ISBE Infrastructure for Systems Biology Europe

WP9

Technology and Science Watch

ISBE WP9 REPORT

Continuous technology forecasting report

D 9.2

Committee members as co-authors

Rüdiger Ettrich

Dimitris Thanos

Sarah Butcher

Marcela Kotrcova

Natalia Stanford

Carole Goble

Angela Oberthuer

Thomas Hoefer

Project ref. no.	INFRA-2012-2.2.4: 312455
Project title	ISBE – Infrastructure for Systems Biology Europe
	R= Report
Contractual date of delivery	Month 21
Actual date of delivery	Month 21
Deliverable number	D9.2
Deliverable title	Continuous technology forecasting report
Dissemination Level	PU
Status & version	Version 1
Number of pages	25
WP relevant to deliverable	WP9
Lead Participant	C4SYS, prof. RNDr. Rüdiger Ettrich, PhD.
Author(s)	Rüdiger Ettrich, Sarah Butcher, Dimitris Thanos, Marcela Kotrcova, Natalia Stanford, Carole Goble, Angela Oberthuer, Thomas Hoefer
Project coordinator	Richard Kitney
EC Project Officer	Andreas Holtel

Dissemination level: PU = Public, RE = Restricted to a group specified by the Consortium (including Commission services), PP = Restricted to other programme participants (including Commission Services), CO= Confidential, only for members of the Consortium (including the Commission Services)

Nature of Deliverable: P= Prototype, R= Report, D=Demonstrator, O = Other.

ISBE Infrastructure for Systems Biology Europe

TABLE OF CONTENT

I. INTRODUCTION

This second continuous technology report is the outcome of the joint effort of the WP9 members and the Technology and Science Watch committee appointed by the steering committee appointed in April 2014. While the first report was designed to serve as a guide for building up the infrastructure, this second report takes into account the recommendations of the SAB that sees the possible role of the future infrastructure in fostering systems biology research by using existing experimental facilities to participate in systems biology experiments guided by the modelling canters. This report shall guide the interaction/discussion with the RI's that potentially could provide the required experimental methodologies or technologies that the modelling and data stewardship canters should be able to utilize for designing systems biology experiments. It will not recommend technologies for building up an infrastructure. Additionally to highlighting technologies or science demands coming from the scientific community, we also provide a fine-grained view of the appointed experts within the Science and Technology watch committee for their relevant fields. Thus the current report gives a first global overview of the existing state-of-the-art of molecular systems biology with respect to technology or methodology and possible near future directions. A further report will include also physiology and a broader view of systems biology.

METHODOLOGY

Systems biology research needs collection and processing of data from large numbers of biological experiments using automated procedures and requires the ability to obtain, integrate and analyze complex data sets from multiple experimental sources using interdisciplinary tools.

Technically, the report represents the essence from

- tech literature watch,
- reports of the Science and Technology Watch Committee members (see appendix),
- a series of interviews held with scientists from Europe and the United States that work at the forefront of systems biology and are regularly invited as plenary speakers to systems biology conferences
- data gained from the ISBE-wide survey – currently updated
- data obtained from a broad analysis of recent conference proceedings and abstracts.

In the report we address the following fields:

Modelling, microscopy & image analysis, live single-cell imaging & modelling, mass spectrometry, proteomics, RNAi screens, genomics & sequencing, metabolomics.

A. OBJECTIVES

➔ Examining and evaluating of the existing state-of-the-art of available technologies

➔ Determining whether future technological and scientific developments in the scientific areas of systems biology should integrate these new technologies

Science and Technology Watch Committee

David J. Duffy, Postdoctoral Researcher with Systems Biology, Ireland

Elaine Kenny, Co-founder of SME called Elda Biotech, Ireland

David Dobnik, National institute of biology Ljubljana, Slovenia

Paul Hitchin, Faculty of Natural Sciences, Department of Life Sciences, Imperial College

Ruedi Aebersold, ETH Zurich, Switzerland

Nicolas Schauer, CEO of Metabolomic Discoveries, Potsdam-Golm, Germany

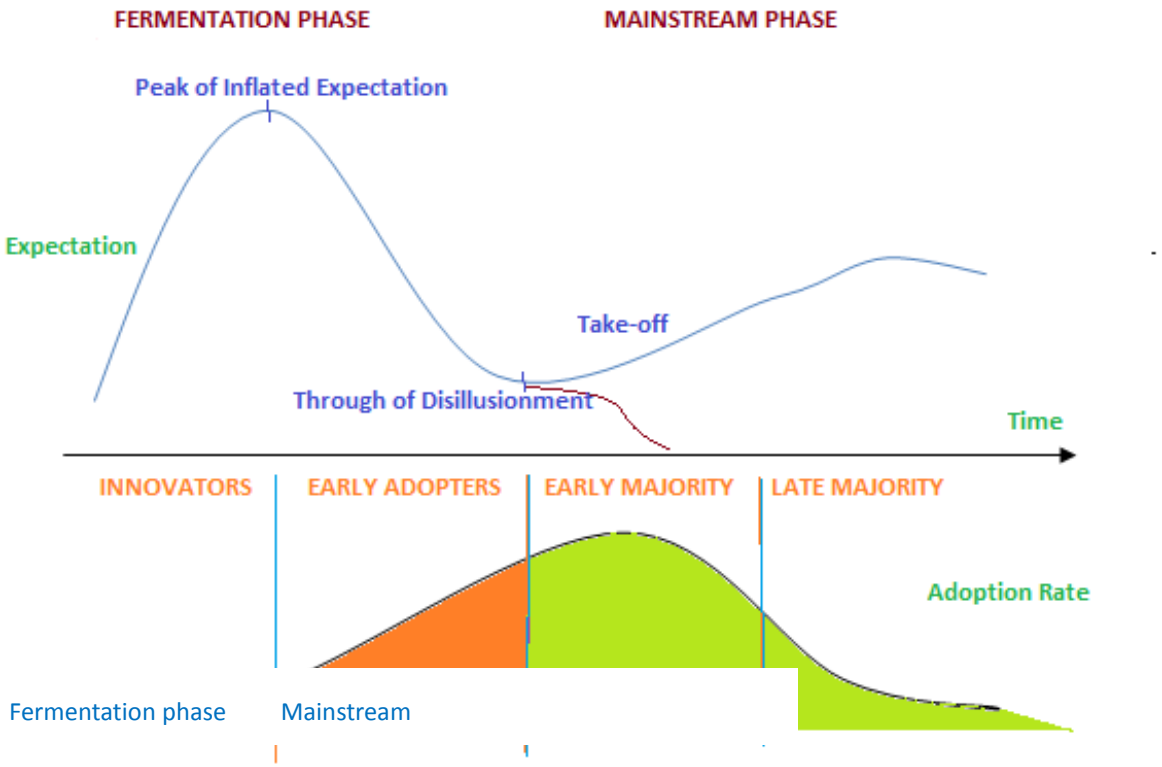
Holger Erfle, Head of the BioQuant RNAi screening Facility, Germany

Martin Spitaler, Facility for Imaging by Light Microscopy [FILM], Imperial College London

Lan K. Nquyen, Systems Biology at Conway Institute, Ireland

II. THE REPORT

Every new technology that finds its way into science follows a sigmoidal curve, first we have a fermentation phase in which only a small number of scientists are using the respective technique, then we see a take-off, when the technique becomes generally accepted and becomes mainstream, followed by the consequent stagnation when the technique reaches maturity.



