INITIAL RESULTS ON KNOWLEDGE DISCOVERY AND DECISION SUPPORT FOR INTRACRANIAL ANEURYSMS

Christoph M. Friedrich, Martin Hofmann-Apitius
Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Department of Bioinformatics
Schloss Birlinghoven 53754 Sankt Augustin Germany

Robert Dunlop
Infermed Ltd, 25 Bedford Square, London, UK

Ioannis Chronakis
Department of Engineering, Oxford University, UK, Department of Academic Oncology, UCL, UK

Miriam C. J. M. Sturkenboom, Roelof Risselada
Erasmus MC, Medical Informatics, 3000 CA Rotterdam, Netherlands

Baldo Oliva, Ferran Sanz
Research Unit on Biomedical Informatics (GRIB) IMIM/UPF, C/Dr. Aiguader, 88, 08003 - Barcelona, Spain

Keywords: Knowledge Discovery, Decision Support, Intracranial Aneurysm.

Abstract: Intracranial Aneurysms are bulbous expansions of the intracranial vessels, that may rupture and lead to subarachnoid haemorrhage, which can result in severe disability or death of the affected person. The prediction of the individual rupture risk of a patient based on information from images, haemodynamic simulations, clinical parameters and genetic markers is one of the aims of the European Integrated Project @neurIST. The predicted rupture risk is meant to support decision making on clinical treatment. We will present initial results on Knowledge Discovery through a combination of text-mining, data integration from public bioinformatics data sources, and database mining. Additionally, we provide first results for decision support through knowledge based clinical guidelines and Bayesian networks.

1 INTRODUCTION

The advent of improved medical imaging facilities and their routine use in clinical practice increases the number of accidentally detected asymptomatic Intracranial Aneurysms (IA). Intracranial Aneurysms are bulbous expansions of the intracranial vessels, that may rupture and lead to intracranial bleeding (subarachnoid haemorrhage), which can result in severe disability or death of the affected person. In (Rinkel et al., 1998) the prevalence of the disease for adults without risk factors for subarachnoid haemorrhage is reported with approximately 2 % and the annual risk rate for a rupture with 0.7 %. This relatively high prevalence with low incidence of the dangerous event leads to controversial discussions on treatment decisions. In general there are three treatment options:

1. do not treat the asymptomatic aneurysm with low risk
2. conduct neurosurgical clipping
3. deploy a platinum coil via endovascular intervention

One of the targets of the European Integrated Project @neurIST[1] is to support decision making on IA treatment options by building a distributed environment for healthcare. This environment will allow access to patient related information from images,

[1] http://www.aneurist.org

M. Friedrich C., Hofmann-Apitius M., Dunlop R., Chronakis I., C. J. M. Sturkenboom M., Risselada R., Oliva B. and Sanz F. (2008).
INITIAL RESULTS ON KNOWLEDGE DISCOVERY AND DECISION SUPPORT FOR INTRACRANIAL ANEURYSMS.
In *Proceedings of the First International Conference on Health Informatics*, pages 265-272

haemodynamic simulations, clinical parameters, genetic markers and epidemiological data.

Additionally, a set of application suites will be developed, that are based on this infrastructure and directly support the goal of improving clinical decision making. A draft architecture of the distributed system is described in (Arbona et al., 2007). Considering the target of this paper, two application suites are of interest, @neuRisk, a decision support system based on clinical guidelines and @neuLink, a research oriented application targeted at linking genetics to disease. The @neuLink suite supports Knowledge Discovery for the detection of genetic risk factors.

2 INITIAL RESULTS

In the following we will give examples and preliminary results of Decision Support and Knowledge Discovery that have been developed during the first year of the project.

2.1 Decision Support based on the Pro*forma* Language and the REACT Application

In this section we will describe the work carried out in the development of the first @neuRisk prototype. This version of the prototype employs a mixed quantitative and qualitative approach to provide risk assessment and decision support in the treatment of cerebral aneurysms. The knowledge base for the approach is derived from two trials: The International Study on Unruptured Intracranial Aneurysms (ISUIA) (Wiebers et al., 2003) and International Subarachnoid Aneurysm Trial (ISAT) (Molyneux et al., 2005). For demonstration purposes, some additional test data were also included to show how future research results could be incorporated to the final @neuRisk suite.

