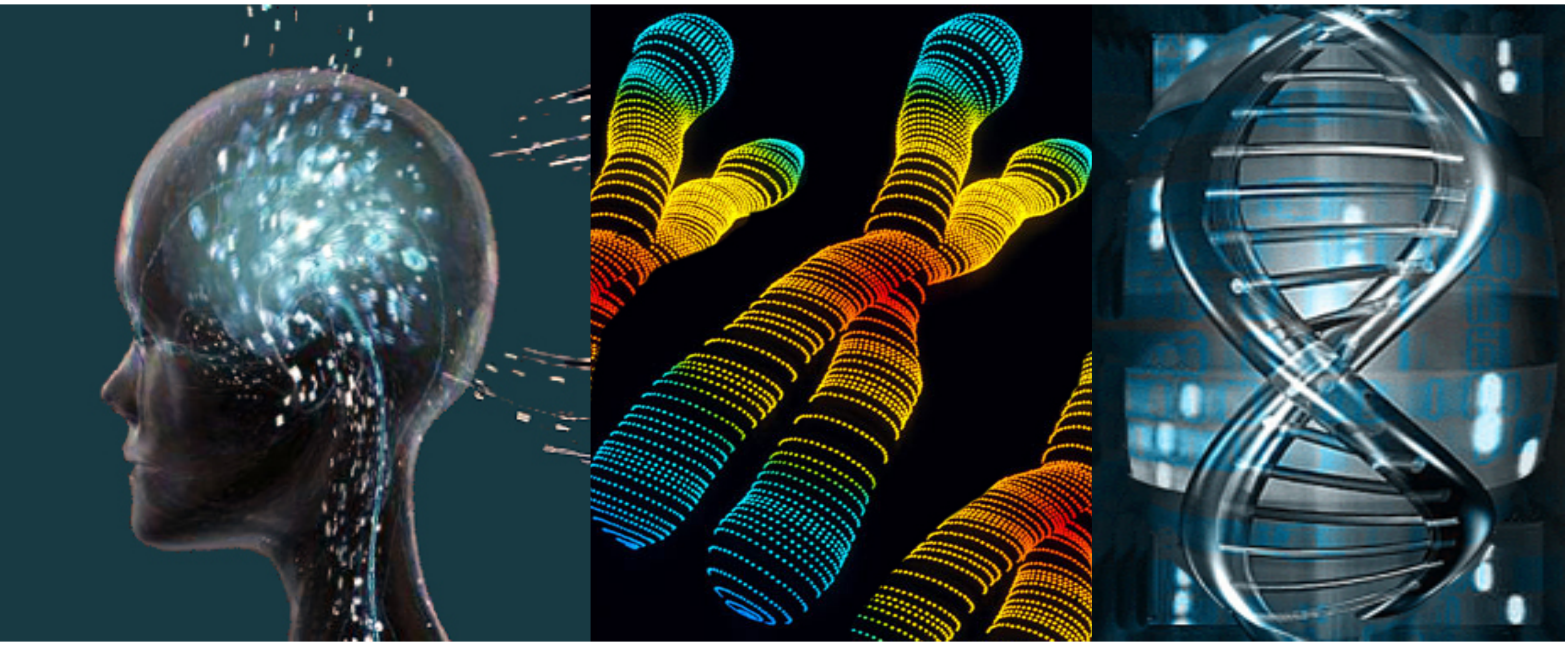
INBIOMEDvision
Promoting and Monitoring Biomedical Informatics in Europe



Strategic Report for Translational Systems Biology and Bioinformatics 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 explores the translational role of bioinformatics and systems biology in drug discovery and clinical medicine. It represents the outcome of 2.5 hours of intensive discussions within a Think Tank convened at the Hospitalet de Llobregat, Barcelona, Spain, on 17 October 2011 under the auspices of the EU-funded INBIOMEDvision Consortium (ICT-270107; http://www.inbiomedvision.eu).

It represents a consensus among the 23 invited experts who participated. These were drawn from a wide range of backgrounds, including academic researchers in bioinformatics and systems biology, translational researchers, industry representatives, and clinicians. The Think Tank was co-chaired by Dr. Nuria Lopez Bigas (Universitat Pompeu Fabra, Barcelona, Spain) and Dr. Nour Shublaq (University College London, UK). The names and affiliations of all the experts are listed in the Appendix.

The report is divided into two sections, each of which consists of a summary of the main discussion of the topic and "boxes" summarising invited statements made to the Think Tank by two expert participants in each case. The first section describes the current status and expected future developments in translational bioinformatics and systems biology. The role of bioinformatics and, particularly, systems and network biology is discussed in connection with drug discovery, with an emphasis on the multi-disciplinary nature of the endeavour and the value of collaboration between academia and industry; and with genomics and personalised medicine. Participants speculated about the types of modelling that are most likely to produce useful results quickly, the so-called "low-hanging fruit". The second section describes the importance of developing standards and methods for validation of data and models as one important way, but by no means the only one of building trust in the clinical community for the systems biology and bioinformatics models that they are being encouraged to use.

