The IIOAB Journal

integrative omics

DATA ANNOTATION AND RELATIONS MODELING FOR INTEGRATED OMICS IN CLINICAL RESEARCH

Arno Lukas[1], **and Bernd Mayer** [1,2,*]

[1] *emergentec biodevelopment GmbH, Rathausstrasse 5/3, 1010 Vienna, Austria*

[2] *Institute for Theoretical Chemistry, University of Vienna, Waehringerstrasse 17, 1090 Vienna, Austria*

Received on: 03-April-2010; **Revised on:** 19-May-2010; **Accepted on:** 21-June-2010; **Published on:** 22-June-2010.

[*]**Corresponding author:** Email: bernd.mayer@emergentec.com Tel: +43-14034966; Fax: +43-14034966

ABSTRACT

Omics has massively permeated translational clinical research with numerous diseases being covered by Omics studies from the genome to the metabolome level. Integrating these disease specific Omics tracks appears a logical next step for building the fundament of Systems Biology and Systems Medicine. Here, coherence of individual Omics tracks regarding clinical hypothesis, samples and clinical descriptors, and finally data handling and integration become pivotal. We present a data integration, annotation and relations modeling concept for heterogeneous Omics data and workflows. With molecular features at the center of all Omics we link the result profiles from different Omics tracks characterizing a specific disease phenotype to a common human molecular reference network for allowing a seamless integration and subsequent support in interpretation of Omics screening results.

Our concept rests on data structures for representing objects specified by metadata and content. For handling diverse Omics tracks a flexible structure for content is proposed allowing data representation at different levels of granularity as demanded by the type of Omics and specific type of data. Content on the molecular level includes deep annotation of molecular features on gene and protein level. Based on this annotation pair-wise relations between molecular objects are, traversing the molecular annotation into a network of relations (molecular feature graph). Such a relation network is also built on the Omics data level, combining explicit relations derived from study setup and implicit relations generated by mining metadata and content (Omics data graph).

Finally both graphs are merged utilizing the molecular feature level as common denominator, enabling a persistent integration and subsequently interpretation of Omics profiling results in the realm of a given clinical hypothesis. We present a case study on integrating transcriptomics and proteomics data on chronic kidney disease for demonstrating the feasibility of this concept.

.

Keywords: *integration; networks; standards; omics; graphs*

[I] INTRODUCTION

With sequencing of the human genome a major cataloguing milestone was reached in 2001 [1], followed by rapid development of Omics tracks spanning from the genome to the metabolome level. A summary statistics on the various Omes is provided at the Gerstein lab (http://bioinfo.mbb.yale.edu/what-is-it/omes/omes.html), clearly indicating the maturity of genomics efforts when compared to the other Omes. Omics has in the meantime entered clinical sciences aimed at elucidating the pathophysiology of diseases, thereby providing the basis for identifying biomarkers serving for novel diagnostics and therapy [2,3]. Specific profiling has already been forwarded to clinical application, e.g. for assessing breast cancer utilizing a profile of about 70 features [4]. Numerous prevalent diseases have been studied on the various Omics levels, and first efforts were introduced for consolidating this body of knowledge in open access data repositories. Usually these repositories are Omics-specific as e.g. ArrayExpress for transcriptomics [5] or PRIDE for proteomics data [6], or Omics profiles are consolidated on the level of genes (gene-centric) as in Genecards (http://www.genecards.org) [7]. For some etiologies also disease-specific databases have been established, with Oncomine as an example for consolidating cancer transcriptomics data [8]. Platforms integrating various Omics levels, however, are less common, although being perfectly in line with approaches in Systems Biology [9], in the meantime already expanding at least conceptually towards Systems Medicine [10]. Aim of these concepts is broad integration of Omics tracks being embedded in clinical data space and sample descriptors, with the ultimate goal of providing a quantitative representation of disease (outcome)- specific molecular processes.

Distinct specifications have to be met in Omics in particular including: i) a quantitative assessment of molecular objects, and ii) approaching the totality of objects at some layer of cellular organization. Advancements in miniaturization, improved readout technologies, and parallelization of established technologies have significantly contributed to the accuracy of quantitative measurement procedures. However, major shortcomings remain with cataloguing efforts for determining the totality of some sort. Here, genomics may come closest to completeness, presently experiencing a further boost resting on next generation sequencing technologies at least in principal allowing an unbiased decoding of entire genomes [11]. However, for all other Omes limitations have to be recognized, and even the notion of a "gene" came under some scrutiny, [12] particularly when evaluating results of the ENCODE consortium [13]. Gene expression array data in most cases still focus on protein coding genes, may include some resolution on the level of splice variants, but only in rare cases expand to assessing miRNAs or more generally ncRNAs. The totality of the proteome (and to some extent also of the metabolome) is under question on a theoretical level, but is rapidly evolving due to parallelized high resolution separation, identification, as well as quantification.

