European Exascale Software Initiative

Working Group report on life Science and Health activities

CONTRACT NO	EESI 261513
INSTRUMENT	CSA (Support and Collaborative Action)
THEMATIC	INFRASTRUCTURE

Start date of project: 1 JUNE 2010 Duration: 18 months

Name of lead contractor for this deliverable: BSC

Name of reviewers for this deliverable: Stephane Requena (GENCI)

Abstract: This report is a synthesis of the work and main outcomes identified by the Work Group dedicated to Life Science and Health activities.

Revision R2.0

Release N° 1

Due date of deliverable:	Mai 31, 2011
Submission date:	July 01, 2011

Internal release N° 2

Due date of deliverable:	November 30, 2011
Submission date:	December 14 2011
Publication date:	December 22 2011

Project co-funded by the European Commission within the Seventh Framework Programme (FP7/2007-2013)		
Dissemination Level		
PU	Public	X

Table of Contents

1. SCIENTIFIC AND TECHNICAL PERIMETER OF THE WG 3

2. ECONOMICAL AND SOCIAL IMPACT 5

2.1 EXASCALE THRESHOLD 5
2.2 ECONOMICAL BENEFITS 6

3. SCIENTIFIC AND TECHNICAL HURDLES 8

3.1 GRAND CHALLENGES 8
3.1.1 Simulation of Human Brain 9
3.1.2 Understanding the blueprint of life 10
3.1.3 Genomic Medicine 11
3.1.4 Molecular Simulation 12
3.1.4.1 Multi-scale simulation 13
3.2 NEEDS OF EDUCATION AND TRAINING 13

4. BUILDING EXASCALE (2012-2020) 15

4.1 FUNDING OF E-COMMUNITIES 15
4.2 EXASCALE SKELETON: PILLAR APPLICATIONS IN LIFE SCIENCES 15
4.3 TRANSVERSE ISSUES 18
4.4 TIMELINE 18
4.5 PROVISIONAL COSTS 19

5. EUROPEAN STRENGTHS AND WEAKNESSES 20

5.1 SOURCES OF COMPETITIVENESS FOR EUROPE (STRENGTHS) 20
5.1.1 ELIXIR 20
5.1.2 EMBL-EBI 20
5.1.3 Participation in research consortiums such as ENCODE, ICGC and iHEC 20
5.1.4 Human Brain Project 20
5.1.5 FET Flagship in Personalized Medicine 21
5.1.6 PRACE 21
5.1.7 Pharmaceutical Industry 21
5.1.8 Computational Biology and Bioinformatics Software 22
5.2 EUROPE WEAKNESSES 22
5.3 POTENTIAL COLLABORATIONS OUTSIDE EUROPE 22
5.3.1 D E Shaw Research 22
5.3.2 AMBER developer team 22
5.3.3 NAMD developer team 23
5.3.4 NVIDIA 23
5.3.5 The Beijing Genomics Institute (BGI) 23
5.4 EXISTING FUNDED PROJECTS AND FUNDING AGENCIES 23

6. CONCLUSIONS 24

1. Authors and affiliation

Name	Organization	Role
Modesto Orozco	Institut de Recerca Biomèdica de Barcelona - Barcelona Supercomputing Center	WP3.4 Chair
Ramon Goñi	Barcelona Supercomputing Center	WP3.4 Collaborator
Janet Thorton	EMBL-EBI	WP3.4 Vice-chair
Adrew Lyall	EMBL-EBI	WP3.4 Collaborator
Helmut Grubmüller	Max Planck Institute	WP3.4 Expert
Paolo Carloni	German Research School	WP3.4 Expert
Erik Lindahl	Stockholm University & Royal Institute of Technology	WP3.4 Expert
Richard Lavery	Université de Lyon	WP3.4 Expert
Charles Laughton	University of Nottingham	WP3.4 Expert
Wofgang Wenzel	Karlsruhe Institute of Technology	WP3.4 Expert
Alfonso Valencia	Centro Nacional de Investigaciones Oncológicas	WP3.4 Expert
Reinhard Schneider	EMBL	WP3.4 Expert
Henry Markram	Ecole Polytechnique Fédérale de Lausanne	WP3.4 Expert
Felix Schürmman	Ecole Polytechnique Fédérale de Lausanne	WP3.4 Collaborator
Manuel Peitsch	Swiss Institute of Bioinformatics	WP3.4 Expert
Nicolas Baurin	Sanofi-Aventis	WP3.4 Expert
Anna Tramontano	University of Roma	WP3.4 Expert
Adrian E. Roitberg	University of Florida	WP3.4 Expert
Julian.Tirado-Rives	Yale University	WP3.4 Expert

