Simposio

BIOINFORMÁTICA APLICADA AL ESTUDIO MOLECULAR DE PARÁSITOS

ToxoDB: an integrated genomic database for the study of *Toxoplasma gondii* and *Neospora caninum*

Jessica C. Kissinger[1] and the Eukaryotic Pathogen Database Team[1,2,3]

[1] Center for Tropical and Emerging Global Diseases, Institute of Bioinformatics and Department of Genetics, University of Georgia, Athens, GA, USA
[2] Department of Genetics, University of Pennsylvania, Philadelphia, PA, USA
[3] Department of Biology, University of Pennsylvania, Philadelphia, PA, USA

Integrated databases resources are integral to the study of biology and of host-pathogen interactions. Tremendous strides in technology have permitted the generation of large, valuable data sets for many life cycle stages of numerous pathogens and/or their hosts. The *Toxoplasma* genome database (http://ToxoDB.org) currently hosts, and integrates, a wide array of diverse, large-scale data sets (1).

ToxoDB contains three fully annotated genome sequences for *Toxoplasma gondii* (GT1, VEG and ME49 strains) and the annotated genome sequence of the closely related *Neospora caninum* (2). Additionally, the database contains all available EST sequence data, 13 microarray experimental datasets for tachyzoites and bradyzoites, 2 SAGE-Tag data sets, 1 tachyzoite RNA-Seq experiment from both *T. gondii* VEG and *N. caninum*, and ChIP-on-chip data sets from precipitation with several different histone antibodies. The database also contains 21 proteomic data sets, SNPs and ortholog and synteny determinations between all *Toxoplasma* strains and *Neospora*.

The database is part of the NIH-funded Eukaryotic Pathogen Database, (http://EuPathDB.org) (3). The integrated data are presented in an intuitive interface that permits strategic data mining across the data sets using our strategy system (4). For example, it is possible to identify all annotated genes that contain a signal peptide, are expressed in tachyzoites, but not bradyzoites, have evidence of protein expression and are conserved in *Neospora* with only a few clicks of the cursor.

The database, which has been in existence since 2003 (5), is a free community resource that is linked to other prominent apicomplexan organisms (*Plasmodium, Thileria, Babesia* and *Cryptosporidium*) as well as other prominent eukaryotic pathogens (*Giardia, Entamoeba, Leishmania, Trypanosoma*, Microsporidia and *Trichomonas*) via the EuPathDB.org portal. The EuPathDB portal permits queries of *Toxoplasma* and any other organism contained within the database simultaneously. For example, it is possible to search for genes with certain properties that are conserved in *Toxoplasma* and *Neospora* but not other Apicomplexa.

References

1. Gajria B, Bahl A, Brestelli J, Dommer J, Fischer S, Gao X, *et al*. ToxoDB: An integrated *Toxoplasma gondii* database resource. Nucleic Acids Res. 2008;36(Database issue):D553-6.http://www.sanger.ac.uk/resources/downloads/protozoa/neospora-caninum.html.
2. Aurrecoechea C, Brestelli J, Brunk BP, Fischer S, Gajria B, Gao X, *et al*. EuPathDB: A portal to eukaryotic pathogen databases. Nucleic Acids Res. 2010;38(Database issue):D415-9.
3. Fischer S, Aurrecoechea C, Brunk BP, Gao X, Harb OS, Kraemer ET, *et al*. The strategies WDK: A graphical search interface and web development kit for functional genomics databases. DATABASE - The Journal of Biological Databases and Curation. In press.
4. Kissinger JC, Gajria B, Li L, Paulsen IT, Roos DS. ToxoDB: Accessing the *Toxoplasma gondii* genome. Nucleic Acids Res. 2003;31:234-6.