For integrative Omics, and here in particular in the medical context, numerous additional factors have to be taken into consideration, centrally including sample specifications [14]. A detailed clinical hypothesis comes in the first place, and from there delineation of strict sample inclusion and exclusion criteria result. Case-control studies are the typical setup in screening, where ideally cases and controls are matched for all parameters with known or suspected impact but the clinical question of interest (outcome). Here either a dedicated prospective sample and data collection has to be established, or a retrospective collection is available. Best sources in the latter case include interventional studies performed under strict quality control. In line with sample specification is assessment of sample size for assuring a well powered study from the statistical perspective for each individual Omics track considered [15]. Omics procedures are applicable for various sample types, most frequently utilizing tissue, blood and urine. Here standardized sample handling and preparation comes into play, where standard operating procedures (SOPs) for storage and preparation have been derived for a number of Omics tracks [16].

In the light of the aforesaid the following issues may be considered as central for integrating heterogeneous Omics profiling results:

1. Thorough definition of the clinical hypothesis
2. Detailed specification of cases and controls for each Omics track
3. Sample size calculations for each specific Omics track
4. SOPs for sample and clinical data handling
5. SOPs for Omics procedures and data generation
6. Standardized reporting covering each Omics workflow

Regarding reporting conventions numerous initiatives have been started, including experiment description as well as execution standards [17], and both are to some extend already followed in results reporting, with MIAME being a well known implementation for transcriptomics [18].

If different Omics tracks follow defined *standards* in reporting and are *in line with a given clinical hypothesis* Omics integration on the level of result profiles becomes feasible. For setting up a cross-Omics results integration two approaches may be followed for data preparation: Public domain driven by consolidating available information on a given clinical hypothesis (e.g. by extracting available profiles on a specific disease from ArrayExpress of PRIDE), or implementation of a dedicated cross-Omics project explicitly focusing on the specific clinical question. The latter approach may even expand towards using samples from the very same patients for conducting the individual Omics tracks, certainly adding to data coherence. Prototypical settings of such initiatives include the research consortia predict-IV focusing on toxicological aspects (http://www.predict-iv.toxi.uni-wuerzburg.de), or SysKid (http://www.syskid.eu) analyzing chronic kidney disease by a Systems Biology approach.

Fulfillment of the technological and procedural requirements discussed so far enable consolidation of heterogeneous Omics feature profiles in a Systems Biology (Medicine) context. The next step in implementing such an approach is providing data management and integration which serves as basis for subsequent analysis, ultimately yielding molecular processes, biomarkers and target candidates linked to the specific disease and outcome. At this step the incomplete molecular cataloguing aspect comes in, adding *annotation* as a major aspect to Omics data management and integration.

We in the following propose a data consolidation and annotation framework specifically aimed at covering integration of diverse Omics result profiles directly linked to a human molecular reference network. We in particular present concepts for explicit as well as implicit relations inference aimed at supporting data interpretation in the realm of a given clinical hypothesis.

[II] MATERIALS AND METHODS

2.1. Object abstraction

The generic component of our concept is an *object*, resembling a data structure holding a unique identifier. Practical notion of an object is kept broad, involving molecular objects and Omics data objects. Omics data objects, in the following referred to as "records", involve any type of machine readable data relevant for or generated in the course of experimental procedures. Typical records include raw data matrices, analysis results (being the core of our integration concept), validation results, or sample specifications. Molecular objects on the other hand are defined as known and well annotated genes or proteins (but conceptually may be expanded for also including RNA, metabolites, etc.). For each object metadata are provided allowing further characterization of the object category. Next, the effective *content* of an object is given. Molecular content involves annotation data e.g. specifying a gene's functional terms or protein interaction data. Omics record content is in a first place characterized by the level of granularity, where content of an individual record may involve large profiling matrices covering an entire Omics screening experiment, may resemble results profiles from case-control studies, or may provide individual molecular features and their specific expression value found in a particular experiment. A third major element is *relations* which put objects (and their content) into context. Relations again follow the data structure concept, where next to a unique identifier metadata are provided. Relation specific metadata mainly include a specification of the type and further edge content as directionality, source (explicitly built or implicitly computed), or evidence level.

2.2. Technical implementation

The Java Enterprise Platform (http://java.sun.com/javaee), utilizing a post-relational approach as data foundation, provides an efficient platform for implementing object oriented concepts as discussed here. This platform supports dynamic data models technically enabled via the Content Repository for Java (JCR), complemented by Glassfish as application server. On the server side the Enterprise Java Bean component architecture seamlessly supports an architectural design for separating application logic and presentation logic. Apache Jackrabbit as a reference implementation of JCR provides further functionality including versioning and full text search. Java Server Faces may be used for implementing the client side.