2. Scientific and technical perimeter of the WG

The WG 3.4 is focused on HPC applications for Life Sciences and Health. The Life Science community is very diverse and there is a large unbalance between the large size of experimentalists/biologists (that strongly depends on computational results) and the small size of computational biologists (that depend heavily on HPC resources). For this reason the work of computational biologists has a "multiplicative" effect on Life Sciences. The goal of computational biology and bioinformatics is to understand the mechanisms of living systems. With the recent advances in this area (e.g. next generation of DNA sequencing instruments) the generated data is becoming larger and more complex. In contrast to other communities there are no universal computer packages and software evolves very fast to adapt to new instruments. The problems faced by scientists working in molecular simulations and genomics are also very different, as are the computer algorithms used. The importance of having fast and flexible access to very large computational resources is crucial in the many fields of Life Sciences and the lack of suitable computers can block entire projects with important consequences for science and the society. Discussions in the panel were lively and very productive.

Opinions with respect to Exascale computing were unanimously favourable, but opinions about single-machine Exaflop computing were more varied. While Exaflop machines are a major requirement for specific issues (e.g. brain simulation), and higher computational power will enable significantly increased accuracy for current high-throughput methods, some highly important subfields in Life Science are still expected to be limited by throughput. Therefore a single-sided focus on only achieving high flop rates in individual runs (rather than application results) could seriously hurt European research in these areas. This discussion is being conducted at the world wide supercomputer community at this time and the panel urges this to be clearly stated in the final document. For comparison, the recent 2010 report from the US President's Council of Advisors of Science and Technology[1] also emphasizes that "*Highly influential comparative rankings of the world's fastest supercomputers are for the most part based on metrics relevant to only some of our national priorities, and must not be regarded as the sole measure of our continued leadership in this essential area. Although it is important that we not fall behind in the development and deployment of HPC systems that address pressing current needs, it is equally important that we not allow either the funding allocated to the procurement of large-scale HPC systems, or undue attention to a simplistic measure of competitiveness, to "crowd out" the fundamental research in computer science and engineering that will be required to develop truly transformational next-generation HPC systems*".

Since there is a wide range of projects requiring Exascale performance in Life Sciences, we have organized our panel of experts into four main areas:

- In SYSTEMS BIOLOGY we are now at the stage of collecting data to build models for complex simulations that will describe in the next future the dynamics of cells and organs that remain unknown. The models that are developed today are stored in databases. Progress is rapid and systems biology will allow to couple the simulations of the models with a biomedical problem (e.g. monitor mutations in a specific genome that can change the activity of a protein). This will require large computational resources and systems biology will benefit from Exaflop capabilities, but aspects related to data management are going to be as important as pure processing capability.
- In GENOMICS research we face problems (e.g. the sequencing of 2,500 genomes of cancer patients) involving the management of massive amounts of data in programs that can require

[1] http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-nitrd-report-2010.pdf , page 17.

hundreds of thousands of processors, but little inter-processor communication. However, the vast amount of data to be managed (and often confidentiality and privacy aspects) hampers the use of cloud or grid-computing initiatives as a general solution. Suitable and flexible access to computer resources is crucial in this area. The genomic subpanel asserts that currently known cornerstones for an Exascale system (number of computer nodes, I/O and memory capacities) are clearly driving the focus only to reach the Exaflop peak performance. For most of the genomics challenges an Exaflop computer that could be even less "balanced" than today's HPC systems would be a substantial barrier to use these machines efficiently. The genomics subpanel and by extension the entire Life Sciences panel wishes to stress their major concerns that their Exascale problem will be difficult to treat with the unbalanced architectures of anticipated Exaflop computers.