• • •

Búsqueda racional de medicamentos anti-*Leishmania* mediante aproximaciones computacionales

Andrés Flórez[1], Rodrigo Ochoa[1], Jairo Espinosa[2], Carlos Muskus[1]

[1] Programa de Estudio y Control de Enfermedades Tropicales-PECET Universidad de Antioquia, Medellín, Colombia
[2] Grupo Automática, Universidad Nacional, Medellín, Colombia

La búsqueda racional de medicamentos anti-*Leishmania* mediante métodos computacionales en el Programa de Estudio y Control de Enfermedades Tropicales, PECET, de la Universidad de Antioquia, está enfocada en: i) la selección e identificación de potenciales blancos proteicos, y ii) la predicción de segundos usos de medicamentos anotados en bases de medicamentos y enfatizando en aquellos medicamentos aprobados para uso en humanos o que están en proceso de aprobación.

Para la identificación o selección de blancos de medicamentos se han empleado dos estrategias: una es mediante la generación de redes de interacción de proteínas y la otra es mediante la búsqueda de homología con PSI-BLAST.

Para la generación de las redes de proteínas, se empleó el proteoma anotado de *Leishmania major* y de las otras dos especies de *Leishmania* secuenciadas. Las proteínas para la construcción de las redes se descargaron de geneDB. La red de interacción se construyó empleando tres métodos validados: PSIMAP, PEIMAP y iPfam, con lo cual se obtuvo una red de alta confianza (>0,70). Parámetros como el grado de intermediación, la conectividad y el esquema de doble puntaje, se emplearon para determinar las proteínas esenciales en la red de *L. major*, que es el genoma mejor anotado actualmente.

Con los tres métodos se identificaron 1.366 nodos y 33.861 interacciones. De éstos, 142 proteínas fueron identificadas como blancos potenciales de medicamentos y filtradas por homología con el proteoma humano. Además, se predijo la función de 263 proteínas hipotéticas. Teniendo en cuenta los parámetros de esencialidad, encontramos que la mayoría de las proteínas esenciales fueron clasificadas como cinasas. Este hallazgo es relevante, dado que las cinasas son proteínas implicadas en muchos procesos importantes o esenciales dentro de una célula. También, se predijo la red proteica del proceso de edición del ARNA y se detectaron como esenciales dos proteínas de las 21 que forman el "editosoma".

La segunda estrategia, enfocada a identificar proteínas esenciales mediante PSI-BLAST, está basada en el raciocinio de que si un medicamento tiene acción demostrada o actúa sobre una proteína y esta proteína presenta gran homoogía con una proteína de *Leishmania*, este medicamento, con algunas consideraciones, podría tener acción también contra la proteína de *Leishmania*.

Para esto, se buscaron los medicamentos depositados en bases de datos cuyo blanco de acción son las proteínas. Estas proteínas se compararon con el proteoma anotado de *Leishmania*. Además, se filtraron todas las proteínas que tenían homología con el proteoma del humano. Usando esta estrategia, se encontraron 147 proteínas homólogas en *Leishmania*, las cuales pueden ser blancos potenciales del respectivo medicamento depositado en la base de datos. De la lista de medicamentos correspondientes a las 147 proteínas, se seleccionaron 15 con base en aspectos como capacidad hidrófoba, vía de administración, costos, etc. Varios de estos medicamentos están siendo evaluados en ensayos *in vitro*.

Se piensa que la respuesta inflamatoria generada durante la infección por el parásito *Leishmania* favorece la aparición de la lesión como tal. Señal de esto es la dificultad para aislar parásitos de lesiones mucocutáneas, tanto en biopsias como en aspirados tomados directamente de la lesión, lo que sugiere la presencia de un bajo número de parásitos a pesar de la agresividad de la lesión. Esta forma clínica de la enfermedad también cursa con un título alto de anticuerpos, apoyando aún más la hipótesis de que la respuesta inflamatoria puede exacerbar la lesión.

