Universidade Estadual de Campinas (Unicamp)
Instituto de Matemática, Estatística e Computação Científica(IMECC)

Desenvolvimento de um sistema para análise de genômica, transcriptômica e proteômica através de redes de interação proteína-proteína

Estagiário: Lucas Miguel de Carvalho

Empresa: Associação Brasileira de Luz Síncrotron (ABTLuS)
Período: Janeiro a Julho de 2011
Função: Bioinformata

Agradecimentos

Primeiramente queria agradecer a oportunidade e confiança que a empresa ABTLuS proporcionou. Obrigado a todos os meus amigos e parentes que estiveram ao meu lado nas questões de dúvidas e aprendizagem. Agradeço ao Ramon Oliveira Vidal e Marcelo Falsarella Carazzolle, da equipe de bioinformática da ABTLuS, por ao longo do projeto, me proporcionarem um grande conhecimento na área de bioinformática e biologia molecular, ao professor Francisco Magalhães Gomes, do Instituto de Matemática, Estatística e Computação Científica (IMECC) da Universidade Estadual de Campinas (Unicamp), por se propor a aceitar a me orientar e ao Laboratório de Genoma e Expressão (LGE - Unicamp).

Não poderia deixar de agradecer a Universidade Estadual de Campinas (Unicamp) pelo grande conhecimento que pude obter desde a minha entrada na mesma, sempre com um ótimo nível de aprendizado com ótimos profissionais.

Introdução e ambiente de trabalho

A função do estágio é proporcionar ao profissional uma caracterização do ambiente da empresa, aonde ele aplica seus conhecimento adquirido ao longo do tempo na universidade. Esse período novo de aprendizado, pode lhe proporcionar melhorias na parte prática de sua carreira, não somente visando o uso da teoría em seus objetivos.

A empresa ABTLuS (Associação Brasileira de Luz Síncrotron) ofereceu essa melhoria na área de bioinformática, abrindo um novo laboratório nacional nesse segmento. Esta é uma área nova de atuação e investimento pelo governo brasileiro, crescendo a cada dia com os avanços tecnológicos e biológicos.

Como objeto de estudo, está sendo utilizado o genoma, transcriptoma e proteoma de três tipos do parasita humano Leishmania, com foco na *L. amazonensis*, um tipo de leishmania, ainda não estudado no Brasil, e que faz parte das doenças negligenciáveis no mundo.

Afim de uma melhor capacitação profissional, a empresa junto com o Laboratório de Genoma e Expressão (LGE – Unicamp) ofereceram um curso com ênfase em algoritmos e técnicas computacionais para montagem e análise de genomas (curso oferecido de 14 a 18 de fevereiro de 2011). No curso foi mostrado a importância das análises *in silico* de genomas nos dias de hoje. Com o uso de ferramentas da informática na análise e interpretação dos dados biológicos pode-se gerar novas hipóteses além de criar estratégia biotecnologicas tanto na área agrícola, quanto na medicina.

Entre os temas abordados no curso estavam o conhecimento em ambiente Linux, uso da linguagem Perl na bioinformática, noções de biologia molecular, tecnologias usadas para sequenciamento de genomas, algoritmos para alinhamento de sequências, montagem de genomas, identificação de genes e mineração de informações no genoma.

Um profissional que atua na área de exatas e se depara com aplicações biológicas no seu dia-a-dia necessita de um direcionamento específico de atuação na área da biologia, e com o curso oferecido, esse direcionamento foi possível.

O que é Bioinformática?

O desenvolvimento da engenharia genética e novas tecnologias da informação durante a última década do século XX, condicionou o aparecimento de uma disciplina que gerou *links* indissolúveis entre ciência da computação e biologia conhecido como bioinformática ou biologia computacional.

Ela está na intersecção das ciências da vida com a ciência da informação. É um campo científico interdisciplinar que visa a pesquisa e desenvolvimento de sistemas que facilitam a compreensão do fluxo de informações a partir de genes de estruturas moleculares, bioquímicos sua função, seu comportamento fisiológico e, finalmente, a sua influência sobre a doença e saúde (Martín Sánchez,1999).

Entre os principais fatores que contribuem para o desenvolvimento desta disciplina, é o grande volume de dados da seqüência gerada por projetos genoma diversas (humanos e as outras agências), as abordagens experimentais com base em novos *biochips* que permitem a obtenção de dados genéticos em alta velocidade, tanto os genomas individuais (mutações, polimorfismos), ou abordagens celular (expressão gênica), bem como o desenvolvimento da *Internet* e *World Wide Web*, que permite o acesso global a bancos de dados de informação biológica .

