

## Research Article

# Comparative Analysis of Apicoplast-Targeted Protein Extension Lengths in Apicomplexan Parasites

**Alexandr V. Seliverstov, Oleg A. Zverkov, Svetlana N. Istomina,  
Sergey A. Pirogov, and Philip S. Kitsis**

*Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute),  
Bolshoy Karetny Pereulok 19, Moscow 127994, Russia*

Correspondence should be addressed to Alexandr V. Seliverstov; [slvstv@iitp.ru](mailto:slvstv@iitp.ru)

Received 10 September 2014; Accepted 25 December 2014

Academic Editor: William H. Piel

Copyright © 2015 Alexandr V. Seliverstov et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In general, the mechanism of protein translocation through the apicoplast membrane requires a specific extension of a functionally important region of the apicoplast-targeted proteins. The corresponding signal peptides were detected in many apicomplexans but not in the majority of apicoplast-targeted proteins in *Toxoplasma gondii*. In *T. gondii* signal peptides are either much diverged or their extension region is processed, which in either case makes the situation different from other studied apicomplexans. We propose a statistic method to compare extensions of the functionally important regions of apicoplast-targeted proteins. More specifically, we provide a comparison of extension lengths of orthologous apicoplast-targeted proteins in apicomplexan parasites. We focus on results obtained for the model species *T. gondii*, *Neospora caninum*, and *Plasmodium falciparum*. With our method, cross species comparisons demonstrate that, in average, apicoplast-targeted protein extensions in *T. gondii* are 1.5-fold longer than in *N. caninum* and 2-fold longer than in *P. falciparum*. Extensions in *P. falciparum* less than 87 residues in size are longer than the corresponding extensions in *N. caninum* and, reversely, are shorter if they exceed 88 residues.

## 1. Introduction

In general, the mechanism of protein translocation through the apicoplast membrane requires a specific extension of a functionally important region of the apicoplast-targeted proteins. In *T. gondii* signal peptides are either much diverged or their extension region is processed, which in either case makes the situation different from other studied apicomplexans. We propose a statistic method to compare extensions of the functionally important regions of apicoplast-targeted proteins. More specifically, we provide a comparison of extension lengths of orthologous apicoplast-targeted proteins in apicomplexan parasites. We ground on the notion that the majority of cyanobacterial proteins lack such extensions (including signal peptides) and consist of only functional sequences.

Sporozoans comprise a monophyletic lineage of apicomplexan parasites. Among them, *Toxoplasma gondii* is an important medical and veterinary pathogen commonly

causing morbidity in HIV patients [1, 2]. The study [3, 4] describes the propagation mechanism of *T. gondii* in various hosts worldwide, including several aquatic mammal species, where it may provoke abortion and lethal systemic disease. The observation that the apicoplast of *T. gondii* significantly varies in shape and protein expression patterns at different life stages of the parasite suggests its important role in virulence; the apicoplast of *T. gondii* is also involved in the pathogen stage conversion and the parasite proliferation [5]. Due to bacterial origin of the apicoplast proteins, they present a natural target for selective treatment in the eukaryotic host. The sporozoans contain a semiautonomous organelle, the apicoplast, acquired by secondary endosymbiosis with ancient red algae; plastid organelles of the red algae originate from cyanobacteria [6–8].

Elucidating the molecular mechanism that underlies the role of apicoplast in the parasite invasion, conversion, and proliferation is important for development of novel therapeutics to control infection and reactivation of the parasite.

Further analysis of unique features of apicoplast-targeted proteins (particularly, regions involved in translocation processes) in *T. gondii* can add to the effective design of drug-based or genetic strategies to control the pathogen development and proliferation. At the reported stage of the study, we analyze only extensions in length of orthologs among apicomplexan parasites.

Note that the coccidian *Cryptosporidium parvum* lacks the apicoplast [9], and the apicoplast in piroplasmids *Babesia bovis* and *Theileria parva* largely differs from that in common coccidians and the haemosporidian *Plasmodium* spp. [10, 11].

