Available online at www.sciencedirect.com
ScienceDirect
Molecular & Biochemical Parasitology 151 (2007) 100–110
MOLECULAR & BIOCHEMICAL PARASITOLOGY

Genome-scale protein expression and structural biology of *Plasmodium falciparum* and related Apicomplexan organisms

Masoud Vedadi [a], Jocelyne Lew [a], Jennifer Artz [a], Mehrnaz Amani [a], Yong Zhao [a], Aiping Dong [a], Gregory A. Wasney [a], Mian Gao [a], Tanya Hills [a], Stephen Brokx [a], Wei Qiu [a], Sujata Sharma [a], Angelina Diassiti [b], Zahoor Alam [a], Michelle Melone [a], Anne Mulichak [c], Amy Wernimont [a], James Bray [d], Peter Loppnau [a], Olga Plotnikova [a], Kate Newberry [a], Emayavaram Sundararajan [a], Simon Houston [a], John Walker [a], Wolfram Tempel [a], Alexey Bochkarev [a], Ivona Kozieradzki [a], Aled Edwards [a], Cheryl Arrowsmith [a], David Roos [e], Kevin Kain [a], Raymond Hui [a,*]

[a] *Structural Genomics Consortium, U. of Toronto, 100 College St. Rm 522B, Toronto, Ont., Canada M5G 1L5*
[b] *McLaughlin-Rotman Centre for Global Health, Toronto General Hospital, Eaton Wing Ground Floor, Room 224, 200 Elizabeth Street, Toronto, Ont., Canada M5G 2C4*
[c] *APS, Argonne National Lab Building 435A, Sector 17, 9700 South Cass Avenue Argonne, IL 60439, United States*
[d] *Structural Genomics Consortium (SGC), U. of Oxford, Oxford, United Kingdom*
[e] *Department of Biology, U. of Pennsylvania, 301 Goddard Laboratories, Philadelphia, PA 19104, United States*

Received 26 April 2006; received in revised form 19 October 2006; accepted 20 October 2006
Available online 13 November 2006

Abstract

Parasites from the protozoan phylum *Apicomplexa* are responsible for diseases, such as malaria, toxoplasmosis and cryptosporidiosis, all of which have significantly higher rates of mortality and morbidity in economically underdeveloped regions of the world. Advances in vaccine development and drug discovery are urgently needed to control these diseases and can be facilitated by production of purified recombinant proteins from Apicomplexan genomes and determination of their 3D structures. To date, both heterologous expression and crystallization of Apicomplexan proteins have seen only limited success. In an effort to explore the effectiveness of producing and crystallizing proteins on a genome-scale using a standardized methodology, over 400 distinct *Plasmodium falciparum* target genes were chosen representing different cellular classes, along with select orthologues from four other *Plasmodium* species as well as *Cryptosporidium parvum* and *Toxoplasma gondii*. From a total of 1008 genes from the seven genomes, 304 (30.2%) produced purified soluble proteins and 97 (9.6%) crystallized, culminating in 36 crystal structures. These results demonstrate that, contrary to previous findings, a standardized platform using *Escherichia coli* can be effective for genome-scale production and crystallography of Apicomplexan proteins. Predictably, orthologous proteins from different Apicomplexan genomes behaved differently in expression, purification and crystallization, although the overall success rates of *Plasmodium* orthologues do not differ significantly. Their differences were effectively exploited to elevate the overall productivity to levels comparable to the most successful ongoing structural genomics projects: 229 of the 468 target genes produced purified soluble protein from one or more organisms, with 80 and 32 of the purified targets, respectively, leading to crystals and ultimately structures from one or more orthologues.

Keywords: Structural genomics; Heterologous protein expression; Apicomplexa; Malaria; Crystallography; Crystallization

1. Introduction

The *Apicomplexa* is a protozoan phylum of obligate intracellular parasites characterized by an apical complex, where organelles, such as micronemes and rhoptries reside. While *Eimeria*, *Neospora*, *Babesia* and *Theileria* are causes of mainly veterinary health concerns, organisms from other apicomplexan

Abbreviations: Pf, *Plasmodium falciparum*; Py, *Plasmodium yoelii*; Pb, *Plasmodium berghei*; Pv, *Plasmodium vivax*; Pk, *Plasmodium knowlesi*; Cp, *Cryptosporidium parvum*; Tg, *Toxoplasma gondii*

* Corresponding author. Tel.: +1 416 946 7182; fax: +1 416 946 0588.
E-mail address: raymond.hui@utoronto.ca (R. Hui).

0166-6851/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.molbiopara.2006.10.011

genera, such as *Plasmodium*, *Cryptosporidium* and *Toxoplasma* are responsible for high rates of morbidity and mortality in humans, particularly in economically underdeveloped regions of the world. *Plasmodium* parasites alone are annually responsible for over 300 million cases of human malaria, resulting in up to 3 million deaths [1]. Cryptosporidiosis and toxoplasmosis – mediated, respectively, by *Cryptosporidium parvum* and *Toxoplasma gondii* – are opportunistic infections and major causes of morbidity and mortality amongst immuno-compromised patients, particularly in those infected with HIV. Co-infection with HIV and malaria is also particularly common in Africa, resulting in severe malaria anemia [2]. Simply put, Apicomplexan parasites are ravaging the parts of the world without the infrastructure to control them.