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

This Think Tank on “Translational Systems Biology and Bioinformatics” considered various major opportunities to integrate systems biology approaches in our understanding of drug discovery and clinical medicine. Conventional approaches to drug design are foundering within the pharmaceutical industry, which is reflected in a substantial reduction in the number of novel molecules entering the clinic over the past decade. A large number of candidate drugs fail in clinical trials. In the mean time, within the academic sector considerable progress has been made in unravelling the role of complex networks and pathways involving proteins and nucleic acids, typically at the sub-cellular level, in controlling key living processes. Such networks can be very complex but are gradually being elucidated by a combination of experimental (data acquisition) and bioinformatics methods. These networks, which at heart often boil down to representations of interlocking biochemical rate processes, are often able to pinpoint new and previously unforeseen interactions which may suggest new targets for drugs, and/or increase our understanding of “off-target interactions”, facilitating the reduction of undesired side effects. This approach may be of use in developing more personalised approaches to drug treatment, for example in the selection of multi-target therapy or in identifying genomic components of disease development or drug response. As such, the methodology has the potential to underpin much of the new discipline of personalised medicine.

Molecular (and sub-cellular) systems biology is a young science, particularly as it is applied to medicine, and there remains a vast amount of research to be done. However, some examples can be articulated that highlight the promise offered by this new approach in drug design. There is still a gap between basic (academic) and applied (industrial) research in this domain. Data and models should be validated and published to accepted standards if they are to be trusted sufficiently to be adopted by clinicians. If translational systems biology is to make a major impact, the whole process of data access (including access to medical records) needs to be transformed to allow more openness and sharing of information between academia, clinical medicine and industry. Some cultural and ethical obstacles to this openness remain, and regulatory and funding agencies need to help overcoming these.

1.2. The importance and impact of this Strategic Report

Bioinformatics and systems biology are related disciplines that have been identified as being likely to provide useful insights for both clinical research and practice as Europe moves towards the goal of "personalised, predictive and preventative" medicine (ICT Results, 2010). However, effective translation of insights from computer simulation and modelling into the clinic, as opposed to into biological or even biomedical research, will require an increased understanding of the interplay between the roles and values of the bioinformatics and the medical informatics communities, as well as between academic and clinical research and the drug development industry.

This report aims to discuss issues involved in translating academic and to some extent industrial research from a basic research setting into practical clinical use. Following a general discussion of the cultures of the respective communities, their similarities and differences, several important specific topics were addressed including the impact of high-throughput, affordable personal genome sequencing (Wright *et al*., 2011); requirements for models and data to be open, validated, reproducible and accessible; and what this implies for the education and training of clinicians.

2. Translational Bioinformatics and Systems Biology

2.1. Successes and failures in translating systems biology into the clinic

Following initial presentations by Andrea Califano and David Selwood (see boxes) the first question to be addressed was how best to translate the insights gained from systems and network biology into clinical medicine. Selwood offered a disappointing picture of drug development today but this is not necessarily a complete view: some drugs have taken a shorter time (perhaps eight years) to reach the clinic. The biotech sector is currently rather better placed than "big pharma", at least in terms of investment. Systems biology is a young field and there are great, perhaps unrealistic expectations of what it can offer drug discovery. If initial results are disappointing there will likely be a decrease in investment, and we will need to make much more of its successes when they occur.

It is important to realise that there are medical applications of systems biology other than drug discovery. Systems biology can help define biomarkers to make sure that the right existing drug or drugs reach the right patient.

Drugs fail for many different reasons, only some of which are chemical and still fewer to do with the failure of systems biology models. Models are always more simplistic than the biological system that they are used to represent. A study will start by identifying the most important relevant variables before doing detailed modelling. This will involve a narrowing stage in which the number of variables is reduced. A good mathematical model even of a complex biological situation will frequently not have more than 10-20 variables and can, it is hoped, be explained in terms of two or three dimensions. While there is a trend for models to become more complex, the best models are the most relevant rather than the more complicated. We need to become better at building human rather than mouse models.

Many participants stressed the fact that systems biology is just beginning, that it should not be over-sold as it then becomes susceptible to the backlash effect, and that its models are improving all the time. There are simple models that are working well now and we can expect more progress in the next five years.

There was a detailed discussion of new models of collaboration between academia and industry in drug discovery, particularly systems-based drug discovery. Each sector has its own advantages. Pharmaceutical companies need to make (or buy in) new molecules to make a profit, but academics, who are maybe closer to the clinic as compared to pharmas, can make some progress by studying combinations of existing molecules.

Basic research done internally is very expensive for pharma, and many companies are now examining new, more collaborative business models. GSK, Pfizer and AstraZeneca are essentially outsourcing their early stage research to academia. They now have interesting models to fund this. This is involving a lot of re-organisation of existing resources as not all necessary expertise is available in-house. It is useful for pharma to be able to move to sponsor interesting external research quickly. One problem that academics have with industrial collaboration is that the research structures of industry and academia are different. Many academics criticise industrial research as "milestone-based“, or "quarter-oriented" without long term orientation. Companies have difficulty with investment over the timescale necessary in basic research, whereas academics are not constrained by timelines and project management; the barriers for them more often involve shortage of funds. There was some disagreement about whether academic departments can have sufficient access to patients for appropriate clinical research; the position is certainly variable and perhaps only the best-funded universities with hospital links are able to do this.