The majority of apicoplast proteins are encoded in the nucleus and only few in its own genome. Most of these proteins can be identified due to their cyanobacterial origin. Transport of nuclear-encoded proteins to the apicoplast in *T. gondii* is significantly less documented experimentally compared to *Plasmodium falciparum*. Among the documented cases is the nuclear-encoded lipoic acid synthetase LipA [12]; other examples are described in [13–15]. A mechanism of protein import into secondary plastids is also described in [16], where many orthologous proteins involved in this process were shown to be presented in the sporozoans *P. falciparum* and *T. gondii*, cryptophyte alga *Guillardia theta*, and diatom *Phaeodactylum tricorutum*. Plastids also possess the bacterial system to translocate folded proteins [17, 18].

A variety of protein localization prediction methods are used to identify apicoplast-targeted proteins. Some of them utilize the notion that translocation across the four membranes surrounding the apicoplast is mediated by an N-terminal bipartite targeting sequence, a special N-terminal signal, and a transit peptide [13]. The algorithm ApicoAP described in [19] predicts apicoplast-targeted proteins containing the signal peptide, because it trains on a learning sample of signal peptide-containing proteins. Other apicoplast-targeted proteins are predicted neither with this algorithm nor with ApicoAMP [20]. The comprehensive ToxoDB database constructed using the SignalP algorithm contains proteins with the information on presence/absence of the signal peptide. According to this database, some nuclear-encoded proteins in *T. gondii* that are experimentally shown to reach the apicoplast should do not contain signal peptides, albeit bearing housekeeping functions in the apicoplast. The methods in application to *Plasmodium* spp. are described, for example, in [19, 21–23]. The PlasmoAP algorithm [24] is designed specifically for *Plasmodium* spp. and is of little applicability to coccidians. Hence, these two widely used databases may be considered of limited use to identify apicoplast-targeted proteins not containing the standard signal peptide in coccidians. We therefore applied a crude technique to compare apicomplexan proteins with their orthologs in a cyanobacterium. Namely, orthology between nuclear-encoded sporozoan proteins and cyanobacterial proteins is used as a basis to suggest the apicoplast-targeted nature of the proteins. As our study relies on statistic estimates, its predictions are hopefully not affected by the chosen parameters of global protein alignment.

In this work, the lengths of sporozoan proteins are compared with each other and with the length of their orthologs in the cyanobacterium *Synechocystis* sp. PCC 6803. We

consider lengths of the sporozoan proteins that extend outside the conserved alignment region, which usually covers the entire cyanobacterial sequence. We focus on results obtained for the model species *T. gondii*, *Neospora caninum* (the two coccidian sporozoans with completed genome projects, as per the end of 2013), and the malaria agent *P. falciparum* from the Haemosporidia.

Based on the total comparison, we conclude that *T. gondii* in most cases contains longer proteins compared to both *N. caninum* and *P. falciparum*. We also surmised that at least some of them undergo processing in the cytoplasm to facilitate transporting into the apicoplast. The extended portions of proteins may also be involved in gene expression regulation at the level of protein-protein interaction.

As an argument, the regulation of plastid-encoded genes *ycf24* and *rps4* affects the general functionality of the apicoplast in *T. gondii* [25]. The expression regulation of *ycf24* (the SufB factor mediating the Fe-S cluster assembly in many nuclear proteins) was suggested to take place in the apicomplexans *Eimeria tenella*, *T. gondii* RH, and *Plasmodium* spp., as well as in *Gracilaria tenuistipitata*, *Porphyra purpurea*, and *Porphyra yezoensis* [25]. The same type of regulation was suggested for *rps4* (ribosomal protein S4) in *T. gondii* RH [25].

## 2. Materials and Methods

Protein data for *T. gondii* and *N. caninum* was extracted from the ToxoDB database (version 8.2), data for *Plasmodium* spp. from the PlasmoDB database (version 9.3), and data for *Synechocystis* sp. PCC 6803 from GenBank, NCBI [26]. ToxoDB and PlasmoDB are specialized, regularly updated, and nonoverlapping databases. Conserved domains were detected according to the Pfam database [27]. The location of regions enriched with a certain amino acid was established using the PROSITE database [28].

