



## The structural molecular biology network of the State of São Paulo, Brazil

JOÃO A.R.G. BARBOSA<sup>1</sup>, LUIS E.S. NETTO<sup>2</sup>, CHUCK S. FARAH<sup>3</sup>  
SERGIO SCHENKMAN<sup>4</sup> and ROGÉRIO MENEGHINI<sup>5</sup>

<sup>1</sup>Centro de Biologia Molecular Estrutural, Laboratório Nacional de Luz Síncrotron,  
Rua Giuseppe Máximo Scolfaro, 10.000, 13084-971 Campinas, SP, Brasil

<sup>2</sup>Departamento de Genética e Processos Evolutivos, Instituto de Biociências  
Universidade de São Paulo, Rua do Matão 2777, 05508-900 São Paulo, SP, Brasil

<sup>3</sup>Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo,  
Av. Prof. Lineu Prestes, 748, 05508-000 São Paulo, SP, Brasil

<sup>4</sup>Escola Paulista de Medicina, Universidade Federal de São Paulo,  
Rua Botucatu, 862, 04023-901 São Paulo, SP, Brasil

<sup>5</sup>Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde  
BIREME, Rua Botucatu, 862, 04023-901 São Paulo, SP, Brasil

*Manuscript received on June 13, 2005; accepted for publication on June 14, 2005;  
contributed by ROGÉRIO MENEGHINI\* AND SERGIO SCHENKMAN\**

### ABSTRACT

This article describes the achievements of the Structural Molecular Biology Network (SMolBNet), a collaborative program of structural molecular biology, centered in the State of São Paulo, Brazil, and supported by São Paulo State Funding Agency (FAPESP). It gathers twenty scientific groups and is coordinated by the scientific staff of the Center of Structural Molecular Biology, at the National Laboratory of Synchrotron Light (LNLS), in Campinas. The SMolBNet program has been aimed at 1) solving the structure of proteins of interest related to the research projects of the groups. In some cases, the choice has been to select proteins of unknown function or of possible novel structure obtained from the sequenced genomes of the FAPESP genomic program; 2) providing the groups with training in all the steps of the protein structure determination: gene cloning, protein expression, protein purification, protein crystallization and structure determination. Having begun in 2001, the program has been successful in both aims. Here, four groups reveal their participation in the program and describe the structural aspects of the proteins they have selected to study.

**Key words:** structural genomics, protein crystallography, nuclear magnetic resonance, protein structure.

### INTRODUCTION

The Structural Molecular Biology Network (SMolBNet), if envisaged as a structural genomics program, has some odd features and is therefore en-

titled to an unconventional scientific article, which does not delve heavily into a detailed analysis of results. This text is divided into several independent texts that are linked by the common emphasis on structural molecular biology. We insist on the use of the term “structural molecular biology” (SMB) to define our efforts instead of the terms “structural ge-

Dedicated to the memory of Prof. Francisco Lara (1925–2004)

\*Member Academia Brasileira de Ciências

Correspondence to: Rogério Meneghini

E-mail: rogmene@bireme.ops-oms.org

nomics” (since many investigators may be involved in SMB without using genomic data) or “structural biology” (in homage to classic structural biologists who operate with optical or electronic microscopes in the micron to millimicron scale whereas SMB deals with atomic resolution).

Structural genomics, in the strict sense, is SMB applied on the genomic scale and as such, if not explanatory, is etymologically adequate. With the advent of genomics, the annotation of thousands of known and putative genes became and continues to become available. This technological breakthrough has provided a great impulse to different fields of biology, including SMB. Before the era of large-scale genome sequencing SMB had already been experiencing enormous progress with the elucidation of the 3D-atomic structure of several hundred proteins thanks to the use of two modern technologies: X-ray crystallography, using X-ray beam-lines of synchrotron radiation facilities, and high field nuclear magnetic resonance (NMR). In this way, a somewhat astonishing perception has come to light: the number of ways proteins may fold into defined 3D atomic structures in the universe of all proteins in the biosphere is limited. As more 3D structures of proteins were solved the rarer it became to find novel folding modes. Therefore one may envisage a future where virtually all new structures solved may be classified into a previously defined family. Although large, the number of folding topologies is probably limited to one or a few thousand (Govindarajan et al. 1999). Furthermore, if the structure of one member of a protein family that share significant primary structure similarity is known, feasible molecular models of all the remaining members may be constructed by homology modeling.

After the advent of the genomics era, the structural genomics approach was launched. In Brazil, the ONSA program (Organization for Nucleotide Sequencing and Analysis, FAPESP) was responsible for the sequencing of the first plant pathogen genome, that of the *Xylella fastidiosa* (*Xf*), which has 2900 predicted genes (Simpson et al. 2000), a significant fraction of which has no known function.

The path of structural genomics goes as follow: define a criterion by which to select specific protein targets, for example proteins of unknown function and/or unknown structure. Then: (i) clone the selected target genes into expression vectors, (ii) express the corresponding proteins in an heterologous system, (iii) purify the proteins, (iv) submit them to crystallization trials, (v) collect X-ray diffraction data usually at a synchrotron radiation facility and (vi) solve the structure. Alternatively, from step (iv) obtain a high concentration protein solution and collect data at an NMR facility. The latter is an attractive option for those many cases in which the proteins are refractory to form crystals. This sequence is a funnel, for each step has a limited chance of success. Therefore, most structural genomics efforts adopt a high throughput approach: if one of the steps of the pipeline poses a difficult obstacle for a particular target, that target is disposed of. Hence, only a small percent of the chosen proteins have their 3D-structure resolved (typically 1-5%). Dozens of worldwide structural genomics projects are underway. (<http://www.rcsb.org/pdb.strucgen.html>).

Of course, not all SMB projects are high throughput. It is often the case that a group is interested in resolving the structure of one or a few proteins related to a specific biological question. Here, there is no thoughtless disposal of the protein targets if difficulties in the pipeline arise. The main purpose is not structural *per se*; rather, solving these structures may shed light on functional aspects of interest.