Internal company policies can also make industry-academic collaborations difficult. Several participants on both the industry and academic sides commented on the amount of bureaucracy involved in setting up collaborations and the differences in organisational ethos. Formal collaboration contracts between organisations such as one cited between Stanford and Pfizer[1] (Ratner, 2011) can help overcome some of these difficulties.

Academic research, or "academic-like" research in industry, involves many different types of expertise, including (but is not restricted to) genomics, medicinal chemistry and clinical research. University-based research typically involves specialised teams but with inter-disciplinary links between them. A question for industry wanting to do pre-clinical research is how to build cross-department collaborations in a similarly flexible way.

Systems biology is clearly linked to personalised medicine, the idea of using the genotype of a patient (or of a tumour), or the phenotype (e.g. the approach adopted by the Virtual Physiological Human), to select the right drug at the right time. This is seen as a top priority in Europe and worldwide. This may make it possible to reposition compounds that have failed in toxicology testing, as a compound that is not effective in a general population might be able to be registered for use in a "stratified" sub-population with a particular genetic profile.

This re-opens the question of the reasons for drugs failing in clinical trials. Toxic side effects are responsible for many such failures and models can be used to try to predict the likely side effects of a drug *in silico* (Pauwels *et al*., 2011) thus saving much of the cost of development. Models have also been used to predict the best combinations of drugs to be used in, for example, anti-retroviral treatment of HIV (Zazzi *et* al., 2011).

The "rescue" of existing or previously failed drugs and the development of new drug combinations is useful, but it is not primarily what systems biology is about. It is more about selecting novel targets and drugs to bind to them, making mechanistic models to understand cause and effect and interactions in biological pathways and then adapting the model system accordingly. This, however, is complicated; genes can have many different functions and be involved in many pathways. We will end up wanting to target a gene that acts in a particular role, and the only way to do this is to target all its partners. Network biology makes this possible. A complete understanding of a network will allow us to target a specific part of that network involving maybe subtle interactions between several genes that are also involved in other pathways. Drug combinations may

[1] See e.g. http://www.lupus.org/webmodules/webarticlesnet/templates/new_newsroomnews.aspx?articleid=1610&zoneid=59

be very useful to target this kind of subtle interaction between genes and the disciplines of proteomics and metabolomics, which are closely linked to systems biology, will be important here too. Therefore, complex systems and network modelling can be important in determining the correct combination of existing drugs.

2.2. Translational bioinformatics and personalised medicine

In the long term – perhaps in twenty years' time – we can expect to have a fairly good idea of all the networks, pathways and potential drug targets involved in human health and disease, and (perhaps beyond then) when whole-genome sequencing has become a routine part of medicine (Wright *et al.,* 2011) we will know how to target effectively each gene in the system. This will make truly personalised medicine possible.

Even now we can see that on an individual level, changes in the regulation of single genes can change the regulation of a network. In many common diseases such as diabetes we are finding through genome-wide association studies (GWAS) thousands of variants that each contribute mildly to the disease associations (see e.g. Pitzalis *et al.* (2008) for Type I diabetes). A further step will be to build a network model showing exactly how all the regulatory factors interact with each other, and differ from one person to the next. Machine learning is likely to be a good approach to use. This type of modelling is complex and we cannot expect doctors to ever understand enough network modelling to do it themselves; we will need to generate tools that they can use in order to recommend a drug or combination of drugs for a particular patient. Classical medicine is already quite "personalised" in that it involves numerous interactions between clinicians and their patients; systems biology is not about removing these interactions as much as about using data and modelling to enhance decision-making.

Networks can complement GWAS and network biology can be used to gather data on and model drug responses, including toxic effects. Gathering the massive amount of data needed for a database of drug responses (Abernethy *et al.*, 2011) and complementing that with the information about the genome when this becomes available will enable links between genotype and phenotype (Shublaq, 2012) to be made based not only on preposition to disease but on the specific response to a drug treatment.

However, there are limitations to genotype-phenotype associations. Even if we collect every single genome in Europe, it is not clear how we will be able to derive all the hoped-for connections. Network biology – building models between the molecular and the cellular levels – is an important part of systems biology, but the term "systems biology" goes wider than this. The INBIOMEDvision project is part of the Virtual Physiological Human, which is concerned with modelling at all levels, from the molecule to the organism and including the environment (as the activity of genes who depends on their environmental context). It involves linking molecules at different levels, working always at the right level to address a particular problem. This may mean simplifying the problem, but making the right simplifications at the right times is one of the key contributions that systems biology can make. Limiting systems biology to sub-cellular network modelling is too narrow. All the data we collect are likely to be valuable, and viewing genotype and phenotype data at the population level will yield important results.