We compared the proteomes of three apicomplexan parasites (*T. gondii* ME49, *N. caninum* Liverpool, and *P. falciparum* 3D7) and the cyanobacterium *Synechocystis* sp. PCC 6803. For each pair of proteomes, pairs of orthologous proteins were computed on the basis of an alignment quality score using the Needleman-Wunsch method and BLOSUM62 matrix [29, 30].

Our method to study the lengths of the apicomplexan protein extensions is as follows. For each cyanobacterial protein with length  $s$  and its two orthologs (from a fixed pair of apicomplexans) with lengths  $a$  and  $a'$ , the point with coordinates  $(a - s, a' - s)$  is computed. In some cases, one or both of the coordinates are negative, which indicates a sporadic case of a shorter length of the sporozoan protein versus the cyanobacteria.

In Figures 1–3, the cases of *N. caninum*-*T. gondii* (*N-T*), *P. falciparum*-*T. gondii* (*P-T*), and *P. falciparum*-*N. caninum* (*P-N*) are analyzed using three sets of points. Each coordinate is the difference in lengths between the sporozoan and cyanobacterial orthologs: *T. gondii* versus *Synechocystis* (*T-S*), *N. caninum* versus *Synechocystis* (*N-S*), and *P. falciparum* versus *Synechocystis* (*P-S*). The sets of points are then statistically analyzed. The following statistic was used to test

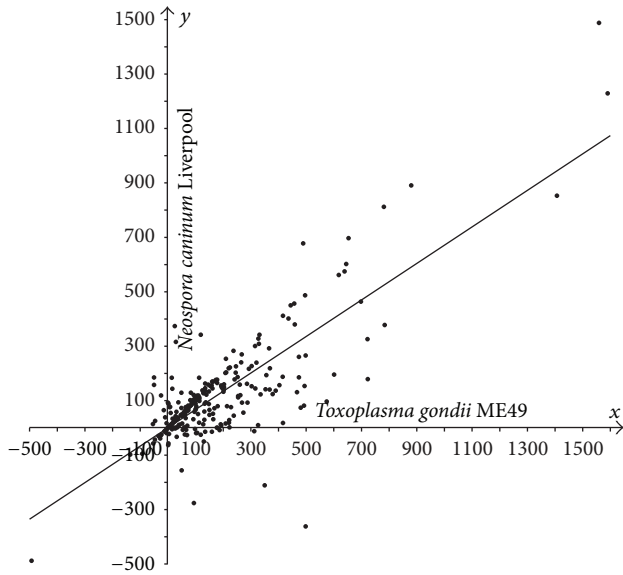


FIGURE 1: Plot of protein length extensions in *T. gondii* ME49 versus *N. caninum* Liverpool relative to their orthologs in *Synechocystis* sp. PCC 6803. Differences in lengths between sporozoan and cyanobacterial orthologous proteins are plotted as follows: *T*-S on the *x*-axis, *N*-S on the *y*-axis.