Qualitative decision support in @neuRisk has been implemented using the PRO*forma* method and tools (Sutton and Fox, 2003) while quantitative decision support was implemented by adapting an existing treatment planning application called REACT (Risk, Events, Actions and their Consequences over Time) (Glasspool et al., 2006).

PRO*forma* is a well established technology, first published in 1996 (Fox et al., 1996) and described in detail in 2000 (Fox and Das, 2000; Sutton and Fox, 2003), is a well established clinical decision support technology, that has been tested in a number of trials with promising results (Fox et al., 2006; Hurt et al., 2003; Patkar et al., 2006). There are two major implementations available, the Tallis implementation from Cancer Research UK[2] and the Arezzo implementation by InferMed Ltd in London[3]. REACT technology is based on PRO*forma* concepts and has been tested in one trial in the area of genetic counselling with encouraging results (Glasspool et al., 2006).

In the qualitative part of the prototype, we used the PRO*forma* language (Sutton and Fox, 2003) to model both the workflow involved in patient management and a set of treatment decisions. The resulted computer-executable guideline application is then enacted by the Tallis PRO*forma* engine. It guides the user through the workflow, provides a set of data capture services to collect data from the various disciplines (clinician, radiologist, geneticist, etc) and finally, it offers support for the treatment decisions. It does this by offering a set of logical arguments (rules) to the clinicians which either support or oppose each of the available treatment actions. The system suggests the most appropriate action but the final decision is taken by the clinician (Fox et al., 2006).

In the quantitative part of the prototype, we used an adapted version of the REACT tool (Glasspool et al., 2006). This tool provides support for planning the treatment of the patient based on the effect that each treatment action has on the risk. The REACT user interface is divided into 4 major parts:

1. The treatment plan,
2. The graph area,
3. The argumentation area and
4. The notification area.

The treatment plan area provides the clinician with a set of available treatment options and a timeline where she/he can schedule them, similar to a Gantt chart. As the user adds or removes events from the treatment plan, the graph area plots the expected consequences in real time (in the @neuRisk prototype these are life expectancy and years gained or lost if an aneurysm were left untreated). This allows the user to explore the space of available options with immediate feedback of the various interactions and consequences. Information directly relevant to each option

[2]Tallis is an implementation of the PRO*forma* engine written in Java which was developed at Cancer Research UK by the Advanced Computation Laboratory (`http://acl.icnet.uk/`) that was led by Prof John Fox. The engine is supported by a suite of tools including an engine, an authoring tool, a tester tool and a web based enacting application (`http://www.cossac.org/tallis.html`).

[3]Arezzo is an implementation of the PRO*forma* engine created by InferMed (`http://www.infermed.com`) and it has been used in a number of commercial products.

pert work, specialisation on one topic and possible selection bias. As an alternative to this manual extraction, we consider text-mining (Jensen et al., 2006). This helps to get an overview on genes possibly involved in a disease and to find potential new genes from publications. We implemented a Find Candidate Genes module in the @neuLink application suite. This part of the application suite is based on two text-mining systems, ProMiner (Hanisch et al., 2005) and OSIRIS (Bonis et al., 2006). ProMiner finds entities (Genes/Proteins, Drugnames, Chromosomal Locations ...) and links them to unique database identifiers, e.g. EntrezGene (Maglott et al., 2005). OSIRIS finds and disambiguates mentions of genetic variations in text to dbSNP (Smigielski et al., 2000) identifiers with a query-expansion approach.

To support the focussed view of the user on relevant information in the disease context, we developed a ranking mechanism based on Relative Entropy (Kullback and Leibler, 1951), also known as Kullback/Leibler divergence. In this ranking mechanism, we use the complete MedLine as reference corpus and contrast it with the specific corpus derived from full-text search.