2.3. “Low hanging fruit” of translational bioinformatics

As most of the discussion had concerned the translation of systems biology into the clinic in the medium and long term, participants concluded with a short discussion of the “low-hanging fruit” that might be of use in the clinic in the next 3-5 years. These are most likely to come from the isolation of sub-networks that need fewer resources to study. Isolating a single, specific sub-system that is involved in a medical condition and looking at all its interactions will be the most appropriate way forward to clinically useful applications on this time-scale.

Statement by Andrea Califano, Columbia University, USA

Research and discoveries in biology, and particularly in modelling and computing, have recently developed exponentially; one could say we have lived in the “dark ages” until the last 5 years or so. We are now beginning to understand that the connection between genotype and phenotype might not be direct, but implemented through existing regulatory networks. These networks are specific to cell types. There is no regulatory network for a complete organism, or even for an organ such as the liver, but there is one for, for example, a hepatocyte (the main type of liver cell). The study of these networks is considered to be part of systems biology.

Networks are beginning to be used to investigate disease - not necessarily to develop new therapies, but to define disease sub-types, for example through biomarkers. However, they are so complex that it is difficult to identify suitable entry points for either therapy or biomarker development. Tackling a disease at the level of an individual genetic change is no longer enough. In cancer, for example, the phenotype of each tumour is controlled through a spectrum of genetic and epigenetic changes. Recently published examples of the use of this broader network approach in cancer include Carro *et al.* (2010), Compagno *et al.* (2009) and Real *et al.* (2009).

Much systems biology research is taking place through collaboration between academia and industry. Columbia University has set up a department of Systems Biology at a cost of $46M. The research carried out there is translational, in that all basic research is translated for use in the clinic. Although translational systems biology research is carried out in many clinical areas, it is particularly well developed in oncology; many specialist meetings have been held in this field and the 2011 meeting of the American Association for Cancer Research (AACR) had a focus on systems biology.

Statement by David Selwood, University College London, UK

It is now worryingly clear that during the last 10 to 15 years, despite advances in gene sequencing and all the developments resulting from the publication of the first draft human genome sequence in 2000, we have become worse at drug discovery. Our new discoveries and technologies may be brilliant, but they are not translating into an increasing number of new drugs coming onto the market: in fact, this number is, if anything, decreasing.

There are many reasons for this, and it is probably not the fault of the technology. We are not yet able to understand the complexity of human systems and the limitations of a gene based approach. For example, we might be able to get a good target to develop a new drug from reviewing the literature. Traditionally, this would be tested independently, in a simple "artificial system". In the systems biology paradigm, the drug target is viewed as part of a series of complicated networks. Computational models of networks will help, but even that is not complex enough. Genetic information is not enough to define a protein. We don't know enough about "RNA editing", and other post-translational modifications which can change activity of a protein. Detailed analysis, including "wet" biochemistry, is needed to build confidence in a target even when the way it fits into a network is known. Understanding how a specific protein works within a human tissue is a fantastically complex process but one that is needed if we are to be able to fully exploit that protein as a drug target.

It already costs at least $800M to develop a drug, which is becoming unaffordable. If developments such as systems and network biology fail to make this process more efficient, increasing numbers of companies will drop out of the drug discovery process.

2.4. Summary

- Systems biology is still a young field and there is a long way to go in developing it and learning to apply it in medicine. There are, however, already some very promising emerging technologies.
- Systems biology is perhaps best applied to drug discovery through partnerships and collaborations between academia and industry, although these can often be difficult. Formal agreements at the institutional level may prove useful.
- Systems biology can be used for repositioning or repurposing older drugs and to understand combination and multi-target therapy; network modelling is a useful approach to take here.
- A network-based approach is also useful for understanding side-effects of drugs and in associating genotype with a disease risk, a drug response or a side-effect profile.

- These are some of the hallmarks of personalised medicine, and bioinformatics and systems biology will play a key role in its development, especially in the areas of predictive medicine and biomarker prediction.
- Genetics is not the only determinant of disease progression or drug response, and the role of the environment should not be neglected.
- Network modelling is already important in understanding the results of genome-wide association studies and will become more so.
- The understanding of gene and protein association networks and sub-networks represents some of the “low-hanging fruit” of systems biology that is likely to yield useful clinical results in the next few years.
- Systems biology, however, is much wider than network or even cell modelling and models should be developed at all levels, from the molecule to the whole organism and beyond.
- The form and format of medical records and their access needs to change to encourage the sharing of information across the informatics, basic research and clinical communities and between industry and the public sector.
- Opening up medical records for research use raises many questions to do with data protection, privacy, patient acceptability, intellectual property, etc. Funding and regulatory agencies should be involved in finding answers to these. This has been the subject of INBIOMEDvision Think Tank on the “Re-Use of Clinical Information for Research” (Shublaq *et al*., 2011).

