Buffalo, 73 High St., Buffalo, NY 14203, USA
[II] Center for Computational Research, Norton Hall Rm. 9, SUNY at Buffalo, Buffalo, NY 14260, USA
[III] Structural Biology and Biochemistry, Research Institute, Hospital for Sick Children, 555 University Avenue, Toronto, Ontario, M5G 1X8, Canada
[IV] Biocrystallography Laboratory, VA Medical Center, University Drive C, Pittsburgh, PA 15240 and Dept. of Pharmacology, University of Pittsburgh, Pittsburgh, PA 15261, USA

Received July 24, 2002; accepted August 29, 2002

Abstract. The direct-methods program *SnB* provides an efficient means for solving protein substructures containing many heavy-atom sites (current record: 160). In order to meet the high-throughput requirements of structural genomics projects, substructure determination needs to be tightly integrated with other aspects of the protein-phasing process. This has been accomplished through the design of a common Java interface, *BnP*, for *SnB* and components of PHASES, a popular and proven program suite that provides all the tools necessary to proceed from substructure refinement to the computation of an unambiguous protein electron-density map. Therefore, *BnP* will facilitate a high degree of automation and enable rapid structure determination by both experienced and novice crystallographers.

Introduction

In the absence of a suitable model for molecular replacement, the determination of a new protein structure is typically a two-step process. If two or more intensity measurements are available for each reflection with differences arising from some property of a small substructure, then the positions of the substructure atoms can be found first and used as a bootstrap to initiate the phasing of the complete structure. Suitable substructures may consist of heavy atoms soaked into a crystal, and the intensity measurements can be made from both unsubstituted (native) and substituted (derivative) crystals. Alternatively, the substructure may consist of anomalous scatterers, such as selenium in the form of selenomethionine, that have been incorporated into the crystal, and measurements of anomalous dispersion can be made at one or more wavelengths.

Heavy-atom substructures can be solved using computational procedures that are based on either Patterson or direct methods. In either case, the positions of the substructure atoms are determined from isomorphous or anomalous difference coefficients. Although either computational method can be used effectively for small substructures (*e.g.*, less than 20 sites), direct methods tend to be faster, and their relative efficiency increases as the size of the substructure increases. So far, the largest substructure that has been shown to be solvable by Patterson-based methods contained 66 Se sites (T. Terwilliger, P. Adams, personal communications) whereas the largest substructure that has been solved by direct methods contained 160 Se sites (von Delft, Blundell, 2002).

Substructure solution by *Shake-and-Bake*

Shake-and-Bake is a powerful algorithmic formulation of direct methods that, given diffraction data to 1.2 Å or better resolution, has made possible the *ab initio* phasing of complete crystal structures containing as many as ∼750 independent non-H atoms no heavier than oxygen (Gessler, *et al.*, 1999) or ∼2000 independent atoms provided that several sulfur or iron atoms are present (Frazão, *et al.*, 1999). The distinctive feature of *Shake-and-Bake* is the repeated and unconditional alternation of reciprocal-space phase refinement (*Shaking*) with a complementary real-space process (*Baking*) that seeks to improve phases by imposing constraints through a physically meaningful interpretation of the electron density (Miller, *et al.*, 1993; Weeks, *et al.*, 1994). This automated recycling technique has proven to be considerably more effective than older direct methods that relied on phase refinement alone.

Shake-and-Bake belongs to the class of phasing methods known as 'multisolution' procedures (Germain, Woolfson, 1968). Multiple trial structures are created by using a random-number generator to assign initial coordinates, and each trial is then subjected to the dual-space improvement process. Phases are refined either by the tangent formula (Karle, Hauptman, 1956) or by constrained minimization of the so-called minimal function (DeTitta, *et al.*, 1994) using the parameter-shift algorithm (Bhuiya, Stanley, 1963). In real space, peak picking is used to impose the

* Correspondence author (e-mail: weeks@hwi.buffalo.edu)

atomicity constraint. The success rate of this process (*i.e.*, the percentage of trial structures that converge to solution) depends on data quality and the size of the structure. Solutions are identified on the basis of the value of a suitable figure of merit such as the minimal function or the crystallographic *R* value. The complete *Shake-and-Bake* algorithm has been described in detail in recent reviews (Weeks, *et al.*, 2001; Sheldrick, *et al.*, 2001).

It has been recognized for some time that the formalism of direct methods carries over to protein substructures when applied to single isomorphous replacement (SIR) (Wilson, 1978) or single anomalous scattering (SAS or SAD) (Mukherjee, *et al.*, 1989) difference data. Multiple isomorphous replacement (MIR) data can be accommodated simply by treating the data separately for each derivative, and multiple anomalous dispersion (MAD) data can be handled by examining the anomalous differences, $|\Delta F_{ANO}|$, for each wavelength individually or by combining them together (along with dispersive differences) in the form of F_A structure factors (Karle, 1989; Hendrickson, 1991). The dispersive differences between two wavelengths of MAD data also can be treated as pseudo-SIR differences. The resolution of the data typically collected for isomorphous replacement or MAD experiments is sufficient for substructure determinations since it is rare for heavy atoms or anomalous scatterers to be closer than 3–4 Å.

The *Shake-and-Bake* procedure has been implemented in the computer program, *SnB*, in a manner convenient for substructures as well as complete structures (Miller, *et al.*, 1994; Weeks, Miller, 1999; Rappleye, *et al.*, 2002). (The *Shake-and-Bake* algorithm has also been implemented independently in the program SHELXD (Schneider, Sheldrick, 2002)). The *SnB* graphical user interface (GUI) controls not only the main phasing program but also the DREAR program suite (Blessing, Smith, 1999, and references therein) that computes the normalized structure-factor magnitudes ($|E|$) and normalized difference structure-factor magnitudes ($|E_\Delta|$) required for direct-methods calculations.