A bioinformática termo é relativamente novo, e apareceu na literatura no início de 1990, quando se começou a estruturar o assim chamado "Projeto Genoma Humano" e o Centro Nacional para Informação em Biotecnologia dos Estados Unidos deu os primeiros passos. No início, sua relação com a Informática Médica (disciplina que se preocupa com a análise e divulgação de dados médicos através da aplicação de computadores) foi devido apenas à semelhança sintática, bem como o uso de computadores por indispensável tanto disciplinas (Hagen JB,2000).

Uso da Linguagem Perl

Perl é uma linguagem *script* interpretada surgida em 1987 para uso na plataforma *Unix*, criado por Lary Wall. O Perl é a principal linguagem usada pelos bioinformatas hoje em dia, por causa da sua fácil manipulação e aplicação a HTML (plataformas CGI).

Quando o projeto genoma humano surgiu (Science,2001), o Perl foi a linguagem de programação mais usada, podendo manipular mais facilmente os dados obtidos. Notavelmente, o Perl é uma linguagem de programação de alta qualidade para manipulação de texto. Embora as ciências biológicas não envolvam uma boa dose de análise numérica agora, a maioria dos dados primários ainda são textos, como as anotações, comentários, referências bibliográficas. As sequências de DNA ainda estão em plataforma texto. Perl é poderosa com o uso de expressões regulares e operadores de manipulação de cadeia, simplificando o trabalho de uma forma inigualável por qualquer outra línguagem moderna.

Os dados biológicos estão muitas vezes incompletos, os campos podem estar faltando, ou os dados foram inseridos manualmente e não cabem no formato esperado, o Perl pode corrigir esses erros facilmente. Expressões regulares podem ser escritas para analizar e corrigir uma série de erros comuns na entrada de dados.

O Perl é orientada a componentes. Ele incentiva as pessoas a escrever o seu software em pequenos módulos, utilizando módulos de biblioteca Perl ou com a ferramenta orientada para *UNIX* clássica abordagem. Programas externos podem ser facilmente incorporadas em um *script* Perl.

Ele é uma linguagem simples e útil para *Web CGI scripting*, e está crescendo em importância, assim, mais laboratórios se voltam para a Web para publicar seus dados (Stein Lincoln D.,acesso em: 28/06/2011).

Em uma iniciativa de criar bibliotecas facilitando a manipulação de dados biológicos pelo Perl foi criado o *Bioperl.* O *Bioperl* é uma coleção de mais de 500 módulos *Perl* para bioinformática que tem sido escrito e mantido por um grupo internacional de voluntários. *Bioperl* é livre (sobre diretos autorais muito restritivos), e seu *website* é http://bioperl.org. Uma característica interessante sobre o *Bioperl* é que

ele é um projeto de código aberto, o que significa que os desenvolvedores interessados são convidados a contribuir, escrevendo código ou de outras formas, e os códigos estão disponíveis para qualquer pessoa interessada (Tisdall, James D,2003).

Método e desenvolvimento do projeto

Descrição do projeto

Em sua primeira parte, o projeto se direcionou em estudar um tipo de leishmania que causa um tipo de leishmaniose silenciosa (assintomática) no ser humano: a *Leishmania amazonensis.* Para isso, o genoma do parasita foi sequenciado na facility de sequenciamento da universidade da Carolina do Norte, toda a montagem do genoma e predição de genes foi realizada pelo Laboratório de Bioinformática do LNBio/ABTLuS. A partir desse ponto se deu o início da análise de dados, sendo ela feita por mim, gerando um *pipeline* de análise de interação do parasita/hospedeiro e redes de interação.

Ao longo do projeto se notou necessidade de estudar como o parasita interage com a célula do hospedeiro, assim, um estudo da interação proteína-proteína entre algumas proteínas do parasita e do ser humano verificaria a ação das proteínas de leishmania no genoma humano.

As interações proteína-proteína desempenham um papel crucial na execução de diversas funções biológicas. Assim, sua descrição completa contribuiria consideravelmente para a interpretação funcional de genomas completamente sequenciado, que são inundados com novos genes de funções imprevisível (Hazbun, T. R., 2001). Em uma interação proteína-proteína, os nós representam as proteínas e arestas representam uma interação física entre duas proteínas (Marc Vidal,2011).

