Comparison of the Compositional Proclivities of the Complete Genomes of *Plasmodium Falciparum* and Human

April Williams, Yuriy Fofanov, and Catherine Putonti

***Abstract*— Pathogens and hosts have a dynamic relationship, one that is ever changing at the molecular level - the pathogen influencing the evolutionary path of the host and the host influencing the evolutionary path of the pathogen. The pathogen's adaptation to a particular host could serve several purposes, e.g. to mimic the host to avoid detection, to take advantage of the host's cellular machinery, to increase virulence, etc. Recognizing these adaptations is far from trivial, particularly when the size of the pathogen's and host's genomes differ by orders of magnitudes. Novel algorithms and data structures have been developed in our laboratory that make it possible to quantify the "distance" (or number of mutations) separating pathogen and host sequences. Through the examination of these distances, we hypothesize that it is possible to monitor pathogen adaptations at the sequence level and further our understanding of the function of the pathogen machinery. Even though the genome of *Plasmodium falciparum*, the agent causing malaria in humans, is complete, the functions of many of the coding regions remain unknown. Herein we present the results of our exhaustive calculations for each of the annotated coding regions in *P. falciparum* and the human genome.**

***Index Terms*—comparative genomics, nucleotide composition, *Plasmodium falciparum*.**

I. Introduction

Evidence of human infection by malaria has been found in writings from antiquity as well as Egyptian mummies more than 3000 years old [1],[2]. Responsible for nearly a million deaths a year world-wide, understanding the pathogenicity, virulence, and transmission of the parasite is of paramount importance [3]. There are four types of malaria that can infect humans - *Plasmodium falciparum, P. vivax, P. malariae,* and *P. ovale* [3]. Malaria is caused by these malaria parasites that are transmitted through the bite of the female *Anopheles* mosquito. In essence, one can consider malaria as having two hosts – the mosquito vector and the infected animal. The lifecycle of the protist in fact occurs in both hosts. The complete genomic sequence and annotation of *P. falciparum*, released in 2002 [4], has subsequently been followed by other *Plasmodium* species [5]-[7]; *P. falciparum*, however, is the most well annotated and studied of the genus.

The means in which a pathogen such as *P. falciparum* can persist within the human host necessitates a means of entry (the mosquito) in addition to molecular mechanisms to interact with the host system while avoiding the host's defenses. The means in which a pathogen adapts at its genetic level to its host organism can take a variety of forms. For instance, many bacterial pathogens have been found to mimic the function of host proteins, including instances of specifically acquiring host gene sequences as well as the development of new factors that produce structures that benefit the pathogen during interactions with the host [8]-[13]. The avoidance of particular sequence compositions is likely in response to the host's RNAi-related pathways of defense which are based upon the binding of specific short (approximately 20-30 nt) RNAs and target sequences. Several mammalian viruses have in fact been found to encode for small microRNAs (miRNAS) in order to manipulate their host [14]-[16].

Despite the wealth of sequence and expression data available for *P. falciparum*, the majority of the genes are annotated as "hypothetical proteins". Curating this genome with the same degree of accuracy as the human and mouse genomes is challenging due to the fact that the *P. falciparum* genome is extremely AT-rich. A composition many of the existing annotation tools are unable to handle. Moreover, the expression of genes and their associated protein functions is plagued by the fact that the parasite resides within two very different host systems. Herein, we present the preliminary results of a rigorous comparison of the *P. falciparum* and human genomes using a suite of computational tools developed within our laboratories. The goal of this study is in fact two-fold. Firstly, based upon the similarities in sequence composition observed between other pathogen-host systems, we aim to use these properties to more accurately identify human-specific genes. In doing so, we specifically focus on genes under selection for a human-like composition as such genes are likely critical to the virulence of the pathogen with this host. Our second goal is to identify small subsequences present within both the pathogen and host which may be miRNAs.

Manuscript received July 26, 2009. This work is partially supported by the Department of Homeland Security Advanced Research Projects Agency (Y.F.). AW supported by a Carbon Fellowship from Loyola University Chicago.