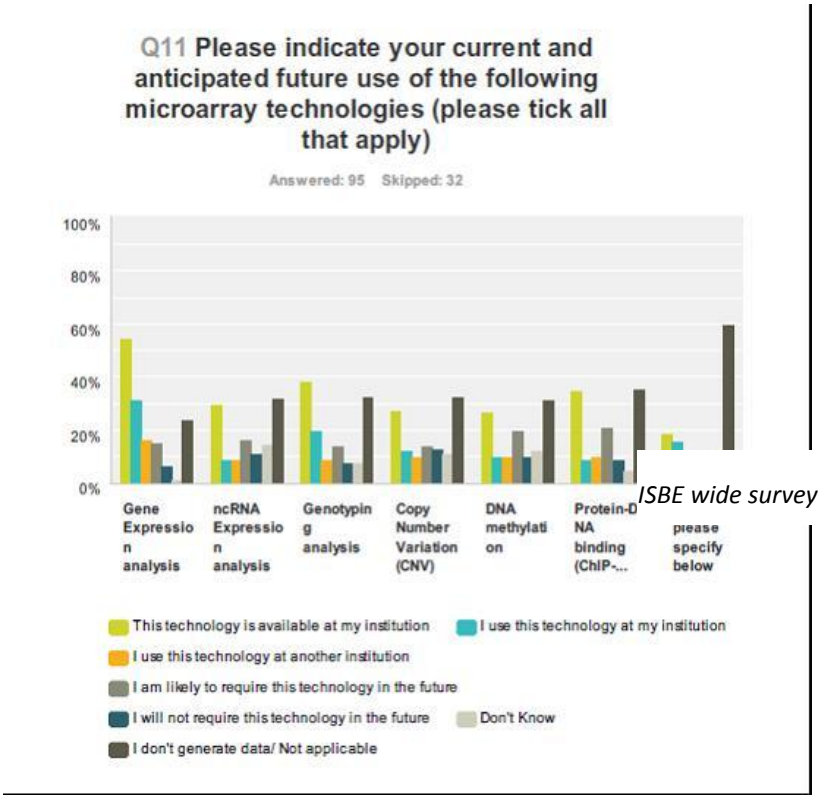
Expectation and Technology Adoption Lifecycle plotted together

This then can be followed by either discontinuity, when the technique is replaced by a newer one, or by an extension, in case the technique is modernized or upgraded. Hereby it is important to note that the revolutionary step, in which a new technique is invented doesn't necessarily change the market and leads to the take-off, but is in general much earlier and is followed by the fermentation that can take a significantly long time. The disruptive step then is not the invention itself but the time when the science is ripe for the new technique and demands it. This makes it extremely difficult to predict which of the new techniques that are developed and reported daily does actually make it to the take-off and will be demanded by the systems biology community. This will therefore be the most difficult task of the science watch committee and our continuous forecasting reports.

A. MICROARRAY TECHNOLOGIES

Microarrays technologies can be widely applied at Gene Discovery, Drug Discovery, Disease Diagnosis and Toxicological Research. High-throughput transcriptomics technologies (like microarrays or RNA-seq) usually require additional confirmation of results. This is mostly done by real-time PCR only on few of the selected candidate genes. Recently, with the development of digital PCR (dPCR) this kind of confirmations could gain on throughput. An example could be the Biomark HD System from Fluidigm, which enables performing real-time PCR reactions for 96 samples, where each sample is tested for 96 assays. Other dPCR platforms available are of course state-of-art technology with respect to absolute quantification, however throughput in terms of samples and assays is not as high as with Biomark HD. With high probability in near future also other platforms for dPCR will become handier for handling a systems biology experimental setup (Dobnik, appendix I).

In reference to our survey, while a large number of **microarray technologies** is available in a majority of institutions that conduct systems biology and in principal would cover the demand fully (Copy number variant, Gene Expression analysis, Genotyping analysis, ChIP chip) for some of the technologies, as is gene expression analysis, more than 16% of survey respondents use this technology at a different institution and additional 15% require the usage in the future. There are a couple of reasons for having the experiments done externally, which are accessibility, costs, bureaucracy (internal), easier to use (some private companies include basic analysis). This might point on a potential mediating role for the infrastructure.

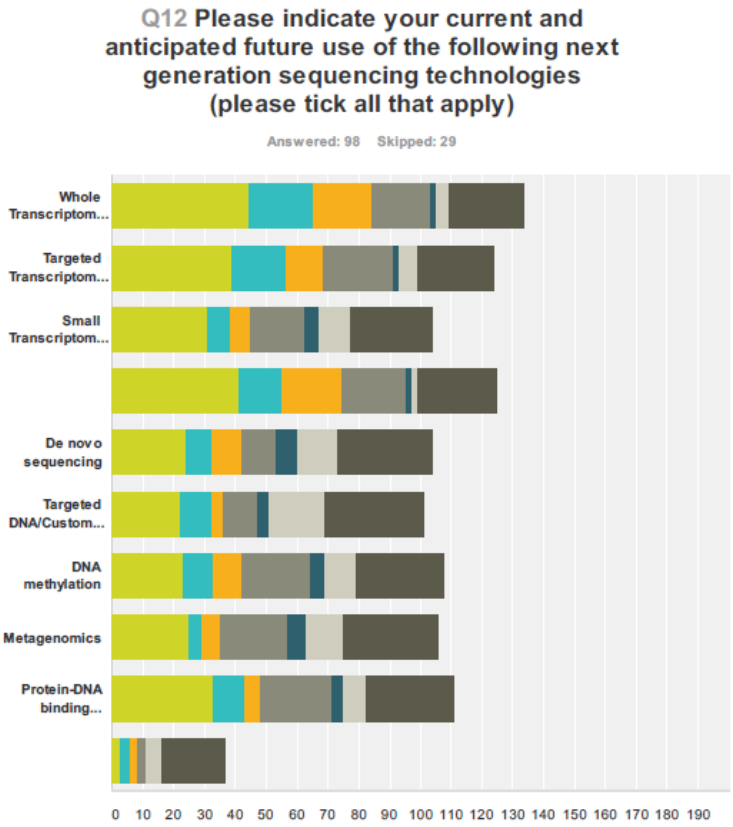


ISBE wide survey

Though currently only a minimum of contributions at systems biology conferences mention **DNA methylation** as their used technique, it is likely to be required by over 20% of the responding scientists in the future and more than 10% use this technology at another institution, however it is available only in less than 26% of the institutions. We can identify this as a potential technique that is not yet used widely but might become a mainstream in the future and should be taken care of by the infrastructure. We can identify the similar situation with **Protein – DNA binding (ChIP chip)**, when it is likely required by over 21% of the responding scientists in the future.