3. Reliability, Reproducibility, Verification and Use of Computational Models

3.1. Mechanisms to close or tighten the loop between research and patient care

The reproducibility and reliability of computational methods, and the need for verification, are clearly very important issues to consider if clinicians and industry partners are to trust and to use these models. On reproducibility, reviewers of papers often ask whether the full methods can be made available. However, it is not always easy or even possible to test these during the peer review process, as, for example, they may be platform dependent. Formal standards have been developed for some methodologies, such as, for example, MIAME (**M**inimum **I**nformation **a**bout a **M**icroarray **E**xperiment) for gene expression studies using microarrays (Brazma *et al.,* 2001). Standardisation of database formats would also be useful.

However attractive this type of standardisation of methodology appears to be, it will be difficult to implement in systems biology modelling. Trying to describe how an algorithm works is immensely time consuming and previous attempts have failed. Making software open source so the code is freely available is one solution, but this may not be feasible. Even if code is freely available, it should be well commented with examples provided for ease of use. One good example is a website for the open source statistical programming language R that includes a manual with sample programs for many common tasks (see e.g. the pyramid plot: http://www.oga-lab.net/RGM2/func.php?rd_id=plotrix:pyramid.plot).

The ENCODE project[2], run by NIH in the USA, has set up a common data analysis and coordination centre. All the data sets must go through the same standard data validation, quality control, and reproducibility procedures, where each replicate has to be submitted independently and those replicates tested against each other based on their reproducibility rate with an assessment score given for each of the data sets. This centralises work that would otherwise be done, and repeated, by different laboratories and groups. The methods and underlying statistics are developed by professionals, each assessment is run automatically, and an assessment given for each dataset. This large-scale project has been found to increase the reproducibility of data analysis.

Ideally, every figure of every publication in the literature that contains data will be associated with the datasets, methods and perhaps code needed to reproduce the figure. This would give access to an increased level of detail for each figure. It is currently possible, but laborious and will not be possible for small laboratories. Working collaboratively and collecting data centrally in a project like ENCODE should allow this to be adopted more widely. The ENCODE standard can be seen as a "quality stamp" but it is a uniquely well-funded organisation and its services are not widely available.

[2] See http://www.ncbi.nlm.nih.gov/geo/info/ENCODE.html and https://www.dtmi.duke.edu/news-publications/research-news/translational-news-archives/nih-approves-new-funding-for-encode

There was some discussion of the extent to which this approach to collecting and analysing experimental data could be applied to regulatory networks and other models. All models, however, are ultimately dependent on experimental data.

Journals also struggle with the issue of how to verify computational results. All scientists in this field learn from experience that many reviewers do not download and test the software that was used, let alone re-analyse data to check the analysis. Many groups are working on issues related to the verification of models, and it will be important to liaise between them. One important issue is getting hold of the data: many fields do not issue guidelines about what raw or processed data should be made available when a paper is published, and standards are only available for a few very popular data types. It might be possible for third parties to be involved in validating data, but the problem of funding still remains: who will pay the third parties to do the validation?

Reproducibility and reliability are not the same as accuracy. Some technologies, such as the Affymetrix chips that are very commonly used for microarray analysis, are recognised to have some bias even though they are very reproducible. It might be useful for journals to introduce a benchmarking system that submitted papers have to go through, although passing this should not be necessarily essential for a paper to be accepted. This might provide a useful comparison between methods. The research community, which includes journal editors, should be responsible for this evaluation.

Much systems biology, however, is so new that standards and benchmarks are difficult to simulate or simply don't exist; the algorithms are still under development. We need to test our own algorithms during the peer review process. Self-assessment of models gives rise to problems in peer review, since journals could be advised not to publish papers if the authors do not provide the code for their algorithms. Some reviewers tend to rate their own algorithms as best in any comparison. It is also possible to over-optimise data and datasets.

However, it is important not to overemphasise assessment and validation. Systems biology models are designed to be used by working biologists and clinicians. In order for this to happen, they must be available to the community in the published literature, and the biologists must choose them based on their utility and ease of use. One simple example of this is BLAST (Altschul *et al.,* 1990) which is the overwhelmingly popular choice of biologists for the similarity scanning of sequence databases.

3.2. Closing the loop between bio- and medical informatics, research and patient care

There was some discussion of similarities and differences between bio- and medical informatics. The main goal of medical informatics – for example, of the people working in the American Medical Informatics Association – is to translate work in bioinformatics and systems biology into clinical practice for patient benefit. Some people working in medical informatics think that these two disciplines have are already merged. This is not necessarily the case, and there is an open question: how are we as a community trying to reproduce all our models using clinical data, medical records and data from real patients. This is the saison d'être of INBIOMEDvision.

One difference between bio- and medical informatics is that medical informatics is not specifically regarded as a research discipline. It is more like "plumbing", in that it is trying to use informatics to make it easier for doctors to make clinical decisions. Doctors want reliable, validated data; they are not usually interested in discovery research. In contrast, bioinformatics is interested in discovery: for example, in using algorithms to discover the gene that controls a particular mechanism. Stanford University, for example, has a Centre for Biomedical Informatics Research which is separate from the medical school. But if bioinformatics is to become clinically relevant, practitioners must always think about the correlation between their discoveries and disease or its treatment. Here it is particularly important to make our findings statistically sound and reproducible.