Protein phasing with PHASES

Heavy-atom substructures determined by *SnB* using isomorphous replacement and/or anomalous scattering information provide the starting point for PHASES, a suite of 44 individual Fortran programs that can do everything else necessary to produce interpretable protein maps (Furey, Swaminathan, 1997). These programs, which can be combined in many ways, communicate through files having a common format. Although all programs in the package can be run standalone, many are often chained together through shell scripts, and template scripts are provided for common iterative procedures. All output is entered into a single running log file so that it is easy to maintain a complete history of the calculations performed.

The main tasks addressed by PHASES include (1) the merging and scaling of native and derivative data, (2) the refinement of heavy-atom positional, thermal, and occupancy parameters against all available isomorphous and anomalous differences, (3) the computation of protein phase angles based on these differences, (4) the improvement or extension of protein phases by solvent flattening with negative density truncation (Wang, 1985), and (5) phase improvement and extension by means of non-crystallographic symmetry (NCS) averaging (Rossmann, Blow, 1963). Partial structure phase combination with a variety of weighting options can be carried out as an aid in structure completion. A fully automated version of the solvent flattening/negative-density truncation procedure includes three solvent-mask iterations with a total of 16 phase-combination cycles. Averaging envelope masks can be created and manipulated in a variety of ways, NCS operators can be refined, and phases combined with those from another source.

In addition to the main computational programs, the PHASES package includes auxiliary programs for porting information to/from other software, for assessing the phasing results, for displaying contoured electron-density or Patterson maps, for editing solvent and averaging masks, and for creating electron-density maps and skeletons for use in other programs. The interactive program MAPVIEW is an especially powerful graphical tool with a menu that permits the user to scroll quickly through map sections, to change direction, and to superimpose sections creating projections. These and other features allow one to identify a region of the cell containing a single molecule, to trace and display averaging masks for every section, and to verify that all points within a mask are unique.

The *BnP* interface

In order to increase automation, the two-step process of substructure determination and complete protein phasing must be combined within a single program. This need has motivated the design of *BnP*, a common graphical user interface (GUI) for *SnB* and components of the PHASES suite. (*BnP* is an acronym for ***B***uffalo a***n***d ***P***ittsburgh, the home cities of the program developers.) The *BnP* GUI is modeled after the familiar *SnB* GUI (Weeks, Miller, 1999) that has been in use for several years. The *BnP* interface, itself written in Java, initiates the writing of shell scripts that then control the actual crystallographic calculations performed by the back-end Fortran executables (Fig. 1).

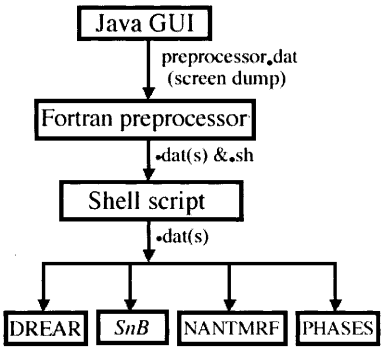

Fig. 1. The *BnP* GUI initiates a crystallographic computation by writing information from the interface screens into a data file for a Fortran preprocessing program. The preprocessor then creates individual data files for all the necessary Fortran crystallographic programs as well as a shell script that will proceed to execute them in the proper sequence. NANTMRF (Smith, 2002) is a program for comparing the sites in different *SnB* trial substructures.

The *BnP* interface consists of several information screens or pages (Table 1), two of which are illustrated in Fig. 2. Different screens are selected by clicking the top buttons, and normally the screens are used in order from left to right. The bottom buttons found on each of the *BnP* screens allow the user to choose between two operating modes, to save and restore the contents of the screens themselves, and to access the on-line documentation. In manual mode, the user can change the values of many parameters, and the major steps in the phasing process are executed sequentially by clicking several buttons on successive interface screens. In automatic mode, the user has access to only a few parameters, and the entire phasing process from data normalization through phase refinement and solvent flattening is chained together and started by clicking a single button on the *Auto Run* screen. Creation of the automatic operating mode has required a fine tuning of the criteria for recognizing when a substructure solution has occurred, the development of procedures for validating sites, and methods for determining whether or not a substructure determined by *SnB* has the proper hand.

Use of the *BnP* package is best illustrated by a three-wavelength MAD phasing example. After crystal-specific information about the structure and its data sets has been entered (Fig. 2a), the interface automatically tabulates the types of difference amplitudes that are available as the input data for substructure determination. The user can then select one or more of these for use in *SnB* (normally, the peak-wavelength anomalous differences would be used first) as well as prepare all difference data for refinement