Em sua segunda parte, a comparação da espécie assintomática (*L. amazonensis*) com outras com sintomas graves em relação a funções biológicas e toxidade das redes de interação, poderia fornecer detalhes do porque a ação da *Leishmania amazonensis* é menos impactante do que os outros tipos. Comparamos a *L. amazonensis* com a *L. major* e *L. infantum* (sendo essas duas tipos de leishmania com efeitos mais severos ao ser humano).

Descrição da doença

Leishmaniose se refere a um espectro de doenças parasitárias causadas por protozoários pertencentes ao gênero Leishmania. As doenças são classificadas como doenças tropicais negligênciadas segundo a Organização Mundial Organização Mundial da Saúde (OMS) e constituem um problema de saúde pública em muitos países do leste da África, no subcontinente indiano e América Latina. A Leishmaniose humana tem uma prevalência de 12 milhões de casos, com uma população estimada em 350 milhões em risco e uma incidência de 2 milhões de novos casos anualmente.

A leishmaniose apresenta um amplo espectro de manifestações clínicas que dependem da espécie do parasita, da resposta imune do hospedeiro e dos fatores ambientais (Murray HW,2005). Por exemplo, no Novo Mundo, a *Leishmania (Leishmania) amazonensis*, *Leishmania (Viannia) guyanensis* e *Leishmania (Viannia) braziliensis* são agentes causadores leishmaniose cutânea e mucocutânea enquanto *L. infantum* e *L. chagasi* são agentes etiológico da leishmaniose visceral americana (Murray HW,2005;Marzochi MC,1994).

Leishmania amazonensis é um modelo interessante de estudo pois está geralmente associada a infecções silenciosas. A detecção e registro desses casos de infecção por esta espécie estão relacionadas a episódios agudos de pessoas que, muitas vezes, foram infectadas meses ou anos atrás. As lesões, em geral, apresentam cura espontânea, exceto em imunocomprometidos.

O parasita, sob forma amastigota, abrigado em um vacúolo parasitóforo comum aos indivíduos da espécie, pode secretar proteínas para o lúmen acidificado do vacúolo. Essas proteínas podem ficar retidas nos vacúolos ou serem exportadas para o citoplasma (citosol) do macrófago (células humanas que fagocitam elementos estranhos ao corpo). Uma vez no citosol do macrófago, as proteínas secretadas pelo parasita podem ser transferidas para o núcleo do macrófago, controlando funções celulares diversas.

Método

As etapas de sequênciamento, montagem do genoma, predição e anotação dos genes mostraram que o genoma de *L. amazonensis* tem um tamanho aproximado de 30 Mb e possui 8100 genes, números muito semelhantes aos das outras Leishmanias já sequenciadas. *L. major* possui um genoma de aproximadamente 32.8 Mb com 8298 genes e o genoma de *L. infantum* possui aproximadamente 32.1 Mb com 8154 genes.

As sequências proteicas dos genes, foram submetidas ao programa TargetP (Olof Emanuelsson, 2007) que identifica a localização celular de uma proteína com base em sequências de amino ácido sinalizadoras de secreção, endereçamento para mitocôndria ou endereçamento para outros compartimentos intracelulares.

Apartir do conjunto de proteínas secretadas de L. amazonensis, L. infantum e L. major, foram feitos agrupamentos de ortólogos entre essas proteínas e o conjunto formado por todas as proteínas humanas. Essa análise foi realizada através do software OrthoMCL (Li Li, 2003) que é capaz de construir agrupamentos de ortólogos e parálogos entre vários organismos utilizando os valores de identidade entre as proteínas usando uma ferramenta computacional chamada BLAST (Altschul, S.F., 1990) e modelos de markov.

Os programas BLAST são ferramentas amplamente utilizadas para buscar sequências com similaridades em bases de dados de proteínas ou DNA. Para comparações de proteína, uma variedade de definições, refinamentos de algorítmos e estatísticas permite que o tempo de execução dos programas BLAST seja diminuído substancialmente, reforçando simultaneamente a sua sensibilidade à semelhanças.

