Chapter 33

Structural Genomics of Pathogenic Protozoa: an Overview

Erkang Fan, David Baker, Stanley Fields, Michael H. Gelb, Frederick S. Buckner, Wesley C. Van Voorhis, Eric Phizicky, Mark Dumont, Christopher Mehlin, Elizabeth Grayhack, Mark Sullivan, Christophe Verlinde, George DeTitta, Deirdre R. Meldrum, Ethan A. Merritt, Thomas Earnest, Michael Soltis, Frank Zucker, Peter J. Myler, Lori Schoenfeld, David Kim, Liz Worthey, Doug LaCount, Marissa Vignali, Jizhen Li, Somnath Mondal, Archna Massey, Brian Carroll, Stacey Gulde, Joseph Luft, Larry DeSoto, Mark Holl, Jonathan Caruthers, Jürgen Bosch, Mark Robien, Tracy Arakaki, Margaret Holmes, Isolde Le Trong, and Wim G. J. Hol

The Structural Genomics of Pathogenic Protozoa (SGPP) Consortium aimed to determine crystal structures of proteins from trypanosomatid and malaria parasites in a high throughput manner. The pipeline of target selection, protein production, crystallization, and structure determination, is sketched. Special emphasis is given to a number of technology developments including domain prediction, the use of "co-crystallants," and capillary crystallization. "Fragment cocktail crystallography" for medical structural genomics is also described.

1. Introduction

The Structural Genomics of Pathogenic Protozoa (SGPP) Consortium (www.sgpp.org) focused on the determination of crystal structures of proteins from major eukaryotic tropical pathogenic protozoa, specifically:

- *Plasmodium* spp., in particular *Plasmodium falciparum*, which causes the most lethal form of malaria, and also *P. vivax*.
- *Leishmania* spp., causing various forms of leishmaniasis throughout the tropics and subtropics
- *Trypanosoma brucei*, the causative agent of sleeping sickness in Africa
- *Trypanosoma cruzi*, responsible for Chagas disease in Latin America

The importance of the pathogens under investigation becomes clear from the alarming statistics from the WHO web site (www.who.int/topics/en/):

- Malaria: Each year, approximately 300 to 500 million malaria infections lead to over 1 million deaths, and about 90% of these occur in Africa, especially among young children. The rapid spread of resistance to

From: *Methods in Molecular Biology, Vol. 426: Structural Proteomics: High-throughput Methods*
Edited by: B. Kobe, M. Guss and T. Huber © Humana Press, Totowa, NJ

antimalarial drugs, coupled with widespread poverty and weak health infrastructure means that mortality from malaria continues to rise in developing countries.

- Leishmaniasis is endemic in 88 countries on five continents with approximately 12 million persons infected. Visceral leishmaniasis, or "kala azar," results in a mortality rate of nearly 100% if left untreated. Since 1993 there has been a significant increase in both the geographical area in which the diseases occur as well as the number of people infected. Particularly worrisome is the tendency of coinfection with HIV and *Leishmania donovani*, the causative agent of visceral leishmaniasis.
- Sleeping sickness, or African trypanosomiasis, is caused by the *T. brucei* parasite and is a daily threat to more than 60 million people. Occurring only in sub-Saharan Africa, it is estimated that 300,000 to 500,000 people have the disease. The disease is fatal if not treated.
- Chagas disease, another form of trypanosomiasis, is caused by the *T. cruzi* parasite and has a wide distribution in Central and South America. Early diagnosis and treatment can be difficult, but are again essential. It is endemic in 21 countries, with 16–18 million people infected and 100 million people at risk.

The available treatments for the three "trypanosomatid" infections are often ineffective, toxic, and/or difficult to administer, and resistance to antimalarial drugs is widespread globally. Research to define new therapeutic targets will aid the discovery of better treatments for these neglected diseases. A structural genomics effort on these protozoan organisms is likely to provide critical information on the precise architecture of potential drug targets. With this in mind, the targets of SGPP are chosen in part on the basis of medical relevance, and partly on the basis of expectations for discovering novel sequence-to-fold relationships as outlined in the Protein Structure Initiative strategy (www.nigms.nih.gov/Initiatives/PSI).

The other major mission of SGPP was the development of new methodologies for high throughput structural genomics projects. All of the organisms targeted by SGPP are eukaryotes, which are well known to be difficult targets for structural biology. Additionally, the genomes of these organisms have only recently been sequenced and are still in the process of annotation. Because of these complicating factors, a substantial effort has been devoted to the development and testing of technologies and procedures for optimizing the various steps in the structural genomics effort.

2. Pipeline Overview

The main pipeline of SGPP consists of target selection, protein production, protein crystallization, x-ray data collection, and structure determination. These pipeline units are supported by an informatics unit that performs data archiving, sharing, and mining (Fig. 33.1). During a period of 4 years, a high-capacity and high throughput structural genomics pipeline was established for obtaining crystal structures of the targeted protozoan parasites, as evident by a significant increase in production seen throughout the SGPP pipeline during the time period from September 2004 through August 2005 (Table 33.1).

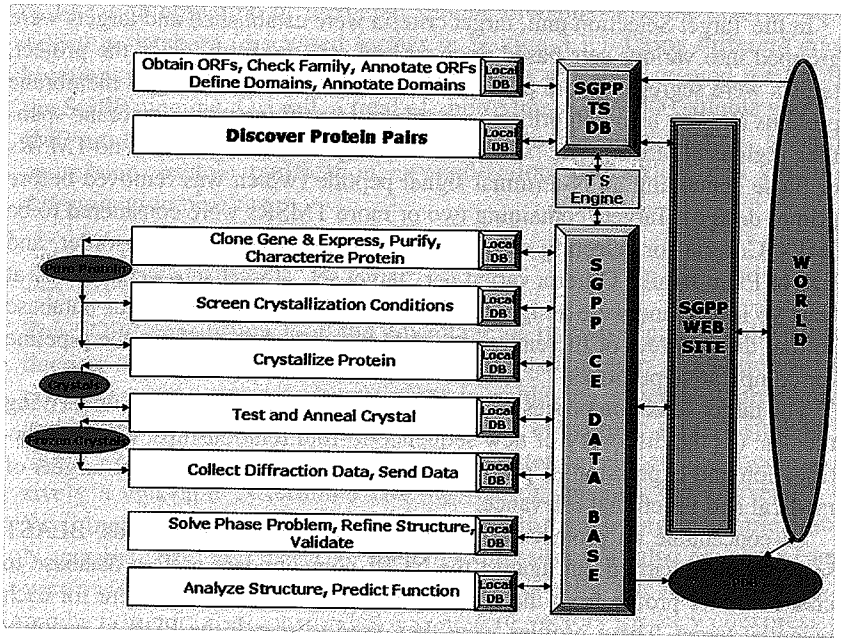

Fig. 33.1 Schematic diagram of the SGPP pipeline and data flow. DB, database; ORF, open reading frame; PDB, protein data bank; TS, target selection (CE, Central).