One of the projects within the Virtual Physiological Human, p-medicine (http://www.p-medicine.eu) is aiming to do the same thing with three sets of clinical trials data for cancer. This involves the integration of a whole range of data sets including but by no means limited to genomics and is trying to use these to make fundamental scientific discoveries. One of the key difficulties with this type of study is with getting hold of sufficient quantities of patient data. Questions of privacy, confidentiality and data protection all need to be addressed here. Procedures for obtaining this data vary from institution to institution, and will be easier in some places than others, but ethical approval will always be needed even for fully anonymised data.

In some bioinformatics fields, such as genomics, there has been enormous progress in recent years in validating methods and making results more reproducible. If these results are to be used in the clinic, however, the judgement of clinicians will always have to be involved. Doctors need interfaces to models that are easy to use, and they need to be confident that the results have been validated. Statisticians and bioinformaticians give results in the form of probabilities, and these are not always easy for doctors to interpret. Another problem is that medical assessments themselves are not always fully reproducible: different clinicians can give different diagnoses based on the same data. Scientists tend to set demands on verification that are higher than what doctors need. Clinicians don't reason in terms of probability and p-values, but in terms of whether they trust the scientists who develop the models they use and the journals that publish them. It would, nevertheless, be useful for clinicians to be given more training in probability and statistics.

There was agreement that the remaining questions about standards and reproducibility can be addressed with the cooperation of journal editors. It should be possible to insist that, for a large majority of systems biology papers published in high profile, high impact journals, only methods and models that are reproducible are acceptable. Medical informatics has different requirements for reproducibility and precision than bioinformatics but the bio- and medical informatics communities are to be brought closer via the efforts of INBIOMEDvision.

Statement by Gustavo Stolovitzky, IBM Computational Biology Center, USA

As scientists, we should always be conscious of the many things we do not know yet and remain aware that it is possible to make mistakes. Some of our distinguished predecessors have made very conspicuous ones, proposing the existence of æther (Aristotle, 3rd century BC); phlogiston (C17) and more recently cold fusion (1989). In order to understand the extent to which our research is correct we need to verify our results using the highest possible standards. The process of verification is often incorrectly thought only to refer to the acceptability of results, and not to their validity.

Peer review is the most common process through which scientific results are verified. Although this is in general considered as objective and non-biased, it cannot be perfect, and it is getting harder, in part because the number of journals is proliferating. There is a particular problem with, for example, a network model that makes thousands of predictions about gene or protein interactions. How do we know that the way the network was created based on experimental data and algorithms applied to analyse it, is correct? It is simply impossible for peer review to test each of those predictions.

One process that has recently assisted peer review is associated with the notion of "collaborative competition” (Meyer *et al*., 2011), as exemplified by CASP, which is used to assess the accuracy of protein structure prediction techniques (Cozzetto *et al*., 2009). The community is asked to predict the structures of defined protein sequences, and the predictions are assessed once the experimental structure is known. This is run every two years and has helped to verify the prediction software and assess how it is developing. Similar competitions are used in other fields, including systems biology; DREAM (Dialog for Reverse Engineering Assessments and Methods, http://the-dream-project.org) has offered “prediction challenges in systems biology” each year since 2006 (see for example (Stolovitzky *et. al*, 2009; Prill *et al*., 2010). This type of assessment sits mid-way between self-assessment and, for example, clinical trials; DREAM has also proven that the concept of crowd-sourcing can yield novel biological insight, for example in postulating new signaling pathways that were previously unrecognised (Prill *et al*., 2011). In summary, the application of crowd-sourcing and collaborative-competition in systems biology in projects such as DREAM can be a means to validate methodologies thereby assisting peer review, to create community generated hypotheses and to nucleate the community of systems biologists around important and yet unsolved problems of the day.

Statement by Jaap Heringa, Free University, The Netherlands

Computer modelling is not intuitive in the same way that human actions can be. We build models, we use them, but we may not understand the details of how they work internally. This is more serious for people who use models others have developed. New data and new concepts are continually arising, and there are consequently a lot of issues to consider, particularly in multi-scale modelling. We need to build as accurate and reliable a model framework as we can, but this is difficult when we lack many details. We are looking to develop personalised models within a timeframe of, say, 10 years and the extent to which this will be possible is not yet clear.

Specific issues that need to be addressed in building complex computational models include:

- The need for technical correctness: how do we prove that the model is really doing what it should be doing?
- The importance of mining the scientific literature.
- The value of a unified framework in which disparate data sources can be integrated and combined with data from the literature.