The central focus of the SMolBNet program was to bring together a wide variety of biomedical and biological research groups with an interest in SMB. In 1999, twenty groups of the State of São Paulo entered the program, almost all of them with no previous experience in SMB. Most of them had in mind the use of SMB to study proteins that were already targeted to be of interest in their specific line of study. However, some were interested in embarking on a structural genomics program, even though it was not directly linked to a traditional line of research, to a great extent influenced by the success of

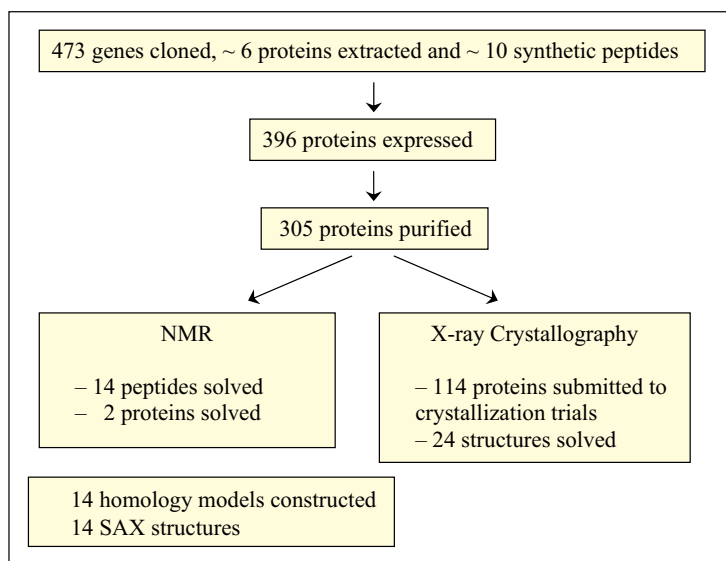


Fig. 1 – The SMolBNet pipeline showing the general achievements at the end of the fourth year.

FAPESP genomics projects.

The program was coordinated by the Center for Structural Molecular Biology (CeBiME) of the Brazilian Synchrotron Light Laboratory (LNLS), in Campinas, São Paulo, which includes several staff investigators with expertise in protein crystallography and protein NMR. The Center facilities include NMR spectrometers, mass spectrometers, light spectrometers, protein purification and crystallization, and molecular biology laboratories, and a crystallography beamline connected to the 1.37 GeV circular electron accelerator that generates synchrotron radiation. In addition, a new MAD crystallography beamline is coming up at the end of 2005. This set of facilities is unique in the southern hemisphere.

The selected groups started working in 1999, under the assistance of the CeBiME staff. Judging from the results accumulated up until now we can conclude that the SMolBNet program was successful in providing an important impetus for the advancement of SMB in Brazil. The pipeline in Figure 1 summarizes the current state of the program at each stage of the production pipeline. The program is intended to terminate by the end of 2006.

Among the five authors contributing to this work, four summarize here the results of their participation. JARGB is a crystallographer and a CeBiME-LNLS staff member. He reports his participation in collaborating with and training other members of the network. Not included in this report is his own work determining the structures of some *Xf* hydrolases and some hypothetical gene products from *Xanthomonas axonopodis* pv. *citri* (*Xac*). CSF is a biochemist who has also chosen to use both NMR and X-ray crystallography to study some *Xac* conserved hypothetical proteins and proteins important for *Xac* pathogenicity. LESN is a biochemist who has been studying functional aspects of antioxidant proteins for some years and has picked several of these proteins from *Saccharomyces cerevisiae* (*Sc*) and *Xf* for crystallographic studies. SS is a biochemical parasitologist interested in the infectious mechanisms employed by protozoa and their insect vectors. He focused his structural investigations on proteins from the protozoa *Trypanosoma cruzi*, *Plasmodium vivax* and *falciparum* and a protein from the insect vector *Triatoma infestans*, using both X-ray crystallography and NMR.

#### CRYSTALLOGRAPHIC EFFORTS IN THE SMolBNet

Within the SMolBNet framework, CeBiME had the challenge of working with the various groups of the network in a way as to provide the expertise necessary to solve peptide or protein structures. Further, this expertise should be passed to the groups, allowing them to gain as much independency as possible.

Crystallization experiments are the first step in trying to obtain a crystallographic structure of a protein. It usually demands that enough quantities (10 mg) of high grade purified and properly folded protein are available. This procedure has a low success rate and is a tough test of the researchers' will. The vast majority of the crystallization experiments were carried out using the hanging- or sitting-drop vapor diffusion methods. Here, the equilibration of two different solutions leads to a very slow depletion of water in the solution that contains the protein and consequently concentrates it. As the protein concentrates, it tends to aggregate in an amorphous or crystalline precipitate. In Figure 2 and Table I it is possible to see the crystals obtained for several proteins. Each of these crystals demanded a unique crystallization condition and represents the mid-point to success mark in the path to structure solution when we consider all the pre-crystallographic work.

The next step was X-ray diffraction data collection at the LNLS beamline. This was a much expected part of the work because the researchers would finally be sure that their crystals would be useful to structure solution and also because they would have contact with the synchrotron instrumentation. From this point onwards there is a clear division in the type of work carried out by the groups: until now they were working in the wet lab as molecular biologists and biochemists, now their work will be carried out on computers. Another issue is that the diffraction experiment is within the domain of physics, and while physicists and chemists might be familiar with it, biologists, who were the majority, felt uncomfortable. The students and/or PIs carried out data processing of the

oscillation images of those crystals that had interpretable diffraction with close and extensive supervision. They were taught not only how to run the programs, but also the ideas behind them: notions of real and reciprocal space, the physics of diffraction, reflection's intensities and indices, the concepts of unit cell, asymmetric unit, space group, integration, scaling, merging, etc.