Foram avaliados proteínas secretadas em comum e específica de cada espécie. As proteínas humanas que formam grupos de ortólogos com as proteínas secretadas de cada uma das espécies de leishmania foram avaliadas. A lista das proteínas humanas expressas em macrófago foi obtida através do banco de dados Gemma (www.bioinformatics.ubc.ca/Gemma) que é um repositório de análises de microarranjo de DNA.

As análises do interactoma foram realizadas através do software Ingenuity Pathway Analysis (Ingenuity Systems,Redwood City, CA) que conecta uma enorme

variedade dados experimentais permitindo uma visualização integrada através interações moleculares e químicas.

Com as proteínas codificadas, geramos redes de interação com no máximo 70 proteínas elegíveis. O Ingenuity Pathway Analysis gera as redes e as ordenam dependendo de seu *score.* O *score* é um valor numérico usado para classificar as redes de acordo com seu grau de relevância para as moléculas elegíveis rede em seu conjunto. O *score* leva em conta o número de moléculas elegíveis na rede e seu tamanho, bem como o número total de moléculas elegíveis analisadas e o número total de moléculas na base de dados que poderiam ser incluídos nas redes. O Índice de rede é baseada na distribuição hipergeométrica e é calculada com o teste exato de Fisher. Por exemplo, suponha que uma rede de 35 moléculas de Fisher e tem um resultado do teste exato de 1.10^{-6} .O score da rede é dado por -log(teste exato de Fisher), que neste caso, o score da rede valeria 6. Isso pode ser interpretado como: "Há um 1 chance em um milhão de começar uma rede com pelo menos o mesmo número de moléculas de rede elegíveis por acaso, quando escolher aleatoriamente 35 moléculas que podem estar em redes da base de dados do Ingenuity ".

O *score* não é uma indicação da qualidade ou relevância biológica da rede, ele simplesmente calcula o "encaixe" aproximado entre cada rede e suas moléculas elegíveis. A informação de expressão ou não em células de macrófagos para cada gene foi inserida na rede utilizando o banco de dados do Gemma (*vide* Apêndice A).

Resultados

A partir da formação dessas redes (*vide* Apêndice A) agrupamos todas as funções moleculares por espécie (Tabela 02) e depois listamos quais funções são comuns e exclusivas para cada espécie (Tabela 03). É interessante notar que funções básicas para a sobrevivência do parasita como metabolismo de carboidrato, transporte de moléculas e metabolismo de ácidos nucleicos são compartilhadas por todas as 3 espécies. São notadas redes com funções específicas em cada uma das espécies como por exemplo: desordem genética em *L. amazonensis*, resposta inflamatória e crescimento celular em *L. major* e degradação e síntese de proteínas por *L. infantum*.

Tabela 01. Secretoma de *L. amazonensis, L. major* e *L. infantum*

	L. amazonensis	L. major	L. infantum
Total de genes	341	435	415
Total de famílias ortólogas com o proteoma humano	86->129 (genes humanos)	98->143 (genes humanos)	91->129 (genes humanos)
Genes exclusivos	19	29	27
Famílias de genes ortólogos em humano em comum com L. amazonensis	-	26	37
Famílias de genes ortólogos em humano em comum com L. major	26	-	71

Tabela 02. Análise do interactoma. Funções das principais redes de cada espécie.

Espécie	#redes	Funções
L. major	1	Metabolismo de carboidratos(1), Transporte molecular(2), Metabolismo de ácido nucléico (3)
	2	Desenvolvimento do sistema hematológico (4), Resposta Antimicrobial (5), Câncer (6)
	3	Ploriferação e crescimento celular (7), Desenvolvimento do sistema hematológico (4), Resposta inflamatória (8)
	4	Metabolismo de amino ácido (9), Bioquímica de moléculas pequenas (10), Metabolismo de vitamina e minerais (11)
	5	Metabolismo de lipídio(12), Transporte molecular (2), Bioquímica de moléculas pequenas (10)
	6	Modificação pós-traducional (13), Morfologia celular (14), Função celular e manutenção (15)