3.3. Summary

- Standardisation of the methodologies and databases used in the construction of models will be useful in order for these to be fully evaluated by (e.g.) journal editors and the reviewers of submitted papers.
- Some sub-disciplines within bioinformatics already employ standards that fulfil a similar purpose, such as the MIAME standards for publication of microarray experiments, and these may provide a useful model.
- However, standardisation will be difficult to implement in systems biology modelling, at least without access to model source codes. The use and development of open source software should be encouraged but cannot be enforced.
- Publishing data and methodologies in enough detail for results to be validated will be laborious particularly for small, poorly funded groups. Centralisation of data validation and quality control may be helpful; the NIH ENCODE project is a good example of this.
- Ensuring the reproducibility and reliability of data is not the same as ensuring its accuracy as it does not take systematic errors into account.
- Many systems biology programs and methodologies are still under development, and validation methods can be expected to evolve along with them.
- There is a risk that too much emphasis is put on standards and validation of models. In the long run, models will be used if the biological and clinical community find them useful, much as BLAST has become one of the most widely used bioinformatics tools. In other words, the best tools will be selected on the basis of their reliability.
- Translating bioinformatics and systems biology into useful clinical tools requires an understanding of the differences between disciplines and communities. While bioinformatics and systems biology are pure research disciplines, medical informatics is more concerned with informatics primarily for patients.
- It is self-evident that, as far as possible, models and results that are to be applied in the clinic need to be sound; more training for clinicians in probability and statistics would be useful in order to help them assess how reliable such tools may be.
- Editors and their editorial boards have responsibility to help ensure that data and models published in their journals are reproducible, reliable and trustworthy.

3.4. Conclusions and Recommendations

- There is considerable potential for systems biology, as there is for bioinformatics, to contribute to drug discovery. This can best be achieved through collaborations between academia and the pharmaceutical and biotechnology industries, and these should be encouraged.
- Applications of systems and network biology that ought to be encouraged include repositioning drugs for new indicators; understanding interactions between genes and disease; and investigating the mechanisms of toxic side effects.
- Systems and network modelling can and should be applied to the development of personalised medicine, particularly as and when personal genome sequencing becomes more widespread.
- The importance of the environment in determining disease progression and drug response should not be neglected.
- Although network and sub-network modelling is likely to provide some of the first clinically useful results, it should be remembered that systems biology is much wider than network biology as usually understood.
- Medical records should provide much useful information for translational bioinformatics and systems biology as long as the data is sharable. These records may need to be modified to encourage this sharing, and ethical issues must be taken into account.
- The development of standards and validation methods for systems biology models should be encouraged.
- Less well-funded research groups will need to be supported to meet these standards.
- Although standards are important, they should not be over-emphasised. In the long term, the value of a translational systems biology model will be based whether and how it is used in the clinic.
- Training for clinicians in statistics, probability and data evaluation should be improved and made more widespread.
- Journal editors and editorial boards have a responsibility to help develop, encourage and support data standards and validation.

REFERENCES

- Abernethy, D. R.; Bai, J. P.; Burkhart, K.; Xie, H. G.; Zhichkin, P. (2011). Integration of Diverse Data Sources for Prediction of Adverse Drug Events. *Clinical Pharmacology & Therapeutics* **90(5)**: 645-646.
- Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215(3)**: 403-410.
- Brazma, A.; Hingamp, P.; Quackenbush, J.; and 21 others (2001). Minimum information about a microarray experiment (MIAME) — toward standards for microarray data. *Nature Genetics* **29**: 365 – 371.
- http://www.mged.org/Workgroups/MIAME/miame.html
- Carro, M. S.; Lim, W. K.; Alvarez, M. J.; Bollo, R. J.; Zhao, X.; Snyder, E. Y.; Sulman, E. P.; Anne, S. L.; Doetsch, F.; Colman, H.; Lasorella, A.; Aldape, K.; Califano, A.; Iavarone, A. (2010). The transcriptional network for mesenchymal transformation of brain tumours. *Nature.* **463(7279)**: 318-325.
- Compagno, M.; Lim, W. K.; Grunn, A.; Nandula, S. V.; Brahmachary, M.; Shen, Q.; Bertoni, F.; Ponzoni, M.; Scandurra, M.; Califano, A.; Bhagat, G.; Chadburn, A.; Dalla-Favera, R.; Pasqualucci, L. (2009). Mutations of multiple genes cause deregulation of NF-kappaB in diffuse large B-cell lymphoma. *Nature.* **459(7247):** 717-721.
- Cozzetto, D.; Kryshtafovych, A.; Tramontano, A. (2009). Evaluation of CASP8 model quality predictions. *Proteins* **77** Suppl 9: 157-166.
- DREAM "competition" – (http://the-dream-project.org)
- ENCODE – (http://www.ncbi.nlm.nih.gov/geo/info/ENCODE.html**;** https://www.dtmi.duke.edu/news-publications/research-news/translational-news-archives/nih-approves-new-funding-for-encode)
- ICT Results (http://ec.europa.eu/ictresults) (2010). A healthy approach: Technology for personalised, preventative healthcare.
- Meyer, P.; Alexopoulos, L. G.; Bonk, T.; Califano, A.; Cho, C.; de la Fuente, A.; de Graaf, D.; Hartemink, A. J.; Hoeng, J.; Ivanov, N. V.; Koeppl, H.; Linding, R.; Marbach, D.; Norel, R.; Peitsch, M. C.; Rice, J. J.; Royyuru, A.; Schacherer, F.; Sprengel, J.; Stolle, K.; Vitkup, D.; Stolovitzky, G. (2011). Verification of systems biology research in the age of collaborative competition, *Nature Biotechnology*. **29(9)**:811-815.
- Pauwels, E.; Stoven, V.; Yamanishi, Y. (2011) Predicting drug side-effect profiles: a chemical fragment-based approach *BMC Bioinformatics* **12**: 169.
- Pitzalis, M.; Zavattari, P.; Murru, R.; *et al.* (2008). Genetic loci linked to type 1 diabetes and multiple sclerosis families in Sardinia. *BMC Med Genet.* **9**: 3.
- Prill, R.; Marbach, D.; Saez-Rodriguez, J.; Sorger, P.; Alexopoulos, L.; Xue, X.; Clarke, N.; Altan-Bonnet, G.; Stolovitzky, G. (2010). Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS One*. **5(2)**:e9202.

