Knowledge environments representing molecular entities for the virtual physiological human

BY MARTIN HOFMANN-APITIUS 1,4,* , JULIANE FLUCK 4, LAURA FURLONG 2, ORIOL FORNES 2, CORINNA KOLÁŘIK 4, SUSANNE HANSER 3, MARTIN BOEKER 3, STEFAN SCHULZ 3, FERRAN SANZ 2, ROMAN KLINGER 4, THEO MEVISSEN 4, TOBIAS GATTERMAYER 4, BALDO OLIVA 2 AND CHRISTOPH M. FRIEDRICH 4

1 Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, Dahlmannstrasse 2, 53113 Bonn, Germany
2 Institut Municipal D’Investigació Mèdica (IMIM), Universitat Pompeu Fabra, C/Doctor Aiguader 88, Barcelona 08003, Catalonia, Spain
3 Universitätsklinikum Freiburg, Hugstetter Straße 49, 79095 Freiburg, Germany
4 Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Institutszentrum Birlinghoven, 53754 Sankt Augustin, Germany

In essence, the virtual physiological human (VPH) is a multiscale representation of human physiology spanning from the molecular level via cellular processes and multicellular organization of tissues to complex organ function. The different scales of the VPH deal with different entities, relationships and processes, and in consequence the models used to describe and simulate biological functions vary significantly. Here, we describe methods and strategies to generate knowledge environments representing molecular entities that can be used for modelling the molecular scale of the VPH. Our strategy to generate knowledge environments representing molecular entities is based on the combination of information extraction from scientific text and the integration of information from biomolecular databases. We introduce @neuLink, a first prototype of an automatically generated, disease-specific knowledge environment combining biomolecular, chemical, genetic and medical information. Finally, we provide a perspective for the future implementation and use of knowledge environments representing molecular entities for the VPH.

Keywords: virtual physiological human; data mining; information extraction

1. Introduction

Knowledge environments use practical, production-quality tools to systematize the consensus knowledge within a scientific domain and facilitate users’ access to that knowledge. (Gough 2002)

* Author for correspondence (martin.hofmann-apitius@scai.fraunhofer.de).

One contribution of 12 to a Theme Issue ‘The virtual physiological human: building a framework for computational biomedicine’.
The above definition of a knowledge environment has been taken from a description of the signalling transduction knowledge environment of the science magazine. In the context discussed here, we will adopt this definition to outline how a knowledge environment representing molecular entities for the virtual physiological human (VPH) can be generated.

In essence, knowledge environments used in the context of the VPH should comprise all relevant objects, attributes and relationships required for multiscale modelling and simulation in the context of the VPH and they should at the same time represent a conceptual model to organize all these data. A knowledge environment for the VPH representing molecular entities should therefore comprise information on all the genes, their allelic variants, their expression in various tissues, their proteins expressed from these transcribed mRNAs and the mode of action of these proteins (including their interaction with other proteins (PPI) and their involvement in signalling pathways, etc.) in a given organ-specific or disease-specific context. Moreover, the molecular entities relevant to the VPH comprise also ligands, metabolites and drugs that are either part of metabolic pathways or that modulate complex biological pathways through binding to essential components of these pathways. It is worth pointing out that a knowledge environment should not only store relevant knowledge in a persistence layer but also comprise all functionalities to present this knowledge to the user and to make it applicable to various data analysis and knowledge discovery methods (Michener et al. 2007).

From the above characterization of a knowledge environment representing molecular entities in the context of the VPH, we can already tell the difference from the existing knowledge environments with a focus on molecular entities such as UniProt (The UniProt Consortium 2007). UniProt comprises highly curated information on proteins in a well-structured format; through extensive referencing of other data sources (e.g. protein structure data; Berman et al. 2007), literature references (Medline; http://www.ncbi.nlm.nih.gov/sites/entrez/) and nucleic acid sequence data (e.g. EMBL; http://www.ebi.ac.uk/embl/), UniProt provides substantial knowledge on proteins to the biomedical community. However, UniProt does not focus on the description of proteins in the system context; although UniProt comprises basic information on protein–protein interactions, detailed information such as information on cell-type-specific protein expression, ligand-binding properties and their characterization as drug targets is incomplete. The role of UniProt entries in pathways is described only through UniProt annotation terms and gene ontology (GO) terms, but these annotation terms provide only a high-level view on the involvement of a protein in a biological process, and thus this functional description in its current state of granularity is not suited to provide a basis for pathway modelling approaches (for comparison, see the representation of molecular interactions in the Reactome database; http://www.reactome.org/). Moreover, UniProt does not provide a user front end that would support the integration of analysis tools and models.