- In MOLECULAR SIMULATION projects, Exaflop capabilities will allow the use of more accurate formalisms and enable molecular simulation for high-throughput applications (e.g. study of larger number of systems). Unfortunately, Exaflop capabilities will not favour the possibility to study longer time scales, since it will not be possible to scale into systems based on hundreds of thousand cores (as the simulated systems typically have less than 1 million atoms). The needs of the molecular simulation field will be better served by a heterogeneous machine, with hierarchical capabilities in terms of number of cores, amount of memory, memory access bandwidth and inter-core communication. This should be contrasted with current ideas regarding a 'flat' machine with peak Exaflop power. Exascale capability will however facilitate biased-sampling techniques, which require parallel computing, enabling *in silico* experiments unreachable today. Examples include the proteome-scale screening of chemical libraries to find new drugs and the study of entire organelles, or even cells, at the molecular level. In most of these cases, parallelization is expected to be hierarchical (e.g. ensemble simulations, multi-scale modelling or a mix of parallelization and high throughput).
- In BIOMEDICINE SIMULATION we envision projects such as the simulation of the brain, organs and tissue modelling, and *in silico* toxicity prediction. In these areas, Exaflop capabilities will be a necessary, but not sufficient, requirement, since the integration of experimental information, human-interaction with calculations and the refinement of underlying physical models will also be instrumental for success. As in the case of Molecular Simulation, multi-scale modelling is one of the major challenges of this area and represents one of the major cross-cutting issues of Exascale systems for Life Sciences.

These four main areas of Life Sciences and Health are strongly related to the pharmaceutical and biotechnology sectors, but also to other economic areas such as Food (agriculture), Environment (biotoxicity) and Energy (biofuels). As the following document will show, the panel of experts agrees that Europe is not in a position to lead the construction of an Exaflop computer, but it has the opportunity to support the first Life Sciences and Health Exascale co-design center.

3. Economical and social impact

Research in Life Sciences generates knowledge with a very clear and direct impact on our society. HPC can be of great importance in research in all areas of the Life Sciences and, specifically, in research related to health and biotechnology. The pharmaceutical sector alone represented a market of roughly 800 B€ in 2010[2]. In the last years, Europe produced approximately 38% of the world's pharmaceutical products, more than the United States and Japan, which contributed 35% and 13%, respectively[3]. Excluding big pharmaceutical companies, there are more than 2,100 biotechnology companies in Europe[4]. Beyond the Pharmaceutical and Biotechnology sectors, other economic areas related to Life Sciences that will greatly benefit from HPC are Food (agriculture), environment (biotoxicity) and Energy (biofuels). These three areas account for more than 40% of the European Commission Cooperation research budget[5] and are considered strategic for the future of Europe. The panel believes that Life Science research will clearly benefit from Exascale initiatives, but, at the same time, considers that the vast investment required to build and to maintain Exascale computers will make little sense if it not also organized to help advance research in Life Sciences.

3.1 Exascale Threshold

The benefits of the continuous development of more powerful computation systems are visible in many areas of Life Sciences. For example, at the beginning of 2000, the Human Genome Project[6] was an international flagship project that took several months of CPU time using a hundred-Gigaflop computer with one terabyte of secondary data storage. Today, genomic sequencing has changed from being a scientific milestone to a powerful tool for the treatment of diseases, in particular because it is able to deliver results in days, while the patients are still under treatment. The Beijing Genomics Institute is capable of sequencing more than a hundred human genomes a week using the Next Generation Sequencing instruments and a 100 Teraflop computer that will migrate in the near future to a 1 Petaflop capability[7]. Today, Genome sequencing technology is ineffective if the data analysis needs to be carried out on a grid or cloud-like distributed computing platform. First, such systems cannot achieve the necessary dataflow, of the order of 20 Petabytes/year, and, second, research involving living patients requires both speed and high security that are lacking in such environments. Lastly, ethical and confidentiality issues handicap distributing patient data across the cloud world. In coming years, sequencing instrument vendors expect to decrease costs by one to two orders of magnitude, with the objective of sequencing a human genome for $1,000. This will make possible to integrate genomic data into clinical trials (that typically involve thousands of human tests) and into the heath systems of European countries, making drug development easier and faster and having a dramatic impact on therapy (it is worth noting again here that Europe's pharmaceutical industry contributes significantly to the region's GPD than is true of the pharmaceutical industries in the U.S. and other nations). We should not forget, however, that all these possibilities can only develop if computer resources can deal with the complexity of the large interconnected datasets that are serving the large community of Life Science. For example, today the EBI (that hosts the major core bio-

[2] Life sciences and Biotechnology: A strategy for Europe. European Commission

[3] European Commission, Research Web site. Available at: http://europa.eu.int/comm/research

[4] Life Sciences and Biotechnology: The Transatlantic Divide. European Union Center of North Carolina EU Briefings, April 2009

[5] http://ec.europa.eu/research/fp7/

[6] International Human Genome Sequencing Consortium. Nature 2001

[7] http://www.genomics.cn/en/platform.php?id=248

resources of Europe) has doubled the storage from 6,000 Terabytes (in 2009) to 11,000 Terabytes (in 2010), and has received on average 4.6 million requests per day (see Figure 2).