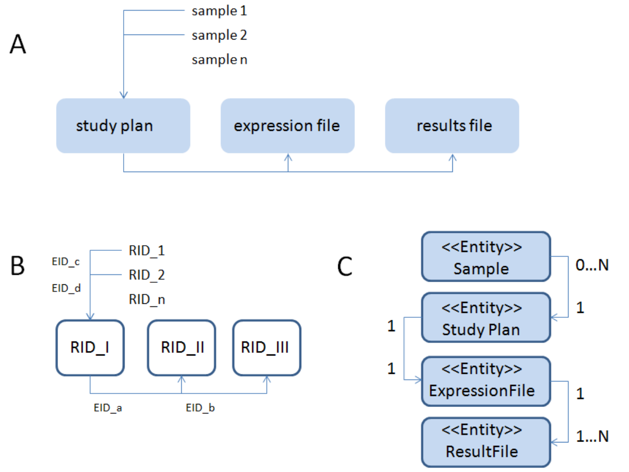
2.3. Public domain sources

Software platforms, modules, as well as molecular content necessary for realizing the technical backbone of the concept presented in this work are available in the public domain. The JCR reference Apache Jackrabbit is found at http://jackrabbit.apache.org, Java Server Faces is found at http://java.sun.com/javaee/javaserverfaces. A manifold of modules for supporting data processing workflows is provided by Taverna (http://www.taverna.org.uk), with the Taverna engine also embedded in Java. Biomart, available at http://www.biomart.org, can be customized for supporting the data management side, and additionally a website can be configured for providing user interfaces. Biomart further allows interfacing via web services for handling large data sets. As objects are represented in their context visualization of resulting networks is essential for supporting interpretation. Gehlenborg et al. [9] recently provided a review on Omics visualization tools, with Cytoscape (http://www.cytoscape.org) as a prominent example. Cytoscape allows an extended definition and display of node (record) types, necessary for visualizing heterogeneous content spanning from clinical sample nodes to molecular feature nodes. Different types of molecular interaction networks are available for download, including procedural interactions from KEGG (http://www.genome.jp/kegg) and PANTHER (http://www.pantherdb.org), physical interactions (both experimentally determined as well as predicted) from the meta-database OPHID (http://ophid.utoronto.ca/ophidv2.201), or interaction networks consolidated from multiple sources as STRING (http://string.embl.de). For assuring coherence on the name space level for molecular reference networks as well as molecular features coming from the various Omics levels a reference namespace has to be selected and regularly updated. Source providing broad coverage of features are found with UNIPROT (http://www.uniprot.org) or NCBI (http://www.ncbi.nlm.nih.gov/refseq).

[III] RESULTS

3.1. Omics record consolidation

The generic object for Omics data consolidation is a record representing data at any given level of detail, e.g. characterizing an entire transcriptomics profile or only a single feature and its associated expression value. For each record metadata may be provided for further characterization of the record content. Furthermore object relations can be built for introducing dependencies between records. [Figure-1] provides an example scheme of the record (node) and relation (edge) concept.



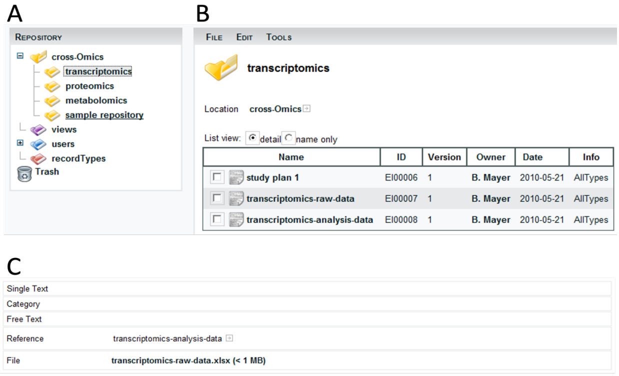
Omics workflows: (A) Schematic setup of an Omics track involving study plan, expression raw data and analysis results data. (B) Formal representation of the workflow as node and edge concept with each object encoded as a data structure holding a unique identifier and a parameter list (C) Representation of the concept in UML (Unified Modeling Language, http://www.uml.org).

Omics procedures follow a generic process as exemplified in Figure 1A. First a study plan is specified defining the case-control setup reflecting the clinical hypothesis, methodology used, etc. Based on the definition of cases and controls samples are linked which effectively undergo screening as specified in the study plan. Equivalent to study plans samples are represented as records (holding sample source, type, amount available, etc. as metadata and content). Frequently samples are organized in dedicated databases and may only be linked into an Omics record management via unique sample identifiers. When retrieving Omics profiles from the public domain the level of detail regarding sample-specific clinical data is frequently sparse and typically limited to clinical categories/stages for the disease.

Executing experimental profiling results in an expression record (e.g. raw data matrix of case and control samples), which after statistical analysis leads to a results record only listing significantly differentially regulated features when comparing case and control group. Typically such a results profile is based on a per-feature statistical test including correction for multiple testing. Although substantial differences in experimental procedures are evident this basic workflow is followed by most Omics tracks assessing continuous concentration values (a fact also becoming evident when comparing MIAME and MIAPE for transcriptomics and proteomics, respectively).