L. amazonensis	1	Desenvolvimento celular (16), Desenvolvimento do sistema esquelético e muscular (17), Morfologia celular (14)
	2	Metabolismo de carboidrato (1), Transporte molecular (2), Metabolismo de ácido nucleico (3)
	3	Doença gastrointestinal (18), Disordem genética (19), Doença metabólica (20)
	4	Ciclo celular (21), Morfologia celular (14), Desenvolvimento celular (16)
	5	Função celular e manutenção (15), Transporte molecular (2), Funções do organismo (22)
L. infantum	1	Desenvolvimento do sistema hematológico (4), Resposta antimicrobial (5), Câncer (6)
	2	Metabolismo de carboidratos (1), Transporte molecular (2), Metabolismo de ácido nucleico (3)
	3	Modificação pós-traducional (13), Degradação de proteína (23), Síntese de proteína (24)
	4	Metabolismo de lipídio (12), Bioquímica de moléculas pequenas (10), Ciclo celular (21)
	5	Metabolismo de lipídio (12), Transporte molecular (18), Bioquímica de moléculas pequenas (10)

Tabela 03. Análise do interactoma. Espécies agrupadas por função em comum.

Funções	Espécies
(1) Metabolismo de carboidratos	major,amazonensis,infantum
(2) Transporte molecular	major,amazonensis,infantum
(3) Metabolismo de ácido nucleico	major,amazonensis,infantum
(4) Desenvolvimento do sistema hematológico	major, infantum

(5) Resposta antimicrobial	major, infantum
(6) Câncer	major, infantum
(7) Ploriferação e crescimento celular	major
(8) Resposta inflamatória	major
(9) Metabolismo de amino ácido	major
(10) Bioquímica de moléculas pequenas	major, infantum
(11) Metabolismo de vitamina e sais minerais	major
(12) Metabolismo de lipídio	major, infantum
(13) Modificação pós-traducional	major, infantum
(14) Morfologia celular	major, amazonensis
(15) Função celular e manutenção	major, amazonensis
(16) Desenvolvimento celular	amazonensis
(17) Desenvolvimento do sistema esquelético e muscular	amazonensis
(18) Doença gastroinstestinal	amazonensis
(19) Disordem genética	amazonensis
(20) Doença metabólica	amazonensis
(21) Ciclo celular	amazonensis, infantum
(22) Funções do organismo	amazonensis
(23) Degradação de proteína	infantum
(24) Síntese de proteína	infantum

Conclusões

Em primeiro momento, o estágio forneceu um suporte adequado ao conhecimento biológico, a área de bioinformática, e as implementações atuais de novas tecnologias para entendimento de doenças.

Este é o primeiro estudo comparativo da interação entre três espécies de leishmania e seu hospedeiro humano incluindo uma das espécie menos infectiva, a *L. amazonensis*. Nessa tentativa de compreender os mecanismos de infecção através das proteínas secretadas do parasita encontramos algumas pistas através da análise das redes de interação formadas pelas proteínas ortólogas em humano. Essa análise análise funcional, focada nos mecanismos de interação parasita *versus* hospedeiro pode levar a um maior intendimento sobre o desenvolvimento da doença.

O estudo de genes em vias específicas numa próxima etapa poderá fornecer novos genes alvos para o combate a doença focando em bloquear as perturbações no sistema celular humano. Também é previsto na próxima etapa do trabalho avaliar as interações entre os genes de leishmania que não tem ortólogos em humano através de análise de domínios e banco de dados públicos de interação como o iPFAM e o PSIMAP.

Apêndice A - Figuras

Redes de interação proteína-proteína usando o sofware Ingenuity:

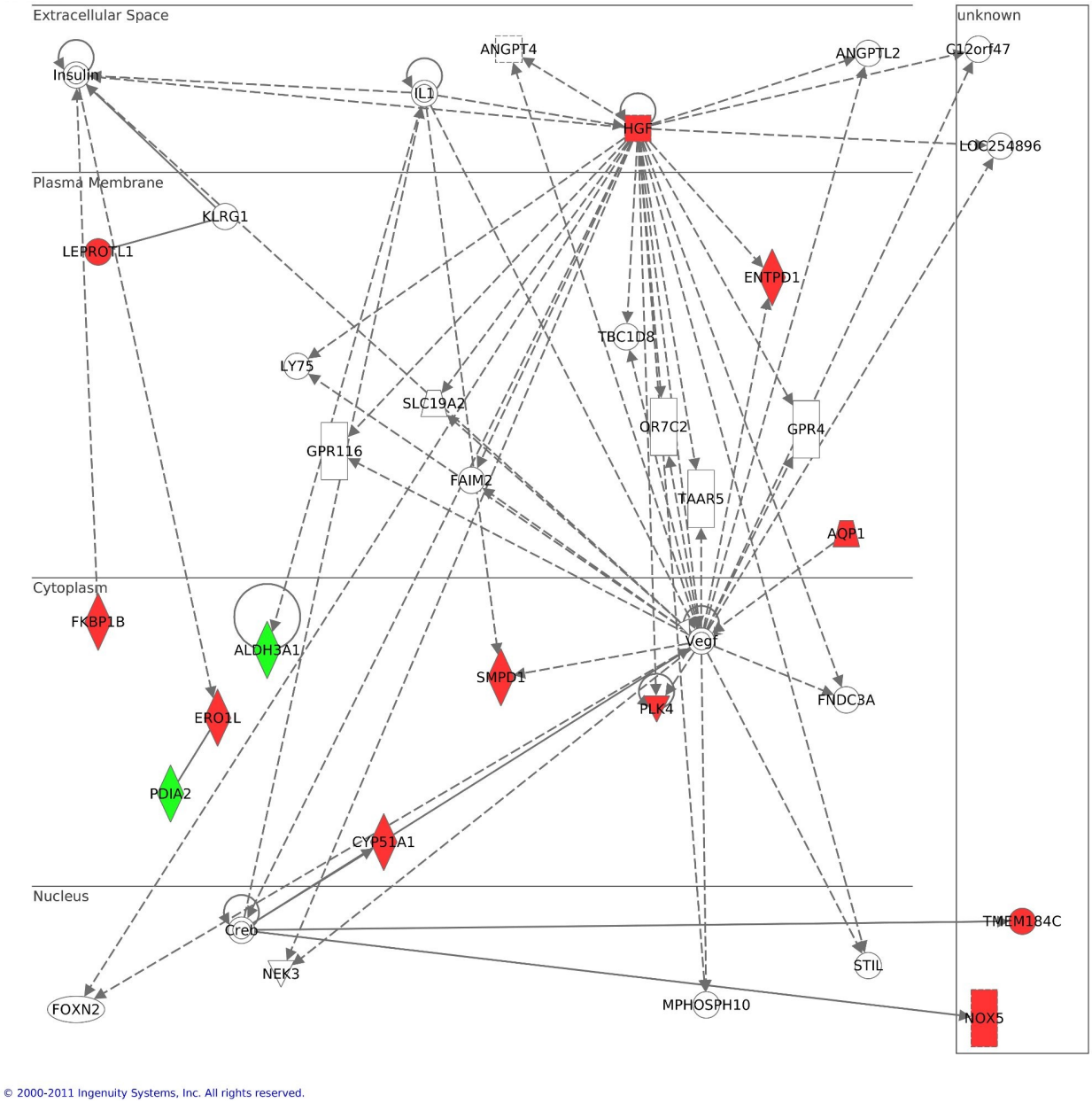

Figura 1 - Exemplo de rede de interação proteína-proteína de *Leishmania amazonensis* separadas por localização celular.