Table 33.1 Progress of all targets in SGPP (membrane proteins not included)

	As of August 2004: total successful targets	As of August 2005: total successful targets	Increase over 1 year (percent)
Cloned	5,471	10,767	97
Purified	404	839	108
Hits in crystal screen	100	194	94
Targets sent to crystallographers	57	123	116
Diffraction quality crystals	40	85	113
Total solved structures	15	40	167

2.1. Target Selection

One challenge for target selection in SGPP was that the genome sequences of the targeted organisms were not complete at the start of SGPP in 2001, with the complete genome sequence of *P. falciparum* 3D7 published in 2002 *(1)*, and the *L. major* Friedlin, *T. brucei* 927, and *T. cruzi* CL-Brener (jointly called the "Tritryp") genomes were published only toward the end of the SGPP project in 2005 *(2–5)*. It was very fortunate that several members of the SGPP consortium were actively involved in the genome sequencing and gene prediction and annotation projects for the trypanosomatids. This allowed the authors to select targets from extensively manually curated gene annotations of these organisms.

In the Target Selection unit, target criteria were established and targets were grouped into various sets based upon several criteria. Early on in the project, targets were segregated into predicted soluble proteins and integral membrane proteins (IMPs), using algorithms capable of identifying transmembrane spanning regions (TMSRs). The "solubles" are composed of targets with no TMSRs or those containing an N-terminal signal peptide (which was removed before primer design). Targets containing two or more TMSRs were considered to be IMPs. Criteria for selecting soluble targets included length, pI, disorder, and size of Pfam family. Criteria varied per "target set." Target sets were usually at least 96 in size and were assigned "set identifiers" in the SGPP status database so that the progress of the entire set could be tracked throughout the pipeline for comparative purposes.

The target selection process for soluble proteins addresses three underlying themes: (1) the discovery of new sequence-to-fold relationships; (2) the identification and expression of protein pairs; and (3) the generation of targets of medical and functional relevance.

To concentrate on proteins that are likely to represent novel folds, BLAST searches were carried out against the NCBI nonredundant protein database to create a score profile from the resulting multiple-sequence alignment for each potential target. These profiles were then used to search the PDB to quantify identity and similarity of the target sequence to proteins of known structure. In general, proteins were excluded from further analysis if the sequence identity was greater than 30%.

Information for protein-pair targets came from two sources. Experimentally, the SGPP consortium collaborated with industrial partners using yeast two-hybrid techniques to analyze numerous interactions in *P. falciparum (6)*. Computationally, from a database of interacting proteins in yeast, the authors' industrial partner provided a list of homologues in Tritryps that are likely to be engaged in pairwise interactions. These results from both sources were filtered for characteristics that would make it difficult to obtain soluble and crystallized protein pairs, such as size, significant stretches of low-complexity regions, and so on. However, the rejection criteria are less stringent than for the single solubles since, for example, low-complexity regions might become ordered in the course of complex formation.

For targets with medical and functional relevance, several approaches were implemented. One consisted of enzyme homologue discovery. Proteins annotated with an EC number in the genome database, those belonging to a cluster of orthologous genes (COG) that also contained a protein with an EC number, or those with PSI-BLAST e-values of 10^{-6} in the BRENDA enzyme database (www.brenda.uni-koeln.de), were selected as a set of "enzyme-like" targets. Second, homologues of known drug targets were identified in the genome database. Third, important protein targets were also solicited from the research community. The SGPP home page (www.sgpp.org) has an area where members of the research community can nominate proteins that they feel should be prioritized for expression, crystallization, and structure determination. Fourth, functional genomics information from publicly available sources, such as the RNAi Database at trypanofan.path.cam.ac.uk/ were consulted to select essential genes that lead to abnormal phenotypes.

For all full-length targets, the SGPP consortium applied computational domain prediction technologies and fed back chunks (defined as a single

or linear combination of predicted soluble domains) into the central target selection database (discussed further in Section 3.1.). Using different combinations of target selection criteria *(7)*, target sets were created for use by the SGPP pipeline. Overall, a total of approximately 19,000 soluble and 949 IMP-targets have been selected for entry into the amplification, cloning, expression, protein purification, crystallization, diffraction, and structure determination pipeline.

2.2. Protein Production

The SGPP consortium took the advantage of its two separate protein production centers for a controlled examination of protein production methodologies, which were critical issues in achieving successful production and crystallization of the generally difficult eukaryotic targets studied in SGPP. The University of Rochester (UR) center developed and used extensively a cleavable N-terminal His_6-tag vector, whereas the University of Washington (UW) center employed a vector encoding a noncleavable N-terminal His_6-tag. The UW protein expression unit also developed new vectors for expression in mammalian and insect cells. The Rochester protein production unit also constructed a system of vectors for easy coexpression of two proteins *(8)*. The results and procedures for cloning and expressing 1,000 malaria genes for protein production has been described by Mehlin et al. *(9)*. It appeared that production of soluble protein is favored by small size, large Pfam family and a pI below 8. In the later years of SGPP, both its protein production centers moved toward attacking the more difficult, medically relevant targets, as it was felt that structural genomics programs in general were neglecting these targets in favor of ones that were more immediately tractable. Technical innovations developed in the SGPP for high throughput protein production included cloning grills *(10)*, ligase-independent cloning vectors (Grayhack et al., unpublished) and flash-freezing of proteins in PCR plates *(11)*.

The organization of SGPP allowed a side-to-side comparison of cleavable vs. noncleavable expression systems. Initially, the two protein centers in SGPP were working on different targets, but lately an effort has been made to clone the same targets into the two vector systems. Although it was initially believed that the gains from this would be marginal, it was of interest to observe that these small sequence differences can have a profound effect on its solubility for some proteins, for example phosphodiesterases (PDEs).