On an abstract level (Figure 1B) a graph representation becomes feasible, holding nodes characterized by record identifiers (RID) and edges specified by edge identifiers (EID, in the example case being directed). Each node and each edge is accompanied by a data structure holding a unique identifier. In the case of nodes metadata and the content as such are stored in the data structure, for edges the node identifiers specifying the connectivity via node IDs as well as metadata (directionality, type of edge, etc.) are provided. For nodes individual content may be represented at arbitrary levels of granularity (spanning from whole profile matrices to single features) depending on subsequent resolution needs in analysis. However, resolution on the level of individual features is mandatory in virtually all analysis procedures. For practicability issues encapsulation of entire profiles, arrays of profiles, or analysis result vectors appears preferable. This approach significantly reduces complexity on the record level and eases upload and management, but still provides access to individual features when using record templates (where e.g. feature and associated expression value reside in defined content locations).

Omics integration naturally demands a combination of profiling efforts, exemplarily shown in **[Figure-2]**. The situation given in Figure 2A is defined by individual study plans I-III, respective screening profiles (e.g. raw data) and results (list of significantly different features on the transcriptome, proteome and metabolome level). Multiple result files may be generated (see also the UML in Figure 1C) e.g. by varying statistical procedures used for analyzing a given case-control group, or by varying the assignment of samples as case and control.
In an ideal setting the studies are fed from a single sample / clinical descriptor repository (feasible for explicitly designed cross-Omics), or have to be extracted to the extent possible if fed from public domain Omics profiling (e.g. gathering available Omics studies regarding a specific clinical hypothesis). Naturally, a dedicated study will provide a more complete and coherent set of records, as these reflect explicitly defined inclusion criteria focusing on a specific clinical hypothesis.



Location cross-Omics

List view: ● detail ○ name only

	Name	ID	Version	Owner	Date	Info
□	study plan 1	EI00006	1	B. Mayer	2010-05-21	AllTypes
□	transcriptomics-raw-data	EI00007	1	B. Mayer	2010-05-21	AllTypes
□	transcriptomics-analysis-data	EI00008	1	B. Mayer	2010-05-21	AllTypes

Single Text	
Category	
Free Text	
Reference	transcriptomics-analysis-data
File	transcriptomics-raw-data.xlsx (< 1 MB)

Fig: 2. Node and edge concept for cross-Omics: (A) Schematic setup of three individual Omics tracks fed by two sample sources. (B) Formal representation of the Omics tracks given in (A) further including two implicit relations (EID_1, EID_2).

Certainly, record types are not limited to the examples given in Figure 1 and 2. Further record types of value include scientific references, standard operating procedures, or experimental validation data (including both, profile validation as well as complementary data e.g. coming from *in-vitro* and *in-vivo* models of the clinical setting), among others.

[Figure-3] provides an implementation example for organizing the corresponding record management. This reference implementation organizes records along specific Omics tracks (Figure 3A), and essential records specifying the study specifications and results (Figure 3B). For each record metadata as well as explicit links between records (in the example linking transcriptomics raw data and analysis data, Figure 3C) can be specified. All relations specified in Figure 1A and 2A are *explicit*, as such defined by the user depositing the records, and reflect the logical structure of Omics procedures. Of central relevance here is that the Omics tracks are driven as independent processes, in a first place only (explicitly) linked if using joint samples (and more generic by focusing on one specific clinical hypothesis).

However, further *implicit* relations are present in the collection of records (Figure 2B). One set of relations may be derived from joint metadata (EID_1) used for characterizing records (e.g. using the same tissue type), and a second set of relations may be derived from the record content as such (EID_2): Software frameworks as Jackrabbit provide full indexing of records for text search, and by this mechanism records can be linked e.g. based on "overall similarity", or specifically by e.g. invoking on joint molecular identifiers in feature lists. Relevant examples include shared gene or protein identifiers for extracting relations e.g. between transcriptomics and proteomics profiles.

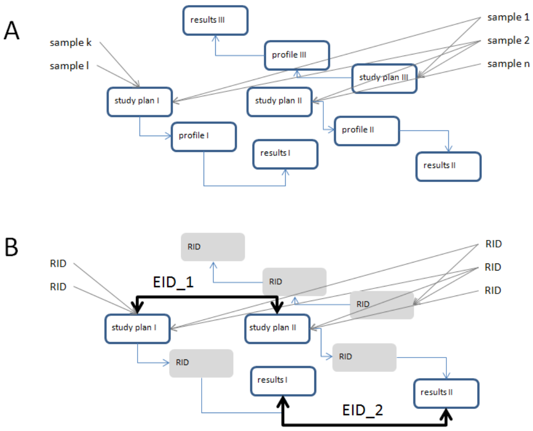


Fig: 3. Example layout for cross-Omics record management: (A) Repository structure involving three Omics tracks and a clinical sample repository. (B) Records assigned to a specific Omics track (see Figure 1A). (C) Example metadata categories for a transcriptomics raw data record including an explicit relation between transcriptomics raw data and analysis data.

Annotation of relations certainly goes far beyond the examples provided here, as numerous project specific term lists may be used. Relevant examples include relations mapping based on disease-feature or feature-drug associations: A results file record may be mined for occurrence of gene or protein identifiers with known links to diseases (e.g. utilizing OMIM data, http://www.ncbi.nlm.nih.gov/omim), and if identified the diseases can be added as metadata to the record. In a next step relations between records can be built based on co-occurrence of disease associations. A comparable procedure may be relevant for known drug-target associations as e.g. provided by STITCH [19]. Yet another relevant procedure is to link scientific publications to records via publication-feature information e.g. mined from MEDLINE [20].