Con base en esta asunción, pensamos que los medicamentos candidatos con actividad anti-*Leishmania*, pero que simultáneamente pueden tener acción antiinflamatoria, podrían ser más efectivos en inducir la cura o resolución de la lesión o la enfermedad. Incluso, si dicho medicamento, además, tiene acción cicatrizante, sería mucho mejor. Haciendo una búsqueda activa en bases de medicamentos, hemos encontrado 64 medicamentos con actividad antiinflamatoria y 17 con actividad antiulcerosa. Se puede predecir una posible acción de estos medicamentos antiinflamatorios contra *Leishmania* mediante herramientas que predicen

la actividad del medicamento con base en su estructura.

Se evaluó la predicción de la actividad con base en la estructura de la anfotericina B, uno de los medicamentos empleados para tratar casos de leishmaniasis y algunas micosis, y se predijo la actividad antifúngica y antiprotozoaria, lo cual sugiere que por medio de esta estrategia podríamos seleccionar medicamentos antiinflamatorios con potencial actividad anti-*Leishmania*.

Otras aproximaciones enfocadas a la búsqueda de medicamentos anti-*Leishmania* incluyen el *docking* masivo y el empleo de herramientas de inteligencia artificial, específicamente máquinas de soporte vectorial y clasificadores bayesianos. El *docking* es una de las herramientas bioinformáticas más prometedoras en la búsqueda de medicamentos contra enfermedades en general. Se han descubierto varios medicamentos contra algunas enfermedades por medio del *docking*.

Hasta la fecha, aproximadamente 158 proteínas de *Leishmania* se han incluido en la base de datos PDB. Las proteínas se han derivado, principalmente, de *L. major* (66) y *L. mexicana* (38) y, en menor proporción, de *L. donovani* (14) *L. infantum* (6) y *L. tarentolae* (5). Dado que estas estructuras tridimensionales están disponibles, la idea es realizar el *docking* empleando la base datos de medicamentos ZINC contra proteínas no redundantes de *Leishmania*, cuyo número hasta la fecha se estima en 70 a 80. Los experimentos de *docking* se correrán empleando el *software Autodock Vina* y *World Community Grid* de IBM.

El proteoma de *Leishmania* está compuesto, aproximadamente, de 8.200 a 8.300 proteínas dependiendo de la especie, pero sólo unas 168 proteínas se han cristalizado y sus estructuras se han depositado en PDB. Este bajo número de proteínas cristalizadas limita el uso del *docking* como una herramienta predictora de medicamentos en aquellos casos en los cuales se identifica un potencial blanco pero carece de estructura en tercera dimensión.

Para tratar de resolver este inconveniente, se han desarrollado herramientas de inteligencia artificial, como las máquinas de soporte vectorial o los clasificadores bayesianos, lo cual evita la necesidad de usar estructuras en tercera dimensión en la búsqueda de medicamentos y, en cambio, hace uso de la estructura primaria o los dominios de las proteínas, entre otros.

En nuestro laboratorio, estamos entrenando máquinas de soporte vectorial y clasificadores bayesianos con medicamentos que reconocen cinasas de cualquier fuente (clasificación positiva). Además, la máquina se entrenará con medicamentos que no reconocen cinasas (clasificación negativa). Una vez entrenada la máquina, el "cinoma" de *Leishmania* será analizado con la herramienta de inteligencia artificial para identificar si un medicamento particular reconoce alguna de las cinasas de *Leishmania*.

Actualmente, estamos haciendo la codificación vectorial de las cinasas con base en los dominios de este grupo de enzimas. El medicamento que sea clasificado o establezca relación con una de las cinasas de *Leishmania*, será evaluado *in vitro* inicialmente y, si los resultados son promisorios, pasará a análisis *in vivo*.

Se presentan las estrategias usadas en estas aproximaciones y empleadas en la búsqueda de medicamentos anti-*Leishmania*, además de los resultados preliminares.