B. NEXT GENERATION SEQUENCING TECHNOLOGIES

Mainly the next generation sequencing technologies (NGS) are used for whole-genome and region sequencing, small RNA discovery, transcriptome analysis, metagenomics, methylation profiling, and genome-wide protein-nucleic acid interaction analysis. Illumina dominates the next generation sequencing market with its suite of systems (MiSeq, HiSeq 2500/ X Ten and Nextseq 500), in 2012 Illumina had a market share of 56%. Life Technologies (Ion Torrent), Roche (454) and Pacific Biosciences (PACBIO RS II) all also provide state-of-the-art systems. As each company provides systems with varying strengths and weakness it would be important for ISBE to provide access to a number of different systems, as the best system depends on the required application. For example, each of the following applications mRNA-seq, exome sequencing, targeted sequencing or novel genome assembly are each best suited to different systems. In addition, the most cost effective system depends on the number of samples being run (e.g. Illumina HiSeq for large numbers and Ion PMG for small sample numbers) and the read depth required. (Duffy, appendix II)



With respect to **next generation sequencing technologies** and its survey results on a first glance it seems that Small Transcriptome analysis, Whole Transcriptome Analysis, Targeted Transcriptome analysis (e.g. mRNA-seq or miRNA-seq), Whole Genome Sequencing and De novo sequencing is available in institutions and covering their demand. What we can identify as potential techniques that are not provided in sufficient amount by home institutions and could be taken care of by the infrastructure are:
Metagenomics which is likely to be required by more than 23% of survey respondents and used by more than 6% at different institutions, and by 4% of respondents at their institutions and **DNA methylation** is likely to be required by almost 25% of the respondents, used by 11% of correspondents at other institutions and is available only in 26% of institutions.

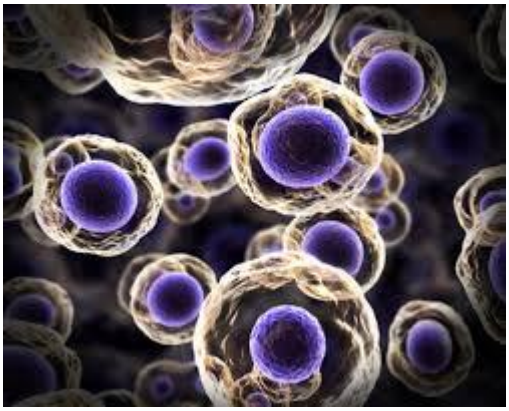
A major current obstacle is seen in the high costs of methylation studies using NGS. However, the history of other techniques has demonstrated that the cost might come down very fast once a technique is established, and so there is a good chance that this will start to change soon for this case, too. At the horizon we already can get a glimpse of third

generation sequencing technologies e.g. Pacific Bioscience systems for this type of sample/study. The methylation status of DNA can be sequenced straight from the DNA with minimal preparation of the DNA sample. This technology is currently prohibitively expensive for individual labs however it may be worth collaboratively outsourcing samples or identifying a central point to which samples can be sent. Since this type of analysis requires vast amounts of genomic data, internal storage of data at research institutions is a severe limitation that could be overcome by NGS machines equipped with the feature for uploading data to a cloud. Therefore the issue of data ownership and safety needs to be addressed; nevertheless the potential possibility of having a central repository/pipeline for data analysis utilizing cloud computing seems attractive. (Kenny, Appendix II)

A recent publication described RNA sequencing in situ (Highly Multiplexed Subcellular RNA Sequencing in Situ, Je Hyuk Lee et al., Science vol. 343, March 2014). The described technique provides a combination of in situ library preparation and sequencing. At time it is limited with amplicon length of 27 bp and the number of sequenced amplicons. The correlation coefficient, when compared to Illumina is still quite low (0.5 – 0.7), however the technique offers the advantage of subcellular localization, so you can see where was the RNA of each of the sequenced amplicons localized. At this point we see this technique at the rise of expectation within the fermentation phase and for sure it is not yet fully tested. (i.e. cannot be included to infrastructure). Nevertheless this kind of techniques might be of great importance in the future. (Dobnik, appendix I).

C. SINGLE CELL TECHNOLOGIES

Currently, a larger number of scientists are really pushing cutting edges on single cell sequencing and on RNA level. Single cell is increasingly getting important for proteomics, RNA, DNA. It seems that though the techniques are available in principle that price is still a big issue and the infrastructure might provide grants that would allow the scientists to pick up more transcripts for a critically lower price because if one wants single cell sequencing of RNA, there are thousands and ten thousands cells. In general, scientist agree that next generation sequencing is not still at the peak of its usage, and near-future development will be again Next gen sequencing based but pushing the cost standard (developing library technologies libraries to get more accurate, how to read a longer bit of DNA, to build better single cell RNA sequencing accurate as possible to capture as many transcripts per single cell). A dream list technology might be a technology that reads of histom mark-up of DNA. In the opinion of most respondents also current technologies – illumino machines, high and low level analysis – have still a solid future.

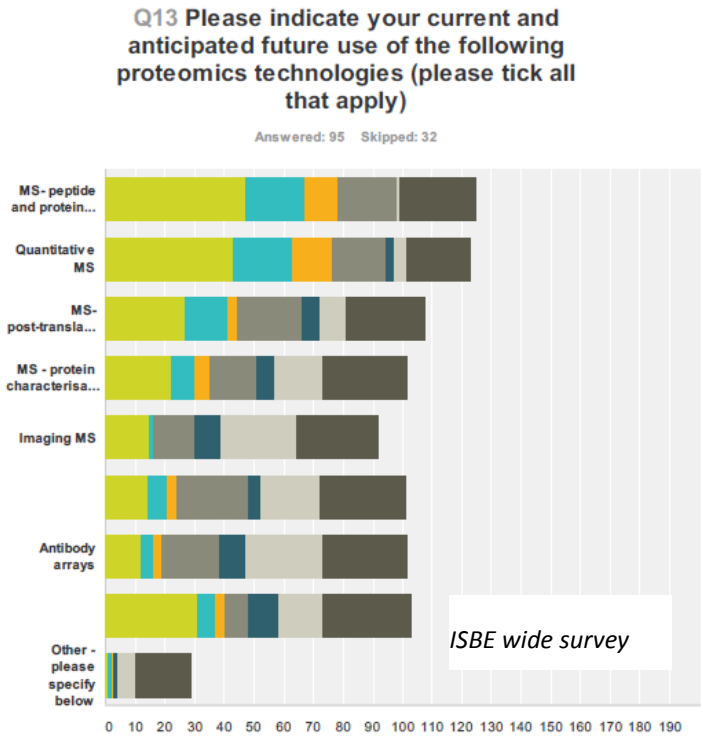


Single cell removal technologies from tissues and cells might be added as a key access technology for single cell analysis. Laser microdissection, like the “LMD6500” and LMD7000 from Leica or different technologies like the “CellSelector from ALS Automated Lab Solutions are state of the art technologies. Due to a high cost of equipment and well trained personal that is required, it would make sense to have this technology available within an infrastructure. For the future emerging technologies would be the solutions combining different technologies in one set-up, like AFM, microinjection and single cell analysis will be important in the near future. Upcoming solutions combining different technologies in one set-up, like AFM, microinjection and single cell analysis will be important in the near future. Those systems allow population based analysis of cellular events after e.g. RNA interference or cDNA over expression and might be by this part of modeling-experimentation feed-back loop. (Erfle, Appendix III)