3.2. Omics feature consolidation

Equivalent to the graph concept for representing Omics workflows also molecular features can be consolidated. In a standard Omics setup a feature denotes a relevant object (gene, transcript, protein, etc.) separating cases and controls utilizing a statistical measure. The typical representation of features including their relations is graphs, with protein-protein interaction networks (PPIs) as well known example [21]. PPIs are usually specific regarding the type of relation, e.g. IntAct networks encode physical (undirected) interactions [22], whereas KEGG represents procedural information also including edge directionality [23]. We derived the human proteome interaction network omicsNET which combines significant annotation with relations modeling. RefSeq (http://www.ncbi.nlm.nih.gov/refseq) is used as reference source for human genes and proteins providing about 25,000 objects (considering a canonical sequence set of genes and proteins). For each gene/protein deep annotation was performed utilizing public domain sources, including tissue specific reference gene expression, various sources for functional annotation as Gene Ontologies, manifold protein interaction data sources and further protein characterization as subcellular location, among others. Additionally transcriptional control on the level of transcription factors and miRNAs was added.

Technically, data structures were used, each holding a unique identifier linking to a gene/protein, and storing the annotation data as content. On the basis of the gene/protein-specific content a pair-wise relation score was computed which may be interpreted as dependency resting on the individual annotation given in the content. For further details on omicsNET we refer to [24]. A schematic layout of the construction principle is given in **[Figure-4]**.

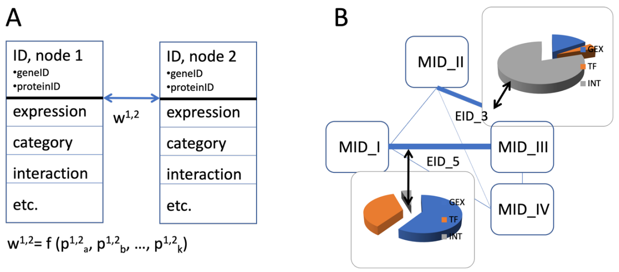


Fig: 4. Concept of a molecular feature annotation graph: (A) Data structures specified by unique identifiers and extended annotation serve as basis for computing dependency weights. (B) Network representation holding molecular nodes and weighted edges, where weights are delineated from given annotation exemplarily shown for two edges (EID_3, EID_5) (individual contributions coming from: GEX: tissue specific gene expression; TF: joint transcription factor binding sites; INT: protein-protein interactions).

Next to consolidated annotation the feature representation given in Figure 4 provides the opportunity for automated relations modeling. All content associated with molecular nodes is parameterized as input for an empirical metafunction *f* which allows computing pair-wise dependencies between molecular features. The metafunction integrates similarity measures as correlation coefficients for tissue specific gene expression profiles, as well as dependency measures as known protein interaction (e.g. coming from Intact or KEGG) for a given pair. The resulting parameter $w^{x,y}$ approximates an aggregate dependency between molecular features, and as this is done for all features a complete matrix and graph results. This graph can now be used for mapping analysis results coming from the various Omics tracks.

3.3. Integrating records and features

Obviously the representation of multiple Omics workflows, but also the system analyzed by Omics as such, namely an extended (although far from being complete) assessment of molecular entities may be represented as nodes (content) and edges (relations).

On this basis an integration of both structures is an obvious next step, as schematically shown in **[Figure-5]**.

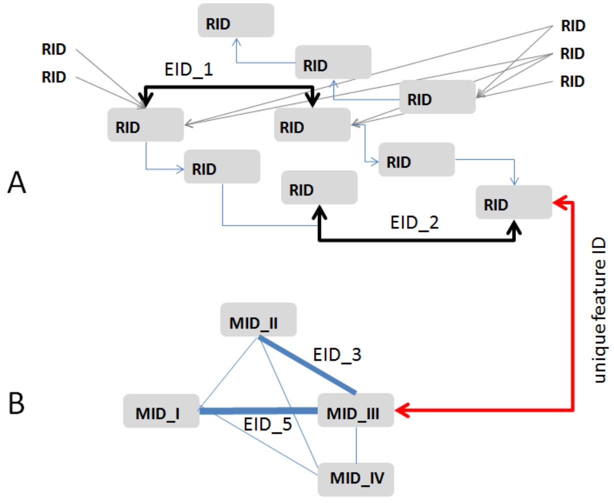


Fig: 5. Integrating Omics results data on a molecular feature graph: (A) Record data structures model (see also Figure 2B) and (B) feature data structure (see also Figure 4B) interlinked on a joint name space level (edge given in red).

Omics operates on molecular name spaces, with gene and protein IDs as the most prevalent reference spaces. Decomposing all data into records with unique identifiers naturally supports building relations between the Omics record structure and the molecular feature structure. From this concept a persistent relations mapping for Omics results integration emerges, embedding sample space, experimental procedure logics, and molecular feature landscape. Features identified as relevant on the record level (stored in Omics result records) have a direct representation on the feature graph and vice versa.