Phase determination/structure solution was usually a straightforward procedure since most of the structures were solved by molecular replacement. The only exception was a hypothetical protein ("f" in Table I) that was solved by multiwavelength anomalous dispersion (MAD). This rather interesting case demanded extra input from the particular group such as growing selenium methionine protein crystals and is described in another part of this manuscript. Structure building and refinement is a procedure that still demands weeks and perhaps months of work. This was accomplished by having the people from the network coming over to the CeBiME for these periods of time. This part helped the students to learn about the protein databank format, atom coordinates, R factor, temperature factor, electron density and many other concepts. Structure analysis would be the final step and as the groups moved towards this, it could be foreseen that difficulties arose when trying to explore completely unrelated structures. Here, there is no clear path to be followed, except for comparing the structures to their homologues when they are present. In most cases some unique features of the proteins will need to be addressed, such as: why is the lysozyme ("b" in Table I) capable of working in a basic pH, what is the information that the hypothetical protein ("f") can give about its function or what makes the infestin 4 ("j") such a good inhibitor of blood coagulation-cascade factor XIIa.

It is clear that throughout this project all of the groups gained experience in diverse techniques used in structural biology. Now that the project has been extended for two years and shall finish in 2006, it is time to consolidate the results and think about how the groups shall move to the next step: become inde-

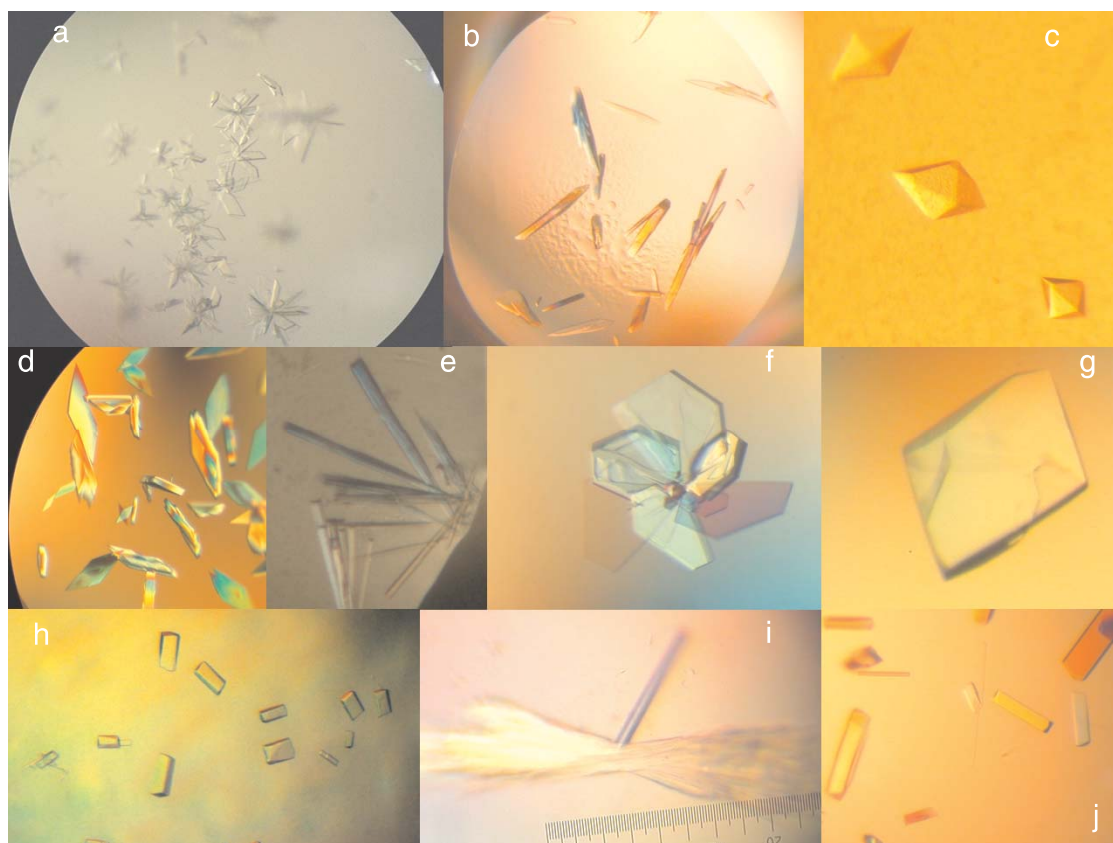


Fig. 2 – Crystals obtained by four different groups of the SMolBNet. The letters “a” to “j” correlate to the information for each individual protein presented in Table I. Their sizes vary approximately from a few  $\mu\text{m}$  to one mm.

pendent of CeBiME. It is not realistic to expect that all of the groups will grow and incorporate structural biology. One solution for everybody’s effort to flourish all its potential lies in new positions opened for structural biologists at the Departments to which SMolBNet groups are affiliated.

#### STRUCTURAL AND FUNCTIONAL ASPECTS OF PEROXIDE DETOXIFICATION PATHWAYS

The mechanisms by which cells protect themselves against the toxic effects of free radicals and related species have been the subject of study in our laboratory (LESN). Cells possess multiple pathways to decompose peroxides such as catalases and thiol dependent peroxidases, including glutathione peroxidases. Emphasis has been given to a new class of antioxidant enzymes named peroxiredoxins, which

includes thioredoxin peroxidases. In spite of the fact that peroxiredoxins and glutathione peroxidases have similar biochemical activity (thiol-dependent peroxidases), they do not share amino acid sequence similarity. Interestingly, however, both glutathione-dependent peroxidases and peroxiredoxins belong to the superfamily thioredoxin that comprises proteins possessing the thioredoxin fold characterized by a central core of four stranded mixed beta-sheet, flanked by three alpha-helices. (Copley et al. 2004).

The yeast *S. cerevisiae*, which has been used as model for higher eukaryotes, possesses five peroxiredoxins among two catalases and three phospholipid glutathione peroxidases. Our studies have demonstrated that although all five peroxiredoxins have the same biochemical activity (thioredoxin peroxidase), their functions are not completely re-

TABLE I

Name of the proteins, organism, maximum diffraction resolution and current research stage of the crystals shown from Figure 1. The letters “a” to “h” on the left of the protein’s column correspond to the letters in Figure 2.