Due to their demonstrated importance to the parasites and the ability for phosphodiesterase inhibitors (e.g., Viagra, Cialis) to be developed clinically, PDEs are high-value, potential drug targets in these organisms. SGPP took a set of 450 PDE variants as targets, which included multiple truncation variants of several of these enzymes, and put them into both the cleavable and uncleavable-tag vector systems, generating a total of 900 protein variants. A total of 13 of them were stable with only three of these soluble *within both* vector systems. (For the sake of this experiment, soluble proteins were only counted if they expressed to high enough levels and were isolated and shipped off for crystal screening trials.) The cleavable tag vector performed marginally better than the uncleavable-tag vector: generating ten as opposed to six soluble proteins. What is truly intriguing about this

work, however, is that a total of ten targets were soluble in one vector system but not in the other.

In addition, factors that affect soluble expression of these large numbers of eukaryotic proteins in *E. coli* were also revealed to some extent. For example, for plasmodium proteins, higher molecular weight, greater protein disorder, more basic isoelectric point, and a lack of homology to *E. coli* proteins all seem to correlate independently with difficulties in expression. In contrast, codon usage, and the percentage of adenosines and thymidines (which is high in *P. falciparum*) did not appear to affect soluble expression significantly *(9)*.

Overall, the two expression centers of SGPP successfully cloned over 10,700 targets out of approximately 11,150 for which primers were designed and ordered during the course of four years, and reached a throughput of producing more than 1,000 soluble eukaryotic protein samples (including repeated production for certain targets) in its final year in sufficient quantities for crystallization experiments.

2.3. Protein Crystallization

The protein crystallization process in SGPP consisted of two stages. The first stage was the high throughout screening (HTS) of initial lead conditions that may produce protein crystals. This was performed by the SGPP unit within the HTS crystallization screening laboratory at the Hauptman-Woodward Medical Research Institute (HWI) (www.hwi.buffalo.edu) *(12)*. The second stage was the follow-up crystal growth, which was performed by the crystallization unit of SGPP at the UW.

The HWI HTS lab uses robotic liquid handling systems extensively. Using a total of 400 µl of sample solution, the liquid handling systems prepare 1,536 well plates that are prepared with oil and 1,536 unique crystallization cocktail solutions. Each plate screens a single protein sample combined with 1,536 unique crystallization cocktail solutions mixed with protein under oil. Results from these micro batch under oil experiments are recorded (saved as images) over a 4-week time course. For SGPP, all images were screened manually for potential leads of crystallization. Cumulatively, the HWI HTS laboratory has been able to identify lead conditions for 194 out of more than 800 unique targets for SGPP, with very often multiple leads per targets.

The Crystal Growth Unit of SGPP followed two parallel paths. One was developing an optimal follow-up for the HTS hits found at HWI. The second was developing a new method for robotic protein crystallization in capillaries, which is briefly described in Section 3. Protein samples received from the SGPP Protein Production Units were immediately characterized systematically by UV spectroscopy, SDS PAGE, native gels, dynamic light scattering, and, for a limited number of cases, limited proteolysis. All the characterization data was archived in a database for future reference and data mining. Cumulatively, the SGPP Crystallization Unit received more than 1,100 protein samples for over 800 different targets. Diffraction quality crystals were obtained for approximately 123 of these targets and delivered to down stream units for data collection and structure determination. Interestingly, the degree of fragmentation in limited proteolysis experiments appeared to be a good predictor of the crystallization

success of protein targets: the more resistant to proteolysis, the greater the probability of crystal growth (O. Kalyuzhniy and W. Hol, unpublished observations).

2.4. Data Collection, Processing, and Structure Determination

SGPP data collection was done at two synchrotron sources: the Advanced Light Source (ALS) at the Lawrence Berkeley National Laboratory and the Stanford Synchrotron Radiation Laboratory (SSRL). Both synchrotron labs developed robotic equipment for crystal handling and loading including automated crystal annealing, streamlining procedures for users including remote data collection, and software for control of data collection, processing, and analysis *(13–15)*. Strategies for improving crystal quality included crystal annealing techniques *(16,17)* and cocrystallization with ligands. SGPP developed a special database (XRAYDB; Bosch et al., unpublished), which was used to track mounted crystals, cryoprotection procedures, results of screens at synchrotrons, and other early data in the structure determination process.

For high throughput data processing HKL2000 *(18)* and the ELVES automated scripts by Holton and Alber *(19)* were used. For structure determination, a variety of modern structural biology techniques were employed including SeMet-based phasing, bromide soak and other heavy atom phasing procedures, molecular replacement phasing and autotracing. Overall, the SGPP consortium collected 549 sets of diffraction data for 123 unique targets, and solved approximately 40 protein structures from the targeted organisms. In addition, a significant effort was made in developing "fragment cocktail crystallography" as described in Section 4.

SGPP explored the use of structural predictions from the ROSETTA program (boinc.bakerlab.org/rosetta) in high throughput structure determination. A specific example is SGPP target Tcru010945AAA from *T. cruzi*. This target has 27% sequence identity to an arginase with known structure (PDB entry 1cev) *(20)*. Attempts to use "1cev" for molecular replacement failed however. SAD phasing based on 9 Se sites yielded a relatively poor initial map, which nevertheless eventually led to a fully refined structure using SOLVE/RESOLVE *(21)* with the help from ROSETTA for prediction of loops and other connecting segments consistent with the partial model. ROSETTA was able to generate candidate loop conformations for the missing portions that agree quite well with the electron density in the map from SAD phasing, and it seems clear that automation of this approach could save much manual effort.

The authors also used ROSETTA to generate a full prediction for this same target Tcru010945AAA starting only from the sequence. This was not fully an *ab initio* prediction, as ROSETTA was able to identify and use suspected homologues, including 1cev, in the PDB. The CCP4 program MOLREP was able to find a molecular replacement solution from this model, and the resulting map is traceable. This is in distinct contrast to the failure to generate a usable map using the 1cev structure directly as a molecular replacement probe. This example does not address the more ambitious goal of solving a crystal structure using a purely *ab initio* prediction, but it does illustrate the value of ROSETTA in reducing the amount of manual work required on the part of a crystallographer.