Finding and disambiguating variation mentions in text with the OSIRIS system, needs a high-quality gene-mention machinery. We therefore combined our text-mining tools and complemented them with a machine learning variation finding engine based on Conditional Random Fields (CRF) (Lafferty et al., 2001). The improved results of this approach have been described in (Klinger et al., 2007).

One of the crucial questions of all Discovery methods is their validation. For the finding and disambiguation of gene mentions in text, we have been able to get an independent assessment of the performance of our approach by participating in the BioCreaTive II assessment (Morgan and Hirschmann, 2007). Our ProMiner system assessed as described in (Fluck et al., 2007), has been ranked 3rd of 21 submissions.

In a second evaluation, we tested wether our system, given the keyword search "intracranial AND aneurysm*", was able to detect the same related suspectibility genes, that have been found by human experts. The review on genetics (Krischek and Inoue, 2006) mentions 18 associated genes in the context of Intracranial Aneurysms. In our evaluation (as of 2007-10-01) (Gattermayer, 2007), we find 16,548 documents in PubMed related to the keyword and 596 documents, that mention 316 different genes/proteins. We find and could disambiguate all 18 genes in publications and rank them to the first 238 hits with 7 hits among the top 16 candidates. See figure 3 for a screenshot of the interface. Among the high-ranked false positives we find frequently used therapeutic proteins like the plasminogen activator (PLAT), but also new true positives like the JAG1 gene, that have not been mentioned in the genetic reviews.

2.5 Generating Protein-Protein Interaction Networks

We used "Protein ineractions and network analysis" (PIANA) (Aragues et al., 2006) to combine data from the Database of Interacting Proteins (DIP) (Salwinski et al., 2004), the MIPS database of interactions (Pagel et al., 2005), the Molecular INTeractions database (MINT) (Chatr-aryamontri et al., 2007), IntAct (Kerrien et al., 2007), the Biomolecular Interactions Database (BIND) (Alfarano et al., 2005), the BioGrid (Stark et al., 2006) and Human Protein Reference Database (HPRD) (Peri et al., 2003) and the human interactions from two recent high-throughput experiments (Rual et al., 2005),(Stelzl et al., 2005). We also provide the interactions obtained from STRING (von Mering et al., 2005) and methods of protein-protein interaction prediction based on sequence/structure patterns (Espadaler et al., 2005), (Cockell et al., 2007). The integration of many different sources of interactions into a single repository allowed us to work with an extensive set of 363,571 interactions between 42,040 different protein sequences.

PIANA represents protein interaction data as a network where the nodes are proteins and the edges interactions between them. In such a network, a set of proteins linked to protein p_j (i.e. physically interacting with p_j) is named "partners of p_j". PIANA builds the network by retrieving partners for a initial set of seed proteins (i.e. the relevant proteins, here referred as "seed proteins") that were obtained from the Find Candidate Genes module in section 2.4. A network is generated for the set of proteins that contains them and their partners. In this network, a protein that is connected to more than one "seed" is referred as a linker-N, with N being the number of seed proteins to which it is connected. Finally, proteins only connected to one seed protein are named leafs. This allowed us to enlarge the interaction network and detect new putatively relevant proteins for the biological pathway.

3 CONCLUSIONS

We have presented initial results on Decision Support, Database Mining and Knowledge Discovery for Intracranial Aneurysms. Due to the lack of patient data,

biomedical data and compute services. *IEEE Trans Nanobioscience*, 6(2):136–141.

Bonis, J., Furlong, L. I., and Sanz, F. (2006). OSIRIS: a tool for retrieving literature about sequence variants. *Bioinformatics*, pages 2567–2569.

Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L., and Cesareni, G. (2007). Mint: the molecular interaction database. *Nucleic Acids Res*, 35(Database issue):D572–D574.

Clarke, M. (2007). Cochrane literature review snapshot v2.0 - aneurysm analysis rupture risk review. Deliverable of the @neurIST project D18, University of Oxford.

Cockell, S. J., Oliva, B., and Jackson, R. M. (2007). Structure-based evaluation of in silico predictions of protein-protein interactions using comparative docking. *Bioinformatics*, 23(5):573–581.