	Protein	Organism	Resolution (Å)	Stage
a	Beta-glicosidase	<i>Spodoptera frugiperda</i>	2.8	Crystallization, data collection
b	Lysozyme	<i>Musca domestica</i>	1.9	Analysis
c	Maltose-binding protein – MalE	<i>Xanthomonas axonopodis</i> pv. <i>citri</i>	~ 4.0	Crystallization, data collection
d	Merozoite surface proteins 3 $\alpha$	<i>Plasmodium</i> sp.	~ 10.0	Crystallization, data collection
e	Xac2355/SufE	<i>Xanthomonas axonopodis</i> pv. <i>citri</i>	1.9	Refinement
f	Hypothetical Xac2396	<i>Xanthomonas axonopodis</i> pv. <i>citri</i>	1.9	Refinement
g	Molibdate-binding protein – ModA	<i>Xanthomonas axonopodis</i> pv. <i>citri</i>	1.7	Refinement
h	Thioredoxin peroxidase III	<i>Saccharomyces cerevisiae</i>	2.2	Crystallization, data collection
i	Infestin 1 – trypsin	<i>Triatoma infestans</i> <i>Bos taurus</i>	2.5	Refinement
j	Infestin 4	<i>Triatoma infestans</i>	1.8	Analysis

dundant. For example, cytosolic thioredoxin peroxidase I (cTPxI/Tsa1/YML028W) is specifically important for the defense of yeast with dysfunctional mitochondria (Demasi et al. 2001), whereas mitochondrial thioredoxin peroxidase I (PrxI/YBL064C) is more active in conditions where yeast obtain ATP preferentially by respiration (Monteiro et al. 2002, Monteiro and Netto 2004). Finally, cytosolic thioredoxin peroxidase II (cTPxII/Tsa2/YDR453C) appears to be an important backup for cTPxI for the defense against organic peroxides, independently of the functional state of mitochondria (Munhoz and Netto 2004).

We are also studying antioxidant proteins from other organisms as a consequence of our participation in *X. fastidiosa* Genome Project. In this regard, we were able to characterize for the first time

the biochemical role of a new class of antioxidant enzymes conserved in several pathogenic bacteria: Ohr (“Organic Hydroperoxide Resistance protein”) is a dithiol-dependent peroxidase (Cussiol et al. 2003).

When we joined SMolBNet, we decided to investigate structural aspects of thiol-dependent peroxidases and related proteins in a hope to better understand mechanisms of peroxide detoxification in cells. Several recombinant proteins from yeast and bacteria were submitted to crystallization trials. Since thiol-dependent peroxidases can assume different redox states, we decided to pre-treat proteins with various oxidants and reductants at different concentrations in an attempt to obtain a homogeneous preparation. This appeared to be a successful approach since several crystals from different pro-

teins were obtained.

In the case of Ohr from *X. fastidiosa*, several crystals were obtained from proteins treated with different redox compounds using the hanging-drop vapor diffusion method in the presence of PEG 4000 as precipitant (Oliveira et al. 2004). The crystal structure was solved by molecular replacement methods and the protein possesses a homodimeric quaternary structure. Two Ohr structures were obtained at two different molecular proportions (16 *tert*-butylhydroperoxide: 1 Ohr and 1 *tert*-butylhydroperoxide: 1 Ohr), which will be described in detail elsewhere. In summary, the protein possesses two domains: the N-terminal domain composed of three antiparallel  $\beta$  strands ( $\beta 1$ ,  $\beta 2$  and  $\beta 3$ ) that forms a  $\beta$  sheet and a short helix (H1) and the C-terminal domain composed of an antiparallel  $\beta$  sheet ( $\beta 4$ - $\beta 6$ ) and three  $\alpha$ -helices (H2-H4). Contrary to glutathione peroxidases and peroxiredoxins, two other classes of thiol-dependent peroxidases, Ohr does not possess the thioredoxin fold. Furthermore, the mechanism by which the reactive cysteine is stabilized appeared to be different among the three classes of thiol-dependent peroxidases. Therefore, Ohr is a new class of antioxidant proteins. The reactive cysteine is in a sulfonate form ( $R-SO_3H^-$ ) in the structure obtained from Ohr treated with very high doses of peroxide, indicating that these enzymes can only be inactivated by very harsh conditions (M.A. Oliveira et al., unpublished data).

Other proteins with redox properties from *S. cerevisiae* also had their structures solved by our group. One of them was glutaredoxin 2, which is the main glutathione-dependent oxidoreductase of yeast. This ubiquitously distributed enzyme is involved in many cellular processes including protein folding, regulation of protein activity and repair of oxidatively damaged proteins (Rietsch and Beckwith 1998). Glutaredoxin 2 structure was solved by molecular replacement using a homologous protein from *Sus scrofa* as a model (Discola et al. 2005). Thioredoxin reductase is a pyridine nucleotide-disulfide oxidoreductase capable to reduce redox active disulfide bond of thioredoxins at the expense

of NADPH. The thioredoxin system (NADPH/thioredoxin reductase/thioredoxin) is involved in the reduction of disulfide bonds in several targets. Crystals of thioredoxin reductase 1 from *S. cerevisiae* were obtained from proteins pretreated with  $H_2O_2$ , using the hanging-drop method. X-ray diffraction data were collected to a maximum resolution of 2.4 Å using a synchrotron radiation source and the structure was also solved by molecular replacement using a homologous protein from *Arabidopsis thaliana* (Oliveira et al., unpublished data). Structures of glutaredoxin 2 and thioredoxin reductase 1 are under refinement and we hope that their elucidation will contribute to the better understanding of their catalytic mechanisms. In summary, participation in the SMoIBNet was very important because it made possible the establishment of structural and functional relationships, which represented a considerable improvement from my previous work.