Molecular entities in the context of the VPH have to be described by their involvement in human physiology and disease processes. This means that in the VPH, molecular entities (genes, proteins, metabolites and drugs) have to be described in the context of healthy physiology and disease states. Therefore, associative relationships existing between molecular entities and, for example, clinical phenotypes have to be represented in knowledge environments covering
the molecular level of the VPH. Obviously, neither UniProt, nor any other biomolecular database, allows for establishing a disease context with reasonable granularity including clinical phenotype descriptions. On the other hand, disease-specific databases (e.g. the AlzGene database; Bertram et al. 2007) typically focus on one disease and do not allow for establishing associative relationships between any given disease, any gene or protein of interest and any known ligand or drug.

Aggregation of all relevant information to describe biological and chemical molecular entities involved in a given disease is therefore one of the goals of the VPH. Moreover, we would like not only to aggregate relevant information but also, at the same time, to make use of the molecular entities and their involvement in molecular pathways, to generate models representing normal physiology at the molecular level or the molecular aetiology of a disease. Consequently, a knowledge environment for the VPH has to provide the functionalities to link the models of molecular processes to the phenotypic (clinical) readout in order to, for example, generate predictive models for human diseases.

The first question that we have to answer on our route towards constructing a knowledge environment for the VPH is: ‘how do we get a hand on all the molecular entities relevant in a given physiological or disease context?’ Obviously, even through the aggregation of information coming from nucleic acid sequence databases, protein sequence databases, protein structure databases and protein interaction databases, we would not be able to link this aggregated information to clinical phenotypes and disease terminologies. Relationships between clinical phenotypes and molecular entities, however, can be found in large numbers in unstructured scientific text.

2. Molecular entity types and technologies for the extraction of information on molecular entities from unstructured knowledge sources

As outlined above, a knowledge environment for the VPH representing molecular entities has to represent information on genes and their alleles, on proteins, their interactions and their involvement in signalling and metabolic pathways as well as on ligands and drugs binding to these proteins and influencing these pathways in an anatomical and (patho-) physiological context. It is obvious that a substantial fraction of the knowledge on these complex relationships is available only in unstructured scientific text and not in databases. Whereas in curated knowledge bases, such as UniProt or Reactome, human expertise is used in a time- and work-intensive process to populate a knowledge base focusing on the protein level, for the VPH we need more information that is only partially available in structured database information. Analysing the state of knowledge on all organs, all diseases, all genes or proteins, all chemical compounds associated with proteins or all drugs that are in use for the treatment of these diseases requires automated procedures as expert curation of a knowledge base covering essentially everything from molecular biology via pharmacology to medicine, which would make the construction of a knowledge environment representing molecular entities for the VPH an unrealistic undertaking.

As an alternative, we suggest to make use of automated information extraction technologies to populate a first version of a knowledge environment for the VPH. While proposing this, we take into account that so far automated approaches
for information extraction do not (yet) deliver the high quality of expert knowledge as we find it in curated databases. Consequently, we are aware that with automated information extraction technologies, we trade the breadth of the scope (‘coverage’) against the correctness (‘precision’) of the individual entry at this point.

(a) Sources for scientific text

One major source for our automated information extraction approach is PubMed (http://www.ncbi.nlm.nih.gov/sites/entrez), the largest biomedical literature databank comprising abstracts of more than 16 million biomedical publications. Medline, the major constituent of PubMed, contains the abstracts of the majority of all biomedical publications; the database actually covers the scientific literature in the field of biomedicine from the 1950s onwards (http://www.ncbi.nlm.nih.gov/entrez/query/static/overview.html). Of course, abstracts do not contain detailed information, and important parts on, for example, experimental procedures are missing, but the number of mentions of genes and proteins, metabolites and drugs, disease designators and clinical phenotypes is surprisingly high. Thus, PubMed provides us with a large resource rich in information on genes, diseases, anatomical context and clinical phenotypes.

Another source for unstructured information is the growing number of open access (full text) literature. Although open access journals are still a minority compared with the overall number of scientific journals, they provide a rapidly growing source for full-text information that can be used as an input for information extraction. Yet another source is patent literature, which comprises substantial information on proteins, their qualification as targets for drugs and information on drug-like molecules binding to these targets.

Information extraction technologies for text are often also called ‘natural language processing’ or ‘text mining’ tools (Ananiadou & Mcnaught 2005; Jensen et al. 2006). This paper will not try to define the subtle differences in the usage of these terms. However, to provide the reader with some information on what text mining technology can do in the area of biomedicine, we will shed some light on one major international contest for the assessment of text mining technology in biology: the BioCreative benchmarking activity.