Druzdzel, M. J. (1999). GeNIe: A development environment for graphical decision-analytic models. In *Proceedings of the 1999 Annual Symposium of the American Medical Informatics Association (AMIA-1999)*, page 1206.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

Espadaler, J., Romero-Isart, O., Jackson, R. M., and Oliva, B. (2005). Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships. *Bioinformatics*, 21(16):3360–3368.

Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. In *Knowledge Discovery and Data Mining*, pages 82–88.

Fluck, J., Mevissen, H. T., Dach, H., Oster, M., and Hofmann-Apitius, M. (2007). ProMiner: recognition of human gene and protein names using regularly updated dictionaries. In (Hirschmann et al., 2007), pages 149–151.

Fox, J. and Das, S. (2000). Safe and sound: Artificial intelligence in hazardous applications. In *Proceedings of AAAI 2000*.

Fox, J., Johns, N., Rahmanzadeh, A., and Thomsen, R. (1996). PROforma: A method and language for specifying clinical guidelines and protocols. In Brender, J., Christensen, J., Scherrer, J.-R., and McNair, P., editors, *Proceedings of the Medical Informatics Europe 1996*, pages 516–520. IOS Press.

Fox, J., Patkar, V., and Thomson, R. (2006). Decision support for health care: the proforma evidence base. *Inform Prim Care*, 14(1):49–54.

Gattermayer, T. (2007). SCAIView: annotation and visualization system for knowledge discovery. Master's thesis, Life Science Informatics at Bonn-Aachen International Center for Information Technology (B-IT); Germany.

Ghinea, N. and van Gelder, J. M. (2004). A probabilistic and interactive decision-analysis system for unruptured intracranial aneurysms. *Neurosurg Focus*, 17(5):E9.

Glasspool, D. W., Fox, J., Oettinger, A., and Smith-Spark, J. H. (2006). Argumentation in decision support for medical care planning for patients and clinicians. In *Proceedings of the AAAI Spring Symposium Series 2006*.

Han, B., Friedrich, C. M., Manthey, R., and Hofmann-Apitius, M. (2006). Bayesian network modeling for the prediction of treatment outcomes of intracranial aneurysms. In Huson, D., Kohlbacher, O., Lupas, A., Nieselt, K., and Zell, A., editors, *Short Papers and Poster Abstracts of the German Conference on Bioinformatics (GCB2006)*, page 122.

Hanisch, D., Fundel, K., Mevissen, H.-T., Zimmer, R., and Fluck, J. (2005). ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics*, 6 (Suppl 1)(S14).

Hirschmann, L., Krallinger, M., and Valencia, A., editors (2007). *Proceedings of the Second BioCreative Challenge Evaluation Workshop*. Centro Nacional de Investigaciones Oncologicas, CNIO.

Hurt, C., Fox, J., Bury, J., and Saha, V. (2003). Computerised advice on drug dosage decisions in childhood leukaemia: a method and a safety strategy. In Dojat, M., Keravnou, E., and Barahona, P., editors, *Proceedings of the 9th Conference on Artificial Intelligence in Medicine in Europe (AIME'03)*, LNAI 2780, pages 158–163. Springer Verlag.

Jensen, L. J., Saric, J., and Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews; Genetics*, 7:119–129.

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

Klinger, R., Furlong, L. I., Friedrich, C. M., Mevissen, H. T., Fluck, J., Sanz, F., and Hofmann-Apitius, M. (2007). Identifying gene specific variants in biomedical text. *Journal of Bioinformatics and Computational Biology*, 5(6):in print.

Krischek, B. and Inoue, I. (2006). The genetics of intracranial aneurysms. *J Hum Genet*, 51(7):587–594.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Ann. Math. Stat.*, 22:79–86.

Lafferty, J., McCallum, A., and Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 282–289. Morgan Kaufmann Publishers.

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2005). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 0:D1–D6.