A. Williams is with the Department of Bioinformatics and the Department of Biology, Loyola University Chicago, Chicago, IL 60660 USA (e-mail: awill11@luc.edu).

Y. Fofanov is with the Department of Computer Science and the Department of Biology and Biochemistry, University of Houston, Houston, TX 77004 USA (e-mail: yfofanov@bioifo.uh.edu).

C. Putonti is with the Department of Bioinformatics, the Department of Biology and the Department of Computer Science, Loyola University Chicago, Chicago, IL 60660 USA (phone: 773-508-3277; fax: 773-508-3646; e-mail: cputonti@luc.edu).

II. Data

The complete genomic sequences and annotations for both the human genome (build 36.2) and *P. falciparum* (build 1.1) as well as the collection of human mRNA sequences were downloaded from NCBI (www.ncbi.nlm.nih.gov). The three sequence collections exhibit the following characteristics:

- Human genome: 22 autosomes and 2 sex chromosomes with a total genome length of 3.4 billion base pairs and an average GC-content of 41%.
- *P. falciparum* genome: 14 chromosomes with a total genome length of 23 million base pairs and an average GC-content of 21%.
- Human mRNA sequences: 46,177 individual mRNA sequences ranging in size from 34 to 101,520 nucleotides in length, with GC-contents ranging from 18-77% (average 50%).

Functionality was developed in C++ for parsing the sequence files according to the annotation files such that the individual coding regions of each genome could be extracted. Sequences of a standard length of 100,000 nucleotides were created with GC-contents between 10 and 80%; these sequences provide a means of assessing the expected appearance of subsequences given the sequences GC composition.

III. Computational Methods

A suite of tools have been developed within our laboratories for efficiently calculating the frequency of all subsequences (*n*-mers) in any sequenced genome within a reasonable time (depending on the length being considered, minutes to a few hours) [17]. Moreover, specialized data structures and algorithms have been designed which make it possible to explicitly calculate the distance or number of base changes necessary to convert one subsequence to another [18]-[19]. Thus, for any two sequences or genomes, e.g. two strains belonging to the same species, we can quickly calculate the distance or number of base changes which have occurred since their last common ancestor. This method follows an alignment-free approach which is well suited for comparing two organisms which differ by orders of magnitude and millions of years since speciation, such as the *P. falciparum* and human genomes. (For a review of alignment-free approaches see [20]). To minimize the run-time necessary to perform such exhaustive calculations, our method maximizes memory usage; the amount of memory needed corresponds not to the size of the sequence(s) or genome(s) being compared but rather the length of the *n*-mer under consideration. Moreover, for each set of comparisons, only one *n*-mer size can be considered at a time.

For any particular pathogen genome, we can define the set of *n*-mers that are present in the host genome as well as those which are absent. Those *n*-mers which are absent can be close to the sequence, e.g. requiring a single base change in order to be "converted" to an *n*-mer present in the sequence. Likewise, an *n*-mer can be distant from the sequence, requiring multiple base changes in order to be converted to an *n*-mer present in the sequence. For a particular *n*-mer *i* in the pathogen sequence, we propose to quantify its distance, $d_{i,n}$, from the host as the number of base changes needed to convert it to the closest *n*-mer present in the host sequence. In order to guarantee that some *n*-mers present within the smaller pathogen sequence are absent in the larger host genome, one must select a value of *n* such that not all of the possible 4^n *n*-mers are present in the host genome. To determine the distance $d_{i,n}$ of a particular *n*-mer *i*, all possible 1, 2, 3, etc. base changes are considered. If any one of the derived *n*-mers are present in the host genome, the distance for the particular *n*-mer *i* is the number of base changes *c* ($d_{i,n} = c$).