D. PROTEOMICS TECHNOLOGIES

Giving the fact that PROTEOMICS is the large-scale study of proteins, particularly their structures and functions the two flavors of proteomic technologies are of critical importance. The first is discovery proteomics, also referred to as shotgun proteomics. This technology is used to identify the components of a biological system. The second proteomic approach, exemplified by targeting proteomics methods, aims at quantifying sets of proteins with high consistency across multiple samples. In systems biology such sample-sets are exemplified by differentially perturbed cells or tissues. Targeting methods include those based on affinity reagents (e.g. reverse arrays) and the mass spectrometric techniques selected/multiple reaction monitoring (S/MRM) and SWATH-MS. (Aebersold, Appendix IV)

Mass spectrometry as based techniques for protein profilling has become widely available in recent years. Nowadays, 2 out of 3 papers in the Nature-Science group are using mass spectrometry and from 2007, when the orbitrap technique became available, mass spectrometry papers have tripled. MS-peptide and protein identification, Quantitative MS, MS posttranslational modifications, Protein and peptide arrays, Antibody arrays and 2-Dimensional electrophoresis for **proteomics** are available in a majority of responding institutions and fully cover their demand. However, there are two techniques: **Protein and eptide arrays** – is available only in 16% of institutions but more than 28% require the usage in the future and **Antibody arrays -** is available only in 14% of institutions but more than 22% require the usage in the future. These techniques should be monitored in the near future and might be worth to implement into the infrastructure.



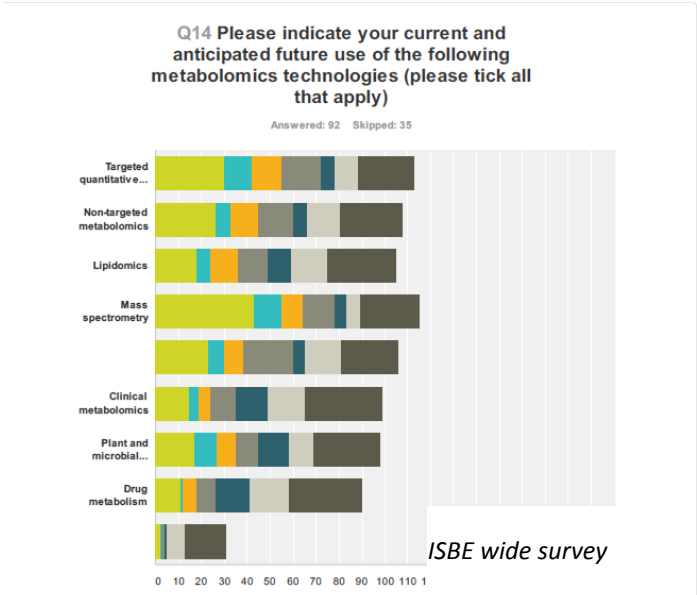
Protein and peptide arrays and antibody arrays might need to be implemented in any infrastructure in the near future. A point that might like to be addressed in future discussions is separating the proteomic technology into 'instrumentation' and 'expertise', since many proteomic technologies can be performed on the MS instrumentation within the infrastructure at an institution, but the data generated often requires both skilled experimentalists and the necessary software to get the most out of a data set. (Hitchin, Appendix IV)

E. METABOLOMICS TECHNOLOGIES

Metabolomics Technologies uses rather complex detection methods that require analytical and extensive data processing. There are two state-of-the-art or key technological approaches in metabolomics. The first employs NMR and the latter mass spectrometry (MS). MS has become mainstream, because of several advantages over NMR. Here the trend is towards high-resolution, accurate mass technology (1-2ppm mass accuracy). Targeted metabolomics focuses on known compounds, while untargeted is analyzing all mass features from one sample. (Schauer, Appendix V)

Targeted quantitative metabolomics, and Mass spectrometry are **metabolomics technologies** available in more than 34% of institutions. We can identify Mass spectrometry (52%) as the mainstream of available technique at institutions, but still only 14% (Mass spectrometry) and 13% (Targeted quantitative metabolomics) of survey respondents use these techniques at their institution. The reasons for this need to be understood and might open the question if the infrastructure should interface these techniques in future. Other techniques (Non-targeted metabolomics, Plant and microbial metabolomics, Highthroughput metabolomics, Clinical metabolomics) are in balance between their availability in their institutions (33%) and being likely to require in the future. **Highthroughput metabolomics** might give opportunity for the infrastructure to take care of because 25% of correspondents is likely to require this technology in the future and 10% use this technology at another institution.

the



ISBE wide survey

Future directions for protein or protein metabolite interaction might be monitoring of the cooperative mechanism. This would be about cooperativity rather than charting interactions one by one, quantitative deregulation implies allosterism, not only interactions happen all the time. Upcoming high-throughput techniques might be lab scale surface plasmon resonance (SPR), or signal machine like electrophoresis machines where you can really measure ass/disassociation constants on the lab scale of a very tiny material. Characterization of protein complexes going back to bench “old fashioned” chemical methods like gel filtration might become important again, however miniaturized and implemented on a micro platform to be used in large-scale screenings.

Based on the current requests and the complexity from the analytical and raw data processing side it may not be beneficial to implement metabolomics into ISBE infrastructure. As seen in gene expression profiling, metabolomics may rather be serviced in from expert companies providing higher throughput, improved quality and shorter turn-back times. Though this is a decision to be taking based on demands, investment, time and level of routine laboratory desired. (Schauer, Appendix V)