(b) Benchmarking of text mining technology in biomedicine

BioCreative stands for critical assessment of information extraction systems in biology (http://biocreative.sourceforge.net/); it is a community-wide effort for evaluating text mining and information extraction systems applied to the biological domain. BioCreative has been initiated by scientists involved in research on text mining systems and in database production and curation. Two international benchmarking activities have been organized so far: the 2004 (Hirschman et al. 2005) and the 2006 (http://biocreative.sourceforge.net/biocreative_2.html) BioCreative critical assessments. The increased importance of text mining in the life sciences can already be deduced from the increase in the number of participants in the competition: whereas 27 groups took part in 2004, approximately 60 groups participated in 2006. The BioCreative organization team provided annotated corpora for training and evaluated precision
(percentage of correct extractions in correspondence to all extracted information) and recall (indicating the percentage of all correct extractions from the possible correct extractions) of the participating systems on separate test corpora. The correct recognition of named entities is one important indicator for the quality of information extraction and semantic retrieval and both critical assessments addressed the detection of gene and protein names in Medline abstracts. The recognition of all mentions of gene or protein names in text was tackled in the gene mention task. Here, most participants used machine learning-based techniques and the most common technique with good performance used in the BioCreative II assessment was conditional random fields (CRF; see below). Successful systems reached performance levels above 86 per cent for precision and recall, which is getting close to the performance of expert annotations.

The gene normalization task evaluated the association of found protein and gene names with their corresponding sequence database entry (e.g. human BCL2 in text should correctly be associated with GeneID: 596 in EntrezGene). For the normalization task, dictionaries extracted from sequence databases like EntrezGene are needed. The gene normalization task in BioCreative I (2004) focused on the three model organisms fly, mouse, yeast and in BioCreative II (2006) on the human organism. Successful systems with a performance above 90 per cent for yeast and approximately 80 per cent (recall and precision) for fly, mouse and human include extensive curation and expansion strategies for the provided dictionaries and handling of gene name ambiguities (e.g. mentions of ‘p21’ which could mean p21ras or p21kip). Whereas these basic tasks demonstrate that gene and protein name recognition works comparably well, other tasks such as the extraction of annotations for proteins in BioCreative I and the detection and extraction of protein interactions in BioCreative II are much more complex. In these tasks, the performance of automated information extraction systems is significantly challenged by the combination of protein name recognition across different organisms, the information extraction of very specific information for database curators (e.g. give me only the protein interaction which is experimentally verified in the corresponding article) and the usage of full text articles and resource restrictions in annotating or evaluating the full text. Even if at present the available text mining solutions cannot be regarded as ready to use for these complicated challenges, database curators are eager to integrate text mining technologies including further developments in their annotation pipeline to speed up database population. Thus, we can conclude that text mining technologies are likely to contribute to the population of expert curated databases in the future. In the context discussed here, the VPH, we foresee that these technologies are ideally suited for the population of knowledge environments representing molecular entities.

3. Recognition of different biologically relevant entities in scientific text and their semantic integration to corresponding data resources

One of the leading technologies for named entity recognition, which was assessed in BioCreative 2004 and 2006, is ProMiner, the text mining system developed by Fraunhofer SCAI (Hanisch et al. 2005; Fluck et al. 2006). ProMiner performed very well in both evaluations and we reached top ranking positions in the 2004
and the 2006 competition for gene normalization. The system uses extensive curation strategies for the generation of its dictionaries, and sophisticated rules for disambiguation of ambiguous and adaptive strategies have been developed for the handling of acronyms.

In the EU project @neurIST (see www.aneurist.org), we used ProMInER technology to identify not only gene and protein names but also biologically active chemical compounds including metabolites and drugs, and medical entities, in a text corpus specific for the disease area of intracranial (cerebral) aneurysms. The advantage of this dictionary-based approach is that for named biomedical entities detected in text the association with database entries in EntrezGene and/or UniProt (for genes and proteins) can be established. For the recognition of chemically active molecules, we used chemical dictionaries with compound names compiled from reference names and synonyms in DrugBank and PubChem. As with biomedical entities, we can use this approach to link chemical entities in text to their respective entries in DrugBank and PubChem. In such a way, the extracted information from text can directly be combined with the available database information or experimental results. The latter is of particular importance if experimental data such as microarray data or proteomics data should be used for the evaluation of quantitative models in the context of the VPH.

Medical terminology is frequently used to describe diseases and also physiological states. Anatomical terms and organ- or cell-type designators are medical terms, too. As outlined above, the goal of the knowledge environment representing molecular entities for the VPH is to identify all molecular entities in any given physiological context or disease condition. This means that we need to be able to put molecular entities in the context of physiology, anatomy and clinical context. Consequently, we need to be able to identify all medical terminology that carries information on physiological processes, anatomical locations and clinical phenotypes. With respect to the text mining approaches available for the population of the knowledge environment on molecular entities for the VPH, this means that we have to be able to identify medical terminology at a generic level (e.g. UMLS [http://www.nlm.nih.gov/research/umls/] and its constituents: ICD-9 or ICD-10 terms; SNOMED terms; Foundational Model of Anatomy; GO; and MeSH). As those data sources are far too large and too comprehensive but do not represent a certain disease context at desirable granularity, we decided to focus on the medical terminology important for the corresponding disease context. For the area of intracranial aneurysm, an aneurysm-specific ontology (see below) has been developed and subsequently enriched with the corresponding terminology from UMLS. This medical terminology has been used as an additional dictionary to search for aneurysm-specific, relevant medical terms in text. The linkage of the terminology used for text mining to the corresponding ontology allows for browsing the text mining results for different aspects of the disease (e.g. anatomic relationships, risk factors or different treatment methods; figure 1).