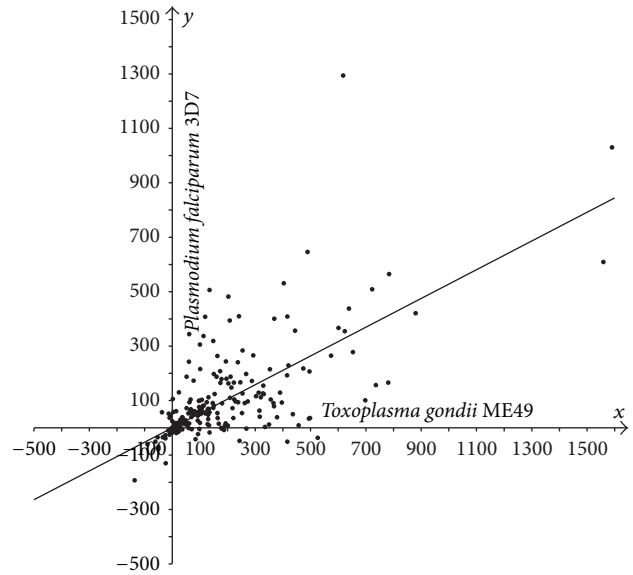


FIGURE 2: Plot of protein length extensions in *T. gondii* ME49 versus *P. falciparum* 3D7 relative to their orthologs in *Synechocystis* sp. PCC 6803. Differences in lengths between sporozoan and cyanobacterial orthologous proteins are plotted as follows: *T*-S on the *x*-axis, *P*-S on the *y*-axis.

the hypotheses that “a constant is better compatible with the set of points  $\{(x_i, y_i) \mid 1 \leq i \leq n\}$  than the nontrivial affine function  $y = kx + b, k \neq 0$ ” and “the linear function  $y = kx$  is better compatible with this set of points than the affine function  $y = kx + b, b \neq 0$ ”:  $F = \sqrt{\frac{((\sum_i (y_i - \hat{y}_0(x_i))^2) / \sum_i (y_i - \hat{y}(x_i))^2) - 1) \cdot (n - 2)}{}}$ , where  $\hat{y}_0$  in the numerator is a constant (mean  $\bar{y}$  over all  $y$ ) or linear regression  $y = kx$  and  $\hat{y}$  in the denominator is affine regression  $y = kx + b$ . This statistic can be explained more clearly: it determines whether there is a correlation between the difference  $y_i - \hat{y}(x_i)$  and  $x_i$ . This statistic is standard and substantiated in [31, 32]. The value of  $F$  was compared against a threshold defined as the Student random variable at significance level  $\alpha, t(n - 2, \alpha)$ . Under the number of degrees of freedom  $n - 2 > 30$ , the Student and standard Gaussian distributions approximate each other, and the threshold  $t(266, 0.05)$  equals 1.96. An analogous statistic was used to test the hypothesis “affine function versus general polynomial of second degree.” The confidence interval radius and the radius of the intercept (further referred to as *radius*) for the affine regression slope as well as the slope coefficient radius for linear regression were calculated in a standard fashion [32]. The Student test statistic  $S$  was used as well [32]. Deming regression and screening singular points were tested as well.

Regions of proteins with a predominance of one amino acid were determined by using the PROSITE program. The distribution of amino acid pairs separated by a fixed distance  $k$  in a given set of amino acid sequences was established using the simple computer program available from <http://lab6.iitp.ru/utlils/aapf/>. Namely, frequencies of all amino acid pairs occurring in the given sequences at

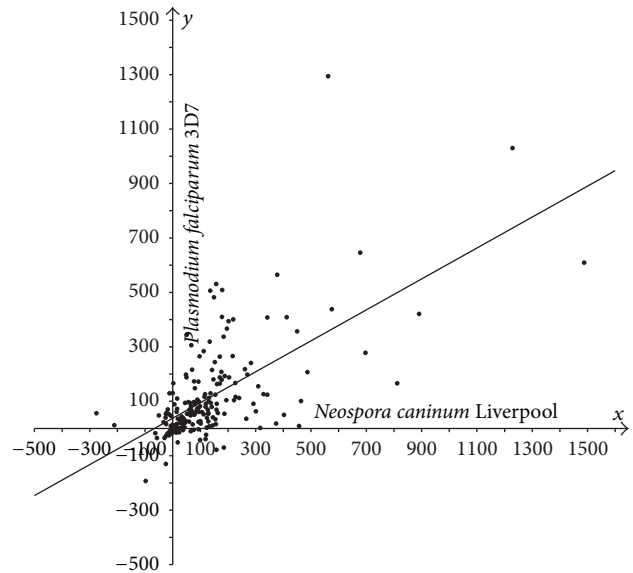


FIGURE 3: Plot of protein length extensions in *N. caninum* Liverpool versus *P. falciparum* 3D7 relative to their orthologs in *Synechocystis* sp. PCC 6803. Differences in lengths between sporozoan and cyanobacterial orthologous proteins are plotted as follows: *N*-S on the *x*-axis, *P*-S on the *y*-axis.