2.5. Informatics

In addition to developing and maintaining several local databases in separate SGPP units, the central informatics unit of SGPP also maintained three major central databases:

1. Target prioritization: Applies criteria selected by investigators to score targets identified by target selection, domain prediction, and yeast two-hybrid. Available criteria included the number of transmembrane helices predicted, Pfam family size, percent identity of nearest PDB and human orthologues, nearest orthologue with diffraction quality crystals in TargetDB (targetdb.pdb.org), sequence length, number and extent of low complexity and disordered regions, amino acid composition, subcellular localization, and curated values such as medical and functional relevance.
2. Target status: Reports overall progress for each ORF and each chunk of an ORF: on the web in a table and in XML to the public; and in shared files for internal data mining by SGPP researchers. This database acquired data directly from Sesame and local databases in each unit from target selection to structure determination.
3. Sesame central tracking database: Tracks each action taken on each target, including repeated steps. Sesame is a three-tiered database system developed at the University of Wisconsin, Madison, by Zsolt Zolnai in the Center for Eukaryotic Structural Genomics (CESG) (www.uwstructuralgenomics.org). A copy of Sesame was maintained in Seattle.

Once sufficient data were collected in the central databases, the informatics unit also performed data mining to answer some critical and interesting questions related to achieving higher success rate for the authors' structural genomics pipeline. For example, data mining on results of HWI's HTS crystallization screening included (F. Zucker et al., unpublished): (1) Polyethylene glycols (PEGs) are the most effective component for protein crystallization in SGPP, both alone and in combinations with salt or other organic chemicals. (2) Proteins with high disorder or hydropathy do not crystallize well in the full screen. No individual condition or class of conditions was found that improved their crystallization. (3) Other predicted properties, such as isoelectric point or charge in a given condition, did not significantly affect crystallization rates. (4) Dynamic light scattering (DLS) was a reasonably good predictor of crystallizability, as expected. Proteins with high polydispersity or multiple DLS peaks did not crystallize well. However, sensitivity to proteolysis was a better indicator (O. Kalyuzhniy and W. Hol, unpublished observations). (5) Protein concentration had a minor and inconsistent effect. (6) Proteins annotated as enzyme-like and proteins found in Pfam crystallized better than other proteins, especially than proteins annotated as hypothetical.

Overall, the central informatics units of SGPP archived and searched information of over 21,000 potential protein targets under consideration, approximately 13,000 active targets, more than 33,000 recorded actions, as well as detailed protocols in various SGPP units.

3. Technology Development

As stated, methodology development was also a major task within SGPP in hope to help achieve high throughputs and high success rates for structural genomics projects. These developmental projects were implemented at

various places throughout the entire SGPP pipeline. They included (in order of pipeline progression from target selection to structure determination and annotation):

- Computational domain parsing and chunk selection
- Experimental yeast two-hybrid selection of protein pairs in *Plasmodium falciparum (6)*
- A "multispecies approach" for target cloning and expression of plasmodium and trypanosomatids, taking in particular advantage of the large number of closely related *Leishmania* species for finding a target-variant that yields soluble protein
- Whole gene synthesis
- Protein pair expression *(8)*
- Integral membrane protein cloning, expression, and purification in *E. coli*
- Single-chain antibody selection, production, and complex formation against target proteins
- "Co-crystallant" design, synthesis, and testing, as well as ligand screening; specific ligands for improving protein cocrystallization
- Capillary crystal-growth robotics
- Crystal annealing, healing, and screening robotics
- Fragment cocktails for crystallography *(33)*

A few of these "special SGPP projects" are selectively described in the following.

3.1. Computational Domain Prediction and Chunk Selection

"Chunk" has a unique meaning in SGPP in that it refers to fragments of a protein that may be solubly expressed. A chunk could be a single domain or a linear combination of several domains of a target protein. The selection of chunks as targets for the SGPP pipeline played a significant role in the target selection process of SGPP, especially due to the fact that the authors were often dealing with hard to express eukaryotic protein targets. In particular, proteins from the targeted organisms frequently contain genes coding for major insertions of amino acids that are absent in other organisms. Two fully automated domain prediction methods, Ginzu and RosettaDOM *(7)*, were implemented in SGPP for chunk selection. Ginzu has been used to predict domains in nearly all SGPP targets using a hierarchical procedure that assigns domains based on homology to known structures and protein families using successively less confident methods. RosettaDOM relies on information only in the query sequence by using the Rosetta *de novo* structure prediction method to build three-dimensional models, and then applying a structure based domain assignment algorithm to parse each model into domains. Domain boundaries that are consistently assigned in the models are predicted to be the actual domain boundaries. Both methods were top performers among automated methods in CASP6 (Critical Assessment of Techniques for Protein Structure Prediction), and performed well even when compared with human predictors *(22)*.

Experimentally, computational chunking has been performed on 253 targets within the SGPP pipeline giving a total of 1,495 chunks. These chunks were subjected to overexpression. Among them, 91 targets with no full-length expression resulted in one or more expressible chunks. For 44 targets that

we were not able to obtain soluble samples of the full-length protein, one or more soluble chunks for each target were purified in sufficient quantities for downstream pipeline use. In summary, SGPP rescued a significant number of targets for further study using chunking. Although the overall expression success rate per chunk is roughly the same with or without chunking for a random target in the genomes of SGPP organisms, the fact that targets might be rescued that did not result in full-length proteins using computational chunking is significant for general biological studies of the parasite proteins. This may be especially important for potential drug development targets of these parasites. For example, within the 253 targets that have been chunked, 57 are potential drug targets, which consist mainly of homologues of known enzymes. Through chunking, 28% (16 targets) gave soluble chunks but not the full-length sample; therefore, for really important protein targets, chunking will significantly enhance the chance of producing soluble parts of those proteins for future studies.

3.2. Design, Synthesis, and Systematic Testing of Co-Crystallants

Protein crystallization has a long history of discovery that very specific additives are indispensable for obtaining diffraction quality crystals for particular proteins. Additives can be inorganic compounds such as zinc chloride or small organic molecules such as phenol. Their beneficial effect can be due to: (1) a change in dielectric, leading to changes in solubility of the protein; (2) the formation of a specific complex with the protein, which can either reduce conformational heterogeneity of the protein or lead to a packing that involves the additive; or (3) the formation of very weak complexes on one face of the crystal, thereby slowing down the growth in that direction and allowing other crystal faces to grow *(23)*. The Biological Macromolecule Crystallization Database (BMCD) *(24)* lists 383 different additives reported in the literature, many of which are already present in crystallization screens. Most of them are ions, alcohols, carbohydrates, and surfactants.

However, there are in principle combinatorial chemistry opportunities to increase the repertoire of additives. Additives that operate through mechanism (2) above, which the authors like to refer to as "co-crystallants", are of great interest because incorporation of elements such as Br or Se in these compounds should, in addition to possibly promoting crystal growth, also allow for MAD/SAD phasing. As a special project in SGPP, the design, synthesis, and testing of a pilot set of such potential co-crystallants were carried out.

