HUMBOLDT-UNIVERSITÄT ZU BERLIN

Diploma Thesis Exposé

Comparing semantically enriched experimental protein networks in colorectal cancer

Author: Marc N. Bux

Tutors: Prof. Dr. Ulf Leser
Philippe Thomas

1 Motivation

With more than one million cases on a global scale and a mortality rate of approximately 33% in the developed world, colorectal cancer (CRC) remains one of the major causes of death in human [1]. Even though numerous improvements have been made to CRC therapy, the overall knowledge on the disease remains poor, resulting in a lack of predictability of treatment benefit and disease outcome. This makes further research all the more crucial considering CRC being the cancer with the second most frequent lethal outcome.
Due to a lack of reliable protein and genetic markers, prognostication is currently limited to mostly clinical staging and pathological analysis of cancer tissue [2]. In order to improve reliability of disease prognosis and treatment prediction, it would be desirable to find new biomarkers that are easily detectable and have a high predictive value. Ideally, the predictions of these markers should also have a sound biological explanation.
The most important example of such a marker in CRC is the KRAS mutation in exon 2 which serves as a negative predictive marker for treatment with epidermal growth factor receptor (EGFR) specific antibodies *cetuximab* [3] and *panitumumab* [4]. Unfortunately, finding new biomarkers like KRAS fulfilling above mentioned criteria is a daunting task. It requires the researcher to have comprehensive knowledge on the molecular structures and processes involved in the disease.
However, as biomedical knowledge is expanding at an exponential rate [5] and the interconnectedness of molecular systems is being revealed, it is becoming more and more difficult for highly specialized researchers to keep track of the broad spectrum of publications on a given research topic. These continuing trends result in an increasing interest in bioinformatic approaches to refine, summarize and process the vast amounts of data.

2 Aim

The aim of this diploma thesis is the development of a method that detects and showcases differences between several protein-protein-interaction (PPI) networks derived from high-throughput experimental data. The purpose of this method is to aid researchers in deriving and asserting hypotheses on molecular interactions that might be common or distinct in different manifestations of CRC. Ultimately, by affecting the design of future experiments, this could lead to the discovery of new prognostic or predictive biomarkers.
We will develop the method described above using KRAS as an example. To this end, two weighted protein networks - so-called "data networks" - for CRC cell lines with positive and negative KRAS mutation status are constructed by computing gene co-expression profiles. In order to reduce noise and include findings from biomedical literature, these networks are then integrated into a major protein network, called the "knowledge network". This network is assembled from a set of publicly available protein databases and enriched through Text Mining applications. Integration of protein networks is performed by mapping proteins on their unique Uniprot identifier and combining edge weights using a probability combination function.
Subsequently, proteins in both integrated networks are ranked using centrality analysis. By comparing the resulting lists of proteins, regions of interest containing major differences between protein ranks are located. These regions of interest are then ranked once again and can be visualized in such a way that allows researchers easy orientation and possibly new hints on the mechanics of the disease.
The idea behind this procedure is that diseases with similar phenotypes are likely to be the consequence of mutations in identical or functionally related genes [6]. Finding similar or different regions of interest in protein networks of CRC cell lines might therefore shed some light in the molecular mechanisms of the disease. This could reveal potential markers or therapeutic targets of the disease. Since complex disorders like cancer cannot be sufficiently described as a list of involved genes, a network-based approach seems a lot more promising to identify potential subnetwork markers [7].

3 Approach

3.1 Experimental Data Networks

A gene expression microarray consists of an array of spots, each containing a large number of single-stranded oligonucleotides (e.g. sections of genes) specific to the spot, so-called probes. In a microarray experiment, mRNA sequences are extracted from a sample and converted to fluorophore-, silver-, or chemiluminescence-labeled cDNA targets. Labeled cDNA targets hybridized with probes are then detected and counted in order to measure the expression levels of genes in the sample.
In this diploma thesis, high-throughput data is employed from a DNA microarray experiment which is conducted for CRC cell lines with positive and negative KRAS mutation status. Gene expression is measured at different points in time. Genes are filtered by variance in expression over time, leaving only genes with a significant expression profile for further examination.
In order to capture co-expression between genes, correlation between remaining gene expression profiles is then determined within investigated cell lines. For genes showing significant positive or negative correlation, gene co-expression is assumed with a probability accord-

ing to their respective correlation coefficient. Based on gene co-expression identified this way weighted protein networks are built for proteins and interactions corresponding to co-expressed genes. This way, two experimental data networks are assembled, one each for KRAS+ and KRAS- cell lines.

