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

IIS - Integrated Interactome System : Uma plataforma web para anotação, análise e visualização de interação de proteínas-genes-drogas através da integração de diversas fontes de dados e ferramentas

Estagiário: Lucas Miguel de Carvalho

Empresa: Associação Brasileira de Luz Síncrotron (ABTLuS)

Período: Agosto a dezembro de 2011

Função: Bioinformata

Agradecimentos

Agradeço ao Marcelo Falsarella Carazzolle e a Gabriela Meirelles, da equipe de bioinformática da ABTLuS, por ao longo do projeto, me proporcionarem um grande conhecimento na área de bioinformática e bioquímica, 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.

Resumo:

O *Integrated Interactome System* (IIS), uma plataforma web baseada em *script* Perl com uma interface amigável para a anotação, análise e visualização de redes de interação entre proteínas. O IIS trabalha em três módulos conectados: (i) O módulo *Submission*, permite a submissão de arquivos gerados por diferentes experimentos de interação proteína-proteína (por exemplo, *two-hybrid*, IP acoplado ao MS/MS), (ii) O modulo *Gene Projects* que pode ser utilizado para analisar dados de *two-hybrid*, reúne os reads, constrói os contigs e singlets comparando-os contra várias bases de dados (por exemplo, Pfam, UniProtKB, HPRD, Gene Ontology, Gemma, DrugBank, entre outros), gerando tabelas e gráficos com anotações automáticas que podem ser curadas manualmente; (iii) o módulo *Interactome*, compara os contigs e singlets (ou simplesmente uma lista de proteínas de interesse) contra um banco de dados de interação proteína-proteína que está sendo implementado: Global Protein-Gene-Drug Interaction Network (GPGDIN, construído pelos bancos de dados BioGRID, Intact, DIP, String, MINT e HPRD). O último módulo produz uma lista de pares de interação que pode ser visualizado pelo Cytoscape, permitindo a visualização de todas as interações (ou subgrupos definidos de interações), análise e correlação das propriedades da rede utilizando as estatísticas de rede topológica (por exemplo, grau médio, coeficiente de clustering, comprimento do caminho característico, diâmetro, entre outros). Desenvolvemos a IIS através da integração de diversas bases de dados e ferramentas computacionais motivado pela necessidade de criar ferramentas adequadas para uma análise sistemática e padronizada de interações físicas, genéticas e químico-genéticas. Nossa plataforma de bioinformática está atualmente sendo validada em dados de humanos utilizando a técnica de two-hybrid, mas em breve será estendido para outros organismos e técnicas experimentais.

- Introdução:

Com o crescente aumento nas tecnologias experimentais de interações proteína-proteína (PPI) a área de biologia de sistemas ganha novas perspectivas possibilitando a realização de análises em larga escala dessas interações conectando-as com os dados experimentais de expressão gênica e mutantes. Estas análises têm levantado novas pistas sobre a função de diversas proteínas, ajudando a desvendar como as redes celulares estão organizadas e facilitando a validação de alvos terapêuticos e concepção de novos medicamentos.

No entanto, tal análise é difícil de ser realizada sem um adequado *pipeline* que agregue diferentes ferramentas computacionais integrando dados provenientes de diversas fontes.

As interações proteína-proteína (PPIs) constituem uma das condições mais importantes para sustentar a vida em organismos. Recentemente, muitos procedimentos experimentais têm sido desenvolvidos para ajudar a elucidar as complexas redes de PPIs que vão desde experimentos com base em análises genômicas escalares [1-4] até abordagens de biologia molecular em uma via específica [5-7]. Às vezes, os custos (financeiros e de pessoal) de tais abordagens experimentais exploratórias são proibitivos sendo que para contornar este problema as análises de bioinformática são freqüentemente utilizadas como um passo preliminar valioso para apontar alvos mais específicos, reduzindo custos e tempo.

Todas as informações analisadas sobre os experimentos de interação proteína-proteína estão muitas vezes disponíveis em diferentes bancos de dados públicos, sendo que alguns deles possuem ferramentas de buscas restritas a um conjunto de dados específicos. A integração desses bancos não é uma tarefa trivial, pois a informação está armazenada de forma não padronizada e com diferentes padrões de nomenclatura das proteínas.