F. IMAGE TECHNOLOGIES

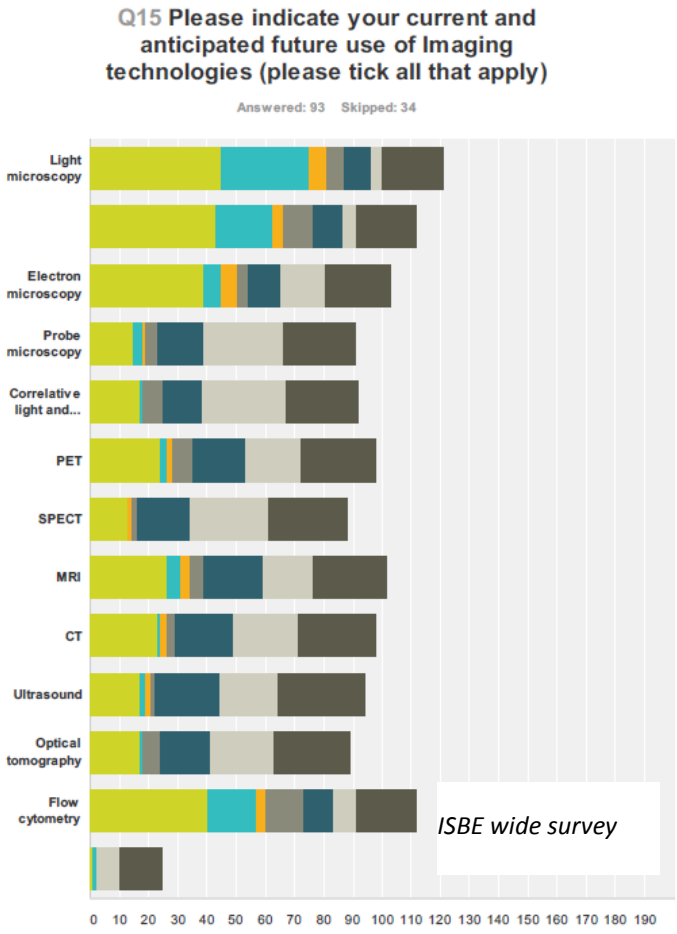
There is a valid point of integrating imaging tools with more conventional approaches to analyze the biological circuits of microorganisms, plants, and animals. The light microscopy methods seams most suited for the systems biology field for a quick validation of proposed models. Due to throughput and ease in experiment planning and running, light microscopy methods are most fitting. (Erfle, Appendix VI)

Imaging stands out amongst the many technologies used in systems biology as being (almost) the only one compatible with live: Rather than a post-mortem analysis of biological systems (e.g. like in sequencing, proteomics, metabolomics, most other –omics), it allows to investigate information flow in its biological context and change in space and time." (Spiteler, Appendix VI)

In imaging the state-of-the-art technologies are the following:

High-resolution:

- **Localisation microscopy (PALM, STORM):** most relevant for systems biology, allowing single-molecule measurements
- **Structured illumination (SIM):** 100nm resolution limit in 3D, but well suited to bridge the gap between single molecules and whole cells (50µm) at reasonable speed (sec)
- **Laser overlay (STED):** high resolution (50nm) at relatively high speed (msec-sec), but at the cost of destructive laser power (mostly incompatible with life)
- **High-speed detectors:** sub-msec time frames



Current ISBE wide survey indicates that most technologies (Light microscopy – 49%, Advanced Light microscopy – 49%, Electron microscopy – 44%) are available in their institutions and cover the demand fully. In comparison with these technologies other **Image technologies** (Probe microscopy, Correlative light and electron, PET, SPECT, MRI, CT, Ultrasound, Optical tomography) are not generally available in the responding institutions but they seem to cover the demand fully, at

least the scientists are currently not aware of a required demand in the future. In microscopy there are couple of techniques that might have taken the revolutionary step (as **2PPM**) and are currently in the fermentation phase, and it will be necessary to monitor if their potential is indeed demanded by future science in systems biology.

Although the feeling is that it would make sense to have these available within an infrastructure, there is a warning sign that all developments (listed in Appendix VI) absolutely decide on well-managed infrastructure „They only work if all developments are integrated, which is beyond most researcher's means (currently lead e.g. by EMBL in Heidelberg, Max Planck-Institute of Molecular Cell Biology and Genetics in Dresden." (Spiteler, Appendix VI)

G. DYNAMIC MODELLING

Key technologies are mathematical modelling and analysis software (SW) for simulation (such as Matlab, Mathematica, Maple, xppaut and diverse systems biology add-ons for those such as the SBPOP PACKAGE (formerly called Systems Biology toolbox). SW will include: Parallel implementations of deterministic and stochastic simulators and analysis tools; model editing/annotation/visualisation tools; standard model exchange language, both textual and graphical. Data integration: standard data description language: tools for integrating proteomics and genomics data from existing databases (DBs) and new experiments.

Experimental perspective: multiplex assays that can measure several intracellular concentrations in one sample, since these facilitate the generation of high-density time-course and perturbation data for model calibration. Single-cell mass spectroscopy (also known as mass cytometry). Imaging flow cytometry. Both, single-cell mass spectroscopy and imaging flow cytometry are not well established and are being constantly improved; hence their importance will increase in the near future.

Rule based or related for large scale modelling, use of multiple data sources: quantitative techniques like proteomics and phosphoproteomics, reverse protein arrays, plus genomics data and single cell measurement techniques take important role. We will need efficient computational methods to extract the information from the ever increasing datasets and DBs.

Clearly Genome-Scale Metabolic models find the widest application in industrial biotechnology in terms of mathematical modelling. Besides this any Metabolomics technology that enables rapid phenotypic characterization is useful - in particular RNAseq and proteomics. Metabolomics used less as it is the most difficult set of data to integrate and use for guiding metabolic engineering.

With the increasingly advanced understanding that diseases like cancer are a manifestation of deregulation of multiple pathways, and with availability of multiplex data on multi-pathways, large-scale mathematical dynamic models that account for pathways cross-talks rather than single pathways should become crucial in the future.

If integrated within the infrastructure, the standards dealing with model sharing, storing, annotating would be particularly important in enabling transparency in the community and speeding up the modeling process.

III. ADDED VALUE OF A STEWARDSHIP INFRASTRUCTURE

Between the respondents and the views represented in various literature reports is a remarkable agreement in the opinion that the big projects are heavily funded for data collection and underfunded on data analysis. **A useful infrastructure for systems biology thus should not provide `huge machines` or data generation facilities but rather complex expertise and stewardship with a strong emphasis on informatics necessary for sharing and analysing data and modelling.** In the context of genomics data basing and the ability to share data seems to be an issue, while instrumentation is relatively available.

A significant challenge in this field is seen is to enable access to existing data. Herby, an infrastructure can help to make experiments by teams of scientists from different fields to get from one experiment data they can use in their specific research area. **Data sharing is generally seen as a critical limitation.**

Better integration of large scale-data and DBs into the modelling process is needed. Most DBs lack the kinetic information, including rate parameters. The standards are importantly, release of raw data and models in standard format upon publications, implementation of easy to use tools for automatic import/export. Single-cell mass spectroscopy allows for multiplexed measurement of up to 100 molecules and phenotypes on the single-cell level with high throughput. The quantitative data obtained with such a measurement can be used to reconstruct topology of signalling networks and their dynamics. Combining flow cytometry with imaging makes it possible to correlate the molecular state of the cell with its morphological changes. Additionally, spatial localisation of molecules can be tracked which provides valuable information for computational modelling of signalling networks. **The key is to make things standard, for exchange and reusability of both models and data.**

Generally, **there is agreement that core facilities in informatics are not meant in terms of storage but in terms of staff people who offer their expertise in limited time projects in the modelling centres** (several months) and will be working on analysing and giving data. It needs specific types of people to do that and the problem could be a tension between giving community service versus their career, which is then an issue that needs to be solved in the implementation of the infrastructure. These people (employees of ISBE) nevertheless cannot be just service personal but need to be embedded in the scientific process, too, not to miss the development and new trends in new technologies usage.