#### STRUCTURAL DETERMINATION OF XANTHOMONAS PROTEINS OF UNKNOWN FUNCTION

The sequencing of the genome of the citrus pathogen *Xanthomonas axonopodis* pv. *citri* (*Xac*) revealed the existence of many macromolecular systems important for pathogenesis (da Silva et al. 2002, Van Sluys et al. 2002, Moreira et al. 2004) and several Brazilian laboratories are studying their functions. In addition to the identification of specific proteins of interest due to their homology with those found in other organisms, the *Xac* genome codes for a large number of proteins that have not yet been characterized in any way and some for which orthologs have not yet been identified (da Silva et al. 2002). Our laboratory began its work in the SMoIBNet project focusing on the determination of the structure of a number of such *Xac* proteins.

The reining philosophy with respect to structural studies of uncharacterized (sic "hypothetical") proteins resides on the following two premises: firstly that this class of proteins may hide previously un-contemplated biological functions and secondly that high resolution structural studies of these proteins may in many cases be the most efficient means

by which these functions are revealed. Several recent examples have demonstrated the feasibility of obtaining functional information through structure (for example see Bhattacharyya et al. 2002). The *Xac* genome codes for 1658 hypothetical and conserved hypothetical proteins (da Silva et al. 2002) and we began by attacking the problem of target selection within this group for structural studies.

#### NMR – BASED ASSAY TO SCREEN TARGETS FOR STRUCTURAL STUDIES

As the success of a structural study will ultimately depend on the protein being well folded and monodisperse, we attempted to obtain information in this respect before large-scale production and purification. To do so, we selected a set of 35 proteins based on size (79–330 residues), methionine content (> 1.1% for future selenomethionine incorporation), and the absence of homologs of known function in the Protein Data Bank. Of these, 31 were successfully expressed and 19 were found to be soluble. We selectively  $^{15}\text{N}$  labeled these proteins during heterologous expression in *E. coli* and, in collaboration with the protein NMR group of Centro Nacional de Ressonância Magnética Nuclear (Universidade Federal do Rio de Janeiro), analyzed the bacterial lysates by rapid  $^1\text{H}$ - $^{15}\text{N}$  HSQC NMR analysis. The chemical shift dispersion, line-width and number of peaks observed in these 2-D spectra allowed us to classify these 19 soluble candidates into 3 groups – good (8 proteins), promising (6) and poor (5) in terms of “foldedness” (Galvão-Botton et al. 2003). The “good and “promising” groups became” our initial objects of study by NMR and X-ray diffraction and our progress in this respect will be outlined briefly below.

#### INITIAL CRYSTALLIZATION STUDIES

Three of the “good” candidates were selected for initial crystallization trials by the sparse matrix sampling approach. Of these, two were found to form crystals. The structure of one, YaeQ (XAC2396), was determined by the multiple wavelength anomalous dispersion (MAD) technique

(Guzzo et al. 2005). The YaeQ structure, determined to a resolution of 1.9 Å has no homologs in the structural databases and represents a new protein fold (C.R. Guzzo et al., unpublished data). The second protein to produce crystals, YajQ (XAC3671), diffracted only to low resolution (4 angstroms) and assays to refine the crystallization conditions are still in progress. The third (XAC1223) has so far resisted crystallization. A fourth good candidate, XAC2355, was recently submitted to crystallization tests with positive results. It diffracted to 1.9 Å at the crystallography beam line of the LNLS, initial experimental phases were calculated by molecular replacement using the recently determined *E. coli* SufE structure as an initial model (pdb 1MZG, 36% identity) and an interpretable electron density map was obtained.

#### INITIAL NMR STUDIES

Two “good” (ApaG/XAC0862 and ClpS/XAC2000) and “one” promising (XACb0070) candidates were initially selected for structure determination by NMR in collaboration with the LNLS protein NMR laboratory led by Alberto Spisni. Through the use of triple resonance experiments the backbone and side chain assignments of all three proteins have been completed (Katsuyama et al. 2003). Through the use of distance and torsion angle restraints obtained from NOEs and  $^3\text{J}$  coupling constants, the solution structure of ApaG was determined. This 124 residue protein adopts a beta cup structure in which the planes of two beta sheets are oriented at a small angle creating a deep central hydrophobic pocket. ApaG has a topology similar, though not identical, to several proteins that utilize this groove to bind to the hydrophobic lipid moieties of lipoproteins: Human GM2-Activator protein, Human PDE – subunit of rod-specific cGMP Phosphodiesterase e rhoG Dissociation Inhibitor (Pertinhez et al. manuscript in preparation). While the function of ApaG homologs in bacteria is not known, ApaG homology domains are encountered in a specific subclass (FBA) of F-box proteins that form part of the E3-ubiquitin ligase complex involved in the recognition of specific



proteins destined for ubiquitination and subsequent degradation (Ilyin et al. 2000).

The XACb0070 protein structure was also solved in a similar manner. This protein is a dimer in solution and has a topology similar to that of proteins of the Arc-MetJ family of DNA binding proteins (A. Gallo et al., unpublished data).

The third protein of this group, ClpS, was an uncharacterized protein at the beginning of this project but has since been characterized as an adapter protein that binds to the N-terminal domain of the ClpA chaperone ATPase that forms part of the ClpAP bacterial proteasome and its structure in complex with the ClpA N-domain has been determined by others. Backbone assignments, chemical shift index analysis and H/D exchange experiments indicate that the ClpS fold topology in solution is very similar to that of the crystal structure (L.M.P. Galvão-Botton et al., unpublished data). By determining the solution structure of ClpS on its own and in complex with ClpA and comparing these structures with that of the ClpA-ClpS crystal we hope to identify important structural changes in ClpS which may take place upon binding.

We are now working on the determination of the structure of XAC1883, a small uncharacterized protein coded by the *Xac Rpf* (regulation of pathogenic factors) locus responsible for the quorum sensing response. Initial multidimensional studies on  $^{13}\text{C}/^{15}\text{N}$ -labelled XAC1883 are being carried out using the 500 MHz NMR spectrophotometer at the Instituto de Química at the Universidade de São Paulo.