As for all relational models querying is naturally supported by the presented concept. However, yet another more powerful type of querying becomes feasible, namely subgraph extraction. An example subgraph is schematically depicted in [Figure-6].

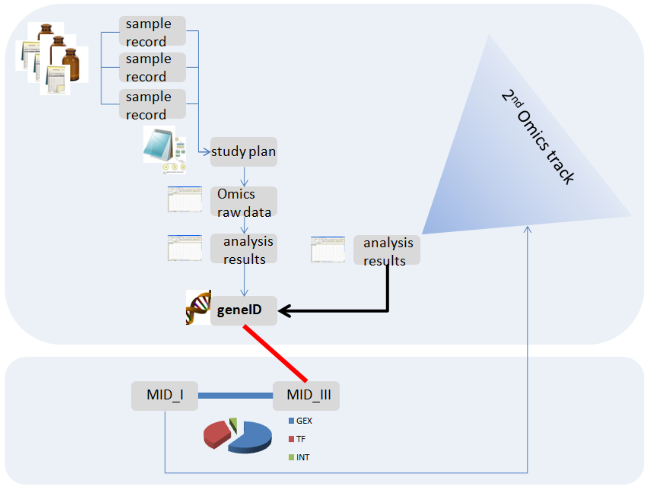


Fig: 6. Navigating in Omics record and molecular feature space: For a particular Omics track explicit relations are provided from sample records to study plan, further to raw data and results data. A specific feature of interest (geneID) given in results data shows an implicit link to a results record from a second Omics track, and on the molecular level equivalency with MID_III, which further shows strong dependency to MID_I (and may link to a result profile coming from a different Omics track).

Merging the record and feature concepts traverses the traditional querying in relational databases into analysis of subgraphs. The example provided in Figure 6 uses a particular feature from an Omics analysis results file as start point. For this feature an implicit link to a second results file coming from a different Omics track is detected which allows tracking the path upstream this second track.
At the same time downstream analysis into the molecular feature space becomes amenable. Here relations rest on computed dependencies based on broad feature annotation. For the example case a strong link to a second molecular feature may be followed which itself eventually may have become evident at some other level (e.g. an associated scientific publication) in a second Omics track.

3.4. Example case

We in the following exemplify the presented concept for Omics profiling of chronic kidney disease (CKD), a disease characterized by progressive loss of kidney function. CKD has been extensively studied on various Omics levels with an impressive consolidation effort on the transcriptomics level provided by the nephromine database (http://www.nephromine.org). Next to diabetic nephropathy (DN) and hypertensive nephrosclerosis other (mainly histopathological) classifications characterize the types of CKD, including IgA nephritis, focal segmental glomerulosclerosis, membraneous glomerulonephritis, and minimal change disease. For these types of CKD specific profiles on transcript and proteome level are available in the public domain, all derived on disease-type specific case-control Omics profiling [25]. Utilizing the cross-Omics integration concept outlined above provides a graph shown in [Figure 7].

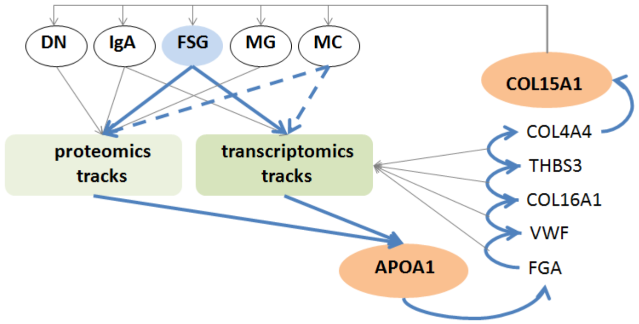


Fig: 7. Integrated Omics for chronic kidney disease: Sample classification is provided by the histopathological representation of the disease (DN: diabetic nephropathy; IgA: IgA nephritis; FSG: focal segmental glomerulosclerosis; MG: membraneous glomerulosclerosis; MC: minimal change disease), each type entering proteomics, transcriptomics, or both. Strong arrows indicate paths linking sample type, Omics track, and molecular feature space (given by gene symbols).

In this example Omics profiling is included for various types of CKD, with feature lists from proteomics characterizing all five representations when compared to matched (healthy) controls, and three conditions characterized by transcriptomics. Implicit linking of Omics result profiles shows APOA1 jointly identified by proteomics and transcriptomics when only considering the CKD type FSG. Analyzing APOA1 on the level of the protein interaction networks a path including the molecular identifiers FGA, VWF, COL16A1, THBS3 and COL4A4 and COL15A1 becomes evident (and all of these are identified as significantly differentially expressed by the transcriptomics track), where COL15A1 is additionally identified as significantly affected for all types of CKD on the basis of proteomics screening results. Consequently, also minimal change disease links into this network. Interesting to note here is that from a clinical perspective minimal change disease presents comparable to prolonged segmental glomerulosclerosis.