There are also many other steps along the drug discovery pipeline that will benefit from advances in supercomputing. For example, the identification of potential drug candidates for identified disease targets will be fuelled by next generation supercomputers. Most lead discovery projects currently involve high-throughput screening (HTS) instruments that can scan 100,000 molecules per day looking for those showing activity against the target. The cost of this technique is very high and the typical success rate is very low. In contrast, Virtual Screening is a computational technique that can scan the ability of a therapeutic target to recognize molecules from a virtual library, extending the chemical search space and dramatically reducing the costs of drug discovery. Current virtual libraries contain around one billion drug-like compounds[8] and they are expected to grow still larger[9]. Only multi-Petascale supercomputers are capable of scanning such chemical spaces, while simultaneously treating a large number of potential targets.

3.2 Economical Benefits

We highlight in this section the major economical benefits of Exascale computing for Life Science and Health applications:

- **Substitution of animal testing.** Animal testing is banned in European Union for cosmetic products, but it is still today a key step in the development of new drugs. The pharmaceutical industry invests in animal testing $150 million per launch[10] of a new molecular entity (there are approximately between 10 and 30 launches per year). Animal testing is also required to evaluate the safety of major chemicals sold in Europe (see European REACH initiative for the next decade[11] that will involve millions of laboratory animals) and its estimated total cost for this initiative is between 1.3 to 9.5 B€. Researchers in medical simulation, systems biology and molecular simulation are working towards the development of *in silico* models that can simulate systems from cells to complex organs such as the brain. These models, when complete, will require Exascale computing and are expected to drastically reduce our dependency of animal testing, and to lower costs.
- **Computer-aided drug design.** Europe has a very competitive industry that launches almost the 40% of the pharmaceutical products on the worldwide market. Advances in genomics, systems biology and molecular simulation are making rational drug design a powerful alternative to trial-and-error methods. Computing systems that are incapable of performing high-power computations in a time frame of weeks would not be profitable for the drug industry. Assuming a 20-year patent, drug companies have an average of 6 years exclusivity to fully recover their $1.2 B investment[12], while maintaining their operating costs. The lack of a latest generation computer system appropriate for this research will displace R&D activities to other countries including the USA, China and Japan, putting European leadership along with the associated GPD contributions and jobs in this field at risk.
- **Personalized medicine** is a concept that will replace the out-dated idea that a single drug is the solution for an entire population. It will develop specific solutions for segments of population characterized by given genetic factors. Thanks to the recent advances in high-throughput genome sequencing, we can already access the full genomic profile of a patient in one week. Currently, sequencing centers require Petabyte systems to store patient data (1Tb

[8] Reymond et al. (2009) J. Am. Chem. Soc.

[9] Bohacek et al (1996) Medical Research Reviews

[10] Paul et al. (2010) Nature Reviews Drug Discovery

[11] http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm

[12] TUFTS University CSDD Outlook 2008

per patient) and data processing is carried out on supercomputers in the 100 Teraflops - 1 Petaflop range. Requirements are expected to increase dramatically as sequencing projects are extended to entire populations. The personalized medicine market is estimated to grow about 10% every year. The core diagnostic and therapeutic segment of the market is comprised primarily of pharmaceutical, medical device and diagnostics companies and the projections for 2015 are for reaching 452 B$[13].

- **Network medicine and Predictive toxicity**. Some diseases are not studied at the gene level (genomics), but in a more complex, pathway context. Drug effects are similarly studied at the systems biology level. Disease-associated networks containing several proteins have been reported as possible causes of some frequent adverse drug effects when their behaviour is perturbed. In addition, the networks contain drug targets linked to specific diseases. Intense research is being carried out to develop models for identifying protein network pathways that will help to understand the undesired effects of drugs and explore how they are related to network connectivity (see Figure 1). The use of complex network medicine is expected to have dramatic impact on therapy in several areas: the discovery of alternative targets; reducing toxicity risks associated with drugs; opening new therapeutic strategies based on the use of "dirty" drugs targeting different proteins; helping to discover new uses for existing drugs.
- **Tissue simulation.** The extensive use of simulation will allow significant improvements in the quality and quantity of research in this area. Simulation will help to integrate knowledge and data on the body, tissues, cells, organelles, and bio-macromolecules, in a common framework that will facilitate the simulation of the impact of factors that perturb the basal situation (drugs, pathology, etc). Simulation will reduce costs, time to market and animal experimentation. In the medium to long term, simulation will have a major impact on public health, providing insights into the cause of diseases and allowing the development of new diagnostic tools and treatments. In parallel, simulations will have a major impact on information technology. Thus, it is expected that understanding the basic mechanisms of cognition, memory, perception etc. will allow the development of completely new forms of energy efficient computation and robotics. The long-term social and economic impact will be potentially immense.