Once a sequence is "annotated" against a particular host for a particular value of *n*, we can calculate the average distance across the entire sequence as:

$$D_n = \frac{\sum_{i=1}^{m} d_{i,n}}{m} \quad (1)$$

where *m* is the genome's (region's) length and $d_{i,n}$ is the number of changes necessary to convert the *n*-mer in the pathogen sequence to an *n*-mer in the host genome. Thus a sequences with a D_n value close to 0 are, at least for the value of *n* under consideration, closer to the host than sequences with a D_n value further from 0.

Figure 1 summarizes the computational method employed in this study. All of the computations were conducted on a 16GB RAM dual-core processor workstation.

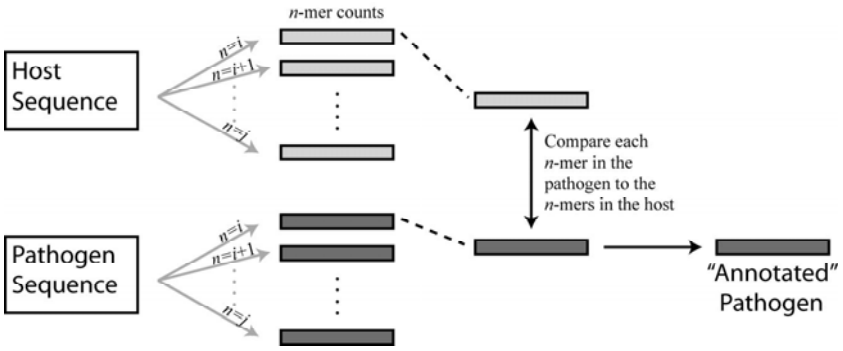

Figure 1. Summary of computational methodology employed to calculate the distances between two sequences.

IV. Results

A. Comparison of P. falciparum Coding Regions to the Complete Human Genome

Each individual *P. falciparum* coding region was compared to the complete human genome. These comparisons were conducted by quantifying the number of base changes necessary to convert every subsequence of length 14-20 present within the particular *P. falciparum* coding sequence to the closest subsequence present in the human genome. We did not consider subsequences less than 14 nucleotides long as the vast majority (>96%) of those possible 4^n *n*-mers are present in the human genome [21]. This metric allows us to identify the following:

- human genes or gene segments which have been integrated into the *P. falciparum* genome.
- *P. falciparum* genes or gene segments which have been integrated into the human genome.
- *P. falciparum* genes selecting for a nucleotide composition similar to the nucleotide composition of the human genome.

For each of the individual *P. falciparum* genes, the average distance was calculated.

In order to qualify the significance of the distances of the individual *P. falciparum* coding regions from the human genome, it is first necessary to determine what degree of sequence similarity can be expected by chance. Thus, we created "synthetic" sequences for each of the *P. falciparum* coding regions with the same GC-content (see methods) and compared each to the complete human genome. Figure 2 details the average distance from the human genome of 18-mer subsequences of the coding regions of each *P. falciparum* chromosome and each randomly generated sequence of similar length and GC-content. As the figure shows, all of the *P. falciparum* chromosomes are significantly closer to human than is expected at random. The individual distances for both the real and synthetic coding sequences were examined further for each chromosome. Figure 3 shows the results for the coding regions within *P. falciparum* chromosome 3; the graphs for the other chromosomes are comparable. Within all of the chromosomes, several coding regions were found to exhibit a composition more similar to the human genome than any of the synthetic sequences; Table I lists some of these coding regions in chromosome 3.

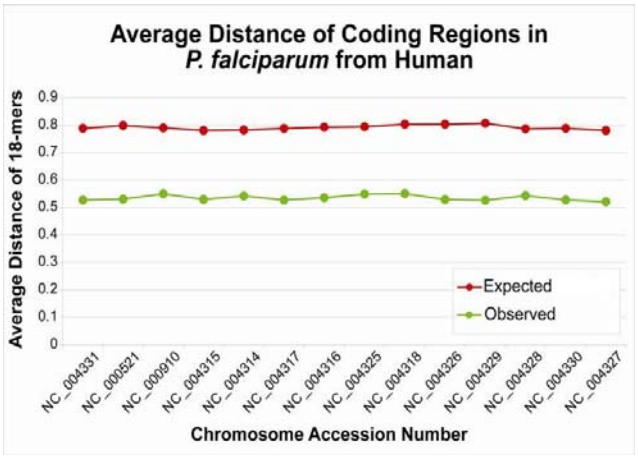

Figure 2. he average distance from the human genome of 18-mer subsequences of the coding regions of each *P. falciparum* chromosome and each randomly generated sequence of similar length and GC-content.

