Decifrando Cisteíno Proteases em *Plasmodium*: Uma Estratégia de Genômica Comparativa e Modelagem Estrutural

Kary A.C.S. Ocaña[1], Marco T. A. Garcia-Zapata[2]

[1]PESC/COPPE - Universidade Federal de Rio de Janeiro, Rio de Janeiro, Brasil
[2]Instituto de Patologia Tropical e Saúde Pública, Universidade Federal de Goiás, Goiânia, Brasil

kary@cos.ufrj.br, zapata@iptsp.ufg.br

1. Introdução

A malária continua sendo um dos mais devastadores problemas de saúde globais em regiões endêmicas. A busca contínua por novos alvos terapêuticos para a malária é uma das principais prioridades, principalmente devido à crescente resistência ao tratamento contra esta doença (Arango *et al*. 2008). Atualmente existem vários esforços em andamento para desenvolver medicamentos quimioterápicos contra a malária. O primeiro passo é identificar um alvo terapêutico adequado. Vários projetos de descoberta da droga se concentram na pesquisa das cisteíno proteases (CP) como alvos fármaco-terapêuticos promissores, uma vez que desempenham um papel importante no ciclo de vida do parasita. Dentre as famílias de CP mais notáveis estão a papaína e a ubiquitina (Rosenthal 2004). A primeira foi identificada como um componente importante no potencial invasivo em vários agentes patógenos humanos e de animais e está relacionada a doenças tais como a artrite, câncer e doenças infecciosas e vasculares. Em relação à segunda, dada a importância da via de ubiquitinação em uma ampla variedade de processos celulares, as adaptações apicomplexa-específicas desta via podem representar novos alvos terapêuticos contra estes parasitas.

Por outro lado, o gerenciamento de experimentos de bioinformática está longe de ser trivial devido ao grande volume de dados biológicos que demanda um ambiente de processamento de alto desempenho (PAD) computacional. Estes experimentos podem ser assistidos por meio de sistemas de *workflows* científicos (Mattoso *et al*. 2010).

O objetivo principal deste artigo é realizar análises de genômica comparativa e modelagem estrutural por meio de sistemas de *workflows* científicos. As análises enfatizam a identificação de sequências ortólogas de CP em *Plasmodium* e a construção de modelos estruturais dessas sequências por meio da modelagem por homologia. Estas análises desempenham um papel importante na descoberta de drogas antimaláricas baseada em estruturas uma vez que possibilitam uma alternativa econômica e viável para gerar modelos estruturais que auxiliem à descoberta de novas drogas. Desta forma, os modelos estruturais obtidos neste estudo estão prontos para serem utilizados em futuras abordagens de genômica estrutural (GS).

2. Materiais e Métodos

Foram executados quatro *workflows* (Figura 1) em paralelo no ambiente de nuvem Amazon EC2, os quais estão em fase de teste e serão disponibilizados no futuro.

(I) **Análise de genômica comparativa**, modelada no SciHmm (Ocaña *et al*. 2011a): os genomas completos de *P. falciparum 3D7, P. knowlesi,* e *P. yoelii yoelii 17XNL* foram obtidos do GOLD (Pagani *et al*. 2012) e (1) organizados por KOG (*EuKaryotic Orthologous Groups*) usando o OrthoSelect (Schreiber *et al*. 2009); (2) HHsearch foi usado para construir os perfis dos KOG e compará-los com os perfis das CP obtidas do MEROPS (Rawlings *et al*. 2011); (3) os *hits* obtidos com HHsearch (*i.e.,* papaína, ubiquitina) foram usados pelo PSIPBLAST (Altschul *et al*. 1997) para buscar proteínas homólogas dessas CP em *Plasmodium* no Refseq do NCBI (Pruitt *et al*. 2009); e (4) os *hits* finais obtidos do PSIBLAST foram verificados manualmente.