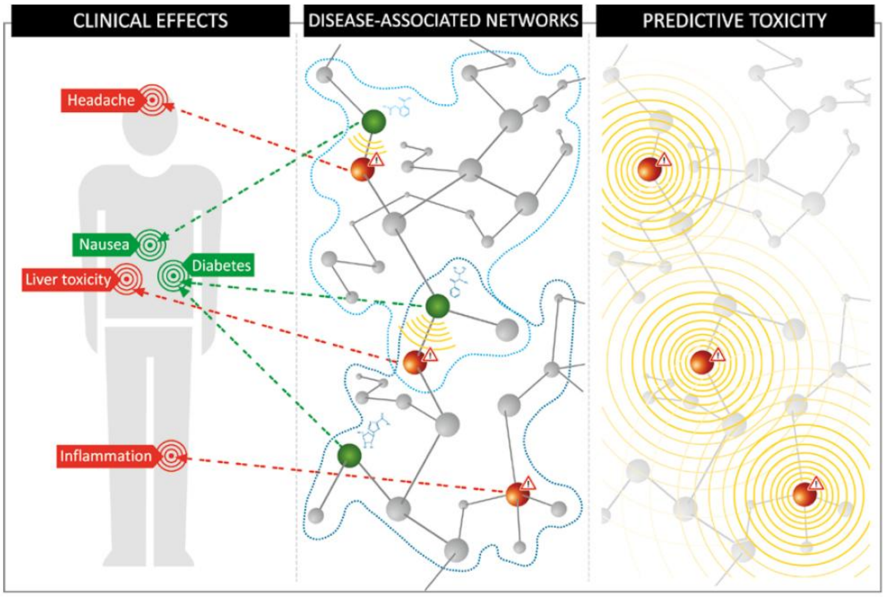

Figure 1: Development of models that can be used to do drug re-profiling and to simulate in silico drug toxicity. Patrick Aloy & Co, IRBB

[13] The Science of Personalized Medicine: Translating the Promise into Practice (2009) PricewaterhouseCoopers

4. Scientific and technical hurdles

4.1 Grand Challenges

Biological data is growing at an incredible rate and, with it, the computational needs in the field. The panel wishes to stress that, in our field, computing “capability” does not simply translate to the number of flops that can be brought to bear on a single project. It requires instead computer systems that can solve biological problems on an appropriate timescale. This point, considered crucial by the panel, becomes very clear when considering studies that can have a direct impact on the health of living patients. The panel considers that flops should be not the only parameter defining HPC capabilities. This, in turn, requires defining exactly what "Exascale resources" means. Efficient data management and fast and flexible interaction with computer resources are, in many fields of Life Sciences, at least as important as theoretical peak power. We must remember that biological data is expected to grow by a factor of 10,000 within the end of the present decade, surpassing Moore’s law (see, for example, the growth of storage in the EBI, Figure 2). Biological data is very heterogeneous (see Table 2), difficult to organize and, in some cases, is subjected to ethical restrictions on its use. Efficient management of biological data to obtain relevant information will require optimized I/O capabilities, efficient structures, post-processing pipelines (quality and validation), multi-Petabyte data sharing systems and, in some cases, significant main memory requirements. The standard protocols for the access to HPC resources are not presently compatible with the needs of research in several areas of Life Sciences, especially concerning human health, where fast data processing has a real impact on patients under clinical treatment. With these technical hurdles in mind, the Life Sciences community is already preparing for the next bio-supercomputing challenges.