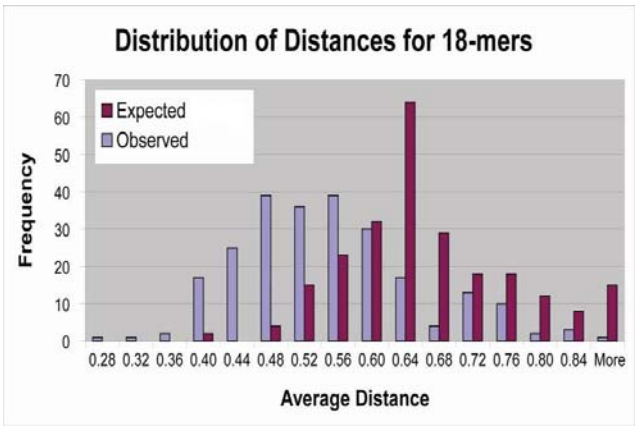

Figure 3. The individual distances for both the real coding sequence of *P. falciparum* chromosome 3 and the synthetic coding sequence.

Table I. Ten *P. falciparum* coding regions of chromosome 3 exhibiting a composition most similar to human.

Gene Name	Product	Length	D_n	GC-Content
MAL3P7.36	F49C12.11-like protein	510	0.288235	20%
PFC0911c	hypothetical protein	1074	0.324953	16%
PFC0582c	hypothetical protein	1999	0.336668	16%
PFC0261c	hypothetical protein, conserved	516	0.337209	20%
PFC1016w	hypothetical protein	1085	0.371429	18%
MAL3P2.13	hypothetical protein	4323	0.371501	18%
MAL3P4.28	60S ribosomal protein L26, putative	982	0.378819	18%
PFC1011c	hypothetical protein	2285	0.382932	18%
MAL3P7.5	hypothetical protein	1416	0.383475	18%
PFC0191c	hypothetical protein	652	0.386503	18%

For each of the possible 4^n *n*-mers present within the *P. falciparum* coding regions, we observed that the vast majority of the *n*-mers with a distance of 0 or 1 were in fact AT-rich sequences, despite the greater average GC-content of the human genome. By scanning the human genome again, we found that the majority of these "close" subsequences were located within the non-coding region, which is known to contain many regulatory elements as well as sequence of unknown function (commonly referred to as "junk" DNA). With the exception of poly-A and poly-T subsequences, there are several possibly interesting subsequences which warrant further investigation to ascertain if they in fact confer some functionality in the immune response of the human to infection by the parasite.

B. Comparison of P. falciparum Coding Regions to Human mRNAs

Each *P. falciparum* chromosomal sequence and the collection of mRNAs were analyzed for the same *n*-mer sizes in order to assess the frequency of appearance of *n*-mers within the coding region exclusively, i.e. avoiding the non-coding for which we found many matches before. As is illustrated in Fig. 4, the probability of the occurrence of AT-rich *n*-mers is significantly greater within the *P. falciparum* chromosomes than in the human; on the contrary, in the human genome the probability of occurrence of *n*-mers with respect to their GC-content follows a normal distribution. Using a sliding window approach, the GC-content along each individual mRNA sequence was assessed; as was expected, the more AT-rich regions appear at the 3' end of the sequence indicating the inclusion of the poly-A tails within many of the mRNA sequences.