The authors incorporated two classes of bromine-containing synthetic co-crystallants in their pilot collection of compounds (Fig. 33.2). The design principle included the consideration of: (1) a water soluble moiety (the guanidine moiety, or one water-soluble function group attached to the imidazolidinone moiety); (2) a bromine-containing moiety for potential MAD/SAD phasing; and (3) a diversification point for incorporation of a variety of chemical substitutions to provide potentially beneficial interactions with protein targets. The synthesis of these two classes compounds was published *(25,26)*, and 44 compounds were prepared for the authors' initial test.

An initial test of co-crystallants with 10 proteins in solutions containing 10% DMSO (v/v) at 10 mM each compound led to heavy precipitation of protein.

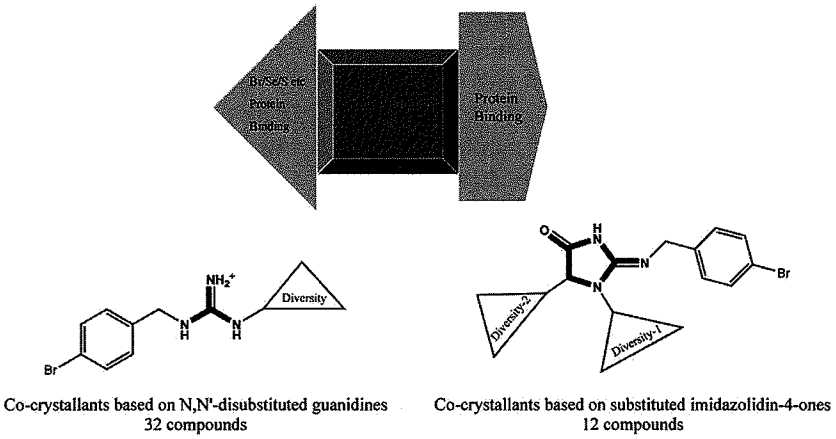

Fig. 33.2 Two classes of synthetic co-crystallants.

A second study investigated the effectiveness of these co-crystallants on crystallization using the 96 crystallization cocktails in the Hampton Research Index screen using a co-crystallant concentration that was lowered from the 10 mM in the initial study to reduce the macromolecules' precipitation. Seven co-crystallants were used at 2 and 0.4 mM concentrations. The extent of macromolecular precipitation was slightly less at 2 mM than at 0.4 mM co-crystallant. The average precipitation for all nine of the proteins studied against the seven co-crystallants decreased from 65% in the initial study to 34% in this study. The extent of precipitation was independent of co-crystallant and dependent upon the macromolecule. After removing the crystallization hits that were identified in the control experiments, there were 36 unique crystallization conditions that appear to require the presence of the co-crystallant. Two other SGPP proteins produced crystalline hits in the presence of co-crystallants.

The crystalline outcomes produced by these two macromolecules and 42% of the lysozyme crystallization conditions all contained the same co-crystallant. It appears that these co-crystallants do have the potential to positively affect the crystallization of protein samples that could otherwise not easily be achieved. However, more systematic studies are needed to fully reveal the pros and cons of using synthetic co-crystallants in a structural genomics setting.

3.3. Capillary Crystal-Growth Robotics

Development of a robotics system for protein crystal growth in capillaries was another special project in SGPP. The rationale for this approach was three-fold:

1. In capillaries, crystallization space can be traversed in a unique manner for optimum crystal growth. One way this can be achieved is by placing air gaps between fluids to permit first free liquid interface diffusion followed by vapor diffusion in glass capillaries. Alternatively, water evaporation through plastic capillary walls provides a second opportunity for optimizing protein crystal growth experiments by removing water

from the crystallization volume, thereby increasing both protein and precipitation concentration simultaneously.

2. The Meldrum group designed and developed a special instrument, the ACAPELLA *(27,28)*, able to deliver into capillaries low (50 nl and possibly down to the 5 nl range) protein and precipitant volumes to yield many experiments per volume of reagent. This is due to the fact that the instrument can precisely control the number of individual 100 pl liquid droplets delivered by piezoelectric dispensing;
3. By growing crystals in plastic capillaries, the potential exists to automate crystal growth and crystal mounting completely. This would involve:

- Taking images for evaluating the progress of crystal growth
- Analyzing and ranking the images by computational procedures
- Designing optimization strategies
- Filling capillaries to obtain diffraction-quality crystals
- Making images of data collection size crystals
- Freezing the crystals in the capillary *in situ*
- Using the images at the synchrotron beam lines to fully automatically center crystals in the x-ray beam

In this ideal approach the crystal would never have to be touched manually and the entire process from protein solution up to and including mounted crystals, crystal centering, and data collection can be fully automated.

The "ACAPELLA-5K" capillary-filling robot (Fig. 33.3) is able to fill 5,000 capillaries in 8 hours with volumes of around 0.5 to 1 µl *(27,28)*. Fully operational capabilities include aspiration of submicroliter volumes in capillaries, followed by delivering even smaller additional volumes of other solutions from several different dispensing piezo electric units. The filled capillaries can be photographed at several time points during this process, the volumes can be mixed in the capillary, and the filled capillaries are stored in holders, which allow easy photography to follow crystal growth. Encouragingly, a few initial tests showed that three entirely different proteins grew as beautifully shaped crystals in capillaries filled by this piece of equipment, which was designed for a different purpose *(27,28)*.

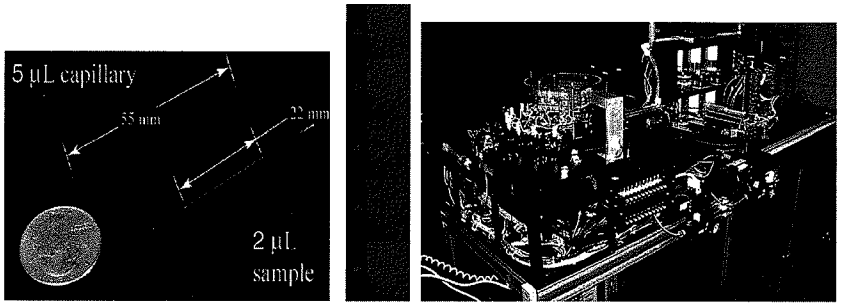

Fig. 33.3 The ACAPELLA capillary crystallization robot. **Left.** 5 µl capillary format (0.2–2 µl sample volumes possible) for high throughput operations in the ACAPELLA instrument. **Middle.** 100 pl droplets being dispensed from piezoelectric reagent dispensers into glass capillaries. **Right.** ACAPELLA-5K general-purpose submicroliter automated fluid handling system *(27)* showing the core processor in the front.