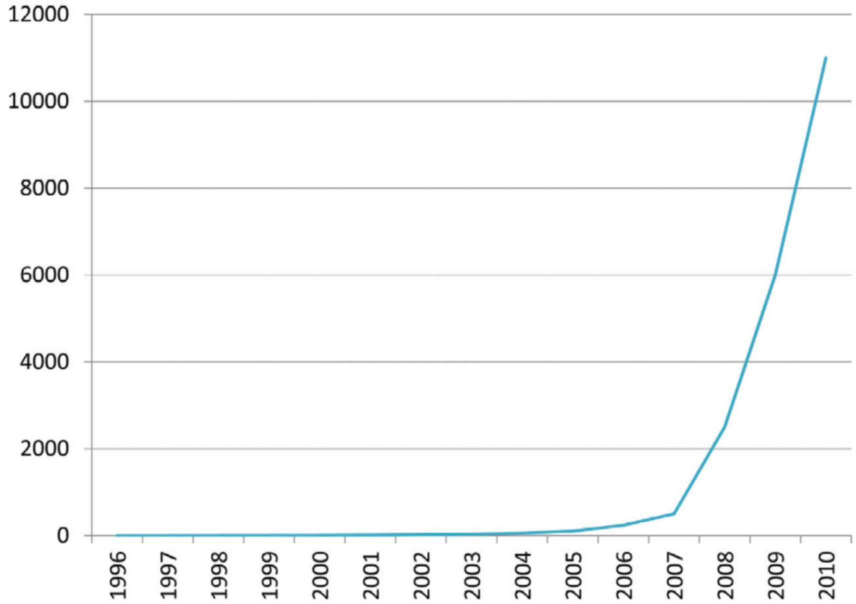

Figure 2: The exponential growth of data storage in EBI (Terabytes). Figure reproduced from the EBI annual report[14]

[14] http://www.ebi.ac.uk/Information/Brochures/pdf/Annual_Report_2010_hi_res.pdf

Table 2. Biological Data size

Data type	Size
Genomes (Eukaryote)	10^2 (genomes)
Genomes (Prokaryote)	10^3 (genomes)
Sequences	10^{11} (base pairs)
Microarray Experiments	10^5 (samples)
Mutations	10^8 (annotations)
Literature	10^7 (scientific articles)
Proteins	10^7 (database entries)
3D Structures	10^4 (database entries)
Chemical compounds	10^7 (database entries)

4.1.1 Simulation of Human Brain

The simulation of complete organs is a frontier of biocomputation (see Figure 3). These simulations are characterized by: (i) a very large, highly heterogeneous state space; (ii) multi-level modelling at the molecular, sub-cellular, cellular, tissue and organ levels; (iii) multiple time scales (from picoseconds to years); (iv) structural plasticity. Handling very large volumes of state data will require new techniques for: (i) data management (in memory databases allowing data to be used for data mining, model building, simulation, visualization, data analysis etc.); (ii) collaborative interactive visualization; (iii) computational steering of simulations; (iv) real-time monitoring of performance; (v) run-time switching and load balancing between models at different levels of abstraction; (vi) coding of parallel tasks and processes. Bandwidth and memory capabilities have for many years been growing more slowly than FLOPs and are far more constrained by energy consumption. (for example, computing 1 FLOP takes about 1 picojoule of energy, but moving the results may require 100 picojoules.). It is currently expected that early Exaflop machines will provide no more than 0.1 exabytes of memory and this may be insufficient for very large simulations (e.g. whole brain simulation). This requires new classes of memory hierarchy. Tissue simulation will benefit from heterogeneous CPUs, i.e. CPUs that combine complex cores (useful for subcellular simulation) with larger numbers of smaller cores (ideal for the cellular level). In current supercomputing, some compute-intensive processes (e.g. visualization, data analysis) are run off-line on specialized machines. This requires large data transfers that will not be practical in Exascale environments. It is therefore important that these processes should be executed *in situ.* More generally, reducing data flow will require new approaches to I/O that avoid large movements of system memory to disk.