(II) **Análise filogenética**, modelada no SciPhy (Ocaña *et al*. 2011b): (1) MAFFT foi usado para construir os alinhamentos múltiplos de sequência (AMS); (2) Readseq foi usado para a conversão

de formato do AMS; (3) a análise filogenética foi baseada no modelo WAG, sugerido pelo ModelGenerator; (4) as matrizes de distância foram calculadas com Tree-Puzzle, as árvores de agrupamento de vizinhos (AV) inferidas com Neighbor, e a confiabilidade de ramificação testada com 1.000 replicações de *bootstrap*, usando SeqBoot e Consense. A análise de Máxima Verossimilhança (MV) foi realizada usando RAxML com 1.000 replicações de *bootstrap*. A inferência Bayesiana (IB) foi realizada usando MrBayes com uma cadeia fria e três aquecidas executadas por 10.000.000 de gerações. A amostragem das árvores foi feita a cada 1.000 gerações e o "*burn in*" foi definido em 25%. (5) As árvores filogenéticas foram visualizadas com MEGA.

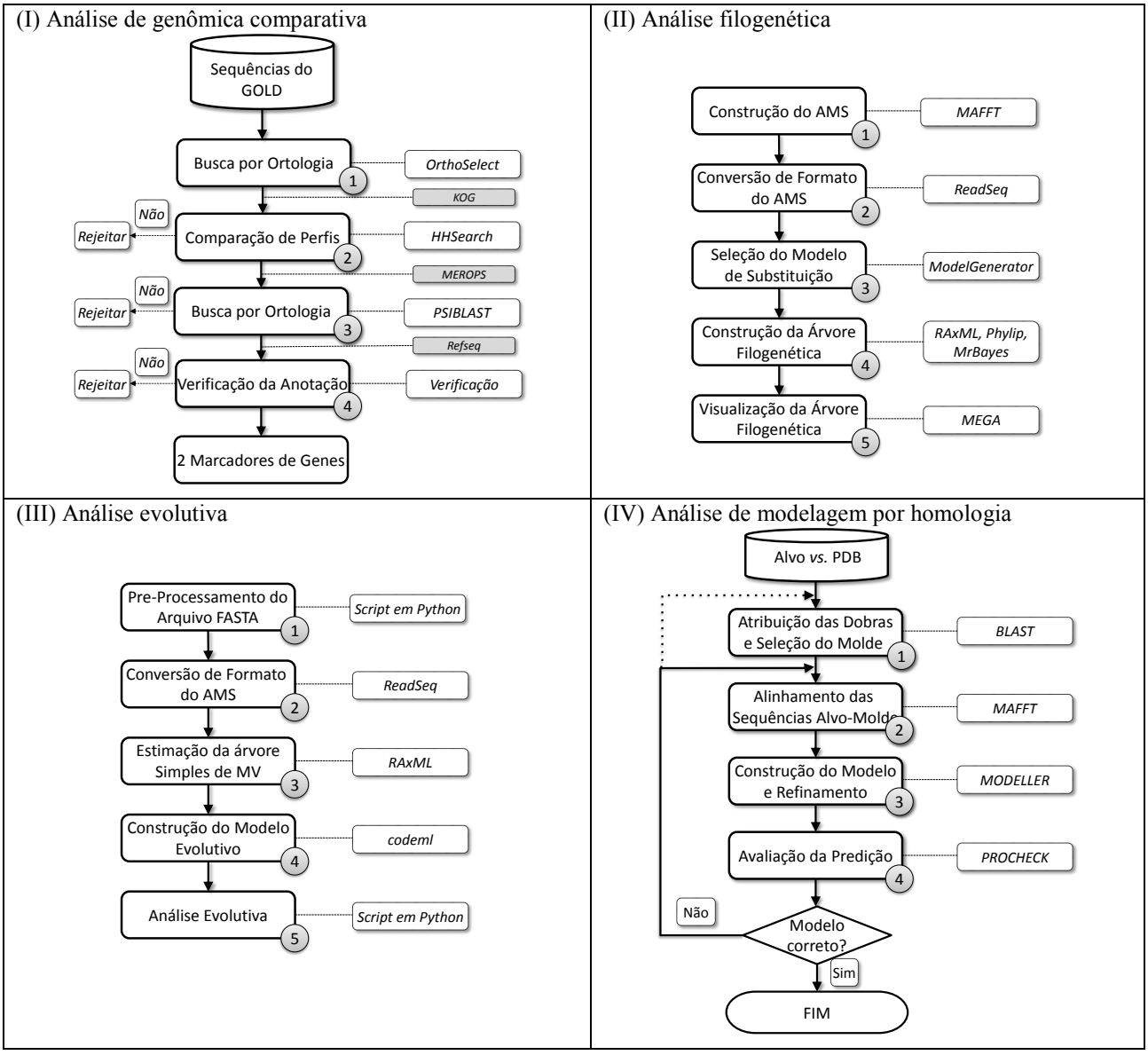

Figura 1. A visão conceitual ilustrando os quatro *workflows* científicos de bioinformática.