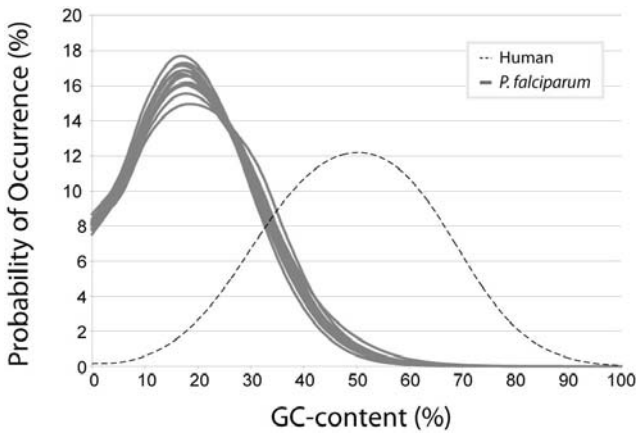

Figure 4. The distribution of the probability of occurrence of *n*-mers with a specific GC-content in the human mRNA sequences and *P. falciparum*.

To more readily identify the occurrence of sequence similarity due to the integration of human coding regions into the *P. falciparum* genome (or vice versa), the entire *P. falciparum* genome was compared to the collection of human mRNA sequences for all 14- through 20-mers. We chose to focus our attention on those subsequences which are present in both the *P. falciparum* genome and human mRNA sequence. For this set of *n*-mers, we rescanned the human mRNA sequences in order to identify the location of each *n*-mer. Although the maximum subsequence length considered was only 20, the location information revealed that, at least in the human mRNA sequence, much longer stretches of subsequences are similar. The most frequently occurring *n*-mers, some more than 5,000 times, are extremely AT-rich and located at the 3' end of the mRNA sequence.

Looking specifically at the *n*-mers present in both organisms, we filtered these results to remove from consideration those *n*-mers which occur within 150bp of the 3' end of the mRNA sequence and those sequences with a GC-content < 20%. This resulted in a dramatic decrease in the *n*-mers to be examined. From this group, we selected several sequences and using the annotations of the *P. falciparum* genome have determined if they are present within the coding or noncoding regions. One such example of sequence homology is sequence conservation which appears between a member of the *P. falciparum* STEVOR multigene family, suspected in playing a crucial role in enabling *P. falciparum*-infected red blood cell surfaces undetectable by the host's immune system [22], and a human membrane protein. This process has identified several candidate loci, including the example cited here, which are currently under investigation by our collaborators.

V. Conclusions

The computational approach proposed here presents a new means of recognizing regions of sequence similarity. As the vast majority of the functionality of the *P. falciparum* genes remains unknown, the results of our comparisons have identified several genes which may be host-specific as they have a nucleotide composition more similar to the human composition than is to be expected by chance. In the sea of A's and T's which make up the *P. falciparum* genome, several regions exhibiting highly similar sequence to regions within the human genome also suggest either ancient acquisitions (be it by the pathogen or the host) of past infections or unknown functional reasons for their presence. Numerous genes within the *P. falciparum* genome have been identified and are currently under investigation by our collaborator.

References

[1] I. W. Sherman, (ed), *Malaria: Parasite Biology, Pathogenesis, and Protection*. Washington D.C.: ASM Press, 1998, ch.1.

[2] R. L. Miller, S. Ikram, G. J. Armelagos, R. Walker, W. B. Harer W, C. J. Shiff, *et al.* (1994, Jan-Feb). Diagnosis of Plasmodium falciparum infections in mummies using the rapid manual ParaSight-F test. *Trans R Soc Trop Med Hyg*. 88(1), 31-2.

[3] World Heath Organization. (2009, January) Malaria. [Online] Available: http://www.who.int/mediacentre/factsheets/fs094/en/index.html

[4] M. J. Gardner, N. Hall, E. Fung,, O. White, M. Berriman, R. W. Hyman, *et al*. (2002, October) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*. 419: 498-511.

[5] A. Pain, U. Böhme, A. E. Berry, K. Mungall, R. D. Finn, A. P. Jackson, *et al.* (2008, October) The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature*. 455, 799-803.

[6] J. M. Carlton, J. H. Adams, J. C. Silva, S. L. Bidwell, H. Lorenzi, E. Caler, *et al.* (2008, October) Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature*. 455, 757-763.