As informações disponíveis nesses bancos de dados podem ser oriundas de metodologias experimentais: duplo híbrido, espectrometria de massa ou imunoprecipitação [9-14]; como também determinadas apenas por ferramentas de bioinformática [15-18]. Além disso, essas previsões raramente consideram a localização subcelular dos interagentes. A função de uma proteína é governada por sua

interação com outras proteínas dentro da célula, mas mesmo que duas proteínas são consistentemente prevista para interagir eles devem estar localizados no mesmo compartimento celular e ao mesmo tempo.

A criação da plataforma IIS (Integrated Interactome System), com seus diversos módulos integrados, tem a finalidade de facilitar a construção de rede de interação proteína-proteína através de buscas com a finalidade final de melhor visualização de PPIs, acompanha o interesse de melhorar a validação das interações entre proteinas/genes.

- Desenvolvimento:

Yeast two-hybrid system (Y2H) [25] é uma técnica de biologia molecular utilizada para descobrir interações proteina-proteina. Neste ensaio, uma proteína de interesse (isca) é fundida ao domínio de ligação de DNA de um fator de transcrição, enquanto outras proteínas (presas) são fundidas ao domínio de ativação. Se as proteínas da presa interagem com a isca, então há ativação dos genes que reportam essa interação. Assim, a seqüência de DNA das presas que estão interagindo são sequenciadas e analisadas.

Após o seqüenciamento das presas os cromatogramas vão para o primeiro módulo do IIS (o módulo *Submission*) onde são processados gerando os reads (arquivos no formato fasta). Estes reads precisam ser submetidos à análise de qualidade e identificação de regiões do vetor. Esta etapa é realizada pelo programa *BD Trimmer* (Bioinformatics 17 (2001), n. 122001, 1093-1104) que identifica regiões de baixa qualidade e vetores. No final do módulo *Submission,* a sequência já tratada é anotada comparando-a, utilizando o *BLASTx,* com bancos de dados de proteínas.

Já no segundo módulo (chamado *Gene Project*) é possível eliminar a redundância dos dados (no caso repetição do mesmo gene) através da clusterização dos reads formando os contigs e singlets. Também é possível realizar a anotação automática desses genes utilizando diversos bancos de dados biológicos (por exemplo: Pfam, UniProtKB, HPRD, Gene Ontology, Gemma,DrugBank, entre outros.), gerando tabelas que pode ser manualmente curadas.

No terceiro módulo (o módulo *Interactome*) os contigs e singlets são comparados contra o banco de dados denominado Global Protein-Gene-Drug Interaction Network (GPGDIN, constituído pelos bancos de dados BioGRID [23], Intact [21] , DIP [19] , String, MINT [22] e HPRD [24]) expandindo a lista dos pares de interação em torno desses genes. Este módulo está ligado ao Cytoscape [26], um software que permite a visualização de todas as interações (ou subgrupos definidos de interações) e a análise e correlação das propriedades da proteína com as estatísticas da rede topológica (por exemplo, grau médio, coeficiente de clusterização, comprimento do caminho característico, diâmetro, etc).

O GPGDIN é um banco de dados geral integrando todas as interações proteína-proteína descrito nos diversos bancos de dados públicos. Nesse banco de dados os pares de interação estão classificados por origem do dado (experimental ou teórica), organismos, artigo que publicou a interação e localização celular.

- Resultado e discussões:

O sistema encontra-se num estágio de desenvolvimento intermediário. As maiores dificuldades encontradas estão relacionadas com a padronização dos bancos de dados públicos (desenvolvimento do GPGDIN), pois os bancos de dados de interação proteína-proteína possuem nomenclaturas especificas de cada banco, e não são indexados de forma única, assim, prejudica a elaboração de um único *script* geral de integração de todos os bancos. O banco de dados uniref90, que agrupa proteínas com até 90% de identidade produzindo um ID único, foi escolhido como sendo o ID de referência para a unificação dos bancos.

Uma grande quantidade de redundâncias também foi percebida ao longo do desenvolvimento, esta redundância foi eliminada agrupando pares de interação com o artigo onde essa interação foi descrita (pubmed ID).

Um primeiro teste foi realizado com três proteínas escolhidas de *Leishmania amazonensis (*HSPA8, CANX e CTNBB1*),* estas proteínas foram escolhidas com base em trabalhos anteriores do grupo no qual a rede de interação foi gerada e analisada

pelo software Ingenuity. Abaixo temos as três redes de interação formadas pelo IIS e visualizadas pelo Cytoscape no qual é possível verificar que uma das proteínas se comporta como um *hub* (proteina que fazem muitas conexões na rede de interação).