3.2 Literature-curated Knowledge Network

The data networks only contain particular sets of assumed protein interactions based on observed gene co-expression profiles. In order to reduce noise and semantically enrich experimental CRC data, a much larger knowledge network has to be consulted. This knowledge network is assembled from major public literature-curated PPI databases: DIP [8] IntAct [9] Mammalian MIPS [10] HPRD [11], MINT [12] and BioGrid [13].
Since these databases are mostly annotated by hand, they do not contain all the protein interactions discovered and published in literature to date. Additional PPIs are therefore added to the knowledge network using a Text Mining approach. Gene names are identified and normalized to their Entrez Gene identifiers using GNAT[1], a named entity recognition framework for biomedical applications [14]. Protein interactions between the retrieved genes are then predicted with jSRE[2], a tool for relation extraction through supervised machine learning techniques [15].
It has recently been reported that literature-curated databases are not as reliable as commonly perceived [16]. Protein interactions retrieved via Text Mining are expected to contain a significant number of false positives as well. Therefore, consensus reliability estimates are computed as proposed in [17] in order to gauge the reliability of data sources used to build the knowledge network.
PPIs are then assigned a weight in the form of a confidence value by combining the reliability values of sources reporting the given PPI. This could be employed using the Noisy-OR function $w = 1 - \prod_i (1 - r_i)$ where r_i is the reliability of a source i. The Noisy-Or function, as displayed in figure 1, has the desirable property that the confidence of an interaction is high when it is mentioned in at least one reliable source yet increases with additional evidence. This is especially relevant as it reduces the probability of false negatives, which are difficult to identify in biologics.

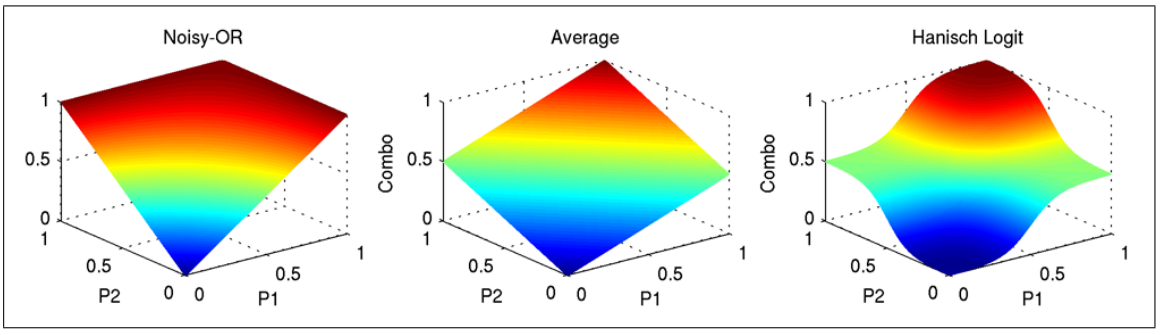

Figure 1: Combination functions favoring agreement among sources. Image taken from [18].