Apicomplexan diseases lack effective treatment. While there is currently no cure for toxoplasmosis and cryptosporidiosis, various anti-malarial drugs exist; however, economic, geopolitical and scientific factors have conspired to allow malaria to re-emerge in the last decade as the leading global cause of child mortality. Scientifically, the challenge is manifold: the paucity of validated drug targets, complexity of Apicomplexan life cycles as well as highly adaptable gene expression mechanisms leading to a unique ability to develop drug resistance and evade immune response.

The urgently needed advances in vaccine development and drug discovery for Apicomplexan diseases can be significantly facilitated by genome-scale production of purified recombinant proteins from *Apicomplexa* and determination of their 3D structures. This structural genomics approach has already been proven effective for a number of genomes. For example, 70% of 424 targets were expressed as soluble proteins in a project on the thermophilic archaeon *Methanobacterium thermoautotrophicum* [3,4]. Of these, 47% and 40%, respectively, of small (<20 kDa) and large proteins yielded 2 mg or more of soluble protein. In addition, 19% and 9%, respectively, of small and large proteins crystallized. A total of 36 structures (8.4%) were derived from these crystals or NMR samples. Similarly, application of structural genomics to 1376 *Thermotoga maritima* clones produced 542 purified targets, 432 crystallized proteins and 24 unique structures [5,6]. Eukaryotic genomes are also being tackled. From a set of 250 *Saccharomyces cerevisiae* proteins, 88% were found to be expressed, with 60 out of 250 or 24% yielding sufficient soluble protein for crystal trial [7]. This culminated in 22 crystallized proteins (8.8%) and 14 structures (5.6%). The *Arabidopsis thaliana* genome has also yielded 496 pure recombinant proteins out of 632 targets, with the MBP fusion tag used to aid solubilization [8]. These projects focused primarily on non-membrane proteins. In a high throughput membrane protein project involving 280 proteins from *Escherichia coli* and *T. maritima*, 30% of the cloned proteins expressed in *E. coli*, leading to 22 pure proteins (7.9%), 2 crystals (0.7%) and 1 structure [9]. Clearly, genome-scale protein expression and structural biology have been successfully implemented in both prokaryotic and eukaryotic organisms.

In contrast, there is no report of successful large-scale protein production or structural biology in any Apicomplexan genomes. Instead, the relevant literature centers around problems of obtaining purified *Plasmodium* proteins, including codon mismatch [10,11] and toxicity of plasmodium proteins [12]. There are also reports of isolated successes using custom techniques, such as specialized expression vectors, codon optimization and refolding [11–13], which may not be effective for a majority of proteins. On a larger scale, two independent pilot projects on expression have, respectively, yielded 13 purified soluble proteins from 368 *Pf* genes [14] and 9 purified proteins from 95 *Pf* genes [10]—a level of success lower even than that achieved with membrane proteins. Most recently and most significantly, Mehlin et al. reported [15] successful expression of soluble proteins from only 63 of 1000 open reading frames from *Plasmodium falciparum*. This study spanned various classes of proteins and yielded instructive findings from comprehensive statistical analysis: (a) smaller proteins, proteins with p*I* lower than 6 and those with relatively higher homology to *E. coli* homologues are more likely to express in soluble form in *E. coli*; (b) codon usage and AT-contents do not affect expression. The general conclusion from previous works is that *E. coli* is not an effective expression host for *Plasmodium* proteins.

Problems in expressing Apicomplexan proteins could be presaged from genomic analysis. The *Pf* genome is the most AT-rich of all genomes sequenced to date [1], at 80%, with the *Py* [16] and *Cp* [17] genomes not far behind at 78% and 70%, respectively. The unusually high AT bias translates into some codons rarely required by *E. coli* proteins, e.g. AGA and AGG [18], highlighting an intrinsic problem with heterologous expression. In addition, *Pf* genes have a mean length of 2.3 kb—1 kb longer than homologues from other organisms [1], with the extra length often featuring unique inserts. Many *Plasmodium* proteins have low complexity regions consisting of long hydrophobic stretches, amino acid repeats or segments highly rich in amino acids encoded by AT-biased codons, notably lysine and asparagine. Furthermore, 10% of the *Pf* proteins are targeted to the apicoplast, with another 10% predicted to be secreted [19,20] to the host. The N-termini of these proteins typically contain sequence motifs regulating their localization. These peptide regions influence localization rather than function but are membrane-like in their effects on expression and folding. Codon bias, size, sequence inserts, signal and transit peptides are all features that can be predicted to affect not only recombinant protein expression, but also crystallization.

Predictably, limited success in recombinant protein expression has translated into a relatively low number of protein structures to date. As of December 31st, 2005, the Protein Data Bank (http://www.pdb.org) included 78 unique structures (sharing less than 90% in sequence identity) from all *Plasmodium* species, 8 from *T. gondii* and 6 from *Cryptosporidium hominis* and *C. parvum* combined. From 2001 to 2004, the number of novel *Plasmodium* structures deposited in the PDB was 6, 8, 10 and 14. The number jumped to 24 in 2005, spurred in part by the release of genomic data.

While the *Py* and *Cp* genomes are close to *Pf* in AT-contents, they differ in other respects. Both *Py* and *Cp* genes are on average shorter [16,17], with fewer proteins exported. Furthermore, other Apicomplexan genomes, including *Plasmodium vivax*, *Plasmodium knowlesi* and *T. gondii* are more balanced in their