This Omics results annotation may be further extended by including genetic studies on CKD [26] identifying uromodulin (UMOD) as affected. UMOD itself is found as differentially regulated by the transcriptomics studies, and shows on the molecular graph level a shortest path to APOA1 (via CRP and APOA2), but also to COL15A1 (via CRP, FN1, and COL5A1).

[IV] DISCUSSION

Omics procedures have reached a level of maturity enabling implementation in standard laboratories, and broad scale application is seen in translational and clinical research. Standards have been derived for most Omics tracks including both experimental design as well as execution, and reproducibility of Omics screening shows satisfactory results. However, integration of results from different Omics tracks and domains, but even of results coming from Omics studies focusing on the very same level of molecular organization experience shortcomings. We consider two main issues as relevant. The first is maintaining strict coherence on the experimental side in particular regarding sample inclusion and processing criteria. Specifically when addressing complex situations as human diseases a strict definition of the clinical hypothesis, associated clinical parameters, and outcome have to be closely shared for individual Omics tracks aimed for integration. For illustrating this issue the clinical presentation of “chronic kidney disease” [27] may be used, which as term includes various causative conditions and on the level of outcome may involve various parameters as levels of albuminuria, creatinine, or glomerular filtration rate. Omics integration for “chronic kidney disease” will certainly provide a far less coherent picture on the molecular level than using studies addressing specific type and specific stage of the disease. Omics procedures following such strict inclusion are certainly less frequently found in the public domain emphasizing the importance of dedicated Omics approaches.

The second major issue is data handling concepts supporting Omics workflows on the entire level of annotation, spanning from the clinical data spectrum to the individual Omics profiles and relevant features resulting from the manifold of different analysis procedures. As mentioned above disease-specific Omics repositories slowly emerge, also including to some extent metadata information as sample specifications on the clinical level. However, most of presently found disease specific repositories in the public domain are too broad in scope, hamper metadata at an adequate level of detail, and mostly include only a specific Omics domain (with transcriptome profiles as the most abundant type).

We in this work present an Omics integration concept covering both, the data spectrum of Omics tracks as well as persistent mapping to molecular annotation. Data management concepts for Omics in a first place need a specification regarding granularity of data representation. Laboratory Information Management Systems (LIMS) have been designed for also covering Omics [28]. However, from the background of LIMS significant standardization of workflows is assumed which for individual Omics tracks appears manageable but for cross-Omics is difficult to maintain (and for repositories built from public domain is merely impossible to achieve). For handling this issue we propose a record concept, formally represented as data structure managing content at arbitrary levels of granularity, where templates serve for standardizing experimental design and execution. This data encapsulation provides easy adaption to expanding scope (e.g. if yet another Omics track becomes available and needs integration), but also allows a representation of the entire Omics workflow including study plans, sample repositories, procedure documentation, raw data files, as well as analysis results and verification data. The proposed Omics annotation concept takes, next to data representation, care of another central aspect, namely relations modeling. Uniquely referenced objects allow explicit definition of relation (as raw data file and associated analysis file(s)), and if implemented in a proper environment provides implicit relations modeling. The latter is of particular relevance on the level of cross-Omics data interpretation.

The combination of Omics procedure annotation and relations modeling traverses the concept into a knowledge representation framework, formally represented as graph with content (nodes) in their context (edges). Such a design naturally enables integration with molecular graphs with genes/proteins being the predominant levels for data interpretation (where e.g. metabolites are mapped to involved enzymes, or SNP data to affected genes including their regulatory regions). Various molecular graphs resting on deep annotation have been derived with omicsNET [24] or STRING [29] as prototypical reference. Merging Omics graphs and molecular graphs enables extended querying utilizing methodologies provided by graph theory [30]. The concept discussed above allows extracting subgraphs and paths linking molecular features to their neighborhood on the molecular, the Omics tracks, and the sample specifications (clinical) level.

[V] CONCLUSION

Omics integration clearly bears the potential of expanding our understanding of complex diseases, and substantial efforts for bridging Omics levels have already been reported [2,9,10,25,31]. However, for building descriptive models characterizing diseases at the interface of clinical specifications and molecular processes in the realm of higher order structures as proposed for formal and biochemical systems [32] more fundamental issues have to be tackled. We consider annotation and relations modeling embedded in flexible data and knowledge management frameworks as a fundament for concise cross-Omics data interpretation on the level of descriptive graphs. Only as cataloguing efforts on the molecular level expand, and the number of different Omics screens on specifically defined clinical etiologies increase, model building in the realm of Systems Biology and Systems Medicine will become amenable.

FINANCIAL DISCLOSURE

This work was financially supported by the European Union FP7 projects predict-IV (project number 202222) and SysKid (project number 241544).

REFERENCES

[1] International Human Genome Sequencing Consortium. [2001] Initial sequencing and analysis of the human genome. *Nature* 409:860–921.

[2] Perco P, Wilflingseder J, Bernthaler A, et al. [2008] Biomarkers for cardiovascular disease and bone metabolism disorders in chronic kidney disease. A systems biology perspective. *J Cell Mol Med* 12:1177–87.

