Vol.2009-MPS-76 No.18
Vol.2009-BIO-19 No.18
2009/12/17

Regular Paper

Evolution of antigenic gene families in the *Plasmodium* genus

Diego Diez, Minoru Kanehisa, Susumu Goto

Kyoto University

Abstract

Antigenic variation is a mechanism employed by different pathogens, like *Plasmodium* among others, to evade the immune response and survive inside the host environment by altering the antigenic properties of surface proteins. Antigenic variation is sometimes mediated through the evolution of multi-gene families, encoding proteins containing hyper-variable regions. The expression of these proteins is switched during the course of infection, confounding the host immune system. The use of multi-gene families is widespread within *Plasmodium*, exemplified by the rif and var gene families in P. falciparum, the causative agent of malaria. We found evidence of a gene family shared among five different *Plasmodium* species, which has undergone differential expansion. In *P. knowlesi* (*kir* gene family) and *P. vivax* (*vir* gene family) they constitute a multi-gene family involved in antigenic variation. Other three species, *P. chabaudi*, *P. yoelii* and *P. berghei*, have *vir*/*kir* single-gene homologous sequences. Sequence and phylogenetic analyses performed on these gene families confirm these findings and reveal clues about the evolution of antigenic gene families in the *Plasmodium* genus.

1. Introduction

Malaria disease is caused by parasites of the genus *Plasmodium*, which infect humans, primates, rodents and birds. *Plasmodium falciparum* alone infects yearly more than 500 million infections, with approximately 1 to 3 million deaths, mainly children under 5 years old[1)]. This burden is directly related to the ability of the parasite to establish persistent infections and evade the immune system of the host. One mechanism used to accomplish this is by changing the antigenic properties of parasite proteins exposed to the host immune system, a process called antigenic variation[2)]. *Plasmodium* parasites use multi-gene families containing hyper-variable regions to generate the sequence diversity needed to generate variability in the antigenic properties of the encoded proteins[3)]. These gene families are evenly distributed among the different plasmodium species and share no sequence similarity among each other[4)–6)]. However, based on multi-character methods, Janssen et al. proposed in 2004 that the different antigenic gene families in Plasmodium comprise a super-family of homolog sequences. The *pir* (*Plasmodium* interspersed repeats) super-family comprises *P. falciparum var*, *rif* and *stevor* gene families, *P. knowlesi kir*, *P. vivax vir*, *P. berghei bir*, *P. yoelii yir* and *P. chabaudi cir*[7)]. Within these gene families, *kir* and *vir* share sequence similarity and constitute a family of homologs sequences in primates (*vir/kir* gene family). In a similar way, *cir*, *bir* and *yir* (*cir/bir/yir* gene family) are homolog sequences in rodents.

We have performed sequence searches on the protein translations of six *Plasmodium* species using HMM profiles for *vir/kir* and found evidence of genes in the rodent parasites with high similarity to some *vir/kir* genes a no similarity to *cir/bir/yir*. Motif analysis demonstrates that these sequences contain features distinct of the *vir/kir* family and suggest that are true homologs of the *vir/kir* family in rodents. This finding challenges the current classification of antigenic variant gene families in *Plasmodium*.

⋆1 The real author is the Editorial Board of the Trans. IPSJ.

2. Materials and methods

Genomes and protein translations were downloaded from PlasmoDB 5.5 for *P. falciparum*, *P. reichenovi*, *P. vivax*, *P. knowlesi*, *P. yoelii*, *P. berghei*, *P. chabaudi* and *P. gallinaceum*[9]. *P. chabaudi* was additionally tested using the current genome assembly (at 8x coverage) in PlasmoDB 6.0. Sequence were detected with HMMER package using the HMM models in Pfam 23[10]. Both fragment and full models were used and the results combined to achieve more sensitivity. The *vir/kir* family was detected using the model Plasmodium_vir (accession: PF05795). Multiple sequence alignments and Neighbor-Joining (NJ) phylogenetic trees were generated using CLUSTALW-MPI 0.13 (corresponding to CLUSTALW 1.82)[11]. Consistency of NJ tree topology was assessed with the bootstrap method, using 1000 replications. Alignment images were created with Jalview[12]. Tree images were created with TreeDyn[?]. The MEME suite (version 4.1) was used to compute motif analyses. Sequence motifs were detected using MEME with a limit of 20 motifs and otherwise default parameters[14]. MAST was used to search for motifs from a motif database (*kir/vir* motifs in *cir/bir/yir* and *vice versa*)[15]. TOMTOM was used to compare both libraries of motifs[16].

