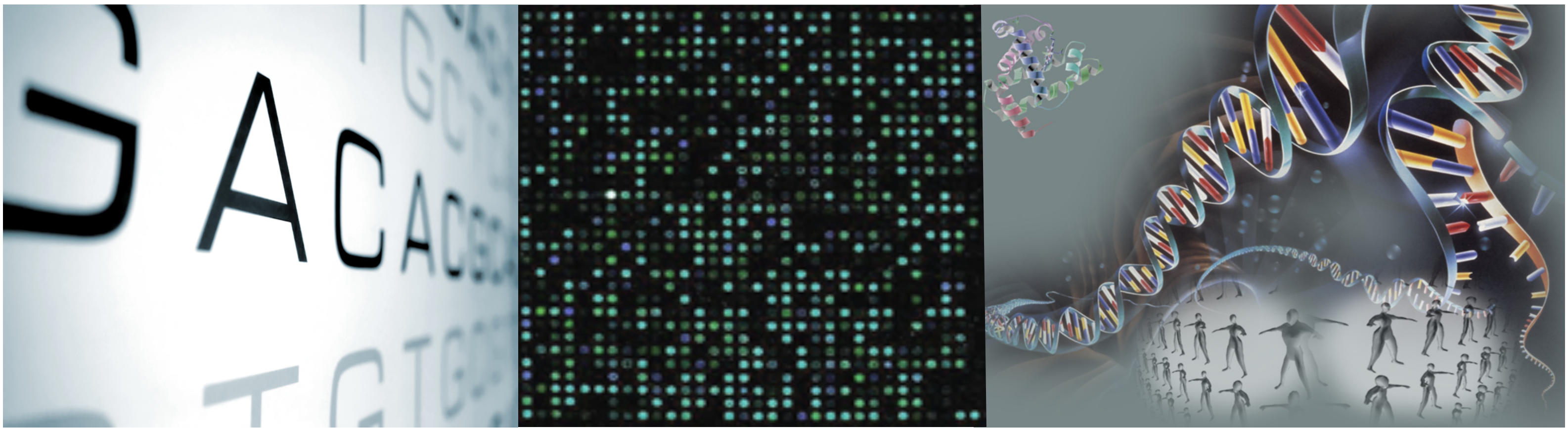
INBIOMEDvision
Promoting and Monitoring Biomedical Informatics in Europe



Strategic Report on Genotype-Phenotype Resources in the European Union

SEVENTH FRAMEWORK PROGRAMME

INBIOMEDvision

January 2012

Executive Summary

Purpose of this document and the endeavour

This strategic report has been prepared in order to assess the opportunities and obstacles that confront us as Europe plans to make full use of the integration of genome-based data resources with resources detailing disease-based and other human phenotypes. It represents the outcome of five hours of intensive discussions within a Think Tank convened at the Thon Hotel, Brussels, Belgium on 5 October 2011 under the auspices of the EU-funded INBIOMEDvision Consortium (ICT-270107; http://www.inbiomedvision.eu).

It represents a consensus among the 17 invited experts who participated. These were drawn from a wide range of backgrounds, including clinicians, engineers, researchers in bio- and medical informatics, industry representatives, and scientists active on translational research projects. The Think Tank was co-chaired by Dr. Scott Boyer (AstraZeneca R&D); Professor Alfonso Valencia (Spanish National Cancer Research Center); and Dr. Nour Shublaq (University College London, UK). The names and affiliations of all the experts are listed in Appendix 1.

The report is divided into two sections. The first section examines, using case studies, the current status of the integration of genotypic and phenotypic data and its use in research. Issues discussed include the differences between the basic biomedical and clinical research communities; the necessity of standardising terminology and the value and use of ontologies; the use of text mining in extracting structured data from health records and the scientific literature; and the use of these resources for patient stratification. The second section focused more practically on how the resources available can be better set up and exploited for clinical use. After exemplifying some resources that are already proving useful and setting out some challenges to their use, the participants ended with a short discussion of the opportunities that will become available as personal genome sequences become cheap enough to be widely available.

I would like to acknowledge the input from the experts, and in particular Dr. Clare Sansom and Ms. Maria Saarela for their assistance in preparing this report. Special thanks to Prof. Peter Coveney for his valuable comments and support throughout.

Dr. Nour Shublaq

University College London

January 2012

Table of Contents

1. Introduction

1.1. Setting the Scene

Today, we are not able to predict the phenotype for an individual based on that person's specific genotype. This report summarises the findings of a Think Tank on "Genotype-Phenotype Resources" that considered the current status and future prospects for databases and models used to predict phenotype from "underlying" genotype, and the value of these predictions for clinical medicine. There is and will continue to be increasing amounts of genotypic information available from clinical studies (e.g. as provided by genome wide association studies (GWAS)), and, similarly, databases of phenotype information at the population level are being made available (see e.g. Pers *et al.* (2011)). However, it is still difficult if not largely impossible to establish any very clear relationships (correlations) between them. Many participants felt that much larger datasets must be collected if such correlations are to be established unambiguously. From this standpoint, data collection and storage must be an integral part of the design of any project. This is likely to continue to cause most problems for rare diseases where the quantities of data available will always be very small and thus prone to unreliability, although clinicians working on rare diseases are likely to be among the early adopters of this technology.

Currently, mining the published literature is easier than mining personal electronic health records for privacy and data protection reasons, although conversely free text is harder than database records to capture and interpret automatically. It should be stated there are problems in the UK to mine the published literature due to copyright laws. Because so much heterogeneous data is needed to establish any form of genotype-phenotype associations, ontologies that link these different forms of data are necessary. Portals to these resources are being developed, by, for example, the US National Institutes of Health (NIH), although these are not always freely accessible. Further investment in both text mining and ontology development is required.

The discussion included both case studies where studies of genotype-phenotype interactions are proving useful (such as in toxicology) and caveats to their development. Among the latter is the role of the environment in the control of gene expression. It is generally assumed that the causal route is mainly from genome to phenome. However, environmental control of gene expression, which is common, is an example of "downward causation" and hence assuming that the genotype controls the phenotype in all cases cannot be correct (Noble, 2008). Nonetheless, the Think Tank concluded that the prospects for the development and application of genotype-phenotype resources are very promising.

1.2. The importance and impact of this Strategic Report

The “genome era” is still less than twenty years old; the first free-living bacterial genome sequences were published in the mid-1990s and the first complete draft of the human sequence less than a decade after that. Now, the publication of even large, complex genomes is almost routine, and a phenomenal amount of information on human genetics, genomics and variation is available for translation into clinical medical applications. There are also many useful resources on disease and other human phenotypes. Linking genotype and phenotype is important in a number of biomedical applications, ranging from early-stage drug discovery to toxicology and clinical trials. In this report we discuss the ways in which genotype and phenotype resources are being linked, how this linked data is used in clinical research and the barriers to its exploitation that still exist. In order to reduce these barriers, recommendations are made for academic and industrial researchers, journal publishers, and funding agencies. A short section towards the end looks forward to an era when personal genome sequencing is so cheap that it is widespread (Wright *et al*., 2011) and suggests that investment in genomic data verification and interpretation is as important as sequencing technology and requires as much investment.

2. State-of-the-art on Genotype-Phenotype Data Integration

The Think Tank started with a general discussion on the need to bridge the gap between the fields of bioinformatics and medical informatics, and difficulties in doing so. Much of this, and the subsequent discussions, were informed by reference to the presentations that had been made at the I-Health conference (http://www.gen2phen.org/i-health2011/), held in Brussels on 3-4 October 2011 immediately preceding this Think Tank. It had been clear that there were frustrations and restrictions involved in bridging the two fields. There is a lack of mutual understanding about what the discipline communities mean when using words such as *ontologies* and *models*, which get used in a slightly different way by the different groups.