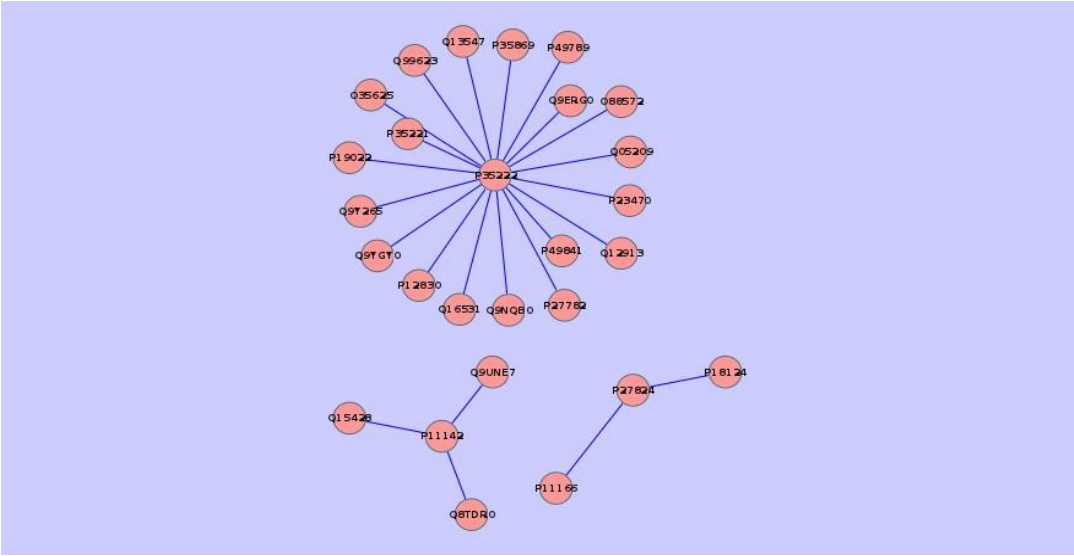

Figura 1. Rede de interação de HSPA8, CANX e CTNBB1

Os pares de interação construídos por essa metodologia podem estar sujeitos a diversos erros e precisam passar por uma etapa de validação. De forma a obter um resultado mais confiável alguns cuidados precisam ser considerados, Proteínas que realmente interagem espera-se que compartilhem do mesmo compartimento celular e que tenham parceiros de interação comuns [28-32].

. Um passo final para a validação da interação entre duas proteínas foi inspirado no artigo de Marcelo M. Brandão (2009) [27], em que consiste aplicar um método numérico que depende do tipo de interação entre elas, sua co-localização e sua localização subcelular.

Os bancos PPIs apresentam essas três características citadas acima, denota por C^3 (*Celular Compartment Classification*), que divide o PPI em 4 classes de acordo com a sua evidência de interação e localização subcelular. O índice de confiabilidade para um PPI relatado pode ser postulada em termos da proporção de parceiros de interação que duas proteínas têm em comum. Para isso está sendo implementado o

cálculo de “peso de similaridade funcional” (FSW) para todas as interações de primeiro nível presente nos bancos de dados.

Duas abordagens matemáticas relacionadas, CD-distance [33] e FSWeight [28], têm sido propostos para avaliar a confiabilidade dos dados de proteína de interação com base no número de vizinhos comuns de duas proteínas. Ambos foram inicialmente projetados para prever funções de proteínas, e ultimamente tem se mostrado um bom desempenho para avaliar a confiabilidade das interações proteína [31]. Wong [34] têm demonstrado que o uso de FSWeight, que estima a força da associação funcional, para retirar as interações não confiável (FSWeight baixo) melhora o desempenho de algoritmos de *clustering*.

Os pares de interação de proteínas que são classificados com alto *score* por este método são susceptíveis de serem verdadeiros positivos. Por outro lado, os pares de proteínas que são classificadas com baixo *score* são susceptíveis de serem falsos positivos. A característica mais interessante do CD-distance e FSWeight é que eles são capazes de classificar a confiabilidade de uma interação entre um par de proteínas utilizando apenas a topologia das interações entre esse par de proteínas e seus vizinhos num raio curto em um grafo de rede [29,35].