3. Results

3.1 Detection of vir/kir sequences in Plasmodium

Vir/kir family members were searched by using HMM profiles in *P. falciparum*, *P. reichenovi*, *P. knowlesi*, *P. vivax*, *P. yoelii*, *P. berghei*, *P. chabaudi* and *P. gallinaceum*. We found, as expected, homologs in *P. vivax* (255) and *P. knowlesi* (64). Sequences in *P. yoelii* (1), *P. berghei* (2) and *P. chabaudi* (1) were also identified (**Table 1**). Sequence numbers in the primate parasites are similar to previously reported values. The rodent parasite sequences, however, have not been reported before, and their existence contrasts with the currently accepted paradigm. Interestingly and differently from the primate genes, these sequences seem to constitute single gene families (except for a presumed duplication in *P. berghei*) and therefore may not be related to antigenic variation. These sequences are annotated as hypothetical proteins in the current genome

assemblies and therefore little extra information can be obtained.

3.2 Phylogenetic analysis of the entire dataset

It is possible that these new sequences are false positives or share some similarity to the *vir/kir* family without being homologs. Another possibility is that the new sequences are indeed *vir/kir* homologs, a result that has consequences to the current classification of antigenic variant gene families in *Plasmodium*. Consequently, to get further insight into the evolutionary relation of these new sequences with the *vir/kir* family we studied their phylogenetic distribution. The 321 protein sequences were aligned using CLUSTALW-MPI with default parameters. A phylogenetic tree was constructed by using the NJ algorithm with 1000 replications of bootstrapping to assess topology reliability (**Fig. 1**). The tree shows in general clear differentiation between *P. vivax* and *P. knowlesi* sequences. There are two big clusters of *P. vivax vir* protein (C1 and C3) and one of *P. knowlesi kir* proteins (C7), plus a small cluster with only two *vir* sequences (C5). There are also some smaller clusters, containing mainly *vir* sequences with a few *kir* sequences included (C2, C4 and C6). Cluster C8 includes sequences from all the different organisms. The 4 rodent parasite sequences cluster with two sequences from the *vir/kir* family (cluster C9), and is supported by high bootstrap values (>900).

3.3 Analysis of the core set alignment

Independent alignment and examination of these 6 sequences shows that they are highly conserved (**Fig. 2**). *P. vivax* and *P. knowlesi* sequences contain a hyper-variable fragment (considerably longer in P. knowlesi) close to the

Table 1 Distribution of *vir/kir* genes in *Plasmodium* species.

Organism	*Host*	*# Seq*
P. vivax	Primate	255
*P. knowles*i	Primate	64
P. berghei	Rodent	2
P. yoelii	Rodent	1
P. chabaudi	Rodent	1