Some people think that there is a separate discipline that bridges the gap between bio- and medical informatics that is still undefined. If so, the research areas that would be included in it might include aspects of modelling and genomics. There was little agreement between the I-Health participants on the importance and role of modelling in the clinic. More discussion is therefore needed to discuss the concepts on a more concrete level, and case studies could act as proof-of-concepts here.

It is worth remembering, however, that not all research is capable of being translated into the clinic. It will not be realistic to create a "pipeline" where it is assumed that all bioinformatics research will eventually feed into medical informatics.

Clinicians, however, need solutions for their patients *now*: a timescale that very rarely works in research. It is important to stimulate a dialogue, but dialogue on a very general level has its weaknesses so that a specific case study approach may be preferred. Changes in and (sometimes exponential) improvements to technology and to the amount of data that is being made available present significant challenges, in that that data must be made readily available, properly validated, and that analysis should be completed within a reasonable time.

2.1. Phenotype data commonly integrated with genotype information

Discussion then moved on to the integration of genomic and phenotype information more specifically. The development of databases integrating information on genetic variants with that on traits or diseases (phenotypes) over the last 20-30 years has been outlined by Thorisson *et al.* (2009). A difference has been noted between the "individual patient (clinical approach)" and the "patient population (research approach)". Disagreements between researchers and clinicians may be due to a lack of clarity as to which of these levels is being discussed. The information will of course eventually be used in individual patient cases, but clinical decisions must always be checked using population based statistics.

Statistical modelling is a powerful tool that must be used to link genotype to phenotype. It is useful to validate findings in a single patient against cohort analysis, but there is a caveat in that (as the claims of personalised medicine make clear) it is possible for an individual patient to have a near unique response. These differences will be much more clearly understood in 50-100 years when clinicians will be using models that are fully parameterised and completely understood.

At the present time, however, there are concerns that research data are not of sufficient quality to be used in healthcare settings. A major introspective question within the omics/biomarker field is, "why is it that over the last decade, we are equipped with better instrumentation and omics technology, and yet we have fewer discoveries of clinically relevant biomarkers"? (Anderson, 2010; NexGen Consortium, 2011). And, even if there are new findings, they become irreproducible and non-predictive in another individual or another cohort. Of course, many potential reasons have been discussed, such as overfitting of data and variability in sample/instrumentation quality (Ning and Lo, 2010). At the same time the instrumentation to obtain '-omics' data from individual patients of sufficiently high quality will not be available for at least ten years as the the cost of generating large-scale prospective bio-banks/trials with high-quality outcome data is enormous (Damia *et al.*, 2011).

Hu and Agarwal (2009) have created a network of interactions between genes, diseases and drugs from a large-scale study of transcriptomes, and used this to discover potential relationships between seemingly unrelated diseases, and then to suggest novel uses of existing drugs (known as "drug repositioning or repurposing"). This is where a marketed drug is approved for a new, non-obvious indication, or a drug that was discontinued in clinical trials, for reasons other than safety, is subsequently successfully developed for a different indication. Drug repurposing offers a promising alternative to drug development and population targeting for drugs that already satisfy basic toxicity, ADME1 and related criteria (Ashburn and Thor, 2004). The same process could be useful for the prediction of adverse events of known or novel drugs. Although repositioning of existing drugs has already begun in early 1990, it has only flourished in the last decade. Biomedical literature mining (BLM) composes the most utilised methodology for the computational discovery of new drugs and/or repurposing of old ones. Such approaches typically begin with an analysis of the literature and aim to reveal indirect relationships among seemingly unconnected biomedical entities such as genes, signalling pathways, physiological processes, and diseases. Networks of associations of these entities allow the uncovering of the molecular mechanisms underlying a disease, better understanding of the biological effects of a drug and the evaluation of its benefit/risk profile. However, drug repositioning uncovers the difficulty of translating research results directly into a clinical setting. As an example, a drug may discovered by BLM to have a positive effect on symptoms of post-menopause or premenstrual symptoms but then, biomedical researchers are asked to investigate the cause of this observed effect. In-silico BLM results should then be tested in relevant cellular and animal models and, eventually, in clinical trials (Deftereos, Andronis *et al.*, 2011). Fluoxetine (Prozac), the first-in-class selective serotonin reuptake inhibitor (SSRI), for the treatment of premenstrual dysphoria (Sarafem, Eli Lilly) is an example of repurposed drugs (Netterwald, 2008) that followed this process of development.

[1] An acronym in pharmacokinetics and pharmacology for absorption, distribution, metabolism, and excretion, and describes the disposition of a pharmaceutical compound within an organism

Another case study of drug repositioning has been used to illustrate the difficulty of translating research results directly into a clinical setting. For decades, the estrogen hormone therapy has been the treatment of choice for relieving menopausal related symptoms (Kolsky, 2011). After 2002, hormone therapy got associated with increased risk of heart disease, blood clots, cognitive impairment, stroke etc (National Institute of Health, 2009; Singer and Wilson, 2009). One antidepressant drug, Escitalopram, was recently discovered in a randomised controlled trial setting to have a positive effect on some symptoms (the hot flashes) of the menopause (Freeman *et al.*, 2011). Researchers, mainly of the NIH Menopausal Symptoms Clinical Research Network are now trying to explain this observed effect and investigate another drug with a positive effect on menopause, Paroxetine (Paxil), so either or both is translated into clinical practice (Kolsky, 2011).

Clinical practice, except, possibly, in academic hospitals, is generally problem-oriented rather than discovery-driven. However, many feel that a closer link from the clinic to research would be beneficial. Researchers should be able to pick up research topics and questions from current clinical practice and then feed the results back, just as basic research findings are now routinely applied in clinical practice.

Publication is another key driver for researchers, and journal editors generally choose papers based on research novelty rather than ability to translate into healthcare. Translational research will not become a priority over basic hypothesis-driven research unless the incentives, goals and drivers for that research change. The goals and drivers of basic researchers are obviously not the same as they are in medicine, even in the clinical research community.

Examples were cited of phenotype-genotype resources that they were using successfully, and explained any difficulties that they had had in setting them up. An example was given of a large project, the US Medicare National Pneumonia Project, involving making decisions about whether or where pneumonia patients should be hospitalised, using genomics to examine biomarkers of the patients' prognosis that might help with this decision (Christ-Crain and Opal, 2010). Complex networks and pathways are involved in an example like this. It is now feasible to check tens of markers in a few hours, and the challenge for modellers is how to present the data to make it easy for clinicians to visualise and interpret the important features. Examples of largely genotype-based, largely phenotype-based and combined resources were given. The EU-ADR project has been exploiting clinical data from electronic healthcare records (EHRs) of over 30 million patients to develop an innovative computerised system to detect adverse drug reactions (ADRs) (http://www.ncbi.nlm.nih.gov/pubmed/19745234, http://www.alert-project.org/). Following this approach it will be of utmost importance studying the impact of the genetic traits in these events, which implies access to patients' information and a large set of genetic studies. Conversely, the Phenotypic Disease Network (PDN) (Hidalgo *et al.*, 2009) is a database of links between disease phenotypes derived form the medical history of over 30 million patients. Despite including no genotypic information, it is useful in quantifying disease correlations and progression. Perhaps of more interest, is a so-called "phenome-interactome network": a prediction tool that uses Bayesian statistics to link protein-protein interaction complexes with human disease (Lage *et al.,* 2007).