Several disease-associated text corpora in the area of intracranial aneurysm and, for comparison, on breast cancer were constructed and analysed in detail for mentions of these biological entities, and a prototype knowledge environment for molecular entities was populated with the results of this analysis (see §3e).
Recognition of allelic gene variants in text and integration of single nucleotide polymorphism information

Allelic variants of genes cannot be easily identified using the dictionary approach. This is partially due to the fact that not all allelic variants are represented in databases, and the identification of information on allelic variants in unstructured text requires special methods that are not dependent on enumeration of entities. Among the different types of allelic variants of genes, single nucleotide polymorphisms (SNPs) are the most studied. Generally, SNPs are represented in text by referring to its position in the genome (more specifically the position in the gene, RNA or protein sequences) and the alleles involved in the sequence change. This differs from the situation for genes and proteins, where specific names are used as identifiers. To identify information on allelic variants of genes in text and to extract mentions of chromosomal localization of genes, we have applied two different technologies, one based on a pattern-based search approach using SNP dictionaries called OSIRIS (Furlong et al. 2008) and the other on a machine learning approach using CRF (Klinger et al. 2007).

Frequently, we find mentions of allelic variants of genes together with descriptions of clinical phenotypes or mentions of disease names in the same sentence or the same section of the text. This information can be used to associate gene allele variants with biological processes or clinical phenotypes.

The automated method used to map mentions of SNPs in scientific text to entries in dbSNP (Sherry et al. 2001), the main repository of SNP data, is based on a pattern matching search approach and the use of SNP dictionaries as described for the OSIRIS system. In this case, the SNP dictionary is composed of the terminology used to describe an SNP and each entry of the dictionary
represents an entry in dbSNP. The terminology comprises different ways used to refer to the alleles and the position of the SNP in the gene or protein sequences. The use of this dictionary allows the normalization or disambiguation of text entities identified as SNPs to entries in biomolecular databases, in this case dbSNP. Normalization of entities extracted from text provides biologically relevant contextual information to data extracted from written resources. For instance, by mapping an SNP identified from text to a dbSNP identifier, all the information collected in the database regarding this specific SNP can be gathered: the gene to which the SNP is mapped, its genome location, the organism, etc.

As a complementary approach to the combined dictionary and query expansion approach taken in OSIRIS, we have chosen to use machine learning to train the computer to recognize mentions of SNPs in text independent of being normalizable or not. This means that mentions of SNPs in text are identified where mapping to entries in dbSNP does not work, because these SNPs are not yet included in dbSNP.

The approach we took is based on CRF, a recently developed machine learning technique, which has been rapidly adopted in the text mining community. Similar approaches to use machine learning for the recognition of gene names or protein names in text or even SNP mentions have been published (Lafferty et al. 2001; McDonald & Pereira 2005; Jin et al. 2006).

The workflow of the system involves two steps: first, several entities are identified and tagged using the CRF, the ProMiner and regular expressions; second, the entities are mapped to dbEntry identifiers (rs numbers).

The entities in the first step are genes (tagged with the help of ProMiner), states (like ‘Ala’ or ‘Pro’), locations (like ‘amino acid 459’) and types (like ‘substitution’ or ‘deletion’) which are tagged with the CRF while directly in the text mentioned rs numbers (like ‘rs 1234567’) are found with regular expressions. In our approach, we trained the CRF using a well-annotated corpus comprising 207 abstracts selected from PubMed. Analysing the performance via bootstrapping, we achieve an $F_1$-measure of 67.9 per cent for entity location, 60.3 per cent for type and 79.2 per cent for state.

In the second step, the normalization is performed using gene, state and location information from the text (figure 2).

An analysis on an independent text corpus of 100 abstracts from Medline showed that normalization of mentions was possible in 142 out of 264 cases, while 216 of them are tagged with the CRF approach. Mapping to the database can fail for different reasons: typos either in the text or in the database as well as simple absence of an SNP in the database.

Taken together, we are able to extract information not only on genes and proteins and the relationship of genes and proteins to diseases (using automated methods) but also on allelic variants of genes. This is important because we complement existing information on molecular interactions in a defined disease context by information on allelic variants mentioned in that very disease context.