Bibliografía

1. AsseAssenov Y, Ramírez F, Schelhorn SE, Lengauer T, Albrecht M, *et al*. Computing topological parameters of biological networks. Bioinformatics. 2008;24:282-4.
2. Batada NN, Hurst LD, Tyers M, *et al*. Evolutionary and physiological importance of hub proteins. PLoS Comput Biol. 2006;2:e88.
3. Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G, *et al*. Support vector machines and kernels for computational biology. PLoS Comput Biol. 2008;4:1-10.
4. Brohée S, Faust K, Lima-Méndez G, Vanderstocken G, van Helden J, *et al*. Network analysis tools: From biological networks to clusters and pathways. Nat Protoc. 2008;3:1616-29.
5. Brohée S, van Helden J. Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics. 2006;7:488.
6. Bulashevska A, Stein M, Jackson D, Eils R, *et al*. Prediction of small molecule binding property of protein domains with Bayesian classifiers based on Markov chains. Computational Biology and Chemistry. 2009;33:457-60.
7. de Azevedo WF Jr, Soares MB. Selection of targets for drug development against protozoan parasites. Curr Drug Targets. 2009;10:193-201.
8. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, *et al*. IntAct: An open source molecular interaction database. Nucleic Acids Res. 2004;32:D452-5.
9. Hood L. Systems biology: Integrating technology, biology, and computation. Mech Ageing Dev. 2003;124:9-16.
10. Jorgensen WL. The many roles of computation in drug discovery. Science. 2004;303:1813-18.

11. Karthikeyan R, Mohamed S, Sridhar V, Nagasuma C, *et al*. Support vector machine classifier for predicting drug binding to P-glycoprotein. J Proteomics Bioinform. 2009;2:193-201.
12. Kasabov N, Pang S. Transductive support vector machines and applications in bioinformatics for promoter recognition. Neural Information Processing. 2004;3:31-8.
13. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK, *et al*. Relating protein pharmacology by ligand chemistry. Nat Biotechnol. 2007;25:197-206.
14. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, *et al*. Predicting new molecular targets for known drugs. Nature. 2009;462:175-81.
15. Lingyi L, Ziliang Q, Yu-Dong C, Yixue Li, *et al*. ECS: An automatic enzyme classifier based on functional domain composition. Computational Biology and Chemistry. 2007;31: 226-32.

• • •

Genomics and bioinformatics applied to the study of the transcriptome of *Eimeria* parasites

J. Novaes[1], M. Ferro[1], R. Y. Abe[2], L. T. R. L. Diniz[1], A. P. S. Manha[1], S. G. Guimarães[1], J. C. M. Mello[1], L. Varuzza[3], C. A. B. Pereira[3], A. M. Durham[2], A. M. B. Madeira[1], A.Gruber[1]

[1] Departamento de Parasitologia, Instituto de Ciências Biomédicas, Universidade de São Paulo, São Paulo, Brasil
[2] Departamento de Ciências da Computação, Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, Brasil

The genus *Eimeria* comprises protozoan parasites that cause severe infections in domestic animals. Seven *Eimeria* species are implicated with coccidiosis of the domestic fowl, an enteric disease that leads to important economic losses in poultry production. The genome of the model species, *Eimeria tenella*, presents a complexity of circa 55-60 MB distributed in 14 chromosomes, with an estimated G+C content of 53% (1).

An international consortium, composed of research groups in the UK, Malaysia and Brazil has been established in 2000, with different tasks being divided among the member groups (2). As a first result of this effort, the complete sequence of the chromosome 1 has been determined (3). This chromosome revealed, as a major feature, a segmentation pattern consisting of gene-rich regions associated with a high content of short tandemly repeated sequences, and regions with poor gene density and low repetitive content. The segments with a highly repetitive content were associated with recombination events, accounting for chromosome size variations observed in molecular karyotypes of different *E. tenella* strains. Preliminary data show that the whole genome of *E. tenella* presents the same segmented structure, and a draft sequence is publicly available at http://www.sanger.ac.uk/Projects/E_tenella/.