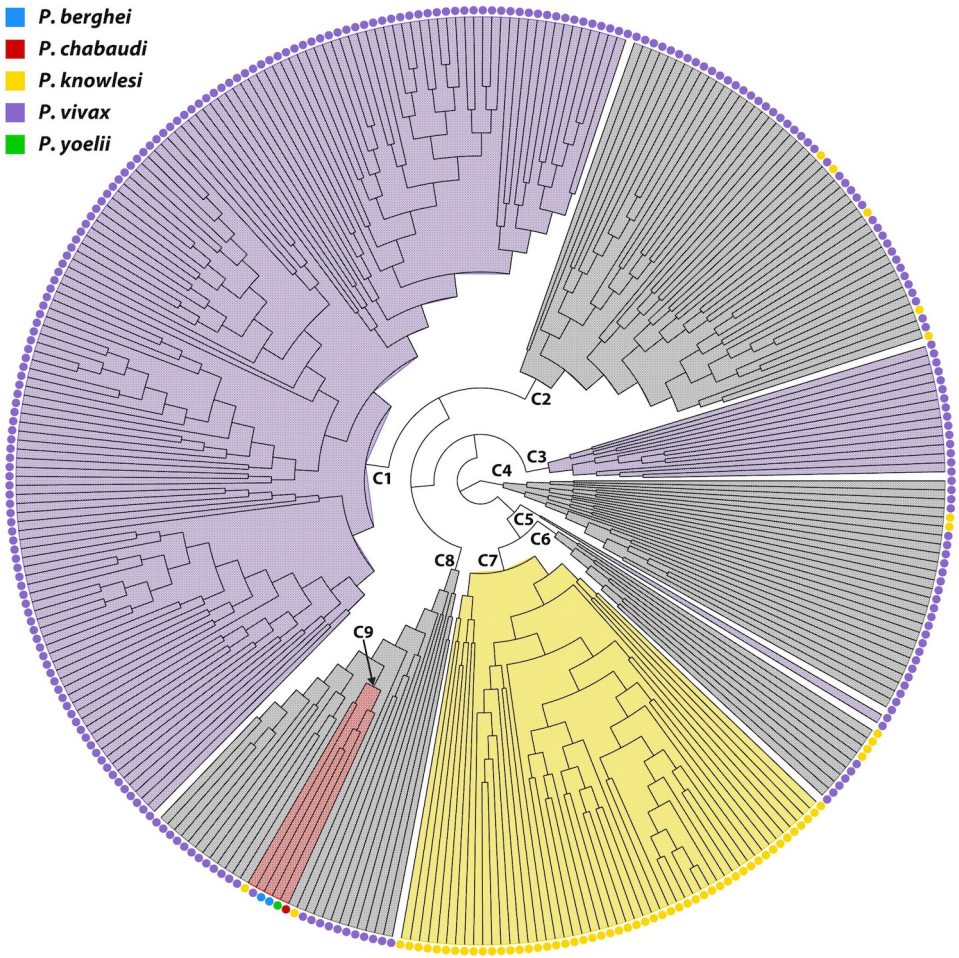

Fig. 1 Tree of *vir/kir* sequences. Clusters are indicated with shades and labels. C9 contains the 6 orthologs sequences.

N-terminal side. The *P. chabaudi* and *P. berghei* sequences are truncated, likely due to the low coverage of the genome assemblies. The sequence from *P. chabaudi* is truncated in the N-terminal, missing important conserved residues, whereas the two *P. berghei* sequences miss part of the C-terminal region. Pairwise similarity shows that the rodent parasite sequences are more than 95% similar (the two *P. berghei* sequences are identical) and the primate parasite sequences are 63% similar. The truncated *P. chabaudi* sequence has a similarity greater than 70% with the rodent parasite sequences, although is only greater than 39% with the primate parasite sequences. Some positions are apparently not conserved, but that maintain conservation at physicochemical properties (e.g. negative, aliphatic, etc). In some cases two adjacent positions will have different but complementary charged aminoacids in the rodent and primate sequences.

Using a recent *P. chabaudi* genome assembly (PlasmoDB 6.0; 8x coverage) a complete sequence is detected (PCAS_010120), which contains the missing conserved fragment, further supporting our homology hypothesis. Interestingly this sequence is annotated as CIR. In addition, we searched the genome sequences of *P. berghei* with the protein sequences in this subset, using BLAST (tblastn). Several contigs matched the sequences and were able to reconstruct part of the C-terminal (data not shown).

3.4 Motif analysis

These results confirm that the rodent and primate sequences in the core cluster are homologs. However, there is the possibility that some of these primate sequences are not *vir/kir* genes, but members of a different gene family sharing some similarity to the *vir/kir* family, and the current HMM profiles cannot differentiate between them (similar to what happens with the *rifin/stevor* gene families). To test this hypothesis we performed a motif analysis to investigate if the rodent sequences contain features shared among all the *vir/kir* genes. Motifs were detected using MEME with a limit of 20 motifs and otherwise default parameters. Motifs found in the core sequences are indicated in Fig. 2 with the percentage of sequences containing the motifs. Motifs V1, V3 and are frequently found in the *vir/kir* population, with 88% and 73% of all sequences having the