(III) **Análise evolutiva**, modelada no SciEvol (Ocaña *et al*. 2012): (1) Um *script* em Python formata o AMS retirando os codões de terminação; (2) Readseq é usado na conversão de formato do AMS; (3) RAxML é usado para estimar a árvore de MV simples; (4) codeml é usado na construção dos modelos evolutivos; e (5) *scripts* em Python são executados para realizar as análises evolutivas. (IV) **Análise de modelagem por homologia**, este *workflow* é baseado nas rotinas de modelagem por homologia apresentadas por (Martí-Renom *et al*. 2000), que constam de quatro etapas principais sequenciais: (1) atribuição das dobras e seleção do molde, usando BLAST; (2) alinhamento das sequências alvo-molde, usando MAFFT; (3) construção e refinamento do modelo, usando MODELLER; e (4) avaliação da predição, usando PROCHECK.

3. Resultados

Um total de 593 KOG foi definido com OrthoSelect comum a *P. falciparum*, *P. knowlesi*, e *P. yoelii*; embora apenas dois KOG (KOG1543 e KOG1863) foram identificados com HHsearch relacionados às CP nessas espécies. Esses KOG pertencem às CP papaína (1 sequência) e ubiquitina (3 sequências). Para incrementar o número de sequências homólogas destas duas famílias, um PSIBLAST adicional foi executado contra todas as espécies de *Plasmodium*, e foram encontradas 13 sequências para a papaína e 7 sequências para a ubiquitina.

As árvores filogenéticas (derivadas de AV, MV, e IB) foram quase idênticas. A Figura 2 mostra as árvores de MV das CP papaína (A) e ubiquitina (B) em *Plasmodium*. Os valores de suporte de *bootstrap* foram consistentes e todas as probabilidades posteriores Bayesianas nos ramos estiveram perto de 1. A topologia da árvore de AV difere ligeiramente das árvores de MV e IB, mas o agrupamento dos clados principais é semelhante. As árvores da papaína (A) e da ubiquitina (B) mostram topologias muito semelhantes (Figura 2). Ambas as árvores apresentam três clados monofiléticos espécie-específicos mais representativos em *Plasmodium*: (i) *P. falciparum* (azul); (ii) *P. berghei*, *P. yoelii*, e *P. chaubadi* (cor de rosa); e *P. vivax*, *P. cymolgi*, e *P. knowlesi* (verde). Além disso, pode ser observado que a árvore da papaína (A) divide-se em subfamílias (*i.e.,* Falcipain, Vivapain, e Bergheipain), que representam as espécies nas quais foram encontradas.