the distance of  $k$  residues (specified in the interval from 0 to 255) are computed and averaged over all sequences. The output is a frequency matrix of amino acid pairs. This matrix can be used to characterize nonstandard types of the putative signal peptide. This way it also appeared impossible to determine specificity of the N-terminus of apicomplast-targeted proteins in *T. gondii*; refer to Figure 4.

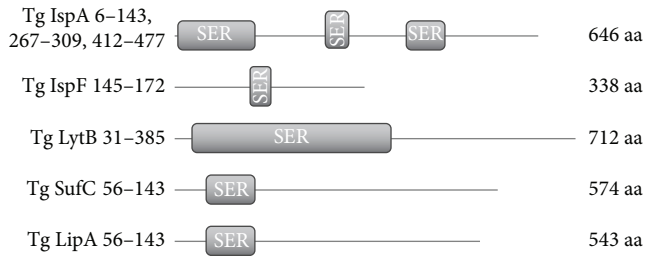


FIGURE 4: Some apicoplast-targeted proteins in *T. gondii* containing serine-rich regions. The region coordinates indicated on the left and the protein lengths on the right. The serine-rich regions are arranged irregularly, which is confirmed on a larger number of proteins for different pairs, triples, and so forth of amino acids.

### 3. Results and Discussion

Orthologs of *Synechocystis* sp. PCC 6803 were identified for 515 of 8319 (~6%) nuclear proteins in *T. gondii*, 560 of 7122 (~8%) nuclear proteins in *N. caninum*, and 390 of 5538 (~7%) nuclear proteins in *P. falciparum*. Only 877 of 3179 (~28%) proteins in *Synechocystis* sp. were found to be orthologous against at least one of the three apicomplexan species (see Supplementary Material available online at <http://dx.doi.org/10.1155/2015/452958>).

The identified orthologs are putative apicoplast-targeted proteins. Among them are proteins with either experimentally shown or anticipated apicoplast affinity, such as the bacterial type RNA polymerase sigma subunit (RpoD), DNA ligase, aminoacyl-tRNA synthetases, cell-cycle-associated protein kinase PRP4, enzymes IspA, IspB, IspE, IspF, IspG (GpcE), and IspH (LytB) of the mevalonate-independent pathway of isoprenoid biosynthesis, sulphur mobilization protein SufC from a Fe-S cluster assembly pathway, and LipA and LipB enzymes of lipoic acid synthesis (refer to the Introduction [5, 12, 14, 15]). In pairwise alignments, the sporozoan and cyanobacterial proteins usually align well at their C-termini, and the cyanobacterial sequence is fully covered by the alignment. In many cases, the N-termini of sporozoan proteins extend outside the alignment (data not shown).

In most cases, sporozoan proteins are longer compared to their bacterial orthologs, Figures 1–3. We demonstrate statistically that the majority of proteins in *T. gondii* are considerably longer compared to their orthologs in *N. caninum* Liverpool and *P. falciparum* 3D7, which was evidenced previously only for selected proteins [12, 33].

The hypotheses “a constant is better than the nontrivial affine function” and “affine function versus general polynomial of second degree” were rejected for every three sets of points shown in Figures 1–3. The hypothesis “the linear function is better than the affine function” was compatible with the first two sets ( $F = 0.15$  and  $F = 1.57$ ; refer to the designation in Materials and Methods section) and was rejected for the third set ( $F = 3.40$ ). Thus, the third set was tested against the hypothesis “the mean over all  $x$ -coordinates coincides with the mean over all  $y$ -coordinates”;

this hypothesis was accepted with the Student test statistic  $S$  at the same significance level  $\alpha$  (with  $S = 1.547$ ) [32].