Mention was made of a largely unsuccessful project that tried to commercialise a system to assess phenotype based on the underlying genotype, in which the users' expectations were too high. On the other hand, 23andme (http://www.23andme.com) has been a big success. Overall, only limited progress in pulling resources that connect phenotype and genotype together is currently possible. The quality of the data is very variable, making it difficult to evaluate the quality of any links made between different types of data. At this moment in time, many in the bioinformatics community agree that we are years away from a generalised and philosophical system that could link genotype and phenotype resources. There are fundamental research problems to be solved in predicting phenotype from genotype before it will be possible to make such a tool available in the clinic.

It is also important to consider that there are cases where the phenotype does not reflect a genotype in any clear way, due, for example, to interactions with the environment. In general, it is unrealistic to expect to correlate genotype with phenotype as there is no one-to-one mapping from genotype to phenotype, and we cannot expect there to be. On one hand, genetic determinism states that an individual's phenotype, including predisposition to disease and response to medication, is wholly determined by his or her genotype (could be referred to as "upward causation"): that, if we only had the knowledge, we could map a person's likely medical history from his or her DNA sequence alone. On the other hand, there are plenty of medically relevant examples which show gene expression being influenced, indeed actually controlled, by an individual's phenotype including the environment. Such "downward causation" (Noble, 2008) is inimical to the conventional molecular biologist's perspective. However, some oncologists speculate that there could be cancers that arise entirely from epigenetic changes in gene expression[2].

Much of the information that has the research community has made available, for example on gene mutations, is not used in the clinic today in any way. Frequently, clinicians do not use or even know the guidelines for the use of research data. Nonetheless, some clinical specialities, such as rare diseases, already make considerable use of genetic research. Researchers have to bear in mind the practical situation in which they are expecting their results to be used in the clinic. Physicians have very little time with each patient and need to be persuaded not only that using genotypic and other data in their practice is beneficial but also that the systems will be easy to learn and use.

Discussion of patient-specific results can be held at two levels, of the individual or of the population. Modelling tends to be applied at the patient population level, but it is important to define the type of patient that is being referred to. Specialist doctors working in genetic medicine and rare diseases will make far more use of research data than those who deal with patients with flu, and it is there that the integration of genotype and phenotype data will be most useful. Doctors who work in research are often able to spend more time with each patient than those in general medicine.

It is important for the research community to introduce new ideas and concepts slowly and earn doctors' trust. In their early forms and stages, knowledge-based systems were

[2] See http://www.ecancermedicalscience.com/blog.asp?postId=174

not easily adopted into clinical practice due to various factors — e.g. lack of adequate technologies, underlying uncertainty in our scientific knowledge, its heuristic-based approach, failure to focus on real clinical problems, and others (Miller *et al.*, 1990). In the case of electronic health records, the development has been slower — in terms of implementation pace — but more successful, in the long term, since its original introduction by scholars such as Weed (1964). Applications that link phenotype and genotype resources will also be adopted in the clinic where effective, but this is bound to take time. There is an analogy with the development of electronic medical records systems: the Regenstreif Institute developed one of the first such systems in 1972, but these did not become widely used until the 2000s. We will be doomed to fail if we try to push too hard to make doctors include all possible genotype and phenotype data in their records; moving step by step will be more beneficial in the long term in order to be able to guarantee the value of such records. The medical informatics community often has more enthusiasm for bridging the gap than the bioinformatics research community that develops the models and methods that they are encouraged to use.

2.2. Ontology standardisation and data input across genotype-phenotype data resources

The word ontology, as used in computer science and related disciplines including bioinformatics, refers to a structured and controlled vocabulary which is used to represent concepts within a specific sub-branch of knowledge. Well-defined ontologies enable different conceptual or theoretical frameworks to be related to one another. Ontologies are generally hierarchical, so a term can be a “parent” and/or a “child” of other terms, and they can be queried rather like databases. The best-known example in the general area of molecular biology and biomedicine is undoubtedly the Gene Ontology (Ashburner *et al*., 2000), which consists of three sets of standard terms covering Cellular Component, Biological Process and Molecular Function. One key advantage of using ontologies is that they can remove the confusion that arises when different but related disciplines (such as bio- and medical informatics) use different terms to refer to the same concept. However, this requires that the ontologies themselves should be standardised and the basic research and medical communities concerned need to decide which ontologies should be used and how. One of the main problems in relation to ontologies is that they are built top-down and thus, do not represent the data a researcher may have in his/her hands. On top of this, we create ontologies based on human-made classes that change with time. It would be beneficial if we could take into account the temporal parameter, which would allow mapping of old ontologies to new ones (e.g. the WHO official classification of disease, ICD-8 vs. ICD-10).

However, “integration of everything” within the biological and clinical environments, even with all concepts defined using ontologies, may ignore the fact that the biological and clinical worlds make different assumptions. Ontologies cannot express common assumptions when there are none. Full integration of the biological and clinical domains of knowledge requires ways to be found to overcome such remaining difficulties. Ultimately, what matters most is the data that these ontologies will be used to annotate and the methods of annotation.

2.3. Genotype-phenotype integration used for patient stratification

Genetic information has been used for a number of years in identifying ("stratifying") subgroups of patients that are likely to respond to a particular treatment or that should be targeted in a clinical trial. This is probably most frequent and most highly developed in the area of oncology; early and very well-known examples include the use of trastuzumab (Herceptin™) for herceptin receptor positive breast cancer only. Several recent clinical trials in oncology have selected patients with solid tumours for inclusion either in the whole trial or in a particular branch based on, for example, their *KRAS* mutation status.

With the almost universal adoption of electronic health records in the developed world an enormous amount of structured or semi-structured information on patient phenotypes is becoming available to the medical community. So far, at least, these resources have been less well exploited for patient stratification than genomic data although papers in this field are beginning to appear. In one interesting recent study (Roque *et al.,* 2011) free text notes and structured diagnosis codes derived from ICD10 were extracted from patient records in a Danish psychiatric hospital setting and used to identify correlations between diseases and thus to stratify patients. It would be possible to use this stratification in clinical trial design. The inevitable full integration of genomic data into electronic health records (see e.g. Hoffman, 2007) is bound to increase the power of this type of information for patient stratification and clinical research even further.

2.4. Ongoing text mining activities on phenotype data resources

Much information on patient phenotypes, in particular, exists in the biomedical literature rather than in electronic health records (EHRs) or in databases. The literature is self-evidently text-based (and often also graphics-based), and both these types of information are harder to extract and manipulate automatically than information that is stored in databases. Automatic extraction is necessary, however, in order for this data to be integrated with the genotypic resources that are typically held in databases. Text mining is one tool that can now be used successfully for automatic extraction of data from the literature. Progress in the use of this technique to "structure [the] unstructured knowledge [within the medical literature]" and to extract pharmacogenomic knowledge has recently been reviewed by Garten *et al.* (2010). Most of the phenotypic data resources are paper-based, so text mining is not always possible.

The way that the medical literature is structured and has developed over the years poses challenges for accurate text mining, and similar challenges are involved with health records, particularly but not only paper-based ones. Different clinicians do not necessarily use the same terms, let alone the same abbreviations, in either patient records or published papers. Doctors often fill out records when they are under pressure so it is not uncommon for these records to contain typographical errors or incomplete and therefore ambiguous sentences. Data protection and privacy policies have to be taken into account when coding information onto health records, as, unfortunately, do legal and insurance liabilities and often financial considerations. All these issues affect the way that clinicians complete their records on a day-to-day basis, and this affects the quantity and quality of information that is available to be mined and combined with genetic information from databases. In a review of intellectual property commissioned by the UK Prime Minister in 2010, Professor Ian Hargreaves (Cardiff University, UK) exemplifies text writing as a technology "which copyright should not inhibit, but does". This raises the problem that in the UK copyright law is an impediment.

It is also often necessary to include phenotypic information from published papers that only exist as graphs or other visual representations (such as images). Clearly a lot of clinical information is most clearly understood in image form. As yet we know that this 2D information can best be represented in and integrated with databases, but more systems are required like "OpenEyes" (http://www.openeyes.org.uk) to make drawing such diagrams easy on a tablet/computer.