[1] http://cbioc.eas.asu.edu/gnat/
[2] http://hlt.fbk.eu/en/technology/jSRE

3.3 Data Integration

It would be desirable to examine the experimental protein networks in the light of existing biomedical knowledge available from literature. Both data networks are therefore independently integrated into the knowledge network. Integration is performed using the Hanisch Logit probability combination function, as discussed in [18] and illustrated in figure 1. Specifically, this is accomplished by computing the average of the logistics functions of edge weights $logit = \frac{logistic(e_i)+logistic(e_j)}{2}$, where $logistic(x) = \frac{1}{1-e^{-s(x-v)}}$. As suggested in [19] the parameter v could be set to the mean of the distributions and s could be set to $\frac{6}{v}$ to allow for a moderate slope.
The Logit combination function features a sinusoidal curve and requires an edge to be present in both networks in order for it to be given a score above 0.5 in the combined network. Also, the combination function is expected to favor protein interactions that are strong in experimental data networks, yet are at least moderately present in the knowledge network. This is largely due to the distribution of correlation being expected to lean more towards 1 than the distribution of combined reliability estimates, as observed in [18].
The relevance of the resulting integrated protein networks to CRC can be evaluated by implementing a similar approach as in [20]. Specifically, established disease proteins related to CRC can be used as seeds and expanded using any of the protein networks. In the sub-network created this way, new candidate disease proteins can be predicted through centrality analysis. By performing leave-one-out cross-validation, CRC-specific relevance of the investigated protein network can be measured as the average rediscovery rate of known disease proteins.

3.4 Network Comparison

After experimental data networks have been integrated into the global knowledge network, protein nodes can be ranked within integrated networks by applying a centrality measure. Centrality analysis could be conducted by performing an iterative random walk on graphs, as described in [21], or by employing the PageRank Algorithm [22], as realized in [20]. This results in a ranked list of proteins for each of the integrated data networks with focal proteins being assigned a higher rank.
The rationale behind this procedure is that genes associated with a particular phenotype or disease are not randomly scattered in the protein network. Instead, they have been observed to cluster together and occur in central network positions, as reported in [23]. Therefore, protein nodes featuring high centrality values are likely relevant to the specific disease.
The aim of this work is to locate regions of interest in protein networks featuring major differences in protein interactions of KRAS+ and KRAS- CRC. Therefore, the two ranked lists of proteins have to be compared to one another. For each protein, the corresponding ranks in both integrated networks are contrasted with one another. This results in another ranked list of proteins capturing the differences of centrality in KRAS+ and KRAS- networks.
The researcher can then browse this ranked list, recognize entities well-known for being distinctive of KRAS+ or KRAS- CRC and encounter unexpected findings. When discovering a particularly interesting entity, its surroundings can be visualized as subgraph of the average of KRAS+ and KRAS- integrated networks in Cytoscape [24]. Visualization allows the analyst orientation in their well-understood biological domain and investigation of unexpected findings. This explorative approach might help in the assertion and generation of novel hypotheses on the mechanics of CRC on the molecular level.