#### FUTURE PROJECTS

The strategy of screening the folded state of a protein in bacterial lysates before purification was successful in identifying proteins with relatively good chances for future success in structural studies, either by NMR or crystallography. While continuing to work on the structure of *Xac* proteins of unknown function, we plan to shift our focus in the future towards the structural analysis of proteins of demonstrated importance for *Xanthomonas*

pathogenicity. Our laboratory is specifically interested in understanding the molecular interactions important for the function of the type III and type IV secretion systems (Alegria et al. 2004, 2005), the regulation of quorum sensing and biofilm formation by Rpf proteins and the functions of alternative sigma factors.

#### STRUCTURE AND FUNCTION OF PROTEINS AND PEPTIDES INVOLVED IN HOST PARASITE INTERACTIONS

One of the major tasks to combat parasite diseases is studying structure-functional relationships of key molecules involved in their transmission. One is Chagas disease, transmitted by the protozoan *Trypanosoma cruzi* and its most successful insect vector, *Triatoma infestans*. The other is human Malaria caused mainly by the protozoa *Plasmodium vivax* and *P. falciparum*. Here we report some of the results we have obtained and what we have learned in participating in this network.

#### STRUCTURE AND FUNCTION OF *Triatoma infestans* PROTEINS

*Triatoma infestans* (Hemiptera: Reduviidae) is an obligate haematophagous insect that transmits *Trypanosoma cruzi*, the agent of Chagas disease. The insect becomes infected when it takes blood containing trypomastigote forms of *T. cruzi*. The trypomastigote turns into epimastigotes, which proliferate in the *T. infestans* midgut. As the nutrients in the insect gut become scarce, the parasite migrates to the posterior end of the gut and transforms into metacyclic trypomastigotes, which are then eliminated with the feces, when the insect takes a new blood meal. These released parasites are capable of contaminating new hosts through the contact with oral and eyes mucosa, skin lesions, or through ingestion. This complex cycle indicates that several *T. infestans* molecules must take part in the parasite life cycle from blood ingestion up to parasite elimination. Our aims are to characterize some of these molecules, particularly those present in the insect saliva and midgut.

Among our studies, we have characterized a 22 kD pore forming protein, named trialysin, from saliva of *T. infestans* (Amino et al. 2002). This protein contains on the N-terminus a sequence that predicts an amphipathic  $\alpha$ -helix and its corresponding synthetic peptide retains cytolytic activity. Our goal has been to understand the lytic mechanism of trialysin as well as the activation mechanism. Therefore we aimed to solve the 3D structure of the precursor and active forms of trialysin as well as the lytic peptides corresponding to the N-terminus of the molecule.

We were unable to generate the recombinant pro-trialysin, or the mature protein for diffraction or NMR studies. No bacterial transformants could be obtained using *E. coli* as host, suggesting that the protein was highly toxic. Removing the first amino acids from the N-terminus only yielded transformants when the constructs lacked the first 52 amino acids. However, the produced protein was inactive and insoluble. We could only detect pro-trialysin expression in fusion with GST. We produced up to three mg of the GST pro-trialysin fusion from one liter of bacteria culture, but the recombinant protein was unstable. We are currently trying to express similar constructs in *Pichia pastoris* to obtain higher mass of soluble protein.

In parallel, we studied the lytic activity of several synthetic peptides corresponding to the N-terminus of trialysin (Martins et al. 2006). We found that although all peptides folded into  $\alpha$ -helices in the presence of SDS or trifluoroethanol (TFE), only the ones close to the N-terminus of mature trialysin were lytic to *T. cruzi*, *E. coli*, and erythrocytes. The most active against all these targets was the peptide named P6. It is a 32mer corresponding to the N-terminus of trialysin. Peptide P7, which only differs from P6 by the absence of the last five amino acids, was less active than P6, though much less hemolytic than P6. P5, lacking the first five residues of P6, was as trypanocidal as P7, but as efficient as P6 to lyse erythrocytes. To understand the lytic mechanism and compare the different activities of these peptides, we solved their struc-

tures by NMR. We found that they fold into similar amphipathic  $\alpha$ -helices in 30% TFE. Based on obtained energy-minimized structures, we noticed the presence of a similar central domain in P5, P6 and P7 composed by the amphipathic helix, and more flexible but similar N-terminal in P6 and P7 and C-terminal domains in P6 and P5. We concluded that presence of both domains might confer stronger activity to P6.

#### STRUCTURE OF TRYPANOSOMA PROTEINS INVOLVED IN RNA PROCESSING AND PROTEIN SYNTHESIS

Trypanosomes have a unique RNA processing mechanism and translation initiation control, which therefore constitute interesting drug targets. Most of the genes are transcribed as polycistronic units that are processed by *trans*-splicing and polyadenylation. All mRNAs receive a 30–40 nt-capped RNA derived from the splice leader gene in the 5' terminal ends (Gull 2001). Both the cap structure and the enzymes involved in the cap synthesis are different in trypanosomes as compared to mammalian enzymes. For example, a guanidyl transferase of *T. brucei* has an extra N-terminal portion (Silva et al. 1998) that seems to carry an adenylate kinase domain. The RNA triphosphatase activity is also catalyzed by a unique enzyme (Ho and Shuman 2001). Therefore we selected as targets for structural studies the RNA adenosine triphosphatase from *T. cruzi* and *T. brucei*. These enzymes have been cloned, expressed in *E. coli*, purified, and submitted to crystallization assays. Three conditions were found promising and are used in refinements.