2.5. Summary

- The integration of genomic and phenotype (particularly disease-related) resources clearly involves both bioinformatics and medical informatics. Although these disciplines are closely related, the two communities work with different assumptions and on different timescales, and care must be taken in attempting to bridge the gap between them.
- The goals and drivers for the biomedical research community revolve around novelty and the need for publication, while the goals for the clinical community revolve around practical validity and ease of use.
- It must be remembered that not every piece of valid biomedical research will have clinical application (particularly immediate clinical application).
- Systems and, in particular network biology rely on large quantities of "omics" data and this is often not yet fully validated or of sufficiently high quality for clinical use.
- There are a number of useful genotype-phenotype resources in current use, but mostly within the research rather than the clinical community. Attempts to commercialise these have so far been largely unsuccessful.
- Genotype-phenotype resources will be incorporated into clinical practice only slowly and researchers will need to earn doctors' trust for their models and tools. Take-up by physicians working on rare diseases is bound to be faster than by those who mainly deal with common conditions.
- Ontologies can play an important part in standardising terminology and reducing confusion, but this will not enable the complete integration of the biomedical and clinical domains of knowledge.
- Data can be extracted from full-text resources such as scientific journals and electronic medical records using text mining; extracting data from pictorial information is much harder but not impossible.
- Both genotype and phenotype information can be used to stratify patients into groups, for example, for recruitment into clinical trials. Combining these two types of resource is likely to be even more powerful (e.g to enhance the levels of stratification).

3. Challenges and Opportunities for a Better Set-up and Exploitation of Genotype-Phenotype Information Resources

3.1. Promising genotype-phenotype resources for clinical research exploitation

Resources were discussed linking genotypic and phenotypic data that are already available for use by researchers and clinicians. One resource exemplified was UniMed, from the Swiss Institute of Bioinformatics (home of the widely used SwissProt and UniProt protein sequence databases (http://research.isb-sib.ch/unimed/; Mottaz *et al.*, 2008). This resource links the UniProt "protein knowledgebase" to two widely used terminologies of disease: ICD-10, and MeSH, a controlled vocabulary used for indexing biomedical databases including PubMed. This was launched in the Geneva University Hospitals (HUG) 18 months ago although not all of its current features have been available for all that time; its practical clinical benefit has not yet been evaluated. Its main use so far has been in helping physicians to generate hypotheses for clinical research studies. The usability of the tool is currently being investigated by representatives of the molecular diagnosis laboratory of the HUGs.

An application scenario of the use of UniMed to produce a Phenotype to Genotype Association Engine, ADN-Prot (Ruch, 2011) is presented in Appendix 2, as an adjunct to this report.

Resources and methods of genotype-to-phenotype data integration are also available for linking genomics to toxicology. Audouze and co-workers (2010) have constructed a human protein-protein association network (P-PAN) that has revealed previously unknown correlations between chemicals and disease, allowing the prediction of further compounds that may be hazardous to health. The eTOX project, funded through the EU's Innovative Medicines Initiative, is using legacy toxicology reports from the pharmaceutical industry, public-domain data and software tools to predict toxicological properties for candidate small-molecule drugs (Steger-Hartmann, 2011; Krallinger *et al.*, 2011); http://www.etoxproject.eu/). SIDER (Kuhn *et al.*, 2010; http://sideeffects.embl.de) is a database linking commercial drugs to terms describing their side effects. In an interesting study described as "a first step towards computational systems medicine", Chang and co-workers (2010) used a combination of structural and network modeling to predict the side effects of the drug torcetrapib, used to treat hypercholesterolemia, in different genetic backgrounds.

Any databases that contain data on new biochemical compounds may form the basis for useful genotype-phenotype resources. Clinical trial data can also be useful, but this most regularly in developing hypotheses for further research. Some topics are more clearly understood than others: the chemical basis of the interaction between a potential drug and its target may be very well understood, but how the human body viewed as a complex system reacts to such a compound is likely to be understood less well. It is still impossible to understand exactly how a compound will react *in vivo* in humans until it is tested experimentally in clinical trials. And, since an individual's reaction to any drug will be determined in part by genetics, both pharmacogenetics and pharmacodynamics must be taken into account, which implies that clinical trial data from many individuals (patients or volunteers) must be captured. This view is not shared by those who view personalised medicine treatment as more than just looking at average/statistical properties.

Making data from clinical trials widely available to researchers depends on obtaining consent from trial participants. This process is time-consuming, and it may be possible only once a drug has been registered. If complete data from a clinical trial is made widely available – which is certainly not impossible – participants may ask for access to data on their own genotypes. Patients are also likely to be interested in identifying trials where there is a close fit between their own genotypes and the trial requirements (Shublaq *et al*., 2011).

Pharmaceutical companies are increasingly interested in designing studies that are focused on participants' genotypes in order to stratify patients, predicting and then testing which cohorts a particular drug is likely to be most successful in treating. The more focused a clinical study can be made, the more likely the trial is to achieve positive results and, just as importantly, the faster the results can be obtained. Companies are already launching trials in specific cohorts of patients; they are primarily interested in those that are likely to respond positively to a treatment and, conversely, in identifying populations of patients that are more at risk of side effects.

Pharma companies are seldom able to share all the data that have been gathered in clinical trials and if they can, there are often significant delays (at least five years) in releasing even partial datasets. Discoveries being made in industry are made available for clinical research, by making those known in the literature, unless the results are negative. Pharmas are only required to share the fact they are running trials, but not the underlying data, which are gathered under a number of different security agreements.

The use of genotypic data in clinical trials is one of the developments that are tending to bring commercial and public sector clinical research closer together. The objectives of the two communities, however, differ. The industrial sector is obviously driven more by the profit motive, but understanding mechanisms of action is still a key driver for both sectors; the problems that pharma are trying to solve today are multi-disciplinary and cannot be solved by any one entity working alone. A closer model for industry-research interaction is likely to be helpful (Shublaq, 2012).

3.2. Impediments to using particular genotype-phenotype resources

One impediment to using genotype-phenotype resources is the difficulty in finding them. There are a lot of resources and publications available, but it is difficult to locate them. Developing tools to mine the literature and the Internet for appropriate and trustworthy research resources would be a useful way forward. The VPH Toolkit (Cooper *et al*., 2010; http://toolkit.vph-noe.eu/) is a repository and knowledge base of tools, methods and services that are proving useful in research related to the Virtual Physiological Human, and this might prove a model for a similar collection for genotype-phenotype resources.

One such example of a text mining tool, "BIRI" (de la Calle *et al*., 2009), which can be combined with other related tools, such as the one reported by Gupta and colleagues

(Gupta *et al.*, 2008). Another more recent text mining tool is “emir2”, currently under development by members of INBIOMEDvision, which is focused on medical informatics applications. While the development of these kinds of public inventories is actually supported by US agencies in the USA, in topics such as ontologies, nanoresources, biological databases, it seems that there should be additional support for funding such projects within the context of the European Union, too.

In the US, the National Institutes of Health (NIH), the main funder of biomedical research there, is providing flexible, ontology based tools to help both funding agencies and researchers to manage these resources, although these are not yet publicly available.

3.3. The $100 genome

There has been much discussion in the biomedical literature about the impact of the rapidly diminishing cost of whole-genome sequencing. The “$1,000 [human] genome” is now almost a reality, and the cost of sequencing is bound to drop even further. It is already possible to obtain a significant amount of low-quality data on a human genome on a microarray for $100. The question of what to do with an enormous amount of whole-genome information – possibly genotypic data for each of a given set of patients – is now a live one. The community needs to think about how best to take advantage of this, since the cost of data analysis is bound to outstrip the cost of data collection. Tracking drug toxicity was suggested as a useful application of widespread genomic data, as it is accessible and doctors are likely to accept it easily (see e.g. Audouze *et al.*, 2010; Kuhn *et al.*, 2010).