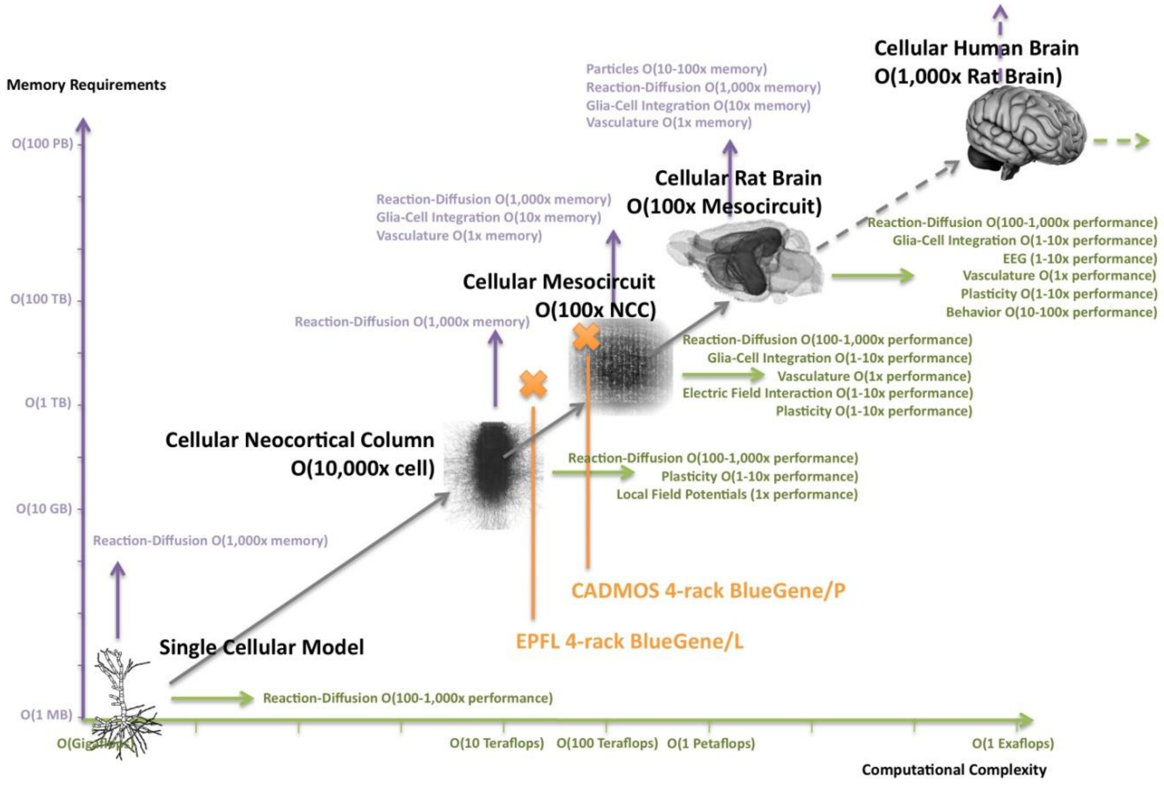

Figure 3: Human Brain Simulation Time line. Felix Schürmann, Herny Markram (Blue Brain Project, EPFL)

4.1.2 Understanding the blueprint of life

One of the main challenges of Systems Biology is the reverse engineering of the biological networks operation in normal cells (of all types) and the identification of inter-cellular communication networks which are responsible of the functioning of multi-cellular organisms. This is a first step towards a full understanding of the impact that external perturbations can cause on biological systems and, in turn, to explain complex human diseases. Current applications dealing with the "Omics" (proteomics, metabolomics, etc) generally require more system memory than intense CPU usage. Extensive information retrieval and database operations constitute the layer underlying systems biology. Problems related to data handling, data integrity and confidentiality are all important in this area. Model reconstruction and engineering will require the integration of different levels of granularity from coarse-grained models to detailed ones. Each specific application will have its own requirements, ranging from easy parallelizable code to highly integrated algorithms. The considerations related with temporal modelling and simulation of fluctuations will add additional levels of complexity (see Figure 4). A central repository of data with distributed hubs across Europe will be a major requirement of Systems Biology. Participation of major bioinformatics initiatives in Europe (such as ELIXIR) in the definition of Exascale requirements is judged as very important by this panel.