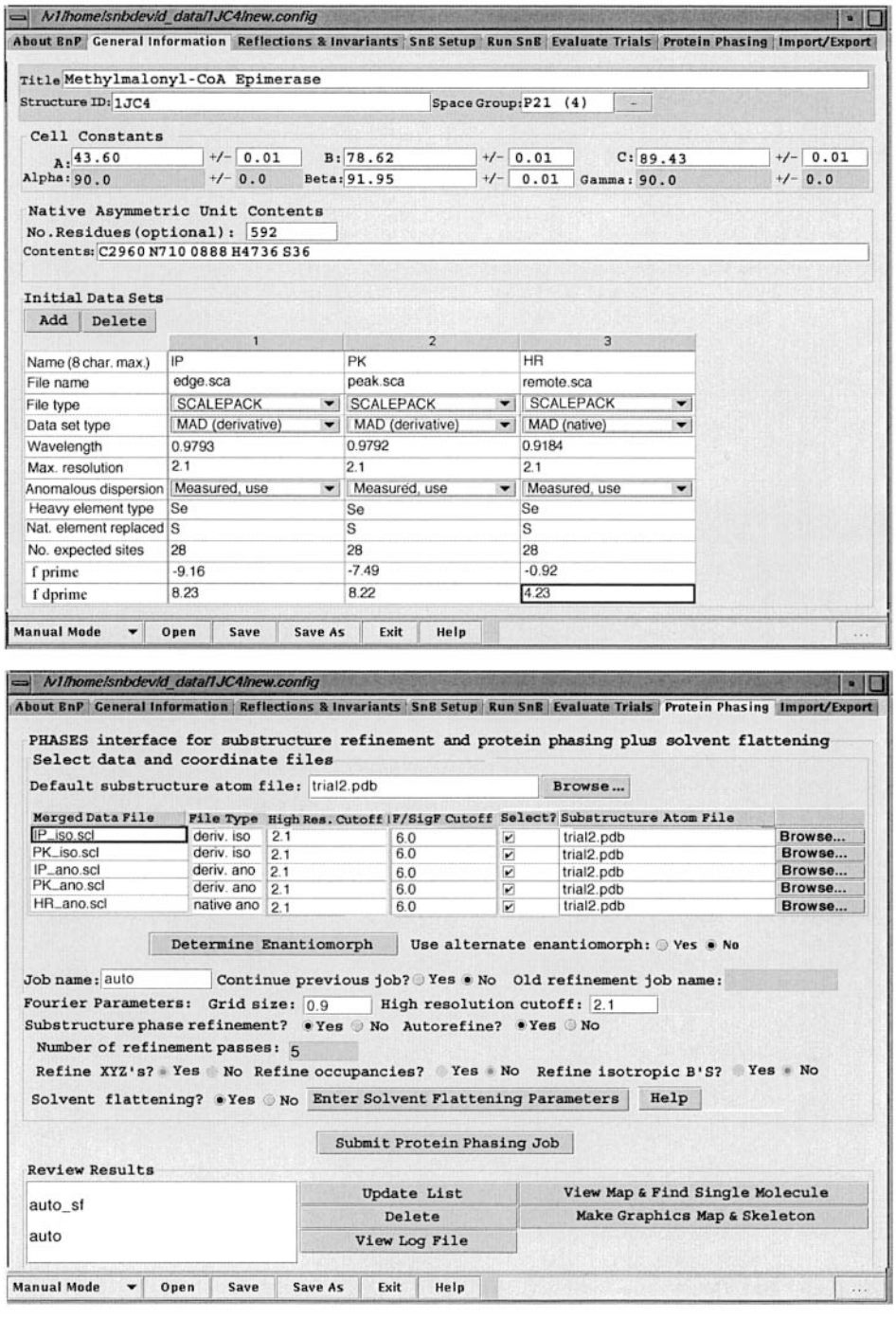

Fig. 2. (**a**) A snapshot of *BnP's General Information* screen for a MAD application to structure MMEPI (McCarthy, *et al.*, 2001). This screen contains all the crystal-specific information that the user must enter. Once this information is supplied, the interface will provide default values for all the operational parameters. Note that the information for all data sets (in this case, all wavelengths) is supplied immediately even if only one data set will be used for the substructure determination. (**b**) A snapshot of the *Protein Phasing* screen. Se sites selected from an *SnB* run have been placed in a PDB file. After determining the correct hand (*Determine Enantiomorph* button), these sites can be refined in the PHASES subprogram PHASIT against all anomalous and isomorphous (dispersive) differences using the autorefinement option. Phase refinement followed by solvent flattening is initiated by clicking the button, *Submit Protein Phasing Job.* When the job is completed, the results may be reviewed by viewing the log file or by viewing a contoured map using the PHASES subprogram MAPVIEW via the *View Map & Find Single Molecule* button. The *Make Graphics Map and Skeleton* button can be used to generate map and skeleton files for use in the program O (Jones, *et al.*, 1991).

Table 1. *BnP* GUI screens or pages.

Screen	Function
General Information	Enter crystal-specific data
Reflections & Invariants	Normalize data; generate three-phase structure invariants; prepare difference files for PHASES
SnB Setup	Enter or check *SnB* parameters
Run *SnB*	Choose multiprocessing options and execute *SnB* job(s)
Evaluate Trials	Identify and compare substructure solutions; improve site model by occupancy refinement
Protein Phasing	Determine enantiomorph; refine substructure; protein phasing; solvent flattening; view map; create skeleton
Import/Export	Communicate with other programs
Auto Run	Automatic mode setup, job submission, and review of results

and protein phasing in PHASES (Fig. 2b). In manual mode, the steps involved in substructure determination are difference data normalization using the DREAR program package, three-phase structure invariant generation, and the processing of trial substructures using the dual-space *Shake-and-Bake* algorithm. In order to maximize throughput during substructure phasing and minimize the real time required to achieve a solution, all the multiprocessing options of *SnB* (Rappleye, *et al*., 2002) are available from the *Run SnB* screen and, if desired, jobs can also be run concurrently using different difference data sets provided that sufficient computing resources are available.

In the following sections, several aspects of *BnP* operation are discussed in detail and illustrated by applications to test data sets (Table 2).