Exome sequencing (the sequencing of the protein-coding regions of genomes, rather than the whole genomes) is faster, cheaper and easier than whole human genome sequencing, and there are already many published whole-exome studies of large numbers of patients with a particular disease (most often a type of cancer; see e.g. Puente *et al.*, 2011; Wei *et al.*, 2011). Exome sequencing is already supplanting microarrays, in that exomes are being read by hi-seq NGS (next generation sequencers) very efficiently now, and generate much more useful data (more detailed and of higher quality) than the microarrays do. Accuracy is based on the number of "reads" performed, and ensuring that errors in the sequencing are kept to a minimal level. A whole genome can be sequenced today in a mere 6 minutes, according to Dr. Ewan Birney (European Bioinformatics Institute). Therefore, we had better record exomes than microarray data where we can. Also, the cheaper and faster the sequencing can be done the more accurate the outcome, since multiple reads can be made much quicker than before, making the exome sequencing technology much more attractive than the much cheaper microarrays.

3.4. Summary

- There are already a number of genotype-phenotype resources available for use by researchers and clinicians, mostly free of charge.
- The area of toxicogenomics is particularly well covered, with several useful resources including the eTOX project having been made available for public use.
- The well-funded eTOX project (from the Innovative Medicines Initiative) is an example of how the EU should be providing infrastructure support for establishing standards and developing ontologies more widely.
- Clinical trials are increasingly using genetic data and it will be very useful to extract both genotypic and phenotypic data from trials for clinical and biomedical research. There are ethical issues here concerned with both privacy and intellectual property, although some of the privacy concerns are likely to be reduced if patients are involved as willing volunteers.
- Patients are likely to be interested in using their own data to plan their treatment, for example in identifying a relevant clinical trial in which they could participate.
- As drug development becomes more complex and inter-disciplinary, pharmaceutical companies are becoming more interested in collaborating with each other and with academia and clinical medicine. Encouraging this collaboration is likely to prove fruitful.
- One of the impediments to the use of genotype-phenotype resources and associated tools is a difficulty in sourcing them and identifying the most suitable ones to use.
- Genome sequencing (and particularly exome sequencing) is rapidly diminishing in cost and time, and it will be important to take advantage of the "$1,000 genome" (let alone the promise of the "$100 genome"). Genome and exome sequencing are already at the forefront of all the world-leading efforts in this space, but more investment in increasing the accuracy and efficiency of high throughput sequencing is required before this technology can replace microarrays, which are cheaper but much less precise.

4. Conclusions and Recommendations

- Members of the basic biomedical, clinical and commercial drug discovery communities should be encouraged to collaborate. Initiatives to support this type of collaboration with or without funding should be encouraged.
- Genotype-phenotype resources will be of value throughout the medical community; they are likely to be taken up most quickly for clinical research and in research-intensive disciplines such as rare diseases. Researchers will need to gain physicians' trust for their models before they will be taken up more widely.
- The use of patient stratification, derived from genotype-, phenotype- or both types of data, is already of value in clinical trials, will become more so, and is bound to improve healthcare provision.
- "Patient Empowerment": Patients should be encouraged to participate in their own treatment and to feel ownership in their data (whether genotypic or phenotypic).
- The EU should provide more infrastructure support for establishing standards and developing ontologies and text mining tools for biomedical and clinical support.
- Models submitted for publication should be more clearly versioned and parameterised.
- One of the major barriers to the use of genotype-phenotype resources in research and the clinic is a perceived difficulty in locating, evaluating and selecting such resources. Tools can be developed to facilitate this, perhaps based on the National Center for Biotechnology Information (NCBI) Phenotype-Genotype Integrator (PheGenI) and the database of Genotypes and Phenotypes (dbGaP), the Virus Pathogen Database and Analysis Resource (ViPR, www.viprbrc.org) sponsored by the National Institute of Allergy and Infectious Diseases, and others that are planned by the NIH in the US.
- Rapid improvements in sequencing technology are not necessarily always accompanied by equally rapid improvements in data analysis techniques. This must be taken into account when planning how best to exploit the coming availability of near ubiquitous genome (or at least exome) sequencing.

REFERENCES

- Anderson N. L. (2010). The clinical plasma proteome: a survey of clinical assays for proteins in plasma and serum. *Clin Chem*. **56**(2): 177-185.
- Ashburn, T. T.; Thor, K. B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery* **3**: 673-683, doi:10.1038/nrd1468.
- Ashburner, M.; Ball, C. A.; Blake, J. A.; and 17 others, for the Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. *Nat. Genet*. **25(1)**: 25-29.
- Audouze, K.; Juncker, A. S.; Roque, FJSSA.; Krysiak-Baltyn, K.; Weinhold, N., *et al*. (2010). Deciphering Diseases and Biological Targets for Environmental Chemicals using Toxico-genomics Networks. *PLoS Comput Biol* **6(5)**: e1000788.
- Chang, R. L.; Xie, L.; Xie, L.; Bourne P. E.; Palsson, B. Ø (2010). Drug Off-Target Effects Predicted Using Structural Analysis in the Context of a Metabolic Network Model. PLoS Comput Biol 6(9), e1000938.
- Christ-Crain and Opal (2010). Clinical review: The role of biomarkers in the diagnosis and management of community- acquired pneumonia. *Critical Care*. **14**:203-214.
- Cooper, J.; Cervenansky, F.; De Fabritiis, G.; Fenner, J.; Friboulet, D.; Giorgino, T.; Manos, S.; Martelli, Y.; Villà-Freixa, J.; Zasada, S.; Lloyd, S.; McCormack, K.; Coveney, P. V. (2010). The Virtual Physiological Human ToolKit. *Phil. Trans. A* **368(1925)**: 3925-36.
- Damia, G.; Broggini, M.; Marsoni, S.; Venturini, S.; and Generali, D. (2011). New Omics Information for Clinical Trial Utility in the Primary Setting. *JNCI Monographs* **43**: 128-133.
- de la Calle, G.; García-Remesal, M.; Chiesa, S.; de la Iglesia, D.; Maojo, V. (2009). BIRI: a new approach for automatically discovering and indexing available public bioinformatics resources from the literature. *BMC Bioinformatics*.**10**:320.
- Deftereos, S. N.; Andronis, C.; Friedla, E. J.; Persidis, A.; Persidis, A. (2011). Drug repurposing and adverse event prediction using high-throughput literature analysis. *Wiley Interdiscip Rev Syst Biol Med*. **3**(3): 323-334. doi: 10.1002/wsbm.147. Epub (2011). Review.
- Freeman, E.; Guthrie, K. A.; Caan, B.; Sternfeld, B.; Cohen, L. S., Joffe, H.; Carpenter, J. S.; Anderson, G. L.; Larson, J. C.; Ensrud, K. E.; Reed, S. D.; Newton, K. M.; Sherman, S.; Sammel, M. D.; LaCroix, A. Z. (2011). Efficacy of Escitalopram for hot flashes in healthy menopausal women: A randomised controlled trial. *JAMA*. **305(3)**: 267-274.
- Garten, Y.; Coulet, A.; Altman, R. B. (2010). Recent progress in automatically extracting information from the pharmacogenomic literature. *Pharmacogenomics* **11(10)**: 1467-1489.
- Gupta, A.; Bug, W.; Marenco, L.; Qian, X.; Condit, C.; Rangarajan, A. and others

(2008). Federated Access to Heterogeneous Information Resources in the Neuroscience Information Framework (NIF). *Neuroinformatics*. **6(3)**:205-217.