In addition to the genome sequencing, a good characterization of the parasite transcriptome may lead to a better understanding of the parasite biology and the development of new control strategies.

With this aim in mind, our group carried out a comparative EST sequencing study of the three most economically relevant *Eimeria* species: *E. tenella*, *E. maxima* and *E. acervulina*. For this purpose, we isolated mRNA from different developmental stages of the parasites, and constructed cDNA libraries using ORESTES method, a technique based on the synthesis of cDNA molecules using arbitrarily primed RT-PCRs (4).

In total, we have generated circa 15,000 high-quality reads for each one of the three *Eimeria* species. All generated cDNA reads were submitted to a multistep pre-processing pipeline using EGene, a bioinformatics platform developed by our group for the generic construction of automated pipelines (5). The 'reads' were assembled with CAP3 and the reconstructed cDNAs submitted to a comprehensive annotation pipeline including ORF determination, similarity searches, protein motif finding, KOG classification and GO mapping. This annotation pipeline was performed using a set of EGene's newly developed components (Durham and Gruber, manuscript in preparation).

In the case of *E. tenella*, in addition to our ORESTES reads, we also employed in the assembly step a set of around 35,000 ESTs publicly available on the GenBank and the Sanger Institute. Altogether, the full ORESTES/EST combined set contained samples from the following developmental stages: unsporulated oocysts, sporulating oocysts (sporoblast phase), sporulated oocysts, sporozoites, and first- and

second-generation merozoites. The annotation data in feature table and GFF3 formats, and all supporting evidences, will be publicly released in a near future at the address http://www.coccidia.icb.usp.br/eimeria_tdb.

The gene expression patterns of the parasite transcriptome were assessed by two different methods. First, we processed the assembled cDNAs using proprietary scripts to generate digital expression profiles. In this approach, the number of EST/ORESTES reads composing each assembled cDNA is assumed to be proportional to the abundance of the corresponding transcript in each one of the respective developmental stages. We estimated the expression levels of all reconstructed transcripts in the different stages, in the three *Eimeria* species. We also performed an expression analysis on sporozoite and second-generation merozoite stages of *E. tenella* using LongSAGE (6), a variation of the original SAGE technique that generates 21-base tags. The LongSAGE libraries were sequenced and the 'reads' were processed using both, EGene and SAGE_Suite, a locally developed package for tag extraction and counting. In total, we have produced more than 35,000 tags, corresponding to 9,516 unique tags.

This number is relatively close to the estimated complexity of circa 8,000 genes found in coccidian (*Toxoplasma* and *Neospora*) transcriptomes. An analysis of the tag profiles revealed that around 66.5% of the SAGE tags present single counts, whereas more than 88% of the unique tags show counts below to five. This result suggests that the eimerian transcriptome is very narrowly expressed in any stage, with some few genes presenting a very high expression, while most of the genes are expressed in low amounts. In both datasets, digital Northern and LongSAGE, differentially expressed genes were identified using Kemp (7), a statistical frequentist exact test that uses a tag-customized critical level which minimizes a linear combination of type I and type II errors.

A comparison of the digital Northern and LongSAGE results in *E. tenella* showed a good agreement, and many differentially expressed genes have been mutually identified in both techniques. In the case of LongSAGE, we detected a total of 270 differentially expressed tags between sporozoites and merozoites, and succeeded to map 199 of these tags onto assembled cDNAs. From this set, a total of 144 tags showed a good agreement with the expression pattern observed on digital Northern.

This result suggests that both techniques, SAGE and digital Northern, present a good correlation for most of the genes. We also carried out a preliminary experimental validation with real-time quantitative RT-PCR using twelve genes selected from sets of differentially and non-differentially expressed genes. In all cases, the expression status observed on LongSAGE and/or digital Northern has been confirmed by real-time PCR. In four tested genes, the expression ratio observed between the tested developmental stages has been experimentally confirmed with a very high correlation.