Hence, the following regressions were justified. For set 1,  $y = 0.6711x$  with radius 0.0468; for set 2,  $y = 0.528x$  with radius 0.0590; for set 3,  $y = 0.5685x + 37.756$  (linear regression rejected with  $F = 3.40$ ) with radii 0.0926 and 21.7521, respectively.

The Deming regression gives approximately the same estimates; screening singular points does not significantly affect the results (data not shown).

So, the following conclusions can be drawn for the apicoplast protein orthologs that have orthologs in the cyanobacterium.

- (1) Protein extensions in *T. gondii* are on average 1.5-fold longer compared to the corresponding extensions in *N. caninum*, with almost 1.0 confidence (Figure 1).
- (2) Protein extensions in *T. gondii* are on average 2-fold longer compared to the corresponding extensions in *P. falciparum*, with high confidence (Figure 2).
- (3) Set 3 (Figure 3) is compatible with the hypothesis that the average of protein extension lengths in *N. caninum* equals that in *P. falciparum*. Extension lengths in *P. falciparum* being less than 87 residues are longer than the corresponding extensions in *N. caninum* and, reversely, are shorter if they exceed 88 residues. In units, the dependency between extension lengths in *P. falciparum* versus *N. caninum* is an affine function  $y = 0.5685x + 37.756$ , where  $y$  runs over extension lengths in *P. falciparum* and  $x$  in *N. caninum*. The affinity, but not the linearity, of the regression testifies on behalf of the difference of *T. gondii* from her immediate species *P. falciparum* and *N. caninum* once again.

Among other specific features of apicoplast-targeted proteins is the abundance of serine-rich regions revealed in analyses with PROSITE (Figure 4). Each of the 3551 proteins in *T. gondii* ME49 possesses at least one 27 amino acid-long region with at least 9 serine residues, and 39 proteins possess at least one region with 27 or more continuous serine residues. Contrary to our expectations, larger-scale searching for serine-rich motifs in *T. gondii* showed their presence in various protein families, thus suggesting a selectively neutral nature of their origin. In other words, serine-rich regions are not specific to N-termini of apicoplast-targeted proteins. The same is also observed for other amino acids. This approach does not allow detecting a novel type of the N-terminal signal.

Earlier preliminary results are reported in [33].

### 4. Conclusions

For apicomplexan parasites, we suggest a statistically based method to compare the extension lengths of orthologous proteins that have orthologs in the cyanobacterium.

With this method, we demonstrate that the majority of cyanobacterium orthologs in *Toxoplasma gondii* are significantly longer compared to those in both *Neospora caninum* and *Plasmodium falciparum*. These proteins commonly lack



signal sequences typical for *Plasmodium* spp. [34]. The corresponding extensions might be essential for regulation of the apicoplast proteins and their translocation into the apicoplast. This notion conforms well with the observation that the apicoplast membrane in *T. gondii* is known to be less permissible, at least against drugs, compared to that in *P. falciparum* (personal communication with Gamaleya Research Institute of Epidemiology and Microbiology). Differences in protein extension lengths between *T. gondii* and other apicomplexan species may suggest different membrane transport mechanisms in these sporozoan groups. Mechanism of regulation and translocation in *T. gondii* may be based on protein processing in the cytoplasm to mature their extended N-termini.

### Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

### Acknowledgments

The authors are deeply respectful to the editor for valuable comments that improved the paper. They also thank L. Rusin for valuable discussions and help with preparing the paper. Research was funded by the Russian Scientific Fund (Project 14-50-00150).