Scoring trial substructures

SnB computes three figures of merit that allow the quality of a trial substructure to be judged and a decision to be made about whether or not a solution has been found. These figures of merit are the minimal function (DeTitta, *et al*., 1994), a crystallographic *R* factor ($R_{cryst} = (\sum ||E_{obs}| - |E_{calc}||)/\sum |E_{obs}|$, and a correlation coefficient (*CC*) between $|E_{obs}|$ and $|E_{calc}|$ (Fujinaga, Read, 1987). The minimal function, R_{min}, is a measure of the mean-square difference between the values of the three-phase structure invariants calculated using a set of trial phases and the expected probabilistic values of the same invariants. Typically, solutions have the smallest values of R_{min} and R_{cryst}, and they have the largest values of *CC*. In manual mode, the *BnP* interface provides several tools for determining whether a solution has occurred. For example, a histogram of the R_{min} values for all trials that have been processed by an *SnB* job can be displayed as illustrated in Fig. 3 for the peak-anomalous difference data for the 30-site test structure, ADOHCY. A clear bimodal distribution of R_{min} values is a strong indication that a so-

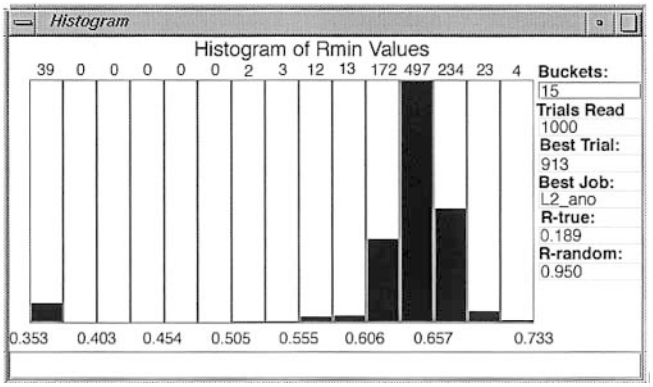

Fig. 3. This bimodal histogram of minimal function (R_{min}) values for 1000 trial substructures for ADOHCY suggests that there are 39 solutions. R_{TRUE} and R_{RANDOM} are theoretical values for true and random phase sets, respectively (Weeks, *et al*., 1994).

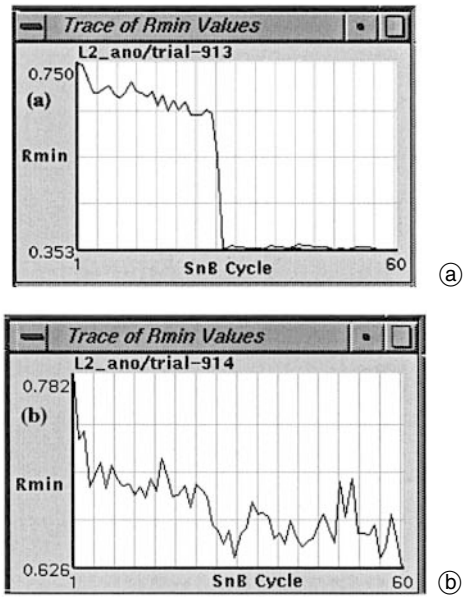

Fig. 4. Plots of the minimal-function value over 60 cycles (**a**) for a solution (trial 913) and (**b**) for a nonsolution (trial 914) for ADOHCY.

Table 2. MAD data sets for selenomethionyl-substituted proteins that were used to test the *BnP* interface.

Protein	Code	PDB#	d (Å)	Actual Se sites	Reference
Replication protein A	RPA	1QUQ	2.5	20	Bochkarev, *et al*., 1999
Methylmalonyl-CoA epimerase	MMEPI	1JC4	2.1	24	McCarthy, *et al*., 2001
Malic enzyme	MALIC	1QR6	2.5	28	Xu, *et al*., 1999
S-adenosylhomocysteine hydrolase	ADOHCY	1A7A	2.8	30	Turner, *et al*., 1998
Pyruvate dehydrogenase: component E1	E1	1L8A	2.6	40	Arjunan, *et al*., 2002
Human HMG-CoA reductase	HMGR	1DQ8	2.33	60	Istvan, *et al*., 2000
Tryparedoxin peroxidase	TRYP	1E2Y	3.2	60	Alphey, *et al*., 2000
2-Aminoethylphosphonate transaminase	AEPT	–	2.55	66	Chen, *et al*., 2000

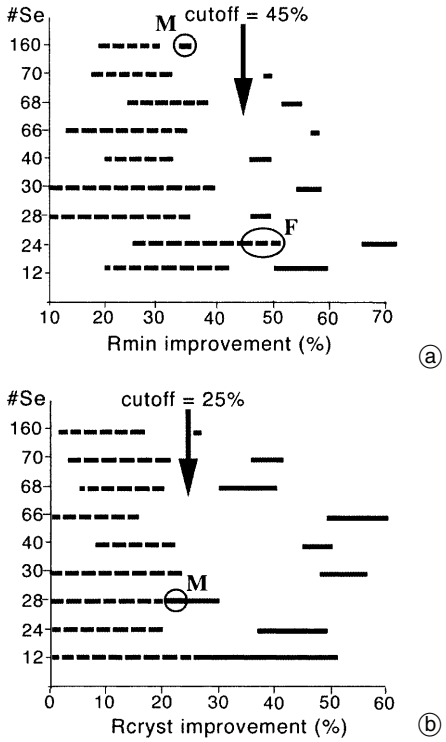

Fig. 5. The ranges of (**a**) $R_{\min}$ and (**b**) R_{cryst} improvement-factor values for solutions (———) and nonsolutions (– – –) for several SeMet substructures. Circles labeled M indicate the solutions that would be missed, and ellipse F indicates the false solutions that would be found if the specified cutoff values were used to distinguish solutions from nonsolutions.

lution has, in fact, been found. Confirmation that this is true for trial 913 in the Fig. 3 example can be obtained by inspecting a trace of the $R_{\min}$ value as a function of refinement cycle (Fig. 4). Solutions usually show an abrupt decrease in value over a few cycles followed by stability at the lower value.