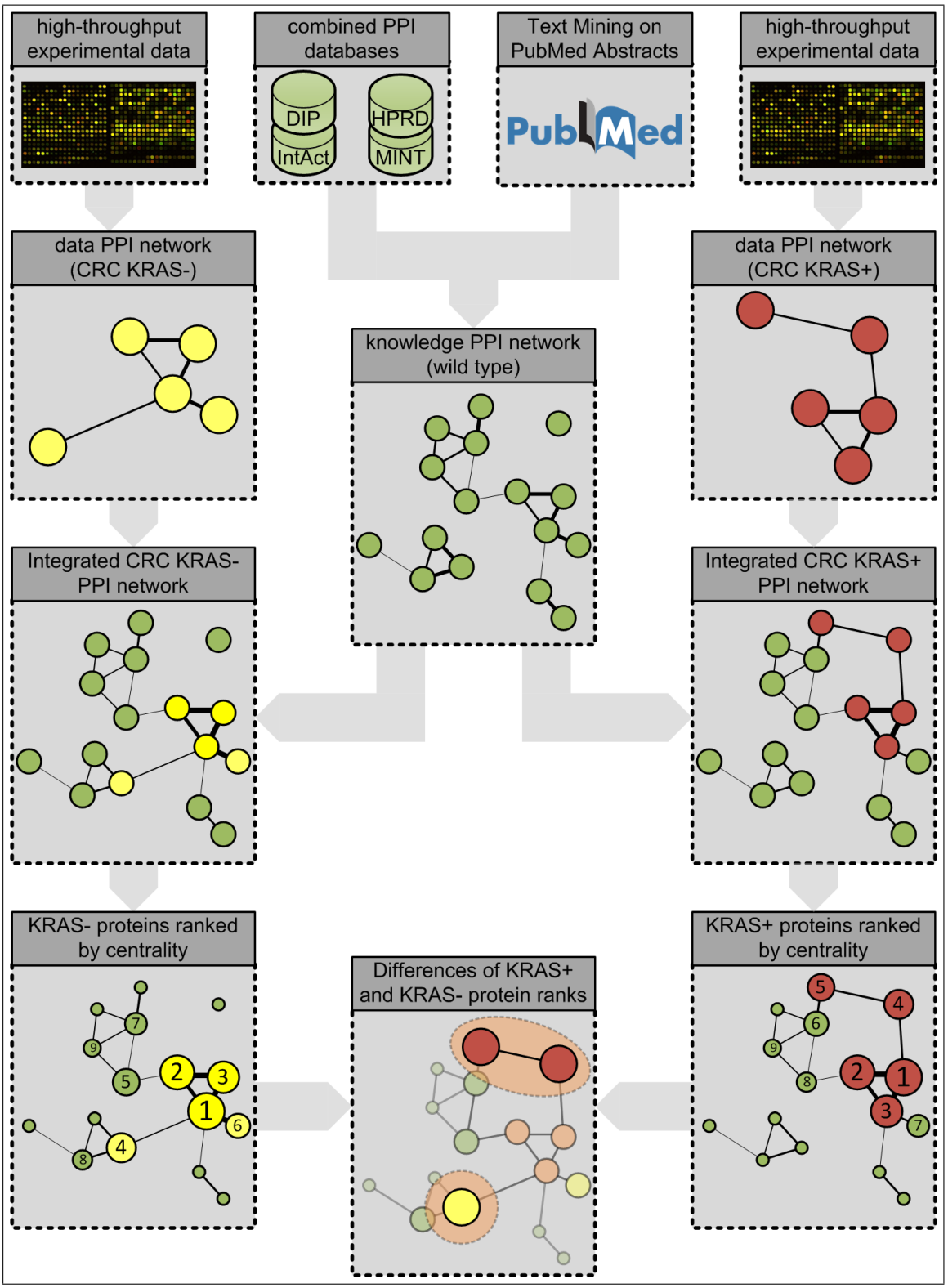

Figure 2: The complete process of building the protein networks and ranking proteins by centrality.

4 Expected Results

At the very minimum, the methods described above should be able to replicate knowledge on KRAS mutation and pathways. As an example, the KRAS signaling cascade is worth mentioning. It is depicted as part of the EGFR signaling pathway in figure 3 and described in [2].
It is expected, that once the introduced methods have been shown to successfully rediscover established knowledge, they will be able to aid scientists in generation and consolidation of hypotheses. This might influence the design of new experiments and, ultimately, the discovery of new CRC-specific biomarkers.
Additionally, the proposed methodology is not limited to the investigation of KRAS+ and KRAS- cell lines. Instead, once abovementioned processes have been validated to benefit CRC research, they can be applied to various other medical fields. For example, protein networks of different cancer cell lines could be compared to the same cell lines treated with certain medications.

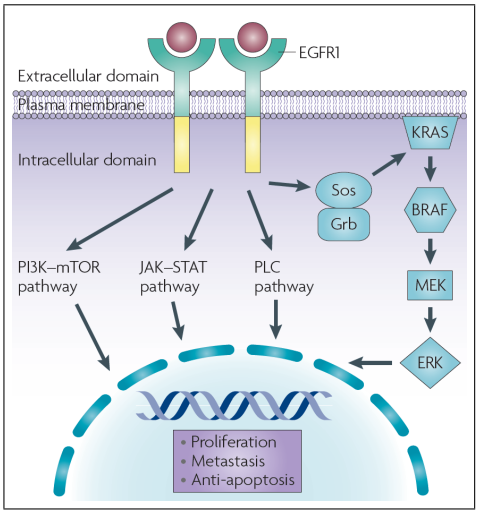

Figure 3: The epidermal growth factor receptor signaling pathway. Image taken from [2].

References

[1] Brian M Wolpin and Robert J Mayer. Systemic treatment of colorectal cancer. *Gastroenterology*, 134(5):1296–1310, May 2008. doi: 10.1053/j.gastro.2008.02.098. URL `http://dx.doi.org/10.1053/j.gastro.2008.02.098`.