O algoritmo FSWeight foi originalmente proposto por Chua [28]. O índice de similaridade de peso funcional em um par de proteínas A e B em um gráfico de interação $(FSW_{A,B})$ é definido como:

$$FSW_{A,B} = \left(\frac{2|N_A \cap N_B|}{|N_A - N_B| + 2|N_A \cap N_B| + \lambda_{A,B}}\right) \times \left(\frac{2|N_A \cap N_B|}{|N_B - N_A| + 2|N_A \cap N_B| + \lambda_{B,A}}\right),$$

Onde,

N_A = conjunto de parceiros de interação de A; N_B= conjunto de parceiros de interação da B; $\lambda_{A,B}$é um peso para penalizar pesos semelhança entre pares de proteína quando qualquer uma das proteínas tem muito poucos parceiros interagindo e é calculado como:

$$\lambda_{A,B} = \max\left(0, N_{avg} - \left(|N_A - N_B| + |N_a \bigcap N_B|\right)\right),$$

Onde,

N_{AVG}= Média de interações feitas por cada proteína.

Para uma próxima etapa final, com este cálculo podemos anotar se duas proteínas compartilham de uma mesma função, e validar a sua interação proposta pelos bancos de dados. Será validado uma interação com FSW maior que 0,2, baseando-se em dados obtidos em [27].

Conclusão:

Tendo em vista a necessidade de entender as interações entre proteínas, o IIS está sendo desenvolvido para facilitar a construção e visualização sendo destinado tanto a novos pesquisadores como aos mais especializados. Sua integração com métodos de validação o deixa mais apto a gerar conexões experimentais válidas, para uma melhor análise sistemática e padronizada de interações físicas, genéticas e quimio-genéticas. O uso computacional para o desenvolvimento do IIS possibilitou um maior uso de dados de PPIs, já que os mesmo em sua forma bruta são difíceis de serem manipulados. Devido ao grande aumento no número de pesquisas sobre bancos de interações, editar um *script* para inserir novas informações fica mais simples, devido a gasto computacionais. Obter um valor numérico de cada conexão gerada pelo IIS atraveś de cálculos matemáticos em uma próxima etapa final poderá fornecer uma validação mais consistente das conexões entre duas proteínas já fornecendo sua ontologia, localização celular e experimentos usados em sua validação teórica ou experimental.

Referências Bibliográficas

1. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, et al.: A protein interaction map of Drosophila melanogaster.Science 2003, 302:1727-1736.

2. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, et al.: A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.Nature 2000, 403:623-627.

3. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, et al.: A human protein-protein interaction network: a resource for annotating the proteome.Cell 2005, 122:957-968.

4. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, et al.: A map of the interactome network of the metazoan C. elegans.Science 2004, 303:540-543.

5. Kormish JD, Sinner D, Zorn AM: Interactions between SOX factors and Wnt/beta-catenin signaling in development and disease.Dev Dyn 2009, 239(1):56-68.

6. Kang HG, Klessig DF: The involvement of the Arabidopsis CRT1 ATPase family in disease resistance protein-mediated signaling.Plant Signal Behav 2008, 3:689-690.

7. Lacatus G, Sunter G: The Arabidopsis PEAPOD2 transcription factor interacts with geminivirus AL2 protein and the coat protein promoter.Virology 2009, 392(2):196-202.

8. Kaiser J: Proteomics. Public-private group maps out initiatives.Science 2002, 296:827.

9. March-Diaz R, Garcia-Dominguez M, Florencio FJ, Reyes JC: SEF, a new protein required for flowering repression in Arabidopsis, interacts with PIE1 and ARP6. Plant Physiol 2007, 143:893-901.

10. Dortay H, Gruhn N, Pfeifer A, Schwerdtner M, Schmulling T, Heyl A: Toward an interaction map of the two-component signaling pathway of Arabidopsis thaliana. J Proteome Res 2008, 7:3649-3660.

11. Dortay H, Mehnert N, Burkle L, Schmulling T, Heyl A: Analysis of protein interactions within the cytokinin-signaling pathway of Arabidopsis thaliana.Febs J 2006, 273:4631-4644.

12. Dray E, Siaud N, Dubois E, Doutriaux MP: Interaction between Arabidopsis Brca2 and its partners Rad51, Dmc1, and Dss1.Plant Physiol 2006, 140:1059-1069.

13. Marrocco K, Zhou Y, Bury E, Dieterle M, Funk M, Genschik P, Krenz M, Stolpe T, Kretsch T: Functional analysis of EID1, an F-box protein involved in phytochrome A-dependent light signal transduction.Plant J 2006, 45:423-438.

14. Ciruela F: Fluorescence-based methods in the study of protein-protein interactions in living cells.Curr Opin Biotechnol 2008, 19:338-343.

15. De Bodt S, Proost S, Vandepoele K, Rouze P, Peer Y: Predicting protein-protein interactions in Arabidopsis thaliana through integration of orthology, gene ontology and co-expression.BMC Genomics 2009, 10:288.