[7] J. M. Carlton, S. V. Angiuoli, B. B. Suh, T. W. Kooij, M. Pertea, J. C. Silva, *et al*. (2002, October) Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature*. 419, 493-494.

[8] Z. A. Hamburger, M. S. Brown, R. R. Isberg,, and P. J. Bjorkman. (1999, October) Crystalstructure of invasin: a bacterial integrin-binding protein. *Science*. 286, 291-295.
[9] W. D. Hardt, L. M. Chen, K. E. Schubel, X. R. Bustelo, and J. E. Galán. (1998, May)*Salmonella typhimurium* encodes an activator of Rho GTPases that induce membrane ruffling and nuclear responses in host cells. *Cell*. 93, 815-826.
[10] R. Holzerlandt, C. Orengo, P. Kellam, and M. M. Albà. (2002, November) Identification of new herpesvirus gene homologs in the human genome. *Genome Res.* 12, 1739-1748.
[11] C. E. Stebbins, and J.E. Galán,. (2000, December) Modulation of host signaling by a bacterial mimic. Structure of the *Salmonella* effector SptP bound to Rac1. *Mol. Cell.* 6, 1449-1460.
[12] C. E. Stebbins, and J. E. Galán. (2001, August) Structural mimicry in bacterial virulence. *Nature.* 412, 701-705.
[13] D. Zhou, M. Mooseker, and J. E. Galán,. (1999, March) Role of the *S. typhimurium* actin-binding protein SipA in bacterial internalization. *Science*. 283, 2092-2095.
[14] S. Pfeffer, M. Zavolan, F. A. Grässer, M. Chien, J. J. Russo, J. Ju,, *et al*. (2004, April) Identification of virus-encoded microRNAs. *Science.* 304, 734-736.
[15] P. Sarnow, C. L. Jopling,, K. L. Norman, S. Schütz, and K. A. Wehner. (2006, September) MicroRNAs: expression, avoidance and subversion by vertebrate viruses. *Nat. Rev. Microbiol.* 4, 651-659.
[16] D. J. Obbard, K. H. Gordon, A. H. Buck, F. M. Jiggins. (2009, January) The evolution of RNAi as a defence against viruses and transposable elements. *Philos Trans R Soc Lond B Biol Sci*. 364(1513), 99-115.
[17] V. Fofanov, C. Putonti, S. Chumakov, B. M. Pettitt, Y. Fofanov. Fast algorithm for the analysis of the presence of short oligonucleotide subsequences in genomic sequences. [Online] *Technical Report: UH-CS-05-11*. Available: http://www.cs.uh.edu/Preprints/preprint/uh-cs-05-11.pdf.
[18] C. Putonti, S. Chumakov, R. Mitra, G. E. Fox, R. C. Willson, Y. Fofanov. (2006, January) Human-blind probes and primers for Dengue virus identification: Exhaustive analysis of subsequences present in the human and 83 Dengue genome sequences. *FEBS J.* 273(2), 398-408.
[19] C. Reed, V. Fofanov, C. Putonti, S. Chumakov, T. Slezak, Y. Fofanov. (2007, October) Effect of the mutation rate and background size on the quality of pathogen identification. *Bioinformatics*. 23(20), 2665-2671.
[20] S. Vinga and J. Almeida. (2003, March) Alignment-free sequence comparison-a review. *Bioinformatics.* 19(4), 513-23.
[21] C. Putonti,, V. Fofanov, Y. Fofanov. "Computational Methods for Estimating the Size and Identifying the Members of Microbial Communities," In: Du, D.-Z. (ed.) *Proceedings of the IASTED International Conference on Computational and Systems Biology*, Dallas, Texas, 2006: pp. 540-033.
[22] M. Niang, Yan, X. Yam and P. R. Preiser. (2009, February) The Plasmodium falciparum STEVOR multigene family mediates antigenic variation of the infected erythrocyte. *PLoS Pathog*. 5(2), e1000307.