Technology must be widely accessible, and the ISBE nodes should be able to negotiate or mediate access to experimental facilities provides by either other RIs or participating institutions. It is not sustainable if high performances analysis can be done only in max 5 labs, nowhere else and thus the analysis relies on specific scientific collaboration only. Some private institutions offer experiments with data generation and analysis nowadays to their communities – EMBO [there are some genomic core facilities], EMBL. If you are not the member of these institutions, it is really difficult to get access to their data and these paid surveys are also quite expensive. **The availability of standardized data that is readily available to the SB community is a step towards a resource to test the robustness of models in a variety of experimental conditions.**

IV. REVIEW FROM TECHNOLOGY AND SCIENCE WATCH COMMITTEE

Appendix I.

Microarray Technologies

By David Dobnik, National institute of biology Ljubljana, Slovenia

High-throughput transcriptomic technologies (like **microarrays** or RNA-seq) usually require additional confirmation of results. This is mostly done by real-time PCR only on few of the selected candidate genes. Recently, with the development of digital PCR (dPCR) this kind of confirmations could gain on throughput. Specifically I have in mind the Biomark HD System from Fluidigm, which enables performing the real-time PCR reaction for 96 samples, where each sample is tested for 96 assays. Other dPCR platforms available are of course the state-of-art technology when speaking of absolute quantification, however the throughput in terms of samples and assays is not as high as with Biomark HD. It could happen that in near future also other platforms for dPCR would become handier for handling the systems biology experimental setup. The availability of this machine within an infrastructure would make sense, if such analyses would prove to be needed.

Appendix II.

Next generation sequencing technologies

By David J. Duffy, Postdoctoral Researcher with Systems Biology Ireland

1. What existing technologies are considered as state-of-the-art or key technology in the specific fields of systems biology?

Illumina dominates the next generation sequencing (NGS) market with its suite of systems (MiSeq, HiSeq 2500/ X Ten and Nextseq 500), in 2012 Illumina had a market share of 56%. Life Technologies (Ion Torrent), Roche (454) and Pacific Biosciences (PACBIO RS II) all also provide state-of-the-art systems. As each company provides systems with varying strengths and weakness it would be important for ISBE to provide access to a number of different systems, as the best system depends on the required application. For example, each of the following applications mRNA-seq, exome sequencing, targeted sequencing or novel genome assembly are each best suited to different systems. In addition, the most cost effective system depends on the number of samples being run (e.g. Illumina HiSeq for large numbers and Ion PMG for small sample numbers) and the read depth required.

2. Would it make sense to have these available within an infrastructure?

While commercial companies and some academic institutes do currently supply aces to these systems it would make sense to make them available in the infrastructure, especially if they are coupled to the supply of state-of-the-art bioinformatic and data handling support. The purchase of commercial bioinformatics and data storage/handling are prohibitively expensive and often the quality is quite limited. Given the nature of Systems Biology projects it is useful to be able to have on-going discussions and collaborations with bioinformaticians, as opposed to the purchase of a one-time only, locked analysis.

In addition, access should be provided to clinical diagnostic grade sequencers such as the MiSeqDx (FDA approved) and soon to be released Ion Proton Dx. Access to such equipment will be key to the application of Systems Medicine (a key emerging branch of Systems Biology) approaches to the clinic. Without diagnostic grade instruments sequencing results can be used for research only,

rather than being directly applicable to individual patients and facilitating the advent of personalised medicine.

3. What emerging technologies will be important in the near future?

The current NGS technologies continue to be upgraded and improved with incremental advances being made. Some of these advances require no further investment, such as the release of improved software and sequencing flowcells/chips. However, incrementally improved systems are also released (primarily by Illumina). Therefore, funds should be budgeted to allow the periodical updating of ISBE equipment, as opposed to only investing in current systems. Also funds should be benchmarked for the next NGS systems which will become available in the short to medium term. Nanopore sequencing (Oxford Nanopore Technologies) currently appears to be the closest technology to market, of a new generation of sequencing technologies. Nanopore sequencing will over a number of revolutionary improvements over current NGS systems. For instance, read lengths of up to tens of kilobases and the ability to sequence RNA directly (no cDNA conversion or PCR enrichment).

4. How can these new technologies be integrated into Systems biology and how an infrastructure might help with this?

NGS technologies are currently integrated into Systems Biology, but this currently happens primarily at a more haphazard local level. An improved infrastructure could make this integration over an EU wide level, providing the required solutions and saving time, effort and money all of which are currently duplicated by every Systems Biology lab who conduct NGS experiments. Any ISBE initiative in this area should also provide open access to standardized reported NGS results using an intuitive interface, to enable the maximum use of the generated data by having it interrogatable by any researcher. The integration of infrastructure to facilitate cost effective sequencing, data management and bioinformatics analyses (both at the initial primary research level and later stage meta-analyses) would be hugely beneficial to future Systems Biology research in Europe.

Appendix II. continues

Next generation sequencing technologies

By Elaine Kenny, Co-founder of SME called Elda Biotech, Ireland

Whilst all of the institutions have access to next generation sequencing (NGS) technologies it's unclear where specific expertise reside. Identifying the key expertise of institutes will also identify partner/collaborator groups for relevant projects and vastly reduce the cost associated with poor data generation. Unlike microarray technology for example the sample preparation protocol employed can have a huge impact on the quality and type of data generated. Whilst there's no need to centralise this technology, possibly a central repository for how samples are prepared (library prep protocol/kits/adjustments etc) for each of the studies would help in the interpretation of the results. This is certainly something that could be provided quite easily by the infrastructure, where basic but often very important information about each experiment is stored.

The report mentioned that the study of DNA methylation was identified as likely required in future by 23% of respondents to the questionnaire. Currently methylation studies using NGS can be quite expensive to run, however as the cost comes down this will certainly start to change. It's also worth looking at some of the third generation sequencing technologies e.g. Pacific Bioscience systems for this type of sample/study. The methylation status of DNA can be sequenced straight from the DNA with minimal preparation of the DNA sample. This technology is currently prohibitively expensive for individual labs however it may be worth collaboratively outsourcing samples or identifying a central point to which samples can be sent.

With the generation of vast amounts of genomic data using this technology data storage and analysis is always going to be an issue. Some of the newer NGS machines come equipped capable of uploading data to the cloud. I see cloud computing becoming quite vital in NGS projects and it has certainly increased our collaborative ability. The use and reliance on it will continue, especially in the research environment. Many users like to have ownership of their data and their analysis; however it's worth looking at the possibility of having a central repository/pipeline for data analysis utilizing cloud computing. The key to making such a thing work however would be the turnaround time of analysis. The idea being that all NGS data could be uploaded and QC passed/checked to ensure a minimum standard is met.

Appendix III.

Single Cell Technologies

By Holger Erfle, Head of the BioQuant RNAi screening Facility

The report tries to address in each area the following:

1. **What existing technologies are considered as state-of-the-art or key technology in the specific fields of systems biology?**

   Single cell removal technologies from tissues and cells might be added as a key access technology for single cell analysis. Laser microdissection, like the "LMD6500" and LMD7000 from Leica or different technologies like the "CellSelector from ALS Automated Lab Solutions are state of the art technologies.