- Prill, R. J.; Saez-Rodriguez, J.; Alexopoulos, L. G.; Sorger, P. K.; Stolovitzky, G. (2011). Crowdsourcing network inference: the DREAM predictive signaling network challenge. *Science Signaling*. **4(189)**:mr7.
- Ratner, M. (2011). Pfizer reaches out to academia – again. *Nature Biotechnology*. **29**: 3-4.
- Real, P. J.; Tosello, V.; Palomero, T.; and 12 others (2009). Gamma-secretase inhibitors reverse glucocorticoid resistance in T cell acute lymphoblastic leukemia. *Nat Med.* **15(1):** 50-58.
- Shublaq, N. (2012). 'Strategic Report on Genotype-Phenotype Resources 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)
- Stolovitzky, G., Prill, R. J.; Califano, A. (2009). Lessons from the DREAM2 Challenges. *Ann N Y Acad Sci*. **1158**:159-195.
- 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)
- Zazzi, M.; Kaiser, R.; Sönnerborg, A.; Struck, D.; Altmann, A.; Prosperi, M.; Rosen-Zvi, M.; Petroczi, A.; Peres, Y.; Schülter, E.; Boucher, C. A.; Brun-Vezinet, F.; Harrigan, P. R.; Morris, L.; Obermeier, M.; Perno, C. F.; Phanuphak, P.; Pillay, D.; Shafer, R. W.; Vandamme, A. M.; van Laethem, K.; Wensing, A. M.; Lengauer, T.; Incardona, F. (2011). Prediction of response to antiretroviral therapy by human experts and by the EuResist data-driven expert system (the EVE study). *HIV Med*. **12(4)**:211-218. doi: 10.1111/j.1468-1293.2010.00871.x. (2010) Epub. *PubMed PMID*: 20731728.

APPENDIX

Venue: *Pisa meeting room, Hotel Hesperia Tower, Bellvitge, Barcelona, Spain*

Participants	Institutional Affiliations
Dr. Nuria Lopez Bigas – co-chair	Research Programme on Biomedical Informatics, Universitat Pompeu Fabra, Spain
Dr. Nour Shublaq – co-chair	Centre for Computational Science, University College London, U.K.
Dr. Martha Bulyk	Division of Genetics, Department of Medicine at Harvard Medical School, and Brigham & Women's Hospital, Boston, U.S.
Prof. Andrea Califano	Columbia University, U.S.
Dr. Raymond Cho	Genetics, University of California Medical Center, U.S.
Prof. Peter Coveney	Centre for Computational Science, University College London, U.K.
Dr. Diana de la Iglesia	Universidad Politécnica de Madrid, Spain
Prof. Dr. Jaap Heringa	Bioinformatics Section, Department of Computer Science, Free University, The Netherlands
Dr. Manolis Kellis	Computer Science and Electrical Engineering Department, Massachusetts Institute of Technology, U.S.
Dr. Irene Kouskoumvekaki	Systems Biology, Technical University of Denmark
Dr. Victoria López Alonso	Carlos III Health Institute, Madrid, Spain
Dr. Miguel Ángel Mayer	Universitat Pompeu Fabra, Spain
Prof. Yves Moreau	Engineering, Catholic University of Leuven, Belgium
Dr. Erik van Mulligen	Biosemantics Group, Erasmus University Medical Center, The Netherlands
Dr. Sylvia Plevritis	School of Medicine, Stanford University, U.S.
Ms. Maria Saarela	Fundació Institut Mar d'Investigacions Mèdiques, Spain

Dr. Julio Saez-Rodriguez	European Bioinformatics Institute, U.K., and EMBL Genome Biology Unit, Germany
Prof. David Selwood	Medicinal Chemistry, University College London, U.K.
Dr. Gustavo Stolovitzky	IBM Computational Biology Center, U.S.
Dr. Gary Stormo	Department of Genetics, School of Medicine, Washington University, U.S.
Mrs. Sandra Pla	Fundació Institut Mar d'Investigacions Mèdiques, Spain
Dr. Dennis Vitkup	Center for Computational Systems Biology and Bioinformatics, Columbia University, U.S.