(b) Recognition of chemical named entities in scientific text

As outlined above, recognition of chemical named entities is highly relevant for a knowledge environment representing molecular entities for the VPH as we would like to take into account all information on natural ligands and all existing
activators and inhibitors of proteins. However, automated recognition of chemical names is a non-trivial problem. Some chemical named entities can be detected and normalized applying the dictionary approach (Uramoto et al. 2004), whereas large parts of the IUPAC expressions encoding chemicals cannot be normalized and thus require a different approach (Wren 2006). Using dictionaries compiled from reference names and synonyms in different chemical knowledge databases such as DrugBank (Wishart et al. 2006), PubChem (http://pubchem.ncbi.nlm.nih.gov/) and ChEBI (http://www.ebi.ac.uk/chebi/), we have been able to recognize a substantial fraction (78%) of the trivial names in scientific text. When combined with a pattern search (so-called Hearst patterns; Hearst 1992), we could make use of this approach to substantially improve the annotation of chemical entities in DrugBank (Kolářik et al. 2007).

IUPAC expressions, however, have been reliably identified using again the machine learning approach based on CRF (Klinger et al. in press). As IUPAC expressions constitute a significant portion of the chemical named entities in scientific text, the ability to recognize IUPAC expressions automatically and reliably is of utmost importance for the population of the VPH knowledge environment representing molecular entities.

The characteristic of IUPAC names is that their number is in principle countable infinite because they are assembled by different parts. One approach to deal with that is to formulate grammatical expressions. Formulating chemical
knowledge in that way can be tedious, so our approach is to let the CRF learn typical structures in a name. That demands a system with high context information in the text which can be represented in different ways in the CRF. To recognize IUPAC and IUPAC-like names, a Medline training and test set was annotated for these chemical names. The training corpus was used to develop an IUPAC name recognizer. The evaluation on the sampled test corpus of 1000 abstracts from Medline shows a high F measure of 85.6 per cent with a precision of 86.5 per cent and a recall of 84.8 per cent. We will apply this approach in combination with dictionary-based approaches for the recognition of trivial names to extend the chemical knowledge in the VPH knowledge environment. An open question in this context is the normalization of synonymous expressions and the conversion of trivial names and nomenclature names to chemical structures. Recent reports made at conferences (Eigner-Pitto 2007; Murray-Rust 2008) have indicated that currently the ‘name-to-structure’ problem cannot be regarded as a solved problem.

(c) Generating dedicated ontologies for physiological states or diseases

As already mentioned above, medical terminology available in disease classification systems and anatomy ontologies is not sufficiently detailed to represent dedicated knowledge about a specific physiological process or a disease.

In the course of EU project @neurIST, we therefore generated a dedicated ontology that contains all relevant concepts for the description of the clinical phenotype and in addition comprises concepts that are needed to infer risk factors for aneurysm formation and aneurysm rupture. The ontology currently contains approximately 2500 entity types (anatomical, pathological and medical procedural entities as well as biomolecular, epidemiological and haemodynamic entities); it gives both textual and formal (i.e. description logic) definitions for all required entities and will be used as a standardizing instrument throughout the project where ambiguities are frequent (e.g. clinical terms). To provide an adequate terminological coverage, the ontology is furthermore linked to a separate lexical resource that is not part of the ontology proper. This resource provides both preferred terms as well as synonyms and will be at least partly multilingual.

The main sources for the acquisition of relevant entity types have been literature, domain experts and clinical information models. Our modelling approach conforms to the principle of reusing widely acknowledged existing terminologies/ontologies and involved a mapping to the unified medical language system (Bodenreider 2004), a hierarchical classification in terms of is-a (subclass) relationships under the top-level categories of the descriptive ontology for linguistic and cognitive engineering (DOLCE; Masolo et al. 2007) and the connection of associated entity types using other relationships (e.g. part-of, has-location). Besides the classification according to DOLCE, we introduced a second hierarchy intended to represent human knowledge on the disease in different contexts that refer to the scientific areas involved. Using this feature, excerpts of the ontology have been converted into a dictionary used by ProMiner to analyse PubMed for mentions of risk factors for, treatments of and clinical phenotypes associated with intracranial aneurysms. Of course, such a disease-specific ontology provides an ideal conceptual template for the representation of relevant entities and relationships, and indeed the @neurIST ontology comprises all
concepts necessary to link molecular entities to clinical phenotypes. The generation of fine-granular disease ontologies such as the @neurIST ontology is a time-consuming task and thus marks one of the bottlenecks associated with the disease- or physiology-centric view. However, this work is not limited to its application in the knowledge discovery environment described here; the ontology as the result of a standardization and structuring of the conceptual space of the @neurIST project is also designed to support an unambiguous communication (e.g. by mapping the defined ontology types to the entities of the @neurIST clinical reference information model or by providing textual definitions integrated in the clinical data collection tool) and intended to support the access to heterogeneous data; the roles of the ontology in this scenario are data mediation and service binding. The role of the ontology in the context of knowledge retrieval is the provision of versioned dictionaries based on user-defined views on certain areas of interest such as aneurysm types, locations of aneurysm, signs for aneurysm rupture or aneurysm treatment options, enriched with a growing number of synonymous terms used for these entity types.