2. **Would it make sense to have these available within an infrastructure?**

   Yes, it would make sense to have this technology available within an infrastructure as purchase of equipment is costly and running the site demands experienced and well-trained personal.

3. What emerging technologies will be important in the near future?

Upcoming solutions combining different technologies in one set-up, like AFM, microinjection and single cell analysis will be important in the near future.

4. How can these new technologies be integrated into Systems biology and how an infrastructure might help with this?

Those systems allow population based analysis of cellular events after e.g. RNA interference or cDNA over expression and might be by this part of modeling-experimentation feed-back loop.

By David Dobnik, National institute of biology Ljubljana, Slovenia

In regard to the **single cell technologies**, there has been a publication recently describing the RNA sequencing in situ (Highly Multiplexed Subcellular RNA Sequencing in Situ, Je Hyuk Lee et al., Science vol. 343, March 2014). The described technique provides a combination of in situ library preparation and sequencing. At time it is limited with amplicon length of 27 bp and the number of sequenced amplicons. The correlation coefficient, when compared to Illumina is still quite low (0.5 – 0.7), however the technique offers the advantage of subcellular localization, so you can see where was the RNA of each of the sequenced amplicons localized. I see this technique at the rise of expectation in fermentation phase and it is yet not fully tested (i.e. cannot be included to infrastructure). Nevertheless this kind of techniques might be of great importance in the future.

Appendix IV.

Proteomics Technologies

By Paul Hitchin, Faculty of Natural Sciences, Department of Life Sciences, Imperial College

Having read through the report on continuous technology forecasting with respect to Proteomics technologies, I can largely agree with the findings from the survey. It seems that many of the proteomic technologies are in place for researchers at their institutions or are available to use at another institution but the survey has identified two techniques: Protein and peptide arrays and antibody arrays, that might need to be implemented in any infrastructure in the near future. A point that might like to be addressed in future discussions is separating the proteomic technology into 'instrumentation' and 'expertise', since many proteomic technologies can be performed on the MS instrumentation within the infrastructure at an institution, but the data generated often requires both skilled experimentalists and the necessary software to get the most out of a data set.

Proteomics Technologies

By Ruedi Aebersold, ETH Zurich

1. What existing technologies are considered as state-of-the-art or key technology in the specific fields of systems biology?

For systems biology two flavors of proteomic technologies are of critical importance. The first is discovery proteomics, also referred to as shotgun proteomics. This technology is used to identify the components of a biological system. The second proteomic approach, exemplified by targeting proteomics methods, aims at quantifying sets of proteins with high consistency across multiple samples. In systems biology such sample-sets are exemplified by differentially perturbed cells or tissues. Targeting methods include those based on affinity reagents (e.g. reverse arrays) and the mass spectrometric techniques selected/multiple reaction monitoring (S/MRM) and SWATH-MS. Targeting methods generally require the development and validation of specific assays for each targeted protein (e.g. an antibody for immunodetection; a reference fragment ion spectrum for MS based techniques) and the one time development for the community of proteome-wide assay libraries would be a particularly fruitful endeavor for ISBE.

2. Would it make sense to have these available within an infrastructure?

Supporting these techniques as infrastructure platforms would certainly generate a very high impact. This is particularly the case for the above described targeting techniques which would allow a large number of systems biologists to accurately quantify essentially any protein with a high degree of reproducibility across multiple samples, e.g. sample sets representing differentially perturbed cells or tissues. Considering that data driven systems biology studies to date are for the most part based on transcript measurements and the well-known fact that transcripts do neither predict the quantity nor the activity state of proteins, quantitatively accurate protein data would greatly advance the field of systems biology.

If the technology is to be supported by an infrastructure/facility, it will be important to make an integrated technology platform available.

Appendix V

Metabolomics Technologies

By Nicolas Schauer, CEO of Metabolomic Discoveries, Potsdam-Golm, Germany

1. What existing technologies are considered as state-of-the-art or key technology in the specific fields of systems biology?

Two technological approaches in metabolomics exist. The first employs NMR and the latter mass spectrometry (MS). MS has become mainstream, because of several advantages over NMR. Here the trend is towards high-resolution, accurate mass technology (1-2ppm mass accuracy). Targeted metabolomics focuses on known compounds, while untargeted is analyzing all mass features from one sample.

2. Would it make sense to have these available within an infrastructure?

Metabolomics will become of more importance over the next years and an integral part of systems biology. Based on the current requests and the complexity from the analytical and raw data processing side it may not be beneficial to implement metabolomics. As seen in gene expression profiling, metabolomics may rather be serviced in from expert companies providing higher throughput, improved quality and shorter turn-back times. Though this is a decision to be taking based on demands, investment, time and level of routine laboratory desired.

3. What emerging technologies will be important in the near future?
In the future the coupling of MS to ion mobility is most promising, though this is still in its infancy.

4. How can these new technologies be integrated into Systems biology and how an infrastructure might help with this?

Metabolome data provides rich information on top level and thus gives the most insights into biological mechanisms. Data can be easily integrated into systems biology approaches, as KEGG and other identifiers allow pathway and network building and thus provides a close link between in-silico and experimental data.

Apendix VI

Image Technologies

By Holger Erfle, Head of the BioQuant RNAi screening Facility

The report tries to address in each area the following:

1. What existing technologies are considered as state-of-the-art or key technology in the specific fields of systems biology?

Due to throughput and ease in experiment planning and running, light microscopy methods are most suited for the systems biology field for a quick validation of proposed models. In addition, scientists can be relatively easy taught to perform themselves individual experiments.

2. Would it make sense to have these available within an infrastructure?

Yes, it would make sense to have this technology available within an infrastructure as purchase and maintenance of equipment are costly and carrying out experiments and teaching users demand experienced and well-trained personal.

3. What emerging technologies will be important in the near future?

Here one might add super resolution techniques, like STED, PALM or STORM and light-sheet imaging.

4. How can these new technologies be integrated into Systems biology and how an infrastructure might help with this?

Those new technologies allow a quick and easy link between modeling and experimental validation.

Apendix VI continues

Image Technologies

By Martin Spitaler, Facility for Imaging by Light Microscopy [FILM], Imperial College London

Imaging stands out amongst the many technologies used in systems biology as being (almost) the only one compatible with live: Rather than a post-mortem analysis of biological systems (e.g. like in sequencing, proteomics, metabolomics, most other –omics), it allows to investigate information flow in its biological context and change in space and time. Two major limitations have limited its use for systems biology:

- (1) limited resolution in space and time
- (2) difficulties extracting unambiguous information from unstructured data

Both these limits are currently being overcome at dramatic speed (see below), although many logs of improvements will still be required to reach a 'saturation' level, at which no more improvements could be expected (maybe ~µsec for speed, nm for dimensions, especially the combination is still utopia). Which leaves as a new limit the handling and processing of data, both from the logistic point of view (annotation, transfer, visualisation) and hardware / software capacities.