### References

- [1] T. N. Ermak, A. B. Peregudova, V. I. Shakhgildian, and D. B. Goncharov, "Cerebral toxoplasmosis in the pattern of secondary CNS involvements in HIV-infected patients in the Russian federation: clinical and diagnostic features," *Meditsinskaja Parazitologija i Parazitarnye Bolezni*, no. 1, pp. 3–7, 2013.
- [2] H. N. Luma, B. C. N. Tchaleu, Y. N. Mapoure et al., "Toxoplasma encephalitis in HIV/AIDS patients admitted to the Douala general hospital between 2004 and 2009: a cross sectional study," *BMC Research Notes*, vol. 6, article 146, 2013.
- [3] N. Andenmatten, S. Egarter, A. J. Jackson, N. Jullien, J. P. Herman, and M. Meissner, "Conditional genome engineering in *Toxoplasma gondii* uncovers alternative invasion mechanisms," *Nature Methods*, vol. 10, no. 2, pp. 125–127, 2013.
- [4] G. Di Guardo and S. Mazzariol, "*Toxoplasma gondii*: clues from stranded dolphins," *Veterinary Pathology*, vol. 50, no. 5, p. 737, 2013.
- [5] R. J. M. Wilson, K. Rangachari, J. W. Saldanha et al., "Parasite plastids: maintenance and functions," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 358, no. 1429, pp. 155–164, 2003.
- [6] S. Köhler, C. F. Delwiche, P. W. Denny et al., "A plastid of probable green algal origin in Apicomplexan parasites," *Science*, vol. 275, no. 5305, pp. 1485–1489, 1997.
- [7] J. Janouškovec, A. Horák, M. Oborník, J. Lukeš, and P. J. Keeling, "A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 24, pp. 10949–10954, 2010.
- [8] S. Sato, "The apicomplexan plastid and its evolution," *Cellular and Molecular Life Sciences*, vol. 68, no. 8, pp. 1285–1296, 2011.
- [9] G. Zhu, M. J. Marchewka, and J. S. Keithly, "Cryptosporidium parvum appears to lack a plastid genome," *Microbiology*, vol. 146, no. 2, pp. 315–321, 2000.
- [10] K. A. Brayton, A. O. T. Lau, D. R. Herndon et al., "Genome sequence of *Babesia bovis* and comparative analysis of apicomplexan hemoprotozoa," *PLoS Pathogens*, vol. 3, no. 10, pp. 1401–1413, 2007.
- [11] M. J. Gardner, R. Bishop, T. Shah et al., "Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes," *Science*, vol. 309, no. 5731, pp. 134–137, 2005.
- [12] N. Thomsen-Zieger, J. Schachtner, and F. Seeber, "Apicomplexan parasites contain a single lipoic acid synthase located in the plastid," *FEBS Letters*, vol. 547, no. 1–3, pp. 80–86, 2003.
- [13] C. Y. He, B. Striepen, C. H. Pletcher, J. M. Murray, and D. S. Roos, "Targeting and processing of nuclear-encoded apicoplast proteins in plastid segregation mutants of *Toxoplasma gondii*," *Journal of Biological Chemistry*, vol. 276, no. 30, pp. 28436–28442, 2001.
- [14] M. J. Crawford, N. Thomsen-Zieger, M. Ray, J. Schachtner, D. S. Roos, and F. Seeber, "Toxoplasma gondii scavenges host-derived lipoic acid despite its de novo synthesis in the apicoplast," *EMBO Journal*, vol. 25, no. 13, pp. 3214–3222, 2006.
- [15] J. Mazumdar, E. H. Wilson, K. Masek, C. A. Hunter, and B. Striepen, "Apicoplast fatty acid synthesis is essential for organelle biogenesis and parasite survival in *Toxoplasma gondii*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 35, pp. 13192–13197, 2006.
- [16] S. Agrawal and B. Striepen, "More membranes, more proteins: complex protein import mechanisms into secondary plastids," *Protist*, vol. 161, no. 5, pp. 672–687, 2010.
- [17] T. Brüser and C. Sanders, "An alternative model of the twin arginine translocation system," *Microbiological Research*, vol. 158, no. 1, pp. 7–17, 2003.
- [18] D. Mehner, H. Osadnik, H. Lünsdorf, and T. Brüser, "The Tat system for membrane translocation of folded proteins recruits the membrane-stabilizing Psp machinery in *Escherichia coli*," *Journal of Biological Chemistry*, vol. 287, no. 33, pp. 27834–27842, 2012.
- [19] G. Cilingir, S. L. Broschat, and A. O. T. Lau, "ApicoAP: the first computational model for identifying apicoplast-targeted proteins in multiple species of apicomplexa," *PLoS ONE*, vol. 7, no. 5, Article ID e36598, 2012.
- [20] G. Cilingir, A. O. T. Lau, and S. L. Broschat, "ApicoAMP: the first computational model for identifying apicoplast-targeted transmembrane proteins in Apicomplexa," *Journal of Microbiological Methods*, vol. 95, no. 3, pp. 313–319, 2013.
- [21] K. E. Jackson, J. S. Pham, M. Kwek et al., "Dual targeting of aminoacyl-tRNA synthetases to the apicoplast and cytosol in *Plasmodium falciparum*," *International Journal for Parasitology*, vol. 42, no. 2, pp. 177–186, 2012.
- [22] B. Kumar, S. Chaubey, P. Shah et al., "Interaction between sulphur mobilisation proteins SufB and SufC: evidence for an iron-sulphur cluster biogenesis pathway in the apicoplast of *Plasmodium falciparum*," *International Journal for Parasitology*, vol. 41, no. 9, pp. 991–999, 2011.
- [23] E. V. S. R. Ram, A. Kumar, S. Biswas, S. Chaubey, M. I. Siddiqi, and S. Habib, "Nuclear *gyrB* encodes a functional subunit of the *Plasmodium falciparum* gyrase that is involved in apicoplast DNA replication," *Molecular and Biochemical Parasitology*, vol. 154, no. 1, pp. 30–39, 2007.