For efficiency when operating in automatic mode, it would be desirable to recognize a solution based on the figures of merit for a single trial without having to wait for a full histogram to develop. However, although the relative values of each figure of merit for any given substructure clearly distinguish solutions from nonsolutions, the exact figure-of-merit values for solutions may vary for different data sets. Discrimination can be improved somewhat by considering the percentage improvement $[100(R_{\text{init}} - R_{\text{final}})/R_{\text{init}}]$ during the course of *SnB* refinement rather than the final figure-of-merit value itself. Fig. 5 illustrates how the application of appropriate cutoff values to the improvement factors for both $R_{\min}$ and R_{cryst} can ensure that false solutions are never selected for a series of SeMet substructures of varying size. On the other hand, if these conservative criteria are used, solutions could not be recognized automatically for the largest substructure, but this is not a serious problem since manual intervention would easily reveal when a solution had occurred.

Site validation

For each job involving an *N*-site substructure, *SnB* provides an output file of 1.5*N* peak positions for the best trial (based on its final $R_{\min}$ value) sorted in descending order according to the corresponding electron density. It is then necessary to decide which, and how many, of these peaks correspond to actual atoms. The first *N* peaks have the highest probability of being correct, and in some cases this simple guideline is adequate. Alternatively, a conservative approach is to accept the 0.8*N* to 0.9*N* top peaks.

Peaks consistently occurring in several independent trial solutions are the most likely to correspond to real atomic sites. When operating in manual mode, the *BnP* interface provides the useful feature of trial comparison using the program NANTMRF (Smith, 2002), which takes into account the fact that different solutions may have different origins and/or enantiomorphs. Fig. 6 shows sample output from the *Compare Trials* option for structure MMEPI. The sequence of this protein indicates that each of the four monomers in the asymmetric unit could have seven Se sites ($N = 28$), but comparison of the best trial obtained from the peak-wavelength anomalous difference data with three other solutions reveals that only the top 24 peaks consistently have matches. In fact, four SeMet residues were indeed missing in the published structure as deposited in the Protein Data Bank.

The *Improve Model* option, available in both the manual and automatic modes, is a useful feature for eliminating spurious sites. The PHASES subprogram PHASIT is used to perform occupancy refinement against one of the difference data sets, generally the anomalous difference data set having the largest value of f''. The occupancy of spurious peaks will typically fall to very low values when refined against anomalous or isomorphous (dispersive) differences. This refinement is very rapid and has the advan-

Compare Trials Results

SnB Trial	2	12		17		5	
Number of Matches		27		26		25	
Mean distance		0.23		0.16		0.23	
Peak	Select ?	Peak	Distance	Peak	Distance	Peak	Distance
21	☑	11	0.49	12	0.52	13	0.53
22	☑	20	0.17	21	0.13	20	0.09
23	☑	24	0.88	22	0.17	23	0.53
24	☑	23	0.18	24	0.18	24	0.6
25	☐						
26	☐						
27	☐						
28	☐	42	0.28				

Save Close

Fig. 6. Results of the *Compare Trials* option for structure MMEPI. Originally, 28 sites were sought, but when the matches between the top 1.5*N* peaks in four solutions were examined, it became clear that only the top 24 peaks corresponded to actual Se sites. By checking the appropriate selection boxes for the first solution (#2), the user can choose which peaks are to be retained for protein phasing.

Structure	Success rate (%)	Theoretical sites (N)[a]	Actual sites[b]	Correct sites identified	
				N peaks	1.5N peaks
RPA	2.2	20	20	16	16
MMEPI	27.8	28	24	24	24
MALIC	0.8	28	28	27	27
ADOHCY	3.6	32	30	30	30
E1	2.5	42	40	39	39
HMGR	6.4	68	60	50	51
TRYP	0.05	70	60	42	43
AEPT	4.1	66	66	66	66

a: Potential sites based on the amino acid sequence
b: Number of sites reported in the published protein structure

Table 3. *SnB* success rates and site identification for substructure solutions using peak-wavelength anomalous differences.

tage of being hand insensitive. Thus, it need be carried out only once prior to enantiomorph determination.

Table 3 summarizes the success rates (percentages of trial substructures recognized as solutions) for each of the test data sets. Trial comparison and occupancy refinement were used to select the probable sites which were then compared to the true sites. In all cases examined so far, all false sites among the 1.5N *SnB* peaks could be distinguished from the true sites using the occupancy cutoff of 0.2 for refinement against peak-wavelength anomalous differences, and in fact most false sites refine to occupancies less than 0.05.