Molyneux, A. J., Kerr, R. S. C., Yu, L., Clarke, M., Sneade, M., Yarnold, J. A., and Sandercock, P. (2005). International subarachnoid aneurysm trial (ISAT) of neurosurgical clipping versus endovascular coiling in 2143 patients with ruptured intracranial aneurysms: a randomised comparison of effects on survival, dependency, seizures, rebleeding, subgroups, and aneurysm occlusion. *Lancet*, 366:809–817.

Morgan, A. A. and Hirschmann, L. (2007). Overview of BioCreative II gene normalization. In (Hirschmann et al., 2007), pages 17–27.

Nahed, B. V., Bydon, M., Ozturk, A. K., Bilguvar, K., Bayrakli, F., and Gunel, M. (2007). Genetics of intracranial aneurysms. *Neurosurgery*, 60(2):213–25; discussion 225–6.

Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stümpflen, V., Mewes, H.-W., Ruepp, A., and Frishman, D. (2005). The mips mammalian protein-protein interaction database. *Bioinformatics*, 21(6):832–834.

Patkar, V., Hurt, C., Steele, R., Love, S., Purushotham, A., Williams, M., Thomson, R., and Fox, J. (2006). Evidence-based guidelines and decision support services: A discussion and evaluation in triple assessment of suspected breast cancer. *Br J Cancer*, 95(11):1490–1496.

Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T. K. B., Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H. N., Rashmi, B. P., Ramya, M. A., Zhao, Z., Chandrika, K. N., Padma, N., Harsha, H. C., Yatish, A. J., Kavitha, M. P., Menezes, M., Choudhury, D. R., Suresh, S., Ghosh, N., Saravana, R., Chandran, S., Krishna, S., Joy, M., Anand, S. K., Madavan, V., Joseph, A., Wong, G. W., Schiemann, W. P., Constantinescu, S. N., Huang, L., Khosravi-Far, R., Steen, H., Tewari, M., Ghaffari, S., Blobe, G. C., Dang, C. V., Garcia, J. G. N., Pevsner, J., Jensen, O. N., Roepstorff, P., Deshpande, K. S., Chinnaiyan, A. M., Hamosh, A., Chakravarti, A., and Pandey, A. (2003). Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, 13(10):2363–2371.

Rinkel, G. J., Djibuti, M., Algra, A., and van Gijn, J. (1998). Prevalence and risk of rupture of intracranial aneurysms. *Stroke*, 29:251–256.

Risselada, R., van der Lugt, A., Niessen, W. J., and Sturkenboom, M. C. J. M. (2007). Hormonal contraceptives and risk of aneurysmatic subarachnoid haemorrhage. *Basic & Clinical Pharmacology & Toxicology, Special Issue on 8th Congress of the European Association for Clinical Pharmacology and Therapeutics*, 101 (Suppl.):40.

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P., and Vidal, M. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178.

Russel, S. and Norvig, P. (2003). *Artificial Intelligence a modern Approach*. Prentince Hall, 2nd edition.

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(Database issue):D449–D451.

Smigielski, E. M., Sirotkin, K., Ward, M., and Sherry, S. T. (2000). dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Research*, 28(1):352–355.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). Biogrid: a general repository for interaction datasets. *Nucleic Acids Res*, 34(Database issue):D535–D539.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksöz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., and Wanker, E. E. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–968.

Sutton, D. R. and Fox, J. (2003). The syntax and semantics of the proforma guideline modeling language. *J Am Med Inform Assoc*, 10(5):433–443.

van der Lei, J., Duisterhout, J. S., Westerhof, H. P., van der Does, E., Cromme, P. V., Boon, W. M., and van Bemmel, J. H. (1993). The introduction of computer-based patient records in the netherlands. *Ann Intern Med*, 119(10):1036–1041.

von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A., and Bork, P. (2005). String: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*, 33(Database issue):D433–D437.

Wiebers, D., Whisnant, J., Huston, J., Meissner, I., Brown RD, J., Piepgras, G., Thielen, K., Nichols, D., O'Fallon, W., Peacock, J., Jaeger, L., Kassell, N., Kongable-Beckman, G., and Torner, J. (2003). Unruptured intracranial aneurysms: natural history, clinical outcome, and risks of surgical and endovascular treatment. *Lancet*, 362:103–110.