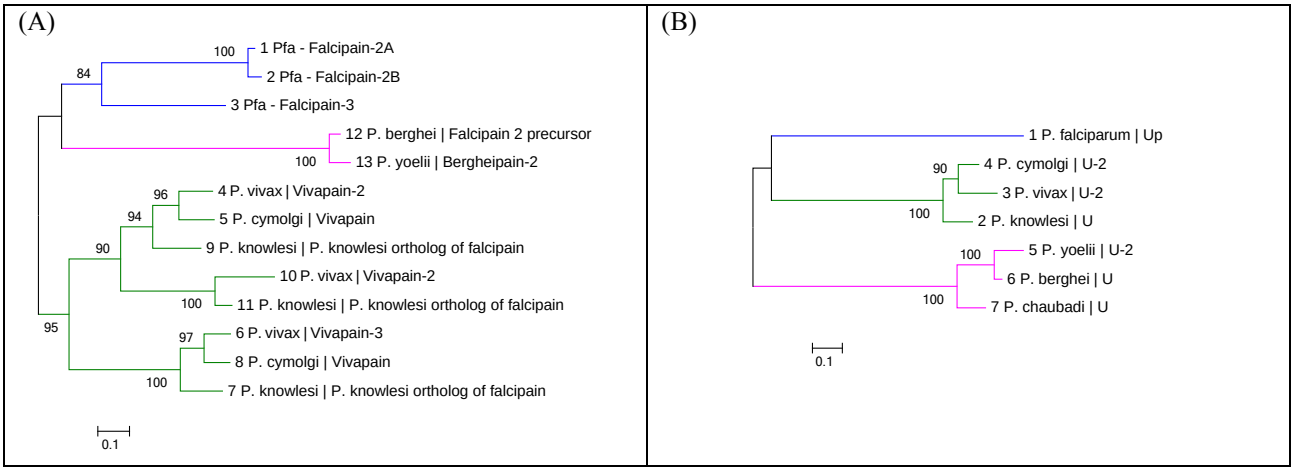

Figura 2. Árvores das (A) papaínas e (B) ubiquitinas em *Plasmodium*. Essas CP estão destacadas em azul para *P. falciparum*; cor de rosa para *P. berghei*, *P. yoelii*, e *P. chaubadi*; e verde para *P. vivax*, *P. cymolgi*, e *P. knowlesi*, respectivamente. RAxML foi usado com 1.000 replicações de *bootstrap*. Valores acima de 80% são mostrados.

As Estimativas de MV (EMV) dos parâmetros e os valores estatísticos (2Δl) do Teste de Razão de Verossimilhança (TRV) foram estimados sob os modelos de códon (*i.e.,* M0, M3, M7, e M8) (Anisimova *et al*. 2001). As comparações dos 2Δl desses modelos para as CP sugerem que: (i) M0 *vs.* M3 indica variabilidade significativa nas sequências entre os sítios: papaína: 49% ($0.61<\omega<1.88$) e ubiquitina: 54% ($0.90<\omega<3.16$) e (ii) M7 *vs.* M8 indica a presença de uma grande proporção de sítios sob seleção positiva de diversificação ($p < 0.0001$), o que foi confirmado pelas EMV que sugere a presença de alguns sítios de aminoácidos sujeitos a seleção positiva. O valor estimado de ω baseado no M8 indica seleção positiva para papaína: 2.21 e ubiquitina: 2.50; porém não foi encontrada evidência significativa de seleção positiva nos sítios ativos dessas CP.

Além disso, os modelos estruturais construídos obtidos com as 20 sequências de CP resultaram em 100 modelos de qualidade para proteínas. Os 20 melhores modelos, ou seja, o melhor para cada uma das 20 sequências de papaína e ubiquitina foram selecionados com base no menor valor do escore de avaliação DOPE. A média da pontuação do valor de GA341 (0,50) do MODELLER bem como a análise com PROCHECK (75% das regiões mais favorecidas) e do MolProbity (80%) das regiões favorecidas, confirmou que estes modelos são razoáveis. Estes 20 modelos irão ser utilizados em outras análises de GS mais aprofundadas *e.g.,* interacção proteína-proteína e acoplamento molecular.

4. Conclusões e Perspectivas

Apresentamos três pontos principais, que compreendem os objetivos que foram especificados neste trabalho. Primeiro, a identificação de sequências ortólogas da papaína e da ubiquitina em *Plasmodium* apoia fortemente a hipótese da origem evolutiva comum destas CP, compartilhando um ancestral apicomplexo comum. A determinação computacional do repertório de KOG fornece não só a identificação e relação destas duas CP, mas também oferece um conjunto exclusivo de sequências CP em *Plasmodium*, que foram usados para construir modelos estruturais, como potenciais candidatos de drogas alvo de inibidores da protease. Em segundo lugar, uma vez que os valores de *bootstrap* mostram uma alta consistência em todas as árvores filogenéticas, isto sugere que a papaína e ubiquitina podem ser alvos potenciais na quimioterapia antimalárica. Terceiro, o nível de variabilidade atuando nas CP em *Plasmodium* indica que a evolução molecular adaptativa desempenha um papel importante no surgimento das novas funções, o que provavelmente contribui para a característica biológica e para a adaptação desses parasitas. Assim também, o comportamento sobre a origem e a divergência dessas CP deve fornecer informações sobre os mecanismos de adaptação do parasita, crescimento, desenvolvimento, infecção, patogênese, e alvo de drogas.