Two targets were also selected for proteins involved in the translation initiation control of trypanosomes. One is the eukaryote translation initiation factor 4E of *T. brucei* (TBeIF4E), which is the cap-binding protein. It has 251 amino acids and was chosen as a target for two reasons: a) it has only 24% similarity to the mammalian counterpart, thus representing an interesting drug target; b) the structures of the mammalian and yeast homologues have been determined (PDBs: 1L8B, 1EJ1, 1AP8), which

might help in the analysis of this protein. TBeIF4E was cloned in pET28a, expressed, purified and submitted to crystallization screenings. The other target was the eIF2 $\alpha$  of *T. brucei*. It is a 419 amino acid protein that contains an extension of ca. 120 residues in the N-terminal relative to all other known eIF2 $\alpha$  sequences. The same extension is present in *T. cruzi* and *Leishmania major*. The complete ORF as well as only the N-terminal half was cloned into pET28a and the protein expressed, purified and submitted to crystallization assays.

#### SURFACE PROTEINS OF MEROZOITE FORMS OF *Plasmodium vivax*

We are also interested in the structure of merozoite surface proteins (MSP) of *Plasmodium sp.*, most of them involved in the erythrocyte and reticulocyte invasion. Merozoites correspond to blood stages of the *Plasmodium* species released by infected cells. These studies can help understanding the biology of these organisms and designing potential agents to block its growth. We have focused on fragments of these major surface proteins, as the entire proteins are too large and insoluble when produced in recombinant form. We expressed fragments corresponding to the C-terminus of MSP3 $\alpha$  and the N and C terminus of MSP3 $\beta$  based on the formation of coiled-coil structure as demonstrated (Galinski et al. 1999). Five recombinant proteins were obtained in fusion with His-tags and purified by affinity chromatography and ion exchange. MSP3 $\alpha$  e MSP3 $\beta$  have been produced in appropriate levels, purified and submitted to crystallization assays.

#### CONCLUSIONS AND PERSPECTIVES

We adopted as a general strategy in SmolBNet to select protein targets exclusively related to our current research lines. In a few cases we were successful; this is the case of the trypsin peptides solved by NMR and the serine protease inhibitors solved by diffraction studies. For most of the chosen targets we failed in obtaining structural data, although we still have promising results. Nevertheless, we acquired experience in all steps of the

process: target selection, protein expression, crystallization, and structure resolution, and the most important achievement was to involve several students, who will keep this knowledge and experience in their scientific formation.

#### ACKNOWLEDGMENTS

The authors wish to thank the Heads of the SMolB-Net groups, technicians, post-doctoral, graduate and undergraduate fellows involved in the Network, whose names would form a long list that cannot be fitted in the space here available. Here, a few successful examples of participating groups are described, what does not mean they are the only ones. It should be noted that CeBiME members other than JARGB also helped researchers in their pre-crystallography steps, in the proteins' biophysical characterization and in solving their structures by NMR spectroscopy or crystallography. It is also necessary to point out that there were a few groups of the SMolBNet that already had structural knowledge and in one way or another helped the other groups.

#### RESUMO

Esse artigo descreve realizações do Programa SMolBNet (Rede de Biologia Molecular Estrutural) do Estado de São Paulo, apoiado pela FAPESP (Fundação de Apoio à Pesquisa do Estado de São Paulo). Ele reúne vinte grupos de pesquisa e é coordenado pelos pesquisadores do Laboratório Nacional de Luz Síncrotron (LNLS), em Campinas. O Programa SMolBNet tem como metas: Elucidar a estrutura tridimensional de proteínas de interesse aos grupos de pesquisa componentes do Programa; Prover os grupos com treinamento em todas as etapas de determinação de estrutura: clonagem gênica, expressão de proteínas, purificação de proteínas, cristalização de proteínas e elucidação de suas estruturas. Tendo começado em 2001, o Programa alcançou sucesso em ambas as metas. Neste artigo, quatro dos grupos descrevem suas participações, e discutem aspectos estruturais das proteínas que eles selecionaram para estudos.

**Palavras-chave:** genômica estrutural, cristalografia de

proteínas, ressonância nuclear magnética, estrutura de proteínas.