[3] Sikaroodi M, Galachiantz Y, Baranova A. [2010] Tumor markers: the potential of "omics" approach. *Curr Mol Med* 10:249–57.

[4] Slodkowska EA, Ross JS. [2009] MammaPrint 70-gene signature: another milestone in personalized medical care for breast cancer patients. *Expert Rev Mol Diagn* 9:417–22.

[5] Parkinson H, Kapushesky M, Kolesnikov N, et al. [2009] ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res* 37:D868–72.

[6] Jones P, Martens L. [2010] Using the PRIDE proteomics identifications database for knowledge discovery and data analysis. *Methods Mol Biol* 604:297–307.

[7] Rebhan M, Chalifa-Caspi V, Prilusky J, et al. [1997] GeneCards: integrating information about genes, proteins and diseases. Trends Genet 13:163.

[8] Rhodes DR, Yu J, Shanker K, et al. [2004] ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* 6:1–6.

[9] Gehlenborg N, O'Donoghue SI, Baliga NS, et al. [2010] Visualization of omics data for systems biology. *Nat Methods* 7:56–68.

[10] Ahn AC, Tewari M, Poon C-S, et al. [2006] The Clinical Applications of a Systems Approach. *PLoS Med* 3:e209.

[11] Mamanova L, Coffey AJ, Scott CE, et al. [2010] Target-enrichment strategies for next-generation sequencing. *Nat Methods* 7:111–8.

[12] Stadler PF, Prohaska SJ, Forst CV, et al. [2009] Defining genes: a computational framework. *Theory Biosci* 128:165–70.

[13] Rosenbloom KR, Dreszer TR, Pheasant M, et al. [2010] ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Res* 38:D620–5.

[14] Taylor CF [2007] Progress in standards for reporting omics data. *Curr Opin Drug Discov Devel* 10:254–63.

[15] Mehta TS, Zakharkin SO, Gadbury GL et al. [2006] Epistemological issues in omics and high-dimensional biology: give the people what they want. *Physiol Genomics* 28:24–32.

[16] Morrison N, Cochrane G, Faruque N, et al. [2006] Concept of Sample in OMICS Technology. *OMICS* 10:127–37.

[17] Ball CA, Brazma A. [2006] MGED standards: work in progress. OMICS 10:138–44.

[18] Quackenbush J. [2009] Data reporting standards: making the things we use better. *Genome Med* 25:111.

[19] Kuhn M, Szklarczyk D, Franceschini A, et al. [2010] STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res* 38:D552-6.

[20] Matos S, Arrais JP, Maia-Rodrigues J, et al. [2010] Concept-based query expansion for retrieving gene related publications from MEDLINE. *BMC Bioinformatics* 11:212.

[21] Alberghina L, Höfer T, Vanoni M. [2009] Molecular networks and system-level properties. *J Biotechnol* 144:224–33.

[22] Aranda B, Achuthan P, Alam-Faruque Y, et al. [2010] The IntAct molecular interaction database in 2010. *Nucleic Acids Res* 38:D525–31.

[23] Kanehisa M, Goto S, Furumichi M, et al. [2010] KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* 38:D355-60.

[24] Bernthaler A, Muehlberger I, Fechete R, et al. [2009] A dependency graph approach for the analysis of differential gene expression profiles. *Mol Biosyst* 5:1720-31.

[25] Perco P, Mühlberger I, Mayer G et al. [2010] Linking transcriptomic and proteomic data on the level of protein interaction networks. *Electrophoresis* 31:1780-1789.

[26] Köttgen A, Pattaro C, Böger CA, et al. [2010] New loci associated with kidney function and chronic kidney disease. *Nat Genet* 42:376-84.

[27] Levey AS, Astor BC, Stevens LA, et al. [2010] Chronic kidney disease, diabetes, and hypertension: what's in a name? *Kidney Int* 78:19-22.

[28] Wishart DS. [2007] Current progress in computational metabolomics. *Brief Bioinform* 8:279-93.

[29] von Mering C, Jensen LJ, Kuhn M, et al. [2006] STRING 7--recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res* 35:D358-62.

[30] Platzer A, Perco P, Lukas A, et al. [2007] Characterization of protein-interaction networks in tumors. *BMC Bioinformatics* 8:224.

[31] Nibbe RK, Koyutürk M, Chance MR. [2010] An integrative -omics approach to identify functional sub-networks in human colorectal cancer. *PLoS Comput Biol* 6:e1000639.

[32] Rasmussen S, Baas NA, Mayer B, et al. [2001] Ansatz for dynamical hierarchies. *Artificial Life* 7: 329–53.

ABOUT AUTHORS:

Arno Lukas, Managing Partner of emergentec biodevelopment GmbH, Vienna, Austria, holds a PhD in Biochemistry. Arno's main research interests involve translational research towards discovery and validation of novel biomarkers in human disease.

Bernd Mayer, Managing Partner of emergentec biodevelopment GmbH , Vienna, Austria, and Associate at the Institute for Theoretical Chemistry, University of Vienna, holds a PhD in Molecular Biology. Bernd's research interests cover bioinformatics concepts and methods.