A series of hardware upgrades were implemented on the ACAPELLA capillary processor instrument platform to assess overall system robustness for high throughput reliability and effectiveness in generating optimal crystallization conditions within plastic capillaries. Although these design improvements worked as intended, there are fundamental limitations in the ability of the piezo dispenser technology to dispense an important set of crystallization reagents. Based on the experimental results from loading capillaries, the hardware architecture is being be reevaluated to address shortcomings discovered in the original architecture.

Encouraging preliminary studies showed that crystals could be flash-frozen *inside* the capillary, completely surrounded by liquid. Recording of excellent data up to 1.8 Å resolution appeared possible for hen egg white lysozyme crystals flash frozen in capillaries. For the *T. brucei* RNA ligase 1 (REL1) catalytic domain *(29)* anomalous differences to 2.5 Å could be recorded of SeMet REL1 crystals (Bosch et al., unpublished results). This shows the great potential of capillary crystal growth combined with "in capillary" flash freezing for *ab initio* structure determination.

4. Medical Structural Genomics

Although the majority of the SGPP studies were geared toward coverage of protein fold space, part of the targets in SGPP were selected for potential applications in drug development. Moreover, we explored an additional direction, called "medical structural genomics," that can have a major impact on drug development for neglected parasitic diseases since it includes discovery of binding modes of small molecule compounds to potential drug target proteins.

Therefore, in the later stages of SGPP, the authors created a library of compounds for carrying out fragment based ligand discovery by x-ray crystallography. This effort is based on an approach initiated with earlier crystallographic studies with small molecule cocktails *(30,31)*. The studies described by Verlinde et al. *(30)* were carried out prior to 1992 at the University of Groningen, The Netherlands. This approach is related to later studies entitled "SAR by NMR" *(32)*, which described how NMR screening can identify small ligands binding at adjacent locations on the protein surface, and how these small ligands, or "fragments," could be linked to obtain high-affinity ligands *(32)*. In brief, assuming one has crystals that allow for access to the binding site of interest, one can soak the crystal with a high concentration cocktail of small but shape-wise diverse fragments prior to x-ray diffraction measurements. Quite often, one of the fragments will bind at or near the target site. Because of the shape diversity, it should then be possible to identify which fragment in the cocktail has bound by inspection of the resulting electron density. Obtaining crystal structures of potential drug targets from these organisms with small organic molecules bound is of even greater importance to drug development efforts than structures of unliganded proteins.

So far, the SGPP consortium has selected about 700 small molecules for creating fragment cocktails using selection criteria as described in the literature *(33)*. These 700 molecules were grouped into about 70 cocktails based on differences in molecular mass and shape. The authors have prepared the first 30 cocktails of

approximately 10 compounds each for "fragment cocktail crystallography." Initial attempts of soaking these cocktails in buffers containing 10% DMSO proved to be quite successful. Soaking three inital SGPP proteins: Lmaj004144AAA, Pfal005984AAA, and Tbru015777AAA in cocktail solutions gave high-resolution datasets that allowed identification of unique ligands in particular cocktails for all three proteins (see www.sgpp.org and www.msgpp.org for additional examples). In the case of Tbru015777AAA, a nucleoside 2-deoxyribosyltransferase from *T. brucei*, several ligands bound to the active site were identified *(33)*.

The success in applying fragment cocktail crystallography in a structural genomics setting will open up new avenues for future drug development efforts against the parasites targeted by the SGPP consortium. This intersection of protein structure space and chemical space is currently pursued in the Medical Structural Genomics of Pathogenic Protozoa (MSGPP) program project (see www.msgpp.org).

5. Summary

In the 4-year period, the SGPP consortium established a high throughput structural genomics pipeline for obtaining protein crystal structures from four major disease-causing protozoa: *Plasmodium falciparum* and three trypanosomatid parasites. A large number of new protein crystal structures were solved for these organisms. Thousands of plasmids (the *P. falciparum* expression constructs have been deposited with the MR4 repository; www.malaria.mr4.org). Hundreds of protein samples, and experimental protocols are also important results assisting in the battle against several neglected parasitic diseases.

Acknowledgments

The authors acknowledge the numerous contributions to SGPP by Ken Stuart, Martin Berg, Keith Hodgson, past SGPP members, advisory committee members, collaborators, industrial partners, and the Protein Structure Initiative (PSI) of NIH (grant GM064655 from NIGMS and NIAID).

References

1. Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S., Paulsen, I. T., James, K., Eisen, J. A., Rutherford, K., Salzberg, S. L., Craig, A., Kyes, S., Chan, M. S., Nene, V., Shallom, S. J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M. W., Vaidya, A. B., Martin, D. M., Fairlamb, A. H., Fraunholz, M. J., Roos, D. S., Ralph, S. A., McFadden, G. I., Cummings, L. M., Subramanian, G. M., Mungall, C., Venter, J. C., Carucci, D. J., Hoffman, S. L., Newbold, C., Davis, R. W., Fraser, C. M., and Barrell, B. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419,** 498–511.
2. Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D. C., Lennard, N. J., Caler, E., Hamlin, N. E., Haas, B., Bohme, U., Hannick, L., Aslett, M. A., Shallom, J., Marcello, L., Hou, L., Wickstead, B., Alsmark, U. C., Arrowsmith, C., Atkin, R. J., Barron, A. J., Bringaud, F., Brooks, K., Carrington, M., Cherevach, I., Chillingworth, T. J., Churcher, C., Clark, L. N., Corton, C. H.,

Cronin, A., Davies, R. M., Doggett, J., Djikeng, A., Feldblyum, T., Field, M. C., Fraser, A., Goodhead, I., Hance, Z., Harper, D., Harris, B. R., Hauser, H., Hostetler, J., Ivens, A., Jagels, K., Johnson, D., Johnson, J., Jones, K., Kerhornou, A. X., Koo, H., Larke, N., Landfear, S., Larkin, C., Leech, V., Line, A., Lord, A., Macleod, A., Mooney, P. J., Moule, S., Martin, D. M., Morgan, G. W., Mungall, K., Norbertczak, H., Ormond, D., Pai, G., Peacock, C. S., Peterson, J., Quail, M. A., Rabbinowitsch, E., Rajandream, M. A., Reitter, C., Salzberg, S. L., Sanders, M., Schobel, S., Sharp, S., Simmonds, M., Simpson, A. J., Tallon, L., Turner, C. M., Tait, A., Tivey, A. R., Van Aken, S., Walker, D., Wanless, D., Wang, S., White, B., White, O., Whitehead, S., Woodward, J., Wortman, J., Adams, M. D., Embley, T. M., Gull, K., Ullu, E., Barry, J. D., Fairlamb, A. H., Opperdoes, F., Barrell, B. G., Donelson, J. E., Hall, N., Fraser, C. M., Melville, S. E., and El-Sayed, N. M. (2005) The genome of the African trypanosome *Trypanosoma brucei. Science* **309,** 416–422.