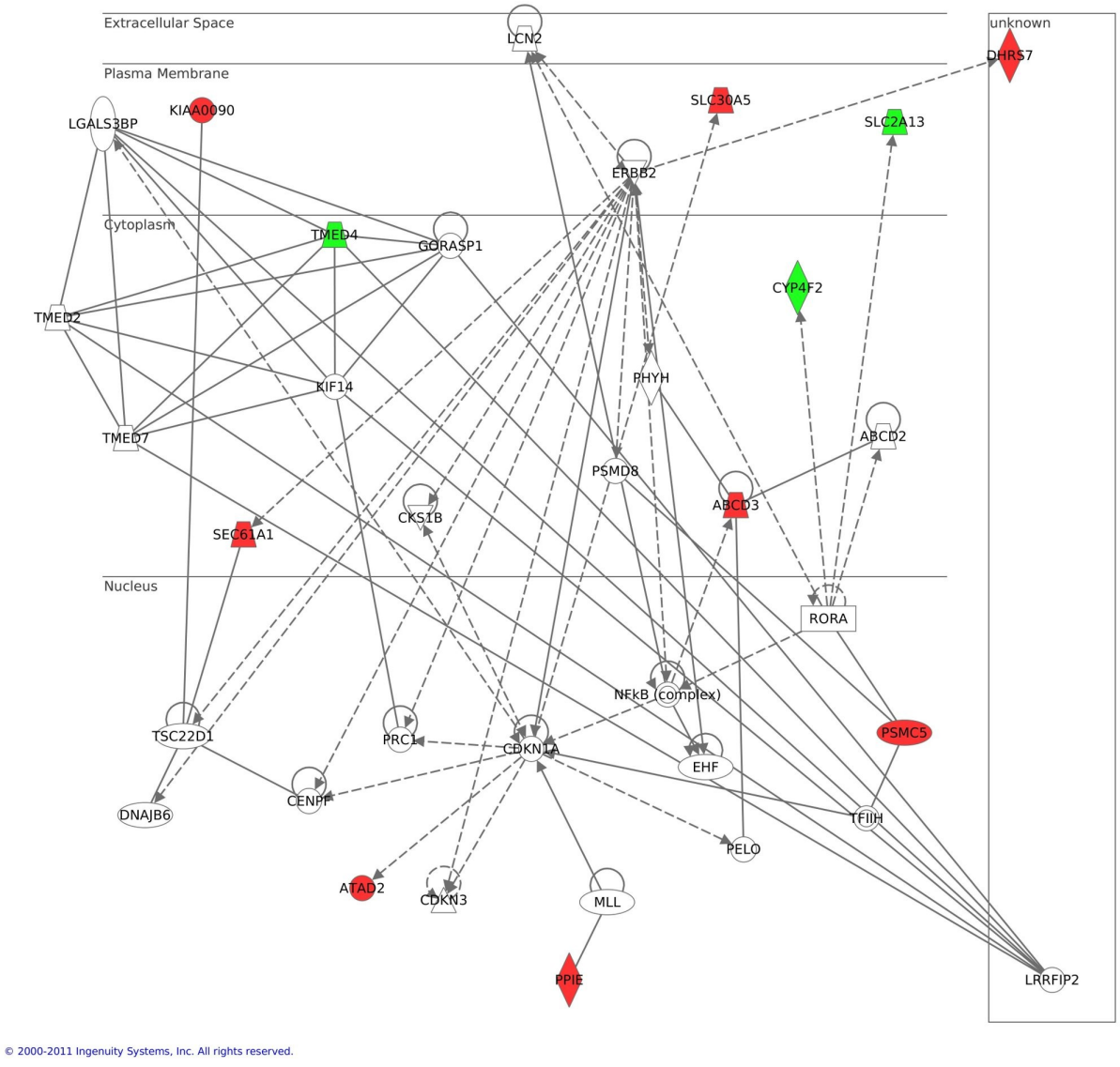

Figura 2 - Exemplo de rede de interação proteína-proteína de *Leishmania infantum* separadas por localização celular.