Enantiomorph determination

SnB substructure determination uses the magnitudes of difference data. Therefore, the probability that any given solution will have the correct hand is 50%, and other information must be used to select the proper hand (enantiomorph). When the *Determine Enantiomorph* option is selected, unrefined protein phases are computed and used to generate protein maps for both enantiomorphs, protein/solvent envelope masks are created for each using the PHASES subprogram BNDRY via the protein-solvent boundary determination algorithm, and the standard deviations of the electron densities in the protein and solvent regions are computed as well as the ratio σ(protein)/σ(solvent). This ratio should be higher for the correct enantiomorph (when anomalous scattering data are included) since atomic sites and gaps between chains within the protein region are expected to show large variations whereas solvent regions should be relatively flat with little variation. This expectation has been found to be the case for all examples tested so far. It should be noted that the standard deviation ratio defined here is similar to, but not identical to, criteria used in programs SOLVE (Terwilliger, Berendzen, 1999) and SHELXE (Sheldrick, 2002). Also, note that this method determines the protein/solvent envelopes by considering only relative mean density heights, and not local fluctuations. The fluctuations are used independently only for enantiomorph discrimination. This overcomes possible correlation problems that can occur with procedures using density fluctuations for protein/solvent envelope construction.

Although previous phase refinement does increase the discriminatory power of the standard deviation ratio, it is clear that the correct hand can also be determined prior to refinement. This is important because the substructure refinement/protein phasing process can be time consuming (particularly for large structures or structures with high resolution data), and it is advantageous to determine the proper hand early in order to avoid wasting time refining/phasing the wrong enantiomorph. Typical results are displayed for the SeMet test data sets in Table 4, and these results show that the ratio is a robust discriminator even when challenged deliberately by *SnB* solutions having both missing and false sites (*i.e.*, skipping the site validation step and blindly accepting the first N peaks from a single solution).

Within the context of *BnP*, it is also easy to verify the correctness of the enantiomorph selection by using the *View Map* option (subprogram MAPVIEW). Fig. 7 shows comparable projections of density for both enantiomorphs for structure MMEPI as well as projections of the same region for the correct hand after refinement and after refinement plus solvent flattening. It is clear that the solvent region is visible even in the unrefined map for the correct hand, and the excellent quality of the map after solvent flattening is apparent.

Structure	Sites used	Sites incorrect	σ(protein)/σ(solvent)			
			Correct hand		Other hand	
			Unrefined	Refined	Unrefined	Refined
RPA	20	4	1.31	1.46	1.12	1.14
MMEPI	28	4	1.47	1.59	1.16	1.20
MALIC	28	1	1.42	1.52	1.17	1.18
ADOHCY	32	2	1.86	2.29	1.13	1.17
E1	42	3	1.55	1.73	1.09	1.11
HMGR	68	17	1.33	1.50	1.10	1.11
TRYP	70	27	1.54	1.61	1.12	1.14
AEPT	66	0	1.81	1.95	1.14	1.18

Table 4. Ratios of the standard deviations of the electron density in the protein and solvent regions for selected *SnB* solutions.

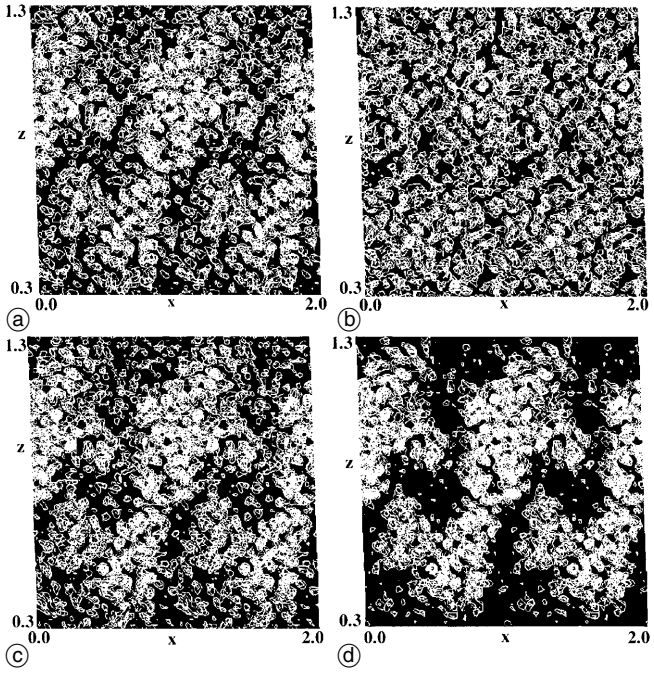

Fig. 7. Projections of the electron density ($y = 0$ to 0.111) for structure MMEPI. (**a**) Correct hand with no phase refinement, (**b**) alternate hand with no phase refinement, (**c**) correct hand after refinement with the autorefine option, and (**d**) correct hand after autorefinement plus solvent flattening.

Protein phasing

Once the proper enantiomorph has been selected, the *BnP* interface provides convenient options for substructure/protein phase refinement using the PHASES subprogram PHASIT. The user can select which difference data files are to be used and the number of refinement passes. Positional, thermal, and occupancy parameters can be refined in any order and in any combination. Additional refinement can be continued at any time by using the output parameter file from one refinement as input to the next job. For convenience, the *Autorefine* option allows the user to select a standard refinement protocol by clicking a single button. This option performs a series of substructure/phase refinement runs optimizing scaling, positional, and thermal parameters. It also determines optimal values for expected lack-of-closure estimates and protein phases. Figure of merit, phasing power, and other statistics are provided for user assessment. If the solvent-flattening option is selected, the autorefine procedure then takes the refined phase set and uses it to carry out a standard solvent-flattening run. This involves automatic construction of multiple solvent masks and 16 cycles of solvent flattening/phase combination iterations. The resulting data are then available for map examination, skeletonization, or export to external programs. The results of applying the autorefine option to the test data are presented in Table 5.