#### REFERENCES

- ALEGRIA MC, DOCENA C, KHATER L, RAMOS CH, DA SILVA AC AND FARAH CS. 2004. New protein-protein interactions identified for the regulatory and structural components and substrates of the type III Secretion system of the phytopathogen *Xanthomonas axonopodis* Pathovar *citri*. *J Bacteriol* 186: 6186–6197.
- ALEGRIA MC, SOUZA DP, ANDRADE MO, DOCENA C, KHATER L, RAMOS CH, DA SILVA AC AND FARAH CS. 2005. Identification of new protein-protein interactions involving the products of the chromosome and plasmid-encoded type IV secretion loci of the phytopathogen *Xanthomonas axonopodis* pv. *citri*. *J Bacteriol* 187: 2315–2325.
- AMINO RRM, MARTINS JP, HIRATA IY, JULIANO MA AND SCHENKMAN S. 2002. Trialysin, a novel pore-forming protein from saliva of hematophagous insects activated by limited proteolysis. *J Biol Chem* 277: 6207–6213.
- BHATTACHARYYA S, HABIBI-NAZHAD B, AMEGBEY G, SLUPSKY CM, YEE A, ARROWSMITH C AND WISHART DS. 2002. Identification of a novel archaeobacterial thioredoxin: determination of function through structure. *Biochemistry* 41: 4760–4770.
- CAMPOS ITN, GUIMARÃES BG, MEDRANO FJ, TANAKA AS AND BARBOSA JARG. 2004. Crystallization, data collection and phasing of infestin 4, a factor XIIa inhibitor. *Acta Crystall D* 60: 2051–2053.
- COPLEY SD, NOVAK WRP AND BABBITT PC. 2004. Divergence of function in the thioredoxin superfamily: Evidence for evolution of peroxiredoxins from a thioredoxin-like ancestor. *Biochemistry* 43: 13981–13995.
- CUSSIOL JRR, ALVES SV, OLIVEIRA MA AND NETTO LES. 2003. Organic hydroperoxide resistance gene encodes a thiol-dependent peroxidase. *J Biol Chem* 278: 11570–11578.
- DA SILVA AC ET AL. 2002. Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. *Nature* 417: 459–463.
- DEMASI APD, PEREIRA GAG AND NETTO LES. 2001. Cytosolic thioredoxin peroxidase I is essential for the antioxidant defense of yeast with dysfunctional mitochondria. *FEBS Letters* 509: 430–434.
- DISCOLA KF, OLIVEIRA MA, MONTEIRO-SILVA G, BARCENA JA, PORRAS P, PADILLA A, NETTO LES AND GUIMARÃES BG. 2005. Crystallization and preliminary X-ray diffraction analysis of glutaredoxin 2 from *Saccharomyces cerevisiae* in different oxidation states. *Acta Crystall F*, F61: 445–447.
- GALINSKI MR, CORREDOR-MEDINA C, POVOA M, CROSBY J, INGRAVALLO P AND BARNWELL JW. 1999. *Plasmodium vivax* merozoite surface protein-3 contains coiled-coil motifs in an alanine-rich central domain. *Mol Biochem Parasitol* 101: 131–147.
- GALVÃO-BOTTON LMP, KATSUYAMA AM, GUZZO CR, ALMEIDA FCL, FARAH CS AND VALENTE AP. 2003. High-throughput screening of structural proteomics targets using NMR. *FEBS Letters* 552: 207–213.
- GOVINDARAJAN S, RECARBARREN R AND GOLDSTEIN RA. 1999. Estimating the Total Number of Protein Folds. *PROTEINS: Structure, Function, and Genetics* 35: 408–414.
- GULL K. 2001. The biology of kinetoplastid parasites: insights and challenges from genomics and post-genomics. *International Journal of Parasitology* 31: 443–452.
- GUZZO CR, NAGEM RAP, GALVÃO-BOTTON LMP, GUIMARÃES BG, MEDRANO FJ, BARBOSA JARG AND FARAH CS. 2005. Expression, purification, crystallization and preliminary X-ray analysis of YaeQ (XAC2396) from *Xanthomonas axonopodis* pv *citri*. *Acta Crystall F*, F61: 493–495.
- HO CK AND SHUMAN S. 2001. *Trypanosoma brucei* RNA triphosphatase. Antiprotozoal drug target and guide to eukaryotic phylogeny. *J Biol Chem* 276: 46182–46186.
- ILYIN GP, RIALLAND M, PIGEON C AND GUGUEN-GUILLOUZO C. 2000. cDNA cloning and expression analysis of new members of the mammalian F-box protein family. *Genomics* 67: 40–47.
- KATSUYAMA AM, CICERO DO, SPISNI A, PACI M, FARAH CS AND PERTINHEZ TA. 2003. <sup>1</sup>H, <sup>15</sup>N and <sup>13</sup>C resonance assignments of the ApaG protein of the phytopathogen *Xanthomonas axonopodis* pv. *citri*. *J. Biomolecular NMR* 29: 423–424.

- MARTINS RM, SFORÇA ML, AMINO R, JULIANO MA, OYAMA JR S, JULIANO L, PERTINHEZ TA, SPISNI A AND SCHENKMAN S. 2006. Lytic activity and structural differences of amphipathic peptides derived from trialysin. *Biochemistry*, in press.
- MONTEIRO G AND NETTO LES. 2004. Glucose repression of PRX1 expression is mediated by Tor1p and Ras2p through inhibition of Msn2/4p in *Saccharomyces cerevisiae*. *FEMS Microbiol Lett* 241: 221–228.
- MONTEIRO G, PEREIRA GAG AND NETTO LES. 2002. Regulation of mitochondrial thioredoxin peroxidase I expression by two different pathways: one dependent on cAMP and the other on heme. *Free Rad Biol Med* 32: 278–288.
- MOREIRA LM, DE SOUZA RF, ALMEIDA JR NF, SETUBAL JC, OLIVEIRA JC, FURLAN LR, FERRO JA AND DA SILVA AC. 2004. Comparative genomics analyses of citrus-associated bacteria. *Annu Rev Phytopathol.* 42: 163–184.
- MUNHOZ DC AND NETTO LES. 2004. Cytosolic thioredoxin peroxidase I and II are important defenses of yeast against organic hydroperoxide insult. Catalases and peroxiredoxins cooperate in the decomposition of H<sub>2</sub>O<sub>2</sub> by yeast. *J Biol Chem* 279: 35219–35227.
- OLIVEIRA MA, NETTO LES, MARTIN FJM, BARBOSA JARG, ALVES SV, CUSSIOL JRR AND GUIMARÃES BG. 2004. Crystallization and preliminary X-ray diffraction analysis of an oxidized state of Ohr from *Xylella fastidiosa*. *Acta Crystall D* 60: 337–339.
- OLIVEIRA MA, DISCOLA KF, ALVES SV, BARBOSA JARG, MEDRANO FJ, NETTO LES AND GUIMARÃES BG. 2005. Crystallization and preliminary X-ray diffraction analysis of NADPH-dependent thioredoxin reductase from *Saccharomyces cerevisiae*. *Acta Crystall F*, F61: 387–390.
- RIETSCH A AND BECKWITH J. 1998. The genetics of disulfide bond metabolism. *Annu Rev Genet* 32: 163–184.
- SILVA E, ULLU E, KOBAYASHI R AND TSCHUDI C. 1998. Trypanosome capping enzymes display a novel two-domain structure. *Mol Cell Biol* 18: 4612.
- SIMPSON AJG ET AL. 2000. The genome sequence of the plant pathogen *Xylella fastidiosa*. *Nature* 406: 151–159.
- VAN SLUYS MA, MONTEIRO-VITORELLO CB, CARMARGO LE, MENCK CF, DA SILVA AC, FERRO JA, OLIVEIRA MC, SETUBAL JC, KITAJIMA JP AND SIMPSON AJ. 2002. Comparative genomic analysis of plant-associated bacteria. *Annu Rev Phytopathol* 40: 169–189.