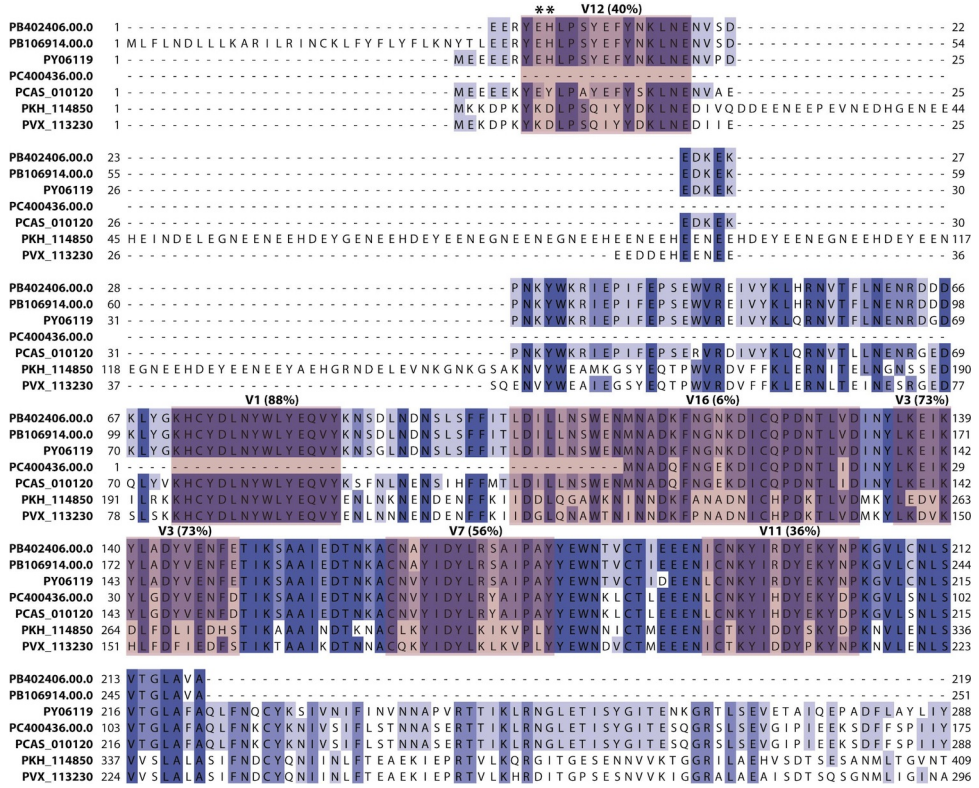

Fig. 2 Alignment of *vir/kir* ortholog sequences present in all studied species.

motifs respectively. Motifs V7, V11 and V12 are moderately frequent, with 56%, 36% and 40% presence. V16 is rare and only 6% of the sequences have it. Overall, this indicates that most of the motifs found in the core sequences are share among the different *vir/kir* genes.

Another possibility is that these motifs are also present in the *cir/bir/yir* gene family. This possibility is unlikely, given the low similarity between the vir/kir and cir/bir/yir families. To test properly this possibility we searched all *cir/bir/yir* sequences in *Plasmodium* and look for occurrences of the *vir/kir* motifs in the *cir/bir/yir* sequences (**Fig. 3**). Only a limited percentage of *cir/bir/yir* sequences contain *vir/kir* motifs. V18 is the most abundant motif (16.6%), followed by V1 (13.4%), V2 (14.9%) and V7 (8.0%). V3 (6.0%), V11 (5.7%), V12 (2.1%) and V16 (1.1%) are present but in very low frequency. Moreover, in most sequences the ordering of motifs is not conserved.

To check if there any of the motifs found in *vir/kir* is similar to any motif in *cir/bir/kir* be computed a motif comparison of both libraries using the