- Hargreaves, I. (2011). *Digital Opportunity: A Review of Intellectual Property and Growth*. Report written for UK government, May 2011.
- Hidalgo, C. A.; Blumm, N.; Barabási, A. L.; Christakis, N. A. (2009). A dynamic network approach for the study of human phenotypes. *PLoS Computational Biology* **5(4)**: e1000353.
- Hoffman, M. A. (2007). The genome-enabled electronic medical record. *J Biomed Inform* **40**: 44–46.
- Hu, G.; Agarwal, P. (2009) Human Disease-Drug Network Based on Genomic Expression Profiles. *PLoS ONE* **4(8)**: e6536.
- Kolsky, K. (2011). New NIH research initiative to test treatments for menopausal symptoms.
- http://biggovcare.com/government-health-news/new-nih-research-initiative-to-test-treatments-for-menopausal-symptoms-2
- Krallinger, M.; Vazquez, M.; Leitner, F.; Salgado, D.; Chatr-aryamontri, A.; Winter, A.; Perfetto, L.; Briganti, L.; Licata, L.; Lannuccelli, M. *et al.* (2011) Text mining for drugs and chemical compounds: methods, tools and applications. *Mol. Inf.* **30**:506–519.
- Kuhn, M.; Campillos, M.; Letunic, I.; Jensen, L. J.; Bork, P. (2010). A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology* **6**: 343. See also http://sideeffects.embl.de/
- Lage, K.; Karlberg, E. O.; Størling, Z. M.; Olason, P. I.; Pedersen, A. G.; Rigina, O.; Hinsby, A. M.; Tümer, Z.; Pociot, F.; Tommerup, N.; Moreau, Y.; Brunak, S. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology* **25(3)**: 309-316.
- Mottaz, A.; Yum, Y. L.; Ruch, P.; Veuthey, A-L. (2008). Mapping proteins to disease terminologies: from UniProt to MeSH. *BMC Bioinformatics* **9** (Suppl 5): S3.
- Ning, M. M. and Lo, E. H. (2010). Opportunities and Challenges in Omics. *Transl Stroke Res*. **1(4)**: 233–237.
- Noble, D. (2008). Genes and causation. *Phil. Trans. R. Soc. A* **366**: 3001-3015.
- Miller, R. A.; Masarie, F. E. Jr. (1990). The demise of the "Greek Oracle" model for medical diagnostic systems. *Methods Inf Med*.**29(1)**:1-2.
- Mottaz, A.; Yip, Y. L.; Ruch, P.; Veuthey, A.-L. (2008). Mapping proteins to disease terminologies: from UniProt to MeSH. *BMC Bioinformatics*. **9(Suppl5)**:S3 http://www.biomedcentral.com/1471-2105/9/S5/S3
- National Institute of Health (2009). Summary of research activities by disease category. Biennial Report of the Director. **2**: 1-292
- http://ospa.od.nih.gov/documents/NIH%20Biennial%20FY0809%20Volume%20II.pdf
- Netterwald, J. (2008). Recycling existing drugs. Drug Discovery & Development magazine: 16-22. http://www.dddmag.com/article-drug-repositioning.asp

- NexGen Consortium (2011). Advancing the next generation (NEXGEN) of risk assessment: The Prototypes Workshop. Report by the Office of Research and Development, National Center for Environmental Assessment, U.S., 1-20.
- http://www.epa.gov/ncea/risk/nexgen/docs/NexGen-Prototypes-Workshop-Summary.pdf
- Pers, T.H.; Hansen, N. T.; Lage, K.; and 16 others (2011). Meta-analysis of heterogeneous data sources for genome-scale identification of risk genes in complex phenotypes *Genetic Epidemiology* **35**: 318–332 (GWAS).
- Roque, F. S.; Jensen, P. B.; Schmock, H.; Dalgaard, M.; Andreatta, M.; Hansen, T.; Søeby, K.; Bredkjær, S.; Juul, A.; Werge, T.; Jensen, L. J.; Brunak, S. (2011). Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Computational Biology* **7(8)**: e1002141 (text mining – extract data from health records, map to disease codes, construct phenotypic profiles).
- Puente, X. S.; Pinyol, M.; Quesada, V.; and 62 others (2011). Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* **475(7354)**: 101-105.
- Shublaq, N. (2012). 'Strategic Report for Translational Systems Biology and Bioinformatics in the European Union'. http://www.inbiomedvision.eu
- Shublaq, N.; Sansom, C. (2011). 'Strategic Report for Re-use of Clinical Information in Research in the EU'. http://www.inbiomedvision.eu
- Singer and Wilson (2009). Menopause, as Brought to You by Big Pharma. *New York Times*. http://tinyurl.com/7ofbho6
- Steger-Hartmann, T. (2011). The IMI eTOX Project. AXLR8-2 Workshop Report on a 'Roadmap to Innovative Toxicity Testing'.
- Thorisson, G. A.; Muilu, J.; Brookes, A. J. (2009). Genotype–phenotype databases: challenges and solutions for the post-genomic era. *Nature Rev Genetics* **10(1):** 9-18.
- Weed, L. L. (1964). Medical records, patient care, and medical education. *Ir J Med Sci*. **462**:271-282.
- Wei, X.; Walia, V.; Lin, J. C.; & 11 others, and the NISC Comparative Sequencing Program (2011). Exome sequencing identifies GRIN2A as frequently mutated in melanoma. *Nature Genetics* **43(5)**: 442-6.
- Wright, C.; Pokorska-Bocci, A.; and co-workers (2011). Next steps in sequencing. Report produced by the PHG Foundation, Cambridge, UK. http://www.phgfoundation.org

APPENDIX 1

Venue: *Narvik meeting room, Thon Hotel Brussels City Centre, Brussels, Belgium*

Participants	**Institutional Affiliations**
Dr. Scott Boyer – *co-chair*	Global Safety Assessment, AstraZeneca R&D
Prof. Alfonso Valencia – *co-chair*	Structural Biology and Biocomputing Programme, Spanish National Cancer Research Center, Spain
Dr. Nour Shublaq – *co-chair*	Centre for Computational Science, University College London, U.K.
Dr. Abel Gonzalez-Perez	Research Unit on Biomedical Informatics, Universitat Pompeu Fabra, Spain
Dr. Erik van Mulligen	Biosemantics Group, Erasmus University Medical Center, The Netherlands
Dr. George Potamias	FORTH (Foundation for Research & Technology), Institute of Computer Science
Dr. Guillermo Lopez-Campos	Medical Bioinformatics Department, Institute of Health Carlos III, Spain
Dr. Irene Kouskoumvekaki	Systems Biology, Technical University of Denmark
Dr. José Luis Oliveira	Department of Electronics, Telecommunications and Informatics, University of Aveiro, Portugal
Prof. Lars Juhl Jensen	NNF Center for Protein Research, University of Copenhagen
Ms. Maria Saarela	Fundació Institut Mar d'Investigacions Mèdiques, Spain
Dr. Montserrat Cases	Research Unit on Biomedical Informatics, Universitat Pompeu Fabra, Spain
Dr. Patrick Ruch	Information Science Department, University of Applied Sciences, HEG
Prof. Peter Ghazal	Division of Pathway Medicine, University of Edinburgh Medicine School, U.K.
Mrs. Sandra Pla	Fundació Institut Mar d'Investigacions Mèdiques, Spain
Dr. Stephane Ballereau	European Institute for Systems Biology & Medicine
Dr. Victor Maojo	Biomedical Informatics Group, Universidad Politécnica de Madrid, Spain

APPENDIX 2 Case study of the use of UniMed to produce a Phenotype to Genotype Association Engine, ADN-Prot

Background

The ADN-Prot (Aide au codage Diagnostique pour la Navigation dans Swiss-Prot) application is a joint project between the University Hospitals of Geneva –the largest healthcare institution in Switzerland; the University of Applied Sciences Geneva and the Swiss-Prot group of the Swiss Institute of Bioinformatics.