Finally, the digital expression profiles were submitted to agglomerative hierarchical clustering analyses using Simcluster (8), and distance trees have been determined. The topologies of the obtained dendrograms were in good agreement to what is known in regard to the biology of the parasites. Thus, expression patterns of unsporulated, sporulating and sporulated oocysts were all clustered into a single group, whereas sporozoites and merozoites were clustered into separate groups. Also, distance trees of the expression profiles clearly showed that stages that are more closely related in the life cycle (e.g sporulated oocysts and sporozoites) present more similar gene expression patterns to one another than to other developmental stages.

To identify pairs of orthologous proteins across the different apicomplexans, we carried out an all-against-all comparison of the translated products of nine Apicomplexa organisms using InParanoid (9), in a total of 36 paired analyses. Next, we merged all pairwise ortholog clusters identified by Inparanoid into multi-species clusters using MultiParanoid (10). This analysis allowed us to identify proteins that are evolutionary conserved across different apicomplexan taxa, and that may potentially exert common functions in the members of the phylum. Also, we were able to identify orthologous groups specific to the genus *Eimeria*.

We now intend to extend the validation of differential expression data to a larger set of genes, and identify genes involved in some particular steps of the parasite life cycle. For instance, the mechanisms and specific genes involved in the sporulation process are still poorly understood. Thus, by using our expression data and selecting the most promising candidates for experimental validation, we expect to better define proteins associated with this essential step of the life cycle. This knowledge may help us to devise new strategies to control this important disease.

References

1. Chapman HD, Shirley MW. The Houghton strain of *Eimeria tenella*: A review of the type strain selected for genome sequencing. Avian Pathol. 2003;32:115-27.
2. Shirley MW, Ivens A, Gruber A, Madeira AM, Wan KL, Dear PH, *et al*. The *Eimeria* genome projects: A sequence of events. Trends Parasitol. 2004;20:199-201.
3. Ling KH, Rajandream MA, Rivailler P, Ivens A, Yap SJ, Madeira AM, *et al*. Sequencing and analysis of chromosome 1 of *Eimeria tenella* reveals a unique segmental organization. Genome Res. 2007;17:311-9.
4. Dias-Neto E, Harrop R, Correa-Oliveira R, Wilson RA, Pena SD, Simpson AJ. Minilibraries constructed from cDNA generated by arbitrarily primed RT-PCR: An alternative to normalized libraries for the generation of ESTs from nanogram quantities of mRNA. Gene. 1997;186:135-42.
5. Durham AM, Kashiwabara AY, Matsunaga FT, Ahagon PH, Rainone F, Varuzza L, *et al*. 2005. EGene: A configurable pipeline generation system for automated sequence analysis. Bioinformatics. 2005;21:2812-3.
6. Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, *et al*. Using the transcriptome to annotate the genome. Nat Biotechnol. 2002;20:508-12.
7. Varuzza L, Pereira CAB. Significance test for comparing digital gene expression profiles: Partial likelihood application. Chilean Journal of Statistics 2010;1: 91-102. Source code available at http://code.google.com/p/kempbasu/.
8. Vêncio RZ, Varuzza L, Pereira CAB, Brentani H, Shmulevich I. Simcluster: Clustering enumeration gene expression data on the simplex space. BMC Bioinformatics. 2007;8:246.
9. Alexeyenko A, Tamas I, Liu G, Sonnhammer EL. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics. 2006;22:e9-15.
10. Ostlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, *et al*. InParanoid 7: New algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. 2010;38 (Database issue):D196-203.

Support

This work was supported by FAPESP (grant 03/14031-3). JN and APSM received a scholarship from CNPq and the work presented herein formed part of their Ph.D. theses. AG, AMBNM and AMD received Productivity Research fellowships from CNPq.

• • •