3. El-Sayed, N. M., Myler, P. J., Bartholomeu, D. C., Nilsson, D., Aggarwal, G., Tran, A. N., Ghedin, E., Worthey, E. A., Delcher, A. L., Blandin, G., Westenberger, S. J., Caler, E., Cerqueira, G. C., Branche, C., Haas, B., Anupama, A., Arner, E., Aslund, L., Attipoe, P., Bontempi, E., Bringaud, F., Burton, P., Cadag, E., Campbell, D. A., Carrington, M., Crabtree, J., Darban, H., da Silveira, J. F., de Jong, P., Edwards, K., Englund, P. T., Fazelina, G., Feldblyum, T., Ferella, M., Frasch, A. C., Gull, K., Horn, D., Hou, L., Huang, Y., Kindlund, E., Klingbeil, M., Kluge, S., Koo, H., Lacerda, D., Levin, M. J., Lorenzi, H., Louie, T., Machado, C. R., McCulloch, R., McKenna, A., Mizuno, Y., Mottram, J. C., Nelson, S., Ochaya, S., Osoegawa, K., Pai, G., Parsons, M., Pentony, M., Pettersson, U., Pop, M., Ramirez, J. L., Rinta, J., Robertson, L., Salzberg, S. L., Sanchez, D. O., Seyler, A., Sharma, R., Shetty, J., Simpson, A. J., Sisk, E., Tammi, M. T., Tarleton, R., Teixeira, S., Van Aken, S., Vogt, C., Ward, P. N., Wickstead, B., Wortman, J., White, O., Fraser, C. M., Stuart, K. D., and Andersson, B. (2005) The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science* **309,** 409–415.
4. El-Sayed, N. M., Myler, P. J., Blandin, G., Berriman, M., Crabtree, J., Aggarwal, G., Caler, E., Renauld, H., Worthey, E. A., Hertz-Fowler, C., Ghedin, E., Peacock, C., Bartholomeu, D. C., Haas, B. J., Tran, A. N., Wortman, J. R., Alsmark, U. C., Angiuoli, S., Anupama, A., Badger, J., Bringaud, F., Cadag, E., Carlton, J. M., Cerqueira, G. C., Creasy, T., Delcher, A. L., Djikeng, A., Embley, T. M., Hauser, C., Ivens, A. C., Kummerfeld, S. K., Pereira-Leal, J. B., Nilsson, D., Peterson, J., Salzberg, S. L., Shallom, J., Silva, J. C., Sundaram, J., Westenberger, S., White, O., Melville, S. E., Donelson, J. E., Andersson, B., Stuart, K. D., and Hall, N. (2005) Comparative genomics of trypanosomatid parasitic protozoa. *Science* **309,** 404–409.
5. Ivens, A. C., Peacock, C. S., Worthey, E. A., Murphy, L., Aggarwal, G., Berriman, M., Sisk, E., Rajandream, M. A., Adlem, E., Aert, R., Anupama, A., Apostolou, Z., Attipoe, P., Bason, N., Bauser, C., Beck, A., Beverley, S. M., Bianchettin, G., Borzym, K., Bothe, G., Bruschi, C. V., Collins, M., Cadag, E., Ciarloni, L., Clayton, C., Coulson, R. M., Cronin, A., Cruz, A. K., Davies, R. M., De Gaudenzi, J., Dobson, D. E., Duesterhoeft, A., Fazelina, G., Fosker, N., Frasch, A. C., Fraser, A., Fuchs, M., Gabel, C., Goble, A., Goffeau, A., Harris, D., Hertz-Fowler, C., Hilbert, H., Horn, D., Huang, Y., Klages, S., Knights, A., Kube, M., Larke, N., Litvin, L., Lord, A., Louie, T., Marra, M., Masuy, D., Matthews, K., Michaeli, S., Mottram, J. C., Muller-Auer, S., Munden, H., Nelson, S., Norbertczak, H., Oliver, K., O'Neil, S., Pentony, M., Pohl, T. M., Price, C., Purnelle, B., Quail, M. A., Rabbinowitsch, E., Reinhardt, R., Rieger, M., Rinta, J., Robben, J., Robertson, L., Ruiz, J. C., Rutter, S., Saunders, D., Schafer, M., Schein, J., Schwartz, D. C., Seeger, K., Seyler, A., Sharp, S., Shin, H., Sivam, D., Squares, R., Squares, S., Tosato, V., Vogt, C., Volckaert, G., Wambutt, R., Warren, T., Wedler, H., Woodward, J., Zhou, S., Zimmermann, W., Smith, D. F., Blackwell, J. M., Stuart, K. D., Barrell, B., and Myler, P. J. (2005) The genome of the kinetoplastic parasite, *Leishmania* major. *Science* **309,** 436–442.