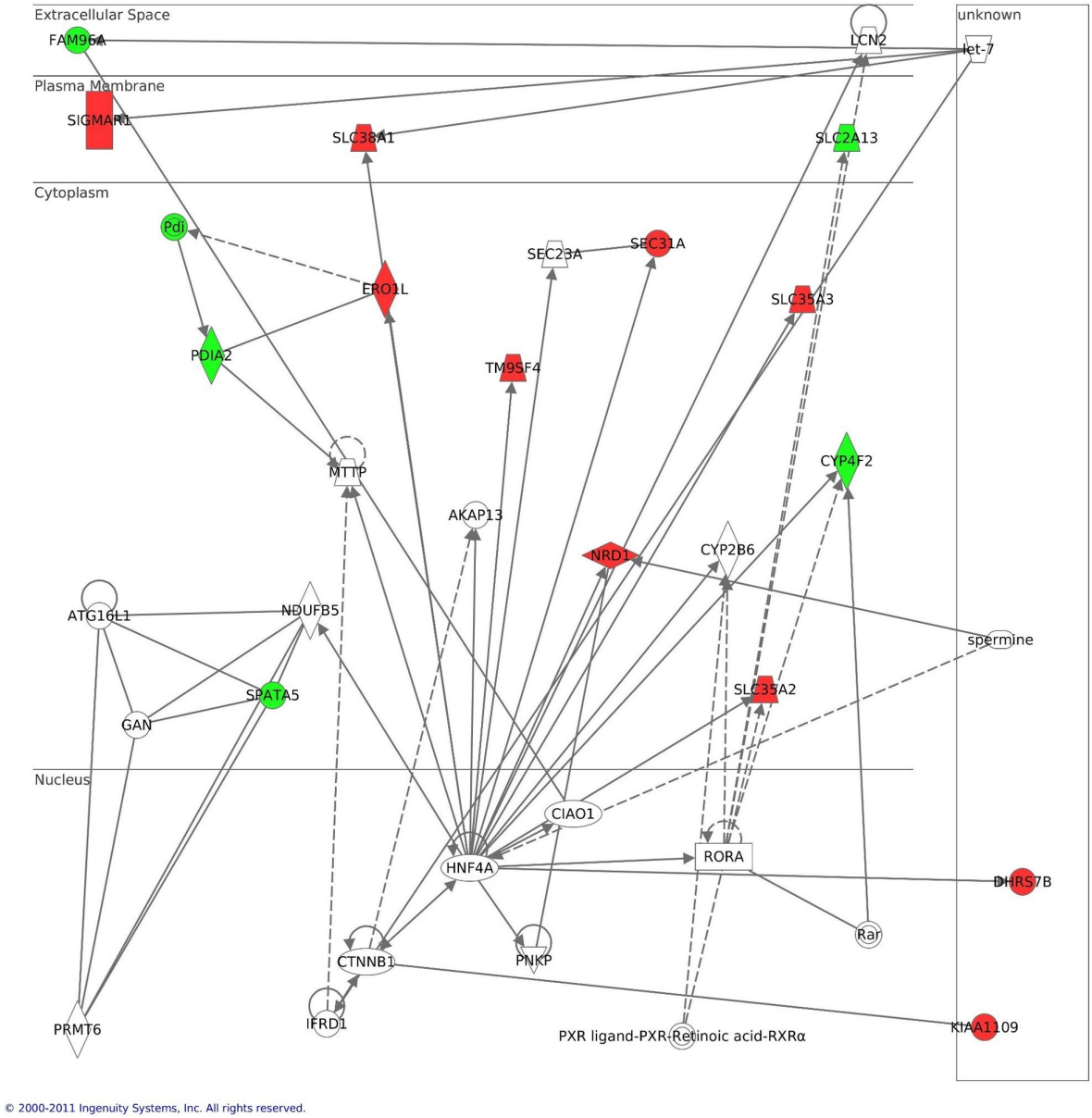

Figura 3 - Exemplo de rede de interação proteína-proteína de *Leishmania major* separadas por localização celular.

Referências Bibliográficas

1. Martín Sánchez F, López Campos G, Ibarrola de Andrés N. La convergencia de la informática médica y la bioinformática: impactos en la práctica clínica y en la educación de los profesionales de la salud. En: Servicios de Salud: ¿estrategias o tecnologías?. Madrid: Editorial Médica Panamericana, 1999.

2. Hagen JB. The origin of bioinformatics. Nat Rev Genet 2000;1(3):231-6.

3. Miller PL. Opportunities at the intersection of bioinformatics and health informatics: a case study. J Am Med Inform Assoc 2000;7(5):431-8.

4. Stein,*Lincoln D*. How Perl Saved the Human Genome Project. Disponível em:<http://www.foo.be/docs/tpj/issues/vol1_2/tpj0102-0001.html>. Acesso em: 28/06/2011.

5. Murray HW, Berman JD, Davies CR, Saravia NG (2005) Advances in leishmaniasis. Lancet 366: 1561–1577.

6. Marzochi MC, Marzochi KB (1994) Tegumentary and visceral leishmaniases in Brazil: emerging anthropozoonosis and possibilities for their control. Cad Saude Publica 10: Suppl 2359–375.

7. Olof Emanuelsson, Søren Brunak, Gunnar von Heijne, Henrik Nielsen, Locating proteins in the cell using TargetP, SignalP, and related tools, Nature Protocols 2, 953-971 (2007).

8. Li Li, Christian J. Stoeckert, Jr.,and David S. Roos, Genome Res. September 2003 13: 2178-2189.

9. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

10. Tisdall, James D.,Mastering Perl for Bioinformatics. 1ª,ed. U.S.A: O'Reilly,2003.

11. Hazbun, T. R., and Fields, S. (2001) Networking proteins in yeast. Proc. Natl. Acad. Sci. U. S. A. 98, 4277– 427.

12. Marc Vidal, Michael E. Cusick, Albert-László Barabási (2011). Interactome Networks and Human Disease.Cell. Vol. 144, Issue 6, pp. 986-998.

13. The Sequence of the Human Genome, Science 16 February 2001:Vol. 291 no. 5507 pp. 1304-1351