- [24] B. J. Foth, S. A. Ralph, C. J. Tonkin et al., "Dissecting apicoplast targeting in the malaria parasite *Plasmodium falciparum*," *Science*, vol. 299, no. 5607, pp. 705–708, 2003.
- [25] T. A. Sadovskaya and A. V. Seliverstov, "Analysis of the 5'-Leader regions of several plastid genes in protozoa of the phylum apicomplexa and red algae," *Molecular Biology*, vol. 43, no. 4, pp. 552–556, 2009.
- [26] T. Kaneko and S. Tabata, "Complete genome structure of the unicellular cyanobacterium *Synechocystis* sp. PCC6803," *Plant and Cell Physiology*, vol. 38, no. 11, pp. 1171–1176, 1997.
- [27] M. Punta, P. C. Coggill, R. Y. Eberhardt et al., "The Pfam protein families database," *Nucleic Acids Research*, vol. 40, no. 1, pp. D290–D301, 2012.
- [28] C. J. A. Sigrist, E. De Castro, L. Cerutti et al., "New and continuing developments at PROSITE," *Nucleic Acids Research*, vol. 41, no. 1, pp. D344–D347, 2013.
- [29] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.
- [30] O. A. Zverkov, A. V. Seliverstov, and V. A. Lyubetsky, "Plastid-encoded protein families specific for narrow taxonomic groups of algae and protozoa," *Molecular Biology*, vol. 46, no. 5, pp. 717–726, 2012.
- [31] J.-R. Barra, *Notions Fondamentales de Statistique Mathématique: Maîtrise de Mathématiques et Applications Fondamentales*, Dunod, Paris, France, 1971.
- [32] G. A. F. Seber, *Linear Regression Analysis*, John Wiley & Sons, New York, NY, USA, 1977.
- [33] N. V. Kobets, D. B. Goncharov, A. V. Seliverstov, O. A. Zverkov, and V. A. Lyubetsky, "Comparative analysis of apicoplast - targeted proteins in *Toxoplasma gondii* and other Apicomplexa species," in *Proceedings of the International Moscow Conference on Computational Molecular Biology (MCCMB '13)*, Moscow, Russia, July 2013.
- [34] L. Lim and G. I. McFadden, "The evolution, metabolism and functions of the apicoplast," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 365, no. 1541, pp. 749–763, 2010.