1. WHAT EXISTING TECHNOLOGIES ARE CONSIDERED AS STATE-OF-THE-ART OR KEY TECHNOLOGY IN THE SPECIFIC FIELDS OF SYSTEMS BIOLOGY?

- High-resolution:
  - **Localisation microscopy (PALM, STORM):** most relevant for systems biology, allowing single-molecule measurements
  - Structured illumination (SIM): 100nm resolution limit in 3D, but well suited to bridge the gap between single molecules and whole cells (50µm) at reasonable speed (sec)
  - **Laser overlay (STED):** high resolution (50nm) at relatively high speed (msec-sec), but at the cost of destructive laser power (mostly incompatible with life)
  - **High-speed detectors:** sub-msec time frames
- Data handling:

- **Data storage and annotation:** first functional solutions (e.g. OMERO), but still rather limited solution (handling of large data, usability, integration with software tools); content-based search in early experimental stage
- **Data analysis:** Pattern recognition becoming established in light microscopy; single-molecule localisation and statistics at the stage of ongoing community-based reviewing and standardisation
- **On-the-fly processing:** slowly moving from developmental stage to early adopters, with emerging support from commercial microscope manufacturers

1. WOULD IT MAKE SENSE TO HAVE THESE AVAILABLE WITHIN AN INFRASTRUCTURE?

All developments above absolutely decide on well-managed infrastructure: They only work if all developments are integrated, which is beyond most researcher's means (currently lead e.g. by EMBL in Heidelberg, Max Planck-Institute of Molecular Cell Biology and Genetics in Dresden).

2. WHAT EMERGING TECHNOLOGIES WILL BE IMPORTANT IN THE NEAR FUTURE?

A) OBERCOMING RESOLUTION LIMITS IN SPACE AND TIME

Two directions will change the way systems biology works: Higher resolution and detector sensitivity allowing single-molecule observations at high speed, and whole-organism (zebrafish, embryos) / whole-organ imaging of physiological process.

Higher resolution in space close to single molecule-scale is now possible thanks to super-resolution microscopy, made possible by a parallel development of novel microscopy techniques, detectors (CCD and CMOS) with single-photon sensitivity and novel fluorescent markers (photoswitchable proteins and chemical fluorophores). These techniques (under the acronyms of PALM and STORM) already allow studying the cellular signalling circuitry at molecular level, or physiological events down to sub-millisecond speed. However, there is still a strong trade-off between spatial and temporal resolution, ranging from 30nm precision of whole-cell cross-sections (10-30µm length) at 3-20min per frame to the other extreme of 50µsec per frame, but of small areas (1µm length) with 1µm precision. At the other end of the spectrum, novel microscopy techniques (2-photon intravital microscopy, light-sheet microscopy) in combination with new markers allow visualisation of mm-scale organs or organism at cellular precision, thereby quantifying interaction between cells and tissues, movements over time, decision trees in development (e.g. embryo development, haematopoiesis) etc.

B) IMPROVEMENTS IN THE EXTRACTION OF INFORMATION FROM UNSTRUCTURED DATA

The main value of microscopy images is it's high-content, multidimensional, unstructured information, but this makes it difficult to translate into computer-readable formats. However, many of these difficulties are based on a slow translation of established technologies from other areas to biological applications: While biologists use mobile phones with smile detection in private life, they still mostly rely on archaic intensity thresholding for object detection in their scientific work.

We are currently witnessing a massive push in this re-adaptation of technologies, be it astronomy algorithms for single-molecule localisation or pattern recognition to track cells in noisy 3D data of whole organs.

(1+2)=(3) INTELLIGENT IMAGE ACQUISITION

A major development currently in an early experimental state (although applied in selected labs for over a decade) will be intelligent acquisition, i.e. rather than generating huge amounts of meaningless data, to incorporate the biological question in the image acquisition. This will drastically improve the data quality while in parallel drastically reducing the data volume, or rather

the ratio data volume per scientific information (the total volume will keep moving on the limits of the technical possibilities). On-the-fly-analysis will work in two ways:

- rather than saving unstructured data, only the information of interest is saved; for example, if studying cell movements in development, the XYZT coordinates of a few thousand cells (Mbytes) would be saved, rather than GBytes of raw images per time point
- low-resolution screening in space and time, then switching to high-resolution mode when encountering an event / object of interest; this will also reduce the amount of data (or increase the number of observable events / objects) by many logs

1. HOW CAN THESE NEW TECHNOLOGIES BE INTEGRATED INTO SYSTEMS BIOLOGY AND HOW AN INFRASTRUCTURE MIGHT HELP WITH THIS?

The main need to integrate them into Systems Biology are:

- standardised annotation of unstructured data:
  - on the hardware side, the solution is on its way with the Open Microscopy Environment (OME) data standard, now supported by most commercial vendors and open-source tools
  - on the sample side, standardised protocols for sample preparation and annotation are still needed; only user education will be able to bridge the gap
  - on the analysis side, standardisation of algorithms is on its way, especially in the super-resolution field, but it will take another few years to find a common sense
- uptake by systems biologists / mathematicians:
  - especially modellers tend to shy away from the unstructured, multi-dimensional nature of microscopy images; in combination with above efforts (improving the data quality), education will be needed to help them understand the huge potential (and some pitfalls) of these technologies

Appendix VII

Dynamic Modelling

By Lan K. Nquyen, Systems Biology at Conway Institute, Ireland

The authors of the report have discussed many key areas that are important for an infrastructure with regards to Dynamic Modeling, both in terms of the modeling techniques and data integration required for the modeling process. I have a number of additional points on both of these aspects:

- With the increasingly advanced understanding that diseases like cancer are a manifestation of deregulation of multiple pathways, and with availability of multiplex data on multi-pathways, large-scale mathematical dynamic models that account for pathways crosstalks rather than single pathways should become crucial in the future. Hence, new computational frameworks that make it easy and time-efficient to build, integrate and maintain these large models are strongly required. These frameworks should allow integration of information like mutational landscape/epigenetics on top of existing network models so that models can be adapted for different cancers and/or patients, thereby pushing modeling towards personalized medicine. Standards dealing with model sharing, storing, annotating would be particularly important in enabling transparency in the community and speeding up the modeling process.
- A key related need is the development of focused databases on kinetic information and protein concentrations that are essential for model calibration and parameter estimation. Good annotation of these databases would be important for modelers to extract cellular-context specific information for model adaptation. Efficient parallel

parameter estimation methods capable of running on clusters (which could be Sharp among the Infrastructure institutions) should be available and integrated into modeling software for access by the community.

- As models are multi-dimensional, novel methods for efficient analysis and visualization of the model dynamics in multi-dimensional settings are crucial for better “global” understanding of the networks being modeled. This would provide a more truthful picture of the network dynamics and facilitate therapeutic strategies. Current analysis methods are limited in this aspect.
- Regarding data as input for modeling process, the report has included the key experimental technologies. In addition, technology that is capable of obtaining (multiplex) data directly on patients sample tissues such as tissue FRET imaging would be particularly useful in the future to adapt models from cell-based towards patient based.

I expect that these techniques would be quite challenging to develop but would be of enormous applicability.