Desta forma, os resultados sugerem que a papaína e a ubiquitina deveriam ser consideradas como potenciais candidatos para a descoberta de drogas com base em (i) a importante via bioquímica no ciclo de vida dos *Plasmodium*, (ii) a específica distribuição filogenética em *Plasmodium*, e (iii) a conservação dos seus sítios ativos, os que apresentan seleção purificadora. No entanto, outros experimentos *in silico*, *in vitro*, e *in vivo* são necessários para reforçar essas hipóteses, bem como uma investigação mais aprofundada em relação à evolução e GS para entender melhor o papel funcional das CP em *Plasmodium*.

Na ausência de estruturas experimentais, a modelagem por homologia desempenha um papel importante no processo de descoberta de drogas baseado em estrutura e tem sido utilizada com sucesso na GS, abrindo uma variedade de aplicações na investigação e na triagem de drogas-alvo. Portanto, a nossa modelagem por homologia *in silico* fornece uma alternativa econômica e viável para gerar modelos razoavelmente precisos que auxiliem na descoberta de drogas, e pode ser extrapolada usando outras enzimas de interesse.

Referências

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J., (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic acids research*, v. 25, n. 17 (set.), p. 3389–3402.

Anisimova, M., Bielawski, J. P., Yang, Z., (2001), "Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution", *Molecular Biology and Evolution*, v. 18, n. 8 (ago.), p. 1585–1592.

Arango, E., Carmona-Fonseca, J., Blair, S., (2008), "In vitro susceptibility of Colombian Plasmodium falciparum isolates to different antimalarial drugs", *Biomédica: Revista Del Instituto Nacional De Salud*, v. 28, n. 2 (jun.), p. 213–223.

Martí-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo, F., Sali, A., (2000), "Comparative protein structure modeling of genes and genomes", *Annual review of biophysics and biomolecular structure*, v. 29, p. 291–325.

Mattoso, M., Werner, C., Travassos, G. H., Braganholo, V., Murta, L., Ogasawara, E., Oliveira, D., Cruz, S. M. S. da, Martinho, W., (2010), "Towards Supporting the Life Cycle of Large-scale Scientific Experiments", *International Journal of Business Process Integration and Management*, v. 5, n. 1, p. 79–92.

Ocaña, K. A. C. S., Oliveira, D. de, Horta, F., Dias, J., Ogasawara, E., Mattoso, M., (2012), "Exploring Molecular Evolution Reconstruction Using a Parallel Cloud-based Scientific Workflow", *Advances in Bioinformatics and Computational Biology*, chapter 7409, Berlin, Heidelberg: Springer, p. 179–191.

Ocaña, K. A. C. S., Oliveira, D., Dias, J., Ogasawara, E., Mattoso, M., (2011a), "Optimizing Phylogenetic Analysis Using SciHmm Cloud-based Scientific Workflow". In: *2011 IEEE Seventh International Conference on e-Science (e-Science)*, p. 190–197, Stockholm, Sweden.

Ocaña, K. A. C. S., Oliveira, D., Ogasawara, E., Dávila, A. M. R., Lima, A. A. B., Mattoso, M., (2011b), "SciPhy: A Cloud-Based Workflow for Phylogenetic Analysis of Drug Targets in Protozoan Genomes", *Advances in Bioinformatics and Computational Biology*, , chapter 6832, Berlin, Heidelberg: Springer, p. 66–70.

Pagani, I., Liolios, K., Jansson, J., Chen, I.-M. A., Smirnova, T., Nosrat, B., Markowitz, V. M., Kyrpides, N. C., (2012), "The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata", *Nucleic acids research*, v. 40, n. Database issue (jan.), p. D571–579.

Pruitt, K. D., Tatusova, T., Klimke, W., Maglott, D. R., (2009), "NCBI Reference Sequences: current status, policy and new initiatives", *Nucleic Acids Research*, v. 37, n. Database issue (jan.), p. D32–D36.

Rawlings, N. D., Barrett, A. J., Bateman, A., (2011), "MEROPS: the database of proteolytic enzymes, their substrates and inhibitors", *Nucleic Acids Research*, v. 40, n. D1 (nov.), p. D343–D350.

Rosenthal, P. J., (2004), "Cysteine proteases of malaria parasites", *International Journal for Parasitology*, v. 34, n. 13-14 (dez.), p. 1489–1499.

Schreiber, F., Pick, K., Erpenbeck, D., Wörheide, G., Morgenstern, B., (2009), "OrthoSelect: a protocol for selecting orthologous groups in phylogenomics", *BMC Bioinformatics*, v. 10, n. 1, p. 219.