(d) Modelling of disease-specific molecular interactions: the protein interactions and network analysis approach

Protein interactions and network analysis (PIANA; Aragues et al. 2006) is a software framework capable of (i) integrating multiple sources of information into a single relational database, (ii) creating and analysing protein interaction networks and (iii) mapping multiple types of biological data onto protein interaction networks. PIANA was created to address nomenclature and integration issues common in protein interaction repositories and network visualization tools. In particular, protein–protein interaction analysis is usually biased by the input sources of data. PIANA is one of the very few protein interaction platforms where all interactions from all external databases can be found for a protein of interest, regardless of the type of identifier used as input or the name given to the protein by the researcher that submitted the interactions. PIANA contains a repository of 2 378 113 interactions from DIP 2007.02.19 (Salwinski et al. 2004), MIPS 2007.04.03 (Mewes et al. 2006), HPRD v. 6.01 (Peri et al. 2004), BIND 2007.04.03 (Alfarano et al. 2005), INTAct 2007.04.23 (Kerrien et al. 2007), BIoGRID v. 2.026 (Stark et al. 2006), MINT 2007.04.05 (Chatr-aryamontri et al. 2007) and predictions from structural relatives (Espadaler et al. 2005a; Cockell et al. 2007). The integration of different sources of interactions into a single database allowed us to work with an extensive set of 110 457 human interactions between 36 900 different protein sequences. This set of human interaction data includes 24 812 interactions from yeast two-hybrid assays, 13 256 interactions from immunoprecipitation methods and 11 174 interactions from affinity chromatography methods.

PIANA represents the protein interaction data as a network where the nodes are proteins and the edges are interactions between the proteins. In such a network, a set of proteins linked to protein \( p_i \) (i.e. physically interacting with \( p_i \)) is named ‘partners of \( p_i \). PIANA builds the network by retrieving direct interaction partners for an initial set of seed proteins (i.e. the proteins of interest). We used PIANA to build an aneurysm protein interaction network (APIN), using, as seeds, the proteins known to belong to a pathway involved in the aneurysm
(i.e. by text mining). Thus, the APIN is composed of the known genes and their direct interaction partners. In this network, we define the aneurysm linker degree (ALD) of a protein as the number of genes implied in the aneurysm to which it is directly connected, excluding the protein itself. The ALD was calculated for each protein and proteins were binned by their ALDs. We assessed the use of protein interaction networks for predicting genes involved in a pathway related with the aneurysm. We hypothesized that proteins whose partners are genes of a pathway involved in aneurysm are likely to be involved as well. This hypothesis was successfully applied previously to predict the protein fold between proteins connected by a linker (Espadaler et al. 2005b), candidate sequence fragments for interactions (iMotifs; Aragues et al. 2007) or new putative cancer genes (Espana et al. 2004; Sanz et al. 2007; Aragues et al. 2008). Accordingly, we wish to score the validity of a gene being implied on the aneurysm using an ALD threshold of $N$. However, our expectations need to be proved for the APIN using the threshold $N$ as parameter. Therefore, we first need to prove the prediction of proteins belonging to a pathway involved in aneurysm, defining as positives those proteins with ALD greater than or equal to $N$. True positives are known genes/proteins involved in an aneurysm pathway among positives. False negatives are known genes of a pathway involved in aneurysm whose ALD is lower than $N$. The positive predictive value is defined as the ratio between true positives and positives. Sensitivity is the ratio between true positives and the sum of false negatives and true positives. The set of proteins that belong to a biological pathway (i.e. signalling pathway) involved in aneurysm is taken from data mining. This set is also referred to as Gold Standard.

In order to calculate a $p$-value for the prediction, we calculate a random APIN, where edges and nodes of a seed protein are randomly chosen from the set of human proteins with known interactions in the total repository of PIANA. The number of interactions assigned to a seed protein is the same as it had originally in order to maintain the topology of the network. The process is repeated 1000 times and a $p$-value of the enrichment of proteins involved in aneurysm is calculated using a Fischer test with respect to the average and the randomly attained ALD.

Finally, a threshold $N$ is chosen in order to obtain the best coverage (sensitivity) when predicting new proteins involved in pathways related to aneurysm with the highest ratio of true positives and the smallest of false positives. This parameter is used to iteratively improve and score the nodes of the APIN. The iteration is performed with the predictions of new seeds from a previous step and initialized with the seed proteins taken from data mining extraction. The score of a node belonging to the APIN is defined as the sum of the scores of its partners divided by $N$ and with an upper limit of 1 (i.e. it ranges between 0 and 1). Hence, a protein involved in a pathway related to aneurysm has a score of 1, while those predicted to be involved in a pathway related to aneurysm become new seeds and modify the APIN: (i) increasing the number of seeds and nodes and (ii) varying the values of ALD.