Objectives

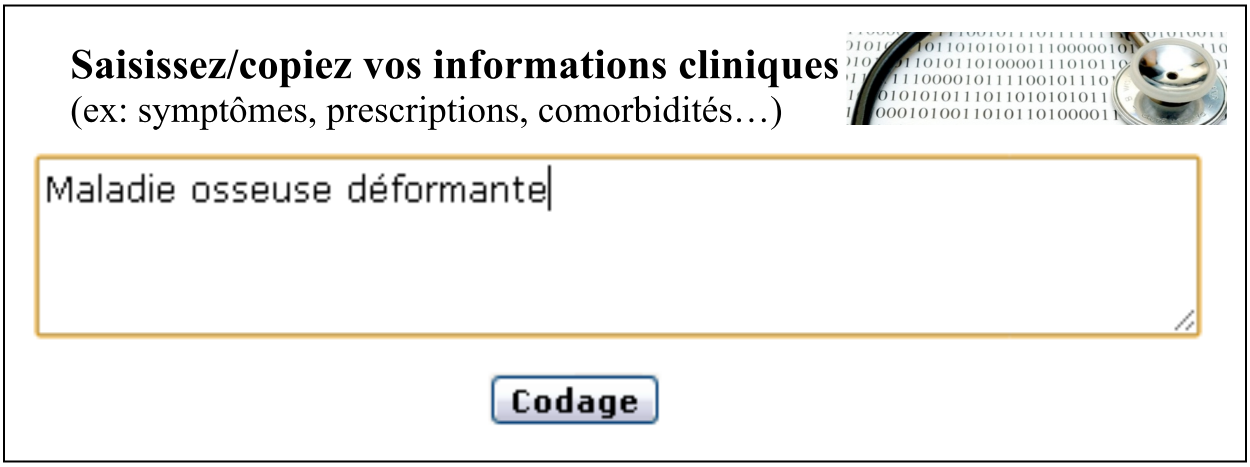
AND-Prot aims at providing a decision support instrument to help physicians and in particular geneticists, to establish navigable associations between any patient data and the Swiss-Prto knowledge bases. Swiss-Prot is used as main entry point to navigate other molecular resources, in particular locus-specific databases such as OMIM, which provides a comprehensive clinical synopsis for genetic diseases and OrphaNet, which adds complementary information (diagnosis tests, references centers, patient associations, clinical trials…) for rare diseases.

Service description

The application gathers the following software and data resources: a set of text mining services (so called ADN) and a transfer table. The UniMed[3] transfer table, whose generation process is described in Mottaz *et al*. (2007) and Mottaz *et al*. (2008), links diagnosis descriptors encoded using the ICD-10 terminology (WHO/French version) and human proteins in Swiss-Prot. The text mining components are standalone web services powered with a multiclass classifier able to associate a ranked list ICD-10 descriptors to any clinical texts. The input text can range from a couple of keywords describing a set of symptoms to a list of clinical reports (pathology reports, discharge letters, pharmacological report about adverse effects…) describing a full episode of care. A typical input output sequence is shown in Figures 1 and 2.

The design and evaluation of the automatic text categorizer is described in Mottaz *et al*. (2008). The case-based categorizer exploits a document/diagnosis database containing more than 100,000 episodes of care and about half a million diagnosis-codes from the University Hospitals of Geneva.

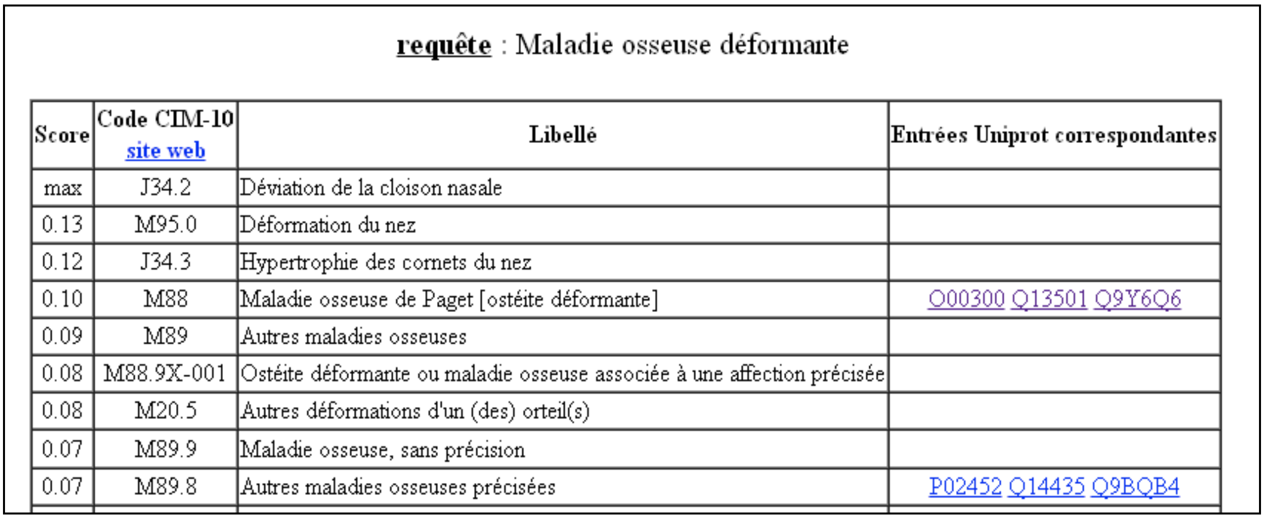
[3] http://research.isb-sib.ch/unimed/



Saisissez/copiez vos informations cliniques
(ex: symptômes, prescriptions, comorbidités…)

Maladie osseuse déformante

Figure 1: Typical example containing a few keywords describing some phenotypic features (in French). This window accepts free-text contents including large document sets or passages of patient records which can be paste in.



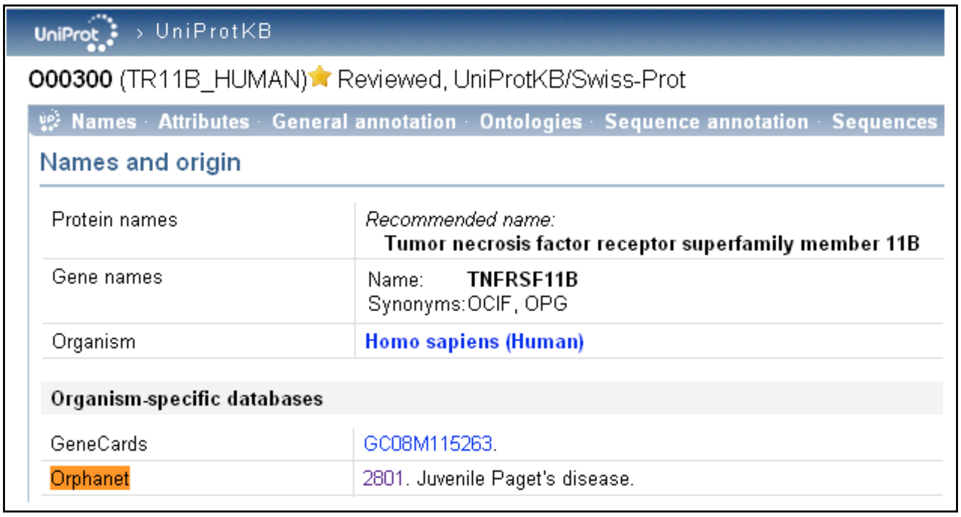
requête : Maladie osseuse déformante

Score	Code CIM-10 site web	Libellé	Entrées Uniprot correspondantes
max	J34.2	Déviation de la cloison nasale	
0.13	M95.0	Déformation du nez	
0.12	J34.3	Hypertrophie des cornets du nez	
0.10	M88	Maladie osseuse de Paget [ostéite déformante]	O00300 Q13501 Q9Y6Q6
0.09	M89	Autres maladies osseuses	
0.08	M88.9X-001	Ostéite déformante ou maladie osseuse associée à une affection précisée	
0.08	M20.5	Autres déformations d'un (des) orteil(s)	
0.07	M89.9	Maladie osseuse, sans précision	
0.07	M89.8	Autres maladies osseuses précisées	P02452 Q14435 Q9BQB4