[2] Axel Walther, Elaine Johnstone, Charles Swanton, Rachel Midgley, Ian Tomlinson, and David Kerr. Genetic prognostic and predictive markers in colorectal cancer. *Nat Rev Cancer*, 9(7):489–499, Jul 2009. URL `http://dx.doi.org/10.1038/nrc2645`.

[3] Christos S Karapetis, Shirin Khambata-Ford, Derek J Jonker, Chris J O'Callaghan, Dongsheng Tu, Niall C Tebbutt, R. John Simes, Haji Chalchal, Jeremy D Shapiro, Sonia Robitaille, Timothy J Price, Lois Shepherd, Heather-Jane Au, Christiane Langer, Malcolm J Moore, and John R Zalcberg. K-ras mutations and benefit from cetuximab

in advanced colorectal cancer. *N Engl J Med*, 359(17):1757–1765, Oct 2008. URL http://dx.doi.org/10.1056/NEJMoa0804385.

[4] Rafael G Amado, Michael Wolf, Marc Peeters, Eric Van Cutsem, Salvatore Siena, Daniel J Freeman, Todd Juan, Robert Sikorski, Sid Suggs, Robert Radinsky, Scott D Patterson, and David D Chang. Wild-type kras is required for panitumumab efficacy in patients with metastatic colorectal cancer. *J Clin Oncol*, 26(10):1626–1634, Apr 2008. URL http://dx.doi.org/10.1200/JCO.2007.14.7116.

[5] Lawrence Hunter and K. Bretonnel Cohen. Biomedical language processing: what's beyond pubmed? *Mol Cell*, 21(5):589–594, Mar 2006. URL http://dx.doi.org/10.1016/j.molcel.2006.02.012.

[6] Anais Baudot, Gonzalo Gomez-Lopez, and Alfonso Valencia. Translational disease interpretation with molecular networks. *Genome Biol*, 10(6):221, 2009. URL http://dx.doi.org/10.1186/gb-2009-10-6-221.

[7] Han-Yu Chuang, Eunjung Lee, Yu-Tsueng Liu, Doheon Lee, and Trey Ideker. Network-based classification of breast cancer metastasis. *Mol Syst Biol*, 3:140, 2007. URL http://dx.doi.org/10.1038/msb4100180.

[8] Lukasz Salwinski, Christopher S Miller, Adam J Smith, Frank K Pettit, James U Bowie, and David Eisenberg. The database of interacting proteins: 2004 update. *Nucleic Acids Res*, 32(Database issue):D449–D451, Jan 2004. URL http://dx.doi.org/10.1093/nar/gkh086.

[9] S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, C. Kohler, J. Khadake, C. Leroy, A. Liban, C. Lieftink, L. Montecchi-Palazzi, S. Orchard, J. Risse, K. Robbe, B. Roechert, D. Thorneycroft, Y. Zhang, R. Apweiler, and H. Hermjakob. Intact–open source resource for molecular interaction data. *Nucleic Acids Res*, 35(Database issue):D561–D565, Jan 2007. URL http://dx.doi.org/10.1093/nar/gkl958.

[10] Philipp Pagel, Stefan Kovac, Matthias Oesterheld, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Goar Frishman, Corinna Montrone, Pekka Mark, Volker Stmpflen, Hans-Werner Mewes, Andreas Ruepp, and Dmitrij Frishman. The mips mammalian protein-protein interaction database. *Bioinformatics*, 21(6):832–834, Mar 2005. URL http://dx.doi.org/10.1093/bioinformatics/bti115.

[11] T. S. Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, Lavanya Balakrishnan, Arivusudar Marimuthu, Sutopa Banerjee, Devi S Somanathan, Aimy Sebastian, Sandhya Rani, Somak Ray, C. J. Harrys Kishore, Sashi Kanth, Mukhtar Ahmed, Manoj K Kashyap, Riaz Mohmood, Y. L. Ramachandra, V. Krishna, B. Abdul Rahiman, Sujatha Mohan, Prathibha Ranganathan, Subhashri Ramabadran, Raghothama Chaerkady, and Akhilesh Pandey. Human protein reference database–2009 update. *Nucleic Acids Res*, 37(Database issue):D767–D772, Jan 2009. URL http://dx.doi.org/10.1093/nar/gkn892.