16. Geisler-Lee J, O'Toole N, Ammar R, Provart NJ, Millar AH, Geisler M: A predicted interactome for Arabidopsis.Plant Physiol 2007, 145:317-329.

17. Lin M, Hu B, Chen L, Sun P, Fan Y, Wu P, Chen X: Computational identification of potential molecular interactions in Arabidopsis.Plant Physiol 2009, 151:34-46.

18. Mosca R, Pons C, Fernandez-Recio J, Aloy P: Pushing structural information into the yeast interactome by high-throughput protein docking experiments.PLoS Comput Biol 2009, 5:e1000490.

19. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The database of interacting proteins: 2004 update.Nucleic Acids Res., 32, D449–D451.

20. Peri,S., Navarro,J.D., Amanchy,R., Kristiansen,T.Z., Jonnalagadda,C.K., Surendranath,V., Niranjan,V., Muthusamy,B., Gandhi,T.K., Gronborg,M. et al. (2003) Development of human protein referencedatabase as an initial platform for approaching systems biology in humans. Genome Res., 13, 2363–2371.

21. Hermjakob,H., Montecchi-Palazzi,L., Lewington,C., Mudali,S.,Kerrien,S., Orchard,S., Vingron,M., Roechert,B., Roepstorff,P.,Valencia,A. et al. (2004) IntAct: an open source molecular interaction database. Nucleic Acids Res., 32, D452–D455.

22. Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G.,Helmer-Citterich,M. and Cesareni,G. (2002) MINT: a Molecularn INTeraction database. FEBS Lett., 513, 135–140.

23. Stark C,Breitkreutz BJ,Reguly T,Boucher L,Breitkreutz A,Tyers M.BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006;34:D535-D539.

24. Mishra G. R.,Suresh M.,Kumaran K.,Kannabiran N.,Suresh S.,Bala P.,Shivakumar .,Anuradha N.,Reddy R.,Raghavan T. M.,Menon S.,Hanumanthu G.,Gupta M.,Upendran S.,Gupta S.,Mahesh M.,Jacob B.,Mathew P.,Chatterjee P.,Arun K. S.,Sharma S.,Chandrika K. N.,Deshpande N.,Palvankar K.,Raghavnath R.,Krishnakanth R.,Karathia H.,Rekha B.,Nayak R.,Vishnupriya G.,Kumar H. G.,Nagini .,Kumar G. S.,Jose R.,Deepthi P.,Mohan S. S.,Gandhi T. K.,Harsha H. C.,Deshpande . .,Sarker M.,Prasad T. S.,Pandey A.(2006) Human protein reference database—2006 update. Nucleic Acids Res. 34, D411– D414.

25. Chien, C. T.; Bartel, P. L.; Sternglanz, R.; Fields, S. The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. Proc. Natl. Acad. Sci. U.S.A. 1991,88, 9578.

26. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498–2504.

27. Brandão MM, Dantas LL, Silva-Filho MC. 2009. AtPIN: Arabidopsis thaliana protein interaction network. BMC Bioinformatics 10: 454.

28. Chua HN, Sung WK, Wong L: Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions.Bioinformatics 2006, 22:1623-1630.

29. Chen J, Hsu W, Lee ML, Ng SK: Increasing confidence of protein interactomes using network topological metrics.Bioinformatics 2006, 22:1998-2004.

30. Gerstein M, Lan N, Jansen R: Proteomics. Integrating interactomes. Science 2002, 295:284-287.

31. Liu G, Wong L, Chua HN: Complex discovery from weighted PPI networks. Bioinformatics 2009, 25:1891-1897.

32. Chua HN, Ning K, Sung WK, Leong HW, Wong L: Using indirect protein-protein interactions for protein complex prediction.J Bioinform Comput Biol 2008, 6:435-466.

33. Brun C, Chevenet F, Martin D, Wojcik J, Guenoche A, Jacq B: Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network.Genome Biol 2003, 5:R6.

34. Wong L: Constructing More Reliable Protein-Protein Interaction Maps. International Symposium on Computational Biology & Bioinformatics; 17-19 January 2008; University of Kerala 2008, 284-297.

35. Chen J, Chua HN, Hsu W, Lee M-L, Ng S-K, Saito R, Sung W-K, Wong L: Increasing confidence of protein-protein interactomes.17th International Conference on Genome Informatics; Yokohama, Japan 2006, 284-297.