This very simplistic approach can be further refined by assigning scores to the original seeds different from 1. This would need a further validation of the limiting threshold used to increase the score of the nodes of the APIN. Further, we can also score the edges with a value of confidence and modulate the relevance of each score in the final sum with the confidence value of the interaction.
Therefore, the score can be written with the formula

\[ \text{score}_j = \sum_{i \in P_j} \delta_{ij} \text{score}_i, \]

where \( P_j \) is the set of partners of \( j \) and \( \delta_{ij} \) is the confidence score for the interaction between \( i \) and \( j \).

A first protein–protein interaction network associated with the clinical phenotype of intracranial aneurysm is shown in figure 3.

(e) Prototypic implementation of a disease-specific knowledge environment representing molecular entities: @neuLink

Based on the integration of information representing molecular entities in scientific text and databases, we have recently created an information system that serves as a prototype for a knowledge environment representing molecular entities for the VPH. The system, named @neuLink, has been designed as a service-oriented information system comprising several databases accessible as web services, a document indexing and search engine based on Lucene (http://lucene.apache.org/java/docs/) plus a presentation layer/front end that supports data analysis and modelling. @neuLink allows defining the physiological or
disease context and querying the entire Medline for genes, proteins, chromosomal locations, allelic variants, chemical compounds and medical terminology associated with any given disease or physiology. Through @neuLink, the researcher can identify genes reported in scientific publications to be involved in a defined disease (either as a biomarker or as a causal factor in the molecular aetiology of the disease) and, moreover, the researcher can rapidly identify those SNPs that have been mentioned in the disease context. Queries are possible such as ‘give me all the drugs that are mentioned in the literature in the context of the disease’ or ‘give me all the chemical compounds mentioned in the context of the proteins selected as relevant for disease XY’. Detailed information on the found entities is given via link-out to the defining databases (figure 4).

The @neuLink system integrates the protein–protein interaction analysis (described in previous sections) started from seed genes that have been selected from unstructured text. The results can be given as network graphs (figure 4) or used for ranking the findings with the disease linkage degree (e.g. aneurysm linkage degree). In figure 4 is shown an example output of the @neuLink knowledge environment. Prototypic queries that can be solved are: ‘Give me all genes and articles related to breast cancer, which mention medication information’, ‘Which is the most associated chromosome in patients with Down syndrome?’, ‘Which gene variations have been mentioned in the context of intracranial aneurysms?’, ‘Show me all proven risk factors for intracranial aneurysms’, etc.

Related systems for the analysis of unstructured information are EBIMed (Rebholz-Schuhmann et al. 2006), Ali Baba (Plake et al. 2006), Fable (http://fable.chop.edu/), GoPubMed (http://www.gopubmed.org/) and iHOP (http://www.ihop-net.org/UniPub/iHOP/). Moreover, there are several commercial solutions.
available in this area, which could not easily be compared with the public solutions mentioned above, as they are not freely available to the authors. In the following, we will therefore briefly discuss some of the publicly available tools.

EBIMed has been developed at the European Bioinformatics Institute and provides a web-based front end that allows querying of a limited number of Medline abstracts. EBIMed combines information retrieval and extraction from Medline. EBIMed finds Medline abstracts in the same way PubMed does. Then, it goes a step beyond and analyses them to offer a complete overview on associations between UniProt protein/gene names, GO annotations, drugs and species. The results are shown in a table that displays all the associations and links to the sentences that support them and to the original abstracts. EBIMed’s analysis is restricted to a limited number of abstracts that are being analysed at a time.

GoPubMed is a search engine used to find biomedical research articles. The site itself has access to over 16 million articles from Medline. Aside from a general search option, one can look for biomedical articles by subject, author, place of publication, as well as when the article was published. Search results are classified according to the GeneOntology and its sub-sections for processes, functions and cellular components.

iHOP is a search engine that analyses Medline for phenotypes, pathologies and gene function. iHOP forms a network of concurring genes and proteins and extends the network through the scientific literature associating further information on phenotypes, pathologies and gene function. Genes and proteins are used as hyperlinks between sentences and abstracts and, consequently, the information in PubMed can be converted into one navigable resource.

None of the above cited search engines allows for integration of internal (e.g. clinical) or experimental data (e.g. microarray data). Thus, we would not call these systems knowledge environments themselves; however, they could of course provide a basis for a knowledge environment if their functionality is made available as a web service.

One of the most crucial problems of knowledge environments is the validation of results. In the context of the European project @neurIST, it has been done in a comparison between an expert’s review and the @neuLink system on the question: ‘Give me all genes related with intracranial aneurysms’. In the evaluation, we tested whether our system, given the keyword search ‘intracranial AND aneurysm’, was able to detect the same related susceptibility genes that have been found by human experts. The review on genetics (Krischek & Inoue 2006) mentions 18 related genes in the context of intracranial aneurysms. In our evaluation (as of 1 October 2007; Gattermayer 2007), we find 16 548 documents in PubMed related to the keyword search 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. Among the high-ranked questionable findings, we observe frequently proteins used in therapeutic treatments like the tissue plasminogen activator, and hypothetically associated but not tested and proven genes. In one case, we found also a new true positive: the JAG1 gene that has not been mentioned in the review on the genetics of intracranial aneurysm but has been associated with the disease recently.