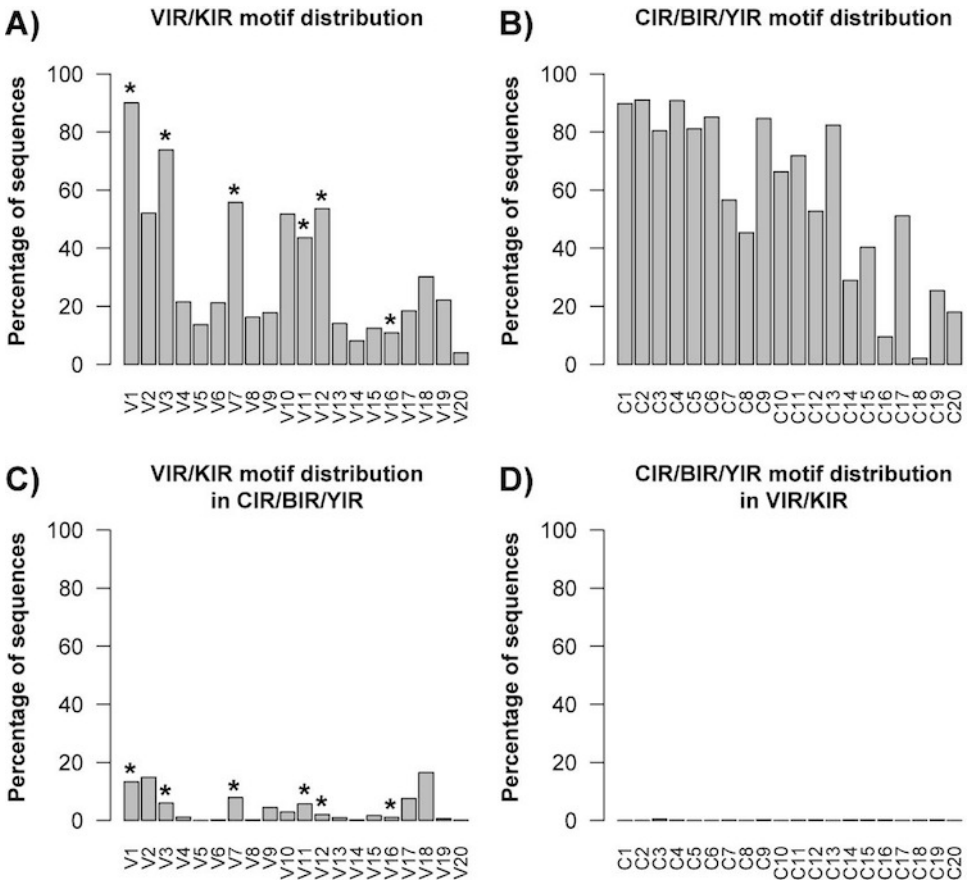

Fig. 3 A) Distribution of motifs found in *vir/kir* sequences. Motifs marked with asterisk are found in the core set. B) Distribution of motids found in *cir/bir/yir* sequences. C) *vir/kir* motifs in *cir/bir/yir* sequences. D) *cir/bir/yir* motifs in *vir/kir* sequences.

tool TOMTOM (MEME suite). The results indicated that motifs found in *cir/bir/yir* are not significantly similar to *vir/kir* motifs (data not shown). This was expected given the low number of *vir/kir* sequences having match to any *cir/bir/yir* motif (Fig. 3D). Overall these results demonstrate that there are only a few motifs shared between the two sequence sets. Only a small subset of *cir/bir/yir* sequences contains some of the *vir/kir* motifs. In most cases these motifs do not maintain the organization found in the *vir/kir* family.

4. Discussion

Antigenic variant gene families are present in most of *Plasmodium* species, where perform functions related to immune evasion and establishment of a persistent infection. However, the role of the *vir*, *kir*, *cir*, *bir* and *yir* gene families in antigenic variation is not yet understood. It is also unclear whether these proteins perform additional functions. The proposal of the pir super-family was convenient to derive a common origin for antigenic variation in Plasmodium species. This hypothesis, however, has some controversy. For example, it is difficult to concur different evolutionary and life cycle characteristics, with a common antigenic variant gene family[8].

On the contrary, our findings indicate that single gene homologs for the *vir/kir* gene family are present in the rodent species. This suggests that different gene families in primates and rodents perform antigenic variation related functions. The fact that rodent parasites have a single copy of *vir/kir* genes suggest that these gene families may be performing additional functions not related to antigenic variation. It is possible, however, that the *vir/kir* genes have sub-functionalized after duplication and expansion. In such a case, the core set of sequences may be performing the original function whereas the multi-copy genes are only used as a decoy for the immune system.