Table 5. Final phasing statistics following site validation and the autorefine option (without solvent flattening). Initial sites found using peak anomalous difference data.

Structure	Sites used	Mean FOM	Mean phasing power
RPA	16	0.638	2.17
MMEPI	24	0.707	3.16
MALIC	27	0.736	2.19
ADOHCY	30	0.858	3.72
E1	39	0.674	2.33
HMGR	51	0.728	3.15
TRYP	43[a]	0.698	2.47
AEPT	66	0.808	4.10

a: Using the dispersive differences between the inflection-point and high-energy-remote data, 58 sites were found. Refinement of these sites gave a mean FOM of 0.707 and mean phasing power of 2.56.

Summary

The *BnP* interface provides an automated pathway for protein phasing beginning with initial substructure sites located by direct methods. This pathway is especially useful for MAD experiments involving large SeMet substructures. The powerful dual-space algorithm implemented in the *SnB* program nearly always yields a good starting substructure containing a substantial fraction of correct sites. Occupancy refinement against the peak anomalous data set then weeds out any spurious sites, the automated enantiomorph determination procedure identifies the correct hand, and full positional and thermal parameter refinement plus solvent flattening yields a set of accurate protein phases.

However, *BnP* is not intended to be used solely as a black box. Manual intervention is conveniently available allowing the user to modify parameters as needed and to examine the results step by step. Additional tools such as the trial-substructure comparison feature are available in manual mode, and more options will be added to later versions of the interface in order to aid the user in choosing correct sites in difficult cases. Future versions will also facilitate use of the NCS-handling routines available in the PHASES program package.

The *BnP* procedures are efficient as shown by applications to the test data sets on a relatively slow 300 mHz SGI R12000 processor. The time required for substructure solution is variable, depending on size and the quality of the data – especially the accuracy of the anomalous signal. The MMEPI substructure solves in a few seconds, most of the other test data sets (RPA, ADOHCY, E1, HMGR, AEPT) solve in an hour or less, and MALIC requires an average of a little more than two hours. For reasons that are unclear, the peak anomalous data for TRYP has a very low success rate, and about three days of computation were required to find a solution. (On the other hand, a solution can be obtained in 15 minutes using the dispersive differences between the inflection-point and high-energy-remote data sets. This is an excellent example of the advantage of a program that permits the various difference data sets to be explored simultaneously in a multiprocessing environment.) The occupancy refinement and enantiomorph determination steps each take a minute or two. The autorefinement option is also variable, requiring two hours for MMEPI, about 12 hours for E1 or HMGR, and 20 hours for TRYP. Solvent flattening requires 1–15 minutes depending on structure size. Thus, in most of the cases described here, the substructure and

proper hand could be determined in an hour, and the full substructure refinement and protein phasing would require, at most, an overnight run.

The *BnP* interface and all of its associated programs are available for a variety of UNIX and LINUX platforms from http://www.hwi.buffalo.edu/BnP/.

Acknowledgments. This research was supported by NIH grant GM-46733 and by computer resources of the Center for Computational Research at SUNY Buffalo. We thank K. Chandrasekhar and P. Lakshminarasimhulu who assisted with the applications.

References

Alphey, M. S.; Bond, C. S.; Tetaud, E.; Fairlamb, A. H.; Hunter, W. N.: The structure of reduced tryparedoxin peroxidase reveals a decamer and insight into reactivity of 2Cys-peroxiredoxins. J. Mol. Biol. **300** (2000) 903–916.

Arjunan, P.; Nemeria, N.; Brunskill, A.; Chandrasekhar, K.; Sax, M.; Yan, Y.; Jordan, F.; Guest, J.; Furey, W.: Structure of the pyruvate dehydrogenase multienzyme complex E1 component from *Escherichia coli* at 1.85 Å resolution. Biochemistry **41** (2002) 5213–5221.

Bhuiya, A. K.; Stanley, E.: The refinement of atomic parameters by direct calculation of the minimum residual. Acta Crystallogr. **16** (1963) 981–984.

Blessing, R. H.; Smith, G. D.: Difference structure factor normalization for heavy-atom or anomalous scattering substructure determinations. J. Appl. Cryst. **32** (1999) 664–670.

Bochkarev, A.; Bochkareva, E.; Frappier, L.; Edwards, A. M.: The crystal structure of the complex of replication protein A subunits RPA32 and RPA14 reveals a mechanism for single-stranded DNA binding. EMBO J. **18** (1999) 4498–4504.

Chen, C. C. H.; Kim, A.; Zhang, H.; Howard, A. J.; Sheldrick, G.; Dunaway-Mariano, D.; Herzberg, O.: Sixty-six Se atoms and 2160 amino acids in the asymmetric unit: structure of AEP-transaminase. Abstract 02. 06. 03, ACA Annual Meeting, St. Paul, MN (2000).

DeTitta, G. T.; Weeks, C. M.; Thuman, P.; Miller, R.; Hauptman, H. A.: Structure solution by minimal function phase refinement and Fourier filtering: theoretical basis. Acta Crystallogr. **A50** (1994) 203–210.

Frazão, C.; Sieker, L.; Sheldrick, G. M.; Lamzin, V.; LeGall, J.; Carrondo, M. A.: Ab initio structure solution of a dimeric cytochrome c3 from *Desulfovibrio gigas* containing disulfide bridges. J. Biol. Inorg. Chem. **4** (1999) 162–165.

Fujinaga, M.; Read, R. J.: Experiences with a new translation-function program. J. Appl. Cryst. **20** (1987) 517–521.