[12] Andrew Chatr-Aryamontri, Arnaud Ceol, Luisa Montecchi Palazzi, Giuliano Nardelli, Maria Victoria Schneider, Luisa Castagnoli, and Gianni Cesareni. Mint: the molecular interaction database. *Nucleic Acids Res*, 35(Database issue):D572–D574, Jan 2007. URL http://dx.doi.org/10.1093/nar/gkl950.

[13] Bobby-Joe Breitkreutz, Chris Stark, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, Michael Livstone, Rose Oughtred, Daniel H Lackner, Jrg Bhler, Valerie Wood, Kara Dolinski, and Mike Tyers. The biogrid interaction database: 2008 update. *Nucleic Acids Res*, 36(Database issue):D637–D640, Jan 2008. URL http://dx.doi.org/10.1093/nar/gkm1001.

[14] Joerg Hakenberg, Conrad Plake, Loic Royer, Hendrik Strobelt, Ulf Leser, and Michael Schroeder. Gene mention normalization and interaction extraction with context models and sentence motifs. *Genome Biol*, 9 Suppl 2:S14, 2008. URL http://dx.doi.org/10.1186/gb-2008-9-s2-s14.

[15] Claudio Giuliano, Alberto Lavelli, and Lorenza Romano. Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proc. of the 11st Conf. of the European Chapter of the Association for Computational Linguistics (EACL'06)*, Trento, Italy, 2006. The Association for Computer Linguistics. ISBN 1-932432-59-0. URL http://acl.ldc.upenn.edu/E/E06/E06-1051.pdf.

[16] Michael E Cusick, Haiyuan Yu, Alex Smolyar, Kavitha Venkatesan, Anne-Ruxandra Carvunis, Nicolas Simonis, Jean-Francois Rual, Heather Borick, Pascal Braun, Matija Dreze, Jean Vandenhaute, Mary Galli, Junshi Yazaki, David E Hill, Joseph R Ecker, Frederick P Roth, and Marc Vidal. Literature-curated protein interaction datasets. *Nat Methods*, 6(1):39–46, Jan 2009. URL http://dx.doi.org/10.1038/nmeth.1284.

[17] Sonia Leach, Aaron Gabow, Lawrence Hunter, and Debra S Goldberg. Assessing and combining reliability of protein interaction sources. *Pac Symp Biocomput*, pages 433–444, 2007. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2517251/pdf/nihms58727.pdf.

[18] Sonia M Leach, Hannah Tipney, Weiguo Feng, William A Baumgartner, Priyanka Kasliwal, Ronald P Schuyler, Trevor Williams, Richard A Spritz, and Lawrence Hunter. Biomedical discovery acceleration, with applications to craniofacial development. *PLoS Comput Biol*, 5(3):e1000215, Mar 2009. URL http://dx.doi.org/10.1371/journal.pcbi.1000215.

[19] Daniel Hanisch, Alexander Zien, Ralf Zimmer, and Thomas Lengauer. Co-clustering of biological networks and gene expression data. *Bioinformatics*, 18 Suppl 1:S145–S154, 2002. URL http://bioinformatics.oxfordjournals.org/content/18/suppl_1/S145.full.pdf.

[20] Samira Jaeger, Gokhan Ertaylan, David van Dijk, Ulf Leser, and Peter Sloot. Inference of surface membrane factors of hiv-1 infection through functional interaction networks. *PLoS One*, 5(10):e13139, 2010. URL http://dx.doi.org/10.1371/journal.pone.0013139.

[21] Sebastian Koehler, Sebastian Bauer, Denise Horn, and Peter N Robinson. Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet*, 82(4):949–958, Apr 2008. URL http://dx.doi.org/10.1016/j.ajhg.2008.02.013.

[22] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Computer Networks and ISDN Systems*, volume 30, pages 107–117, 1998. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.42.3243.

[23] Trey Ideker and Roded Sharan. Protein networks in disease. *Genome Res*, 18(4):644–652, Apr 2008. URL http://dx.doi.org/10.1101/gr.071852.107.

[24] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–2504, Nov 2003. URL http://dx.doi.org/10.1101/gr.1239303.