6. LaCount, D. J., Vignali, M., Chettier, R., Phansalkar, A., Bell, R., Hesselberth, J. R., Schoenfeld, L. W., Ota, I., Sahasrabudhe, S., Kurschner, C., Fields, S., and Hughes, R. E. (2005) A protein interaction network of the malaria parasite *Plasmodium falciparum. Nature* **438,** 103–107.
7. Kim, D. E., Chivian, D., Malmstrom, L., and Baker, D. (2005) Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. *Proteins* **61,** 193–200.
8. Alexandrov, A., Vignali, M., LaCount, D. J., Quartley, E., de Vries, C., De Rosa, D., Babulski, J., Mitchell, S. F., Schoenfeld, L. W., Fields, S., Hol, W. G., Dumont, M. E., Phizicky, E. M., and Grayhack, E. J. (2004) A facile method for high throughput co-expression of protein pairs. *Mol. Cell. Proteomics* **3,** 934–938.
9. Mehlin, C., Boni, E. E., Buckner, F. S., Engel, L., Feist, T., Gelb, M. H., Haji, L., Kim, D., Liu, C., Mueller, N., Myler, P., Reddy, J. T., Sampson, J., Subramanian, E., Van Voorhis, W. C., Worthey, E. A., Zucker, F., and Hol, W. G. J. (2006) Heterologous Expression of Proteins from *Plasmodium falciparum:* results from 1000 genes. *Mol. Biochem. Parasitol.* **148,** 144–160.
10. Mehlin, C., Boni, E. E., Andreyka, J., and Terry, R. W. (2004) Cloning grills: high throughput cloning for structural genomics. *J. Struct. Funct. Genomics* **5,** 59–61.
11. Deng, J., Davies, D. R., Wisedchaisri, G., Wu, M., Hol, W. G. J., and Mehlin, C. (2004) An improved protocol for rapid freezing of protein samples for long-term storage. *Acta Crystallogr. D* **60,** 203–204.
12. Luft, J. R., Collins, R. J., Fehrman, N. A., Lauricella, A. M., Veatch, C. K., and DeTitta, G. T. (2003) A deliberate approach to screening for initial crystallization conditions of biological macromolecules. *J Struct Biol* **142,** 170–179.
13. Rupp, B., Segelke, B. W., Krupka, H. I., Lekin, T., Schafer, J., Zemla, A., Toppani, D., Snell, G., and Earnest, T. (2002) The TB structural genomics consortium crystallization facility: towards automation from protein to electron density. *Acta Crystallogr. D* **58,** 1514–1518.
14. Cork, C., O'Neill, J., Taylor, J., and Earnest, T. (2006) Advanced beamline automation for biological crystallography experiments. *Acta Crystallogr. D* **62,** 852–858.
15. Ellis, P. J., Cohen, A. E., and Soltis, S. M. (2003) Beamstop with integrated X-ray sensor. *J. Synchrotron Radiat.* **10,** 287–288.
16. Yeh, J. I., and Hol, W. G. J. (1998) A flash annealing technique to improve diffraction limits and lower mosaicity in crystals of glycerol kinase. *Acta Crystallogr. D* **54,** 479–480.
17. Harp, J. M., Timm, D. E., and Bunick, G. J. (1998) Macromolecular crystal annealing: overcoming increased mosaicity associated with cryocrystallography. *Acta Crystallogr. D* **54,** 622–628.
18. Otwinowski, Z., and Minor, W. (1997) Processing of x-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326.
19. Holton, J., and Alber, T. (2004) Automated protein crystal structure determination using ELVES. *Proc. Natl. Acad. Sci. USA* **101,** 1537–1542.
20. Bewley, M. C., Jeffrey, P. D., Patchett, M. L., Kanyo, Z. F., and Baker, E. N. (1999) Crystal structures of Bacillus caldovelox arginase in complex with substrate and inhibitors reveal new insights into activation, inhibition and catalysis in the arginase superfamily. *Structure* **7,** 435–448.
21. Terwilliger, T. C. (2003) SOLVE and RESOLVE: automated structure solution and density modification. *Methods Enzymol.* **374,** 22–37.
22. Kryshtafovych, A., Venclovas, C., Fidelis, K., and Moult, J. (2005) Progress over the first decade of CASP experiments. *Proteins* **61,** 225–236.
23. Mikol, V., and Giege, R. (1992) in (Ducruix, A., and Giege, R., eds.), *Crystallization of Nucleic Acids and Proteins*. Oxford University Press, New York, 219–239.
24. Gilliland, G. L., Tung, M., Blakeslee, D. M., and Ladner, J. E. (1994) Biological Macromolecule Crystallization Database, Version 3.0: new features, data and the NASA archive for protein crystal growth data. *Acta Crystallogr. D* **50,** 408–414.

25. Li, J., Zhang, Z., and Fan, E. (2004) Solid-phase synthesis of 1,5-substituted 2-N-Alkylamino)=imidazolidin-4-ones. *Tetrahedron Lett.* **45,** 1267–1269.
26. Li, J., Zhang, G., Zhang, Z., and Fan, E. (2003) TFA sensitive arylsulfonylthiourea-assisted synthesis of N, N'-substituted guanidines. *J. Org. Chem.* **68,** 1611–1614.
27. Meldrum, D. R., Evensen, H. T., Pence, W. H., Moody, S. E., Cunningham, D. L., and Wiktor, P. J. (2000) ACAPELLA-1K, a capillary-based submicroliter automated fluid handling system for genome analysis. *Genome Res.* **10,** 95–104.
28. Arutunian, E. B., Meldrum, D. R., Friedman, N. A., and Moody, S. E. (1998) Flexible software architecture for user-interface and machine control in laboratory automation. *Biotechniques* **25,** 698–702, 704–695.
29. Deng, J., Schnaufer, A., Salavati, R., Stuart, K., and Hol, W. G. J. (2004) High Resolution crystal structure of an editosome enzyme from *Trypanosoma brucei*: RNA editing ligase I. *J. Mol. Biol.* **343,** 601–613.
30. Verlinde, C. L. M. J., Kim, H., Bernstein, B. E., Mande, S. C., and Hol, W. G. J. (1997) Antitrypanosomiasis drug development based on structures of glycolytic enzymes. In *Structure-Based Drug Design* (Veerapandian, P., ed.), pp. 365–394. Marcel Dekker, New York.
31. Nienaber, V. L., Richardson, P. L., Klighofer, V., Bouska, J. J., Giranda, V. L., and Greer, J. (2000) Discovering novel ligands for macromolecules using X-ray crystallographic screening. *Nat. Biotechnol.* **18,** 1105–1108.
32. Shuker, S. B., Hajduk, P. J., Meadows, R. P., and Fesik, S. W. (1996) Discovering high-affinity ligands for proteins: SAR by NMR. *Science* **274,** 1531–1534.
33. Bosch, J., Robien, M. A., Mehlin, C., Boni, E., Riechers, A., Buckner, F. S., Van Voorhis, W. C., Myler, P. J., Worthey, E. A., DeTitta, G., Luft, J. R., Lauricella, A., Gulde, S., Anderson, L. A., Kalyuzhniy, O., Neely, H. M., Ross, J., Earnest, T. N., Soltis, M., Schoenfeld, L., Zucker, F., Merritt, E. A., Fan, E., Verlinde, C. L., and Hol, W. G. (2006) Using fragment cocktail crystallography to assist inhibitor design of *Trypanosoma brucei* nucleoside 2-deoxyribosyltransferase. *J. Med. Chem.* **49,** 5939–5946.