Figure 2: Output for the text submitted in Figure 1. A ranked list of diagnostic codes (CIM-10, i.e. the French ICD-10) and associated terms (libellé) are returned. The top-ranked descriptors arefrequent comorbidities(e.g. *fracture of the nasal septum*) associated to the phenotypes provided as input to the system; see Figure 1 (*deforming bone disease*). At rank #4, we found the first occurrence of a genetic pathology: ICD-10 code M89 (*Paget's disease of bones*), with three proteins entries annotated with the pathology in the UniMed mapping table. Further down in the list additional Swiss-Prot entries are proposed, e.g. for the ICD-10 code M89.9.

Further navigation

The user can click on the Swiss-Prot accession number to view the curated protein description in the UniProt knowledge base (Figure 3), which works as a hub to integrate high-quality information about genetic diseases such as for instance Orphanet records (Figure 4).

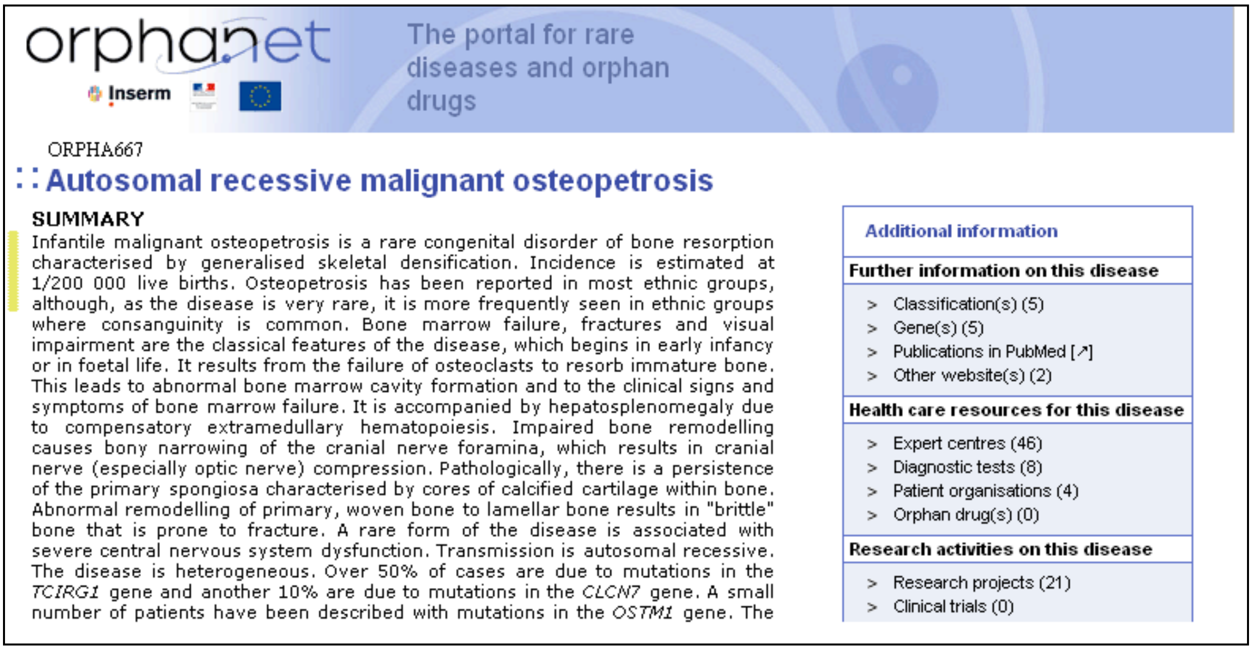

UniProt
› UniProtKB
Names · Attributes · General annotation · Ontologies · Sequence annotation · Sequences

O00300 (TR11B_HUMAN) ★ Reviewed, UniProtKB/Swiss-Prot

Names and origin

Protein names	*Recommended name:* **Tumor necrosis factor receptor superfamily member 11B**
Gene names	Name: **TNFRSF11B** Synonyms: OCIF, OPG
Organism	**Homo sapiens (Human)**
Organism-specific databases	
GeneCards	GC08M115263.
Orphanet	2801. Juvenile Paget's disease.

Figure 3: UniProt entry corresponding to one the Accession number shown in Figure 2 (O00300, *Tumor necrosis factor receptor superfamily member 11B*), referring to the Orphanet entry known as *Juvenile Paget's disease*.



ORPHA667

:: Autosomal recessive malignant osteopetrosis

SUMMARY

Infantile malignant osteopetrosis is a rare congenital disorder of bone resorption characterised by generalised skeletal densification. Incidence is estimated at 1/200 000 live births. Osteopetrosis has been reported in most ethnic groups, although, as the disease is very rare, it is more frequently seen in ethnic groups where consanguinity is common. Bone marrow failure, fractures and visual impairment are the classical features of the disease, which begins in early infancy or in foetal life. It results from the failure of osteoclasts to resorb immature bone. This leads to abnormal bone marrow cavity formation and to the clinical signs and symptoms of bone marrow failure. It is accompanied by hepatosplenomegaly due to compensatory extramedullary hematopoiesis. Impaired bone remodelling causes bony narrowing of the cranial nerve foramina, which results in cranial nerve (especially optic nerve) compression. Pathologically, there is a persistence of the primary spongiosa characterised by cores of calcified cartilage within bone. Abnormal remodelling of primary, woven bone to lamellar bone results in "brittle" bone that is prone to fracture. A rare form of the disease is associated with severe central nervous system dysfunction. Transmission is autosomal recessive. The disease is heterogeneous. Over 50% of cases are due to mutations in the *TCIRG1* gene and another 10% are due to mutations in the *CLCN7* gene. A small number of patients have been described with mutations in the *OSTM1* gene. The

Additional information

Further information on this disease

- Classification(s) (5)
- Gene(s) (5)
- Publications in PubMed [↗]
- Other website(s) (2)

Health care resources for this disease

- Expert centres (46)
- Diagnostic tests (8)
- Patient organisations (4)
- Orphan drug(s) (0)

Research activities on this disease

- Research projects (21)
- Clinical trials (0)

Figure 4: Orphanet record describing the Juvenile Paget's disease with for instance diagnosis tests, when available, and running trials.

References

- Mottaz, Anaïs ; Yip, Yum L. ; Ruch, Patrick ; Veuthey, Anne-Lise. Mapping protein information to disease terminologies. Journal of Integrative Bioinformatics - JIB (ISSN 1613-4516), 4(3), 2007. Special Issue: 4th Integrative Bioinformatics Workshop, Gent, Belgium.
- Mottaz, Anaïs ; Yip, Yum L. ; Ruch, Patrick ; Veuthey, Anne-Lise. Mapping proteins to disease terminologies: from UniProt to MeSH.. BMC Bioinformatics 2008, 9(Suppl 5):S3 (29 April 2008).
- Ruch P, Gobeill J, Tbahriti I, Geissbühler A. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. AMIA Annu Symp Proc. 2008 Nov 6:636-40.