Furey, W.; Swaminathan, S.: PHASES-95: a program package for processing and analyzing diffraction data from macromolecules. Meth. Enzymol. **277** (1997) 590–620.

Germain, G.; Woolfson, M. M.: On the application of phase relationships to complex structures. Acta Crystallogr. **B24** (1968) 91–96.

Gessler, K.; Usón, I.; Takaha, T.; Krauss, N.; Smith, S. M.; Okada, S.; Sheldrick, G. M.; Saenger, W.: V-Amylose at atomic resolution: x-ray structure of a cycloamylose with 26 glucoses. (1999). Proc. Natl. Acad. Sci. USA **96** (1999) 4246–4251.

Hendrickson, W.: Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. Science **254** (1991) 51–58.

Istvan, E. S.; Palnitkar, M.; Buchanan, S. K.; Deisenhofer, J.: Crystal structure of the catalytic portion of human HMG-CoA reductase: insights into regulation of activity and catalysis. EMBO J. **19** (2000) 819–830.

Jones, T. A.; Zou, J. Y.; Cowtan, S. W.; Kjeldgaard, M.: Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Crystallogr. **A47** (1991) 110–119.

Karle, J.: Linear algebraic analyses of structures with one predominant type of anomalous scatterer. Acta Crystallogr. **A45** (1989) 303–307.

Karle, J.; Hauptman, H.: A theory of phase determination for the four types of non-centrosymmetric space groups 1*P*222, 2*P*22, 3$P_1$2, 3$P_2$2. Acta Crystallogr. **9** (1956) 635–651.

McCarthy, A. A.; Baker, H. M.; Shewry, S. C.; Patchett, M. L.; Baker, E. N.: Crystal structure of methylmalonyl-coenzyme A epimerase from *P. shermanii*: a novel enzymatic function on an ancient metal binding scaffold. Structure **9** (2001) 637–646.

Miller, R.; DeTitta, G. T.; Jones, R.; Langs, D. A.; Weeks, C. M.; Hauptman, H. A.: On the application of the minimal principle to solve unknown structures. Science **259** (1993) 1430–1433.

Miller, R.; Gallo, S. M.; Khalak, H. G.; Weeks, C. M.: SnB: crystal structure determination via Shake-and-Bake. J. Appl. Cryst. **27** (1994) 613–621.

Mukherjee, A. K.; Helliwell, J. R.; Main, P.: The use of MULTAN to locate the positions of anomalous scatterers. Acta Crystallogr. **A45** (1989) 715–718.

Rappleye, J.; Innus, M.; Weeks, C. M.; Miller, R.: SnB version 2.2: an example of crystallographic multiprocessing. J. Appl. Cryst. **35** (2002) 374–376.

Rossmann, M. G.; Blow, D. M.: Determination of phases by the conditions of non-crystallographic symmetry. Acta Crystallogr. **16** (1963) 39–44.

Schneider, T. R.; Sheldrick, G. M.: Substructure Solution with SHELXD. Acta Crystallogr. **D58** (2002) 1772–1779.

Sheldrick, G. M.: SHELXD and SHELXE. In notes of the EMBO course on *Automated Macromolecular Structure Solution*, Heidelberg, May (2002).

Sheldrick, G. M.; Hauptman, H. A.; Weeks, C. M.; Miller, R.; Usón, I.: *Ab initio* phasing. In *International Tables for Crystallography* (M. G. Rossmann and E. Arnold, eds.), Vol. F. Kluwer Academic Publishers, Dordrecht (2001) 333–345.

Smith, G. D.: Matching selenium-atom peak positions with a different hand or origin. J. Appl. Cryst. **35** (2002) 368–370.

Terwilliger, T. C.; Berendzen, J.: Discrimination of solvent from protein regions in native Fouriers as a means of evaluating heavy-atom solutions in the MIR and MAD methods. Acta Crystallogr. **D55** (1999) 501–505.

Turner, M. A.; Yuan, C.-S.; Borchardt, R. T.; Hershfield, M. S.; Smith, G. D.; Howell, P. L.: Structure determination of selenomethionyl S-adenosylhomocysteine hydrolase using data at a single wavelength. Nature Structural Biology **5** (1998) 369–375.

von Delft, F.; Blundell, T. L.: The 160 selenium atom substructure of KPHMT. Acta. Crystallogr. **A58** (Supplement) (2002) C239.

Wang, B.-C.: Solvent flattening. Meth. Enzymol. **115** (1985) 90–112.

Weeks, C. M.; DeTitta, G. T.; Hauptman, H. A.; Thuman, P.; Miller, R.: Structure solution by minimal function phase refinement and Fourier filtering: II. implementation and applications. Acta Crystallogr. **A50** (1994) 210–220.

Weeks, C. M.; Miller, R.: The design and implementation of SnB v2.0. J. Appl. Cryst. **32** (1999) 120–124.

Weeks, C. M.; Sheldrick, G. M.; Miller, R.; Usón, I.; Hauptman, H. A.: *Ab initio* phasing by dual-space direct methods. In *Advances in Structure Analysis* (R. Kužel and J. Hašek, eds.), Czech and Slovak Cryst. Assn., Praha (2001) 37–64.

Wilson, K. S.: The application of MULTAN to the analysis of isomorphous derivatives in protein crystallography. Acta Crystallogr. **B34** (1978) 1599–1608.

Xu, Y.; Bhargava, G.; Wu, H.; Loeber, G.; Tong, L.: Crystal structure of human mitochondrial NAD(P)$^+$-dependent malic enzyme: a new class of oxidative decarboxylases. Structure **7** (1999) 877–889.