We believe there is evidence against the pir family paradigm and therefore propose an alternative hypothesis. Antigenic variation is performed in Plasmodium parasites by different, non-homolog, gene families. This function will be performed in concert with their original function, which could be derived from analysis of the rodent single-gene homologs. This scenario has important

implications that affect our understanding on how these species evolved. Additional analysis of other antigenic variant gene families in *Plasmodium* should help elucidate the complex evolutionary history of these parasites.

References

1) Snow, R.W., Guerra, C.A., Noor, A.M., Myint, H.Y., and Hay, S.I., The global distribution of clinical episodes of Plasmodium falciparum malaria. Nature, 2005. 434(7030): p. 214-7.

2) Deitsch, K.W., Moxon, E.R., and Wellems, T.E., Shared themes of antigenic variation and virulence in bacterial, protozoal, and fungal infections. Microbiol Mol Biol Rev, 1997. 61(3): p. 281-93.

3) Ferreira, M.U., Zilversmit, M., and Wunderlic, G., Origins and evolution of antigenic diversity in malaria parasites. Curr Mol Med, 2007. 7(6): p. 588-602.

4) Fernandez-Becerra, C., Yamamoto, M.M., Vencio, R.Z., Lacerda, M., Rosanas-Urgell, A., and del Portillo, H.A., Plasmodium vivax and the importance of the subtelomeric multigene vir superfamily. Trends Parasitol, 2009. 25(1): p. 44-51.

5) Janssen, C.S., Barrett, M.P., Turner, C.M., and Phillips, R.S., A large gene family for putative variant antigens shared by human and rodent malaria parasites. Proc Biol Sci, 2002. 269(1489): p. 431-6.

6) del Portillo, H.A., Fernandez-Becerra, C., Bowman, S., Oliver, K., Preuss, M., Sanchez, C.P., Schneider, N.K., Villalobos, J.M., Rajandream, M.A., Harris, D., Pereira da Silva, L.H., Barrell, B., and Lanzer, M., A superfamily of variant genes encoded in the subtelomeric region of Plasmodium vivax. Nature, 2001. 410(6830): p. 839-42.

7) Janssen, C.S., Phillips, R.S., Turner, C.M., and Barrett, M.P., Plasmodium interspersed repeats: the major multigene superfamily of malaria parasites. Nucleic Acids Res, 2004. 32(19): p. 5712-20.

8) Hall, N., Karras, M., Raine, J.D., Carlton, J.M., Kooij, T.W., Berriman, M., Florens, L., Janssen, C.S., Pain, A., Christophides, G.K., James, K., Rutherford, K., Harris, B., Harris, D., Churcher, C., Quail, M.A., Ormond,

D., Doggett, J., Trueman, H.E., Mendoza, J., Bidwell, S.L., Rajandream, M.A., Carucci, D.J., Yates, J.R., 3rd, Kafatos, F.C., Janse, C.J., Barrell, B., Turner, C.M., Waters, A.P., and Sinden, R.E., A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses. Science, 2005. 307(5706): p. 82-6.

9) Stoeckert, C.J., Jr., Fischer, S., Kissinger, J.C., Heiges, M., Aurrecoechea, C., Gajria, B., and Roos, D.S., PlasmoDB v5: new looks, new genomes. Trends Parasitol, 2006. 22(12): p. 543-6.

10) Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L., and Bateman, A., The Pfam protein families database. Nucleic Acids Res, 2008. 36(Database issue): p. D281-8.

11) Li, K.B., ClustalW-MPI: ClustalW analysis using distributed and parallel computing. Bioinformatics, 2003. 19(12): p. 1585-6.

12) Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J., Jalview Version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics, 2009. 25(9): p. 1189-91.

13) Chevenet, F., Brun, C., Banuls, A.L., Jacq, B., and Christen, R., TreeDyn: towards dynamic graphics and annotations for analyses of trees. BMC Bioinformatics, 2006. 7: p. 439.

14) Bailey, T.L. and Elkan, C., Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol, 1994. 2: p. 28-36.

15) Bailey, T.L. and Gribskov, M., Combining evidence using p-values: application to sequence homology searches. Bioinformatics, 1998. 14(1): p. 48-54.

16) Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S., Quantifying similarity between motifs. Genome Biol, 2007. 8(2): p. R24.