The additional restriction to documents containing the MeSH term ‘genetics’ reduced the amount of genes to 119. Under these conditions, among the 15 top ranking genes selected by relative entropy, we found 12 genes that were described
by Krischek & Inoue (2006). The others are PKD1, APOE and PKD2 that in the literature are clearly suspected to be directly associated with aneurysm. In contrast to expert reviews about the genetics of intracranial aneurysm, the text mining approach presented here is always up to date and provides a superset of genes involved in different aspects of the disease.

4. Conclusion and outlook

Knowledge environments representing molecular entities for the VPH constitute a significant challenge as all types of molecular entities and their relationships have to be related to a physiology or disease context. As we have shown in this paper, such knowledge environments can be constructed by combining information extraction from scientific text and integration of this information with entries in molecular databases. However, @neuLink, the prototypic knowledge environment representing molecular entities for the VPH we present here, is far from being complete.

Some of the information extraction technologies such as the identification of IUPAC names using CRF are not yet ‘production ready’ and it will require substantial research and development to reach the precision and recall required for populating a knowledge environment. Moreover, the integration of modelling and analysis functionalities in the knowledge environment is still in its infancy. An integration of functionalities supporting disease modelling would be desirable. Such modelling capability would probably benefit from being compliant with the systems biology markup language (SBML; http://sbml.org/index.psp); we are currently evaluating in how far we can adopt to SBML and the tools provided by the systems biology community with our approach.

An additional goal of @neuLink and related approaches is the extension of the molecular modelling approach towards multiscale modelling. As the vision of the VPH comprises explicitly the coverage of the entire span from molecules via cells and tissues to organs and the entire body, a knowledge base such as @neuLink would have to support modelling across these scales. Currently, we find mostly cartoon-like knowledge representations when dealing with cellular processes such as signalling or metabolic pathways (http://expasy.org/cgi-bin/search-biochem-index). Such symbolic representations are widely discussed in the area of systems biology (Butcher et al. 2004). When it comes to illustrating what we believe is going on in intercellular processes and tissue remodelling, even these symbolic representations are not suitable any more for the representation of models. Highly dynamic interactions at the physiological level (e.g. the rolling of lymphocytes on endothelial cells) are typically visualized by video microscopy (Radeke et al. 2005), although kinetic models for a limited number of dynamic processes do exist (Heinemann & Panke 2006). True multiscale representation of the role of molecular entities would also have to deal with complexity that requires a high degree of specification on the molecular level. However, the concrete molecular ‘make-up’ of a cell in a defined tissue is largely unknown and technologies to empirically determine the complete proteome in a cell interacting with other cells to form a biological structure involving more than one cell type (e.g. a subendothelial layer) are far from realistic. Plug-ins for Cytoscape such as ‘BINGO’ (Maere et al. 2005) support the interpretation of biological networks in the context of GO annotations and thus
help to extend the analysis from protein–protein interaction networks to the next level, which is the biological process level. To do this contextual interpretation of molecular data in a cell type or tissue-specific way, the relevant information on cell types and tissues (and organs) can be extracted from text along with gene and protein names. However, our experience is that substantial information on cell types and tissues is usually available only from the material and methods section of biomedical publications, and not necessarily contained in Medline abstracts. With the availability of more full text journals, we hope to be able to provide a much higher level of granularity that at least allows us to model molecular entities and their involvement from the single molecule level to the tissue and organ level.

The fact that we could show that our automated approach identifies all genes mentioned in a certain disease context in an expert review underlines that the principal route taken with @neuLink is valid. However, it remains to be demonstrated that this approach works as well in other disease areas (e.g. breast cancer).

One big open question is of course how we plan to support the integration of clinical information in our knowledge environment. Quite a substantial portion of the VPH is based on imaging, image processing and analysis of biological processes represented by three-dimensional models of organs. In the course of EU project @neurIST, we will integrate functionalities for clinical data retrieval in @neuLink; this includes processed imaging data. Moreover, the environment will provide functionalities for the analysis of experimental data such as microarray gene expression analysis data.

It is noteworthy at this point that the availability of a disease-specific ontology with the possibility to generate user-defined views on relevant entities and dictionaries based on these views, which collect the terms used for these entity types provides us with all means to identify mentions of disease-specific risk factors in text; with this approach we will hopefully be able to include molecular entities (e.g. SNPs) in models aiming at assessing the risk of an individual patient. In the disease area of intracranial aneurysm, the risk of aneurysm formation and aneurysm rupture has been modelled by us applying a probabilistic model (Bayesian network; Han et al. 2006). This modelling approach was based on the manual extraction of information from clinical studies. It remains to be shown how we can support such modelling approaches directly within @neuLink, taking knowledge of molecular entities, biomarkers and functional relationships in disease-associated protein–protein interaction networks into account.

References