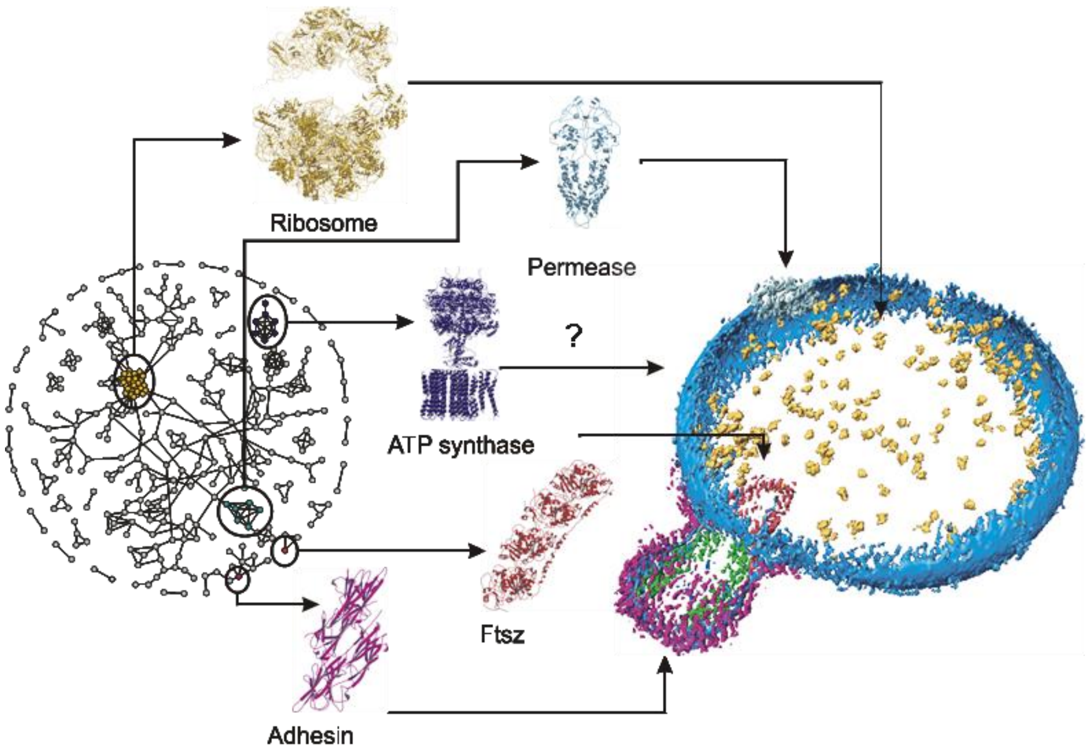

Figure 4: The complex behaviour of the cell cannot be determined or predicted unless a computer model of the cell is constructed and computer simulation is undertaken.

4.1.3 Genomic Medicine

Genetic variability affects how drugs react with each patient, sometimes in a positive manner (increasing the healing effect), sometimes in a negative manner (increasing toxic side effects) or simply by deactivating the drug response and making the treatment useless. Genomic medicine extends into the clinical arena, where it is now a routine to collect DNA sequence data in association with clinical trials and there have already been several cases, albeit so far only at specialized institutes, where whole-genome sequencing has been used as a clinical diagnostic tool. Some small, developed countries are even contemplating sequencing the genomes of their entire populations as part of their healthcare programs (Estonia, Singapore or Iceland). However, genomics generates huge quantities of data (tens of genomes this year, hundreds the next year and hundreds in two years). These data inform all aspects of biological and biomedical research and are valuable at all stages in the scientific process.

Advances in the technologies for data generation that both increase the output and decrease the cost will mean that, over the next decade, the quantity of data being produced will increase by at least a thousand-fold and maybe as much as a million-fold. There are three key aspects that HPC centers will have to deal with (i) data storage, (ii) data transportation and (iii) data confidentiality. On the other hand, while the most popular Genomics software is regularly reviewed and optimized for new systems (e.g. BLAST), a large part of the available Genomics libraries were build since '90's using inefficient script/high level languages (e.g Perl, Java or Python packages). These codes still perform well for current data loads, but may not be ready for the data challenges of the next decade. In order to avoid simplistic views of the problem, it is important to stress that bioinformatics software has been developed under a strong time pressure, given the rapid changes in technology (for example, every new generation of DNA sequencing instruments uses a completely new technological approach that produces data outputs radically different)., several codes are not open-source and can only be optimized by the code owners. Furthermore, they evolve very fast, which generates a serious problem for program optimization. Given the large amount of code available for the community, an interesting alternative could be the development of more efficient compilers for such languages.

4.1.4 Molecular Simulation

Computational simulation of biomolecules provides a unique tool to link our knowledge of the fundamental principles of chemistry and physics to the behaviour of biological systems. Appropriate Exascale resources could revolutionize the science that is possible in this area, allowing molecular simulators to decipher the atomistic clues of the functioning of living organisms. Certain grand challenge problems in this area fit well to the conventional, general-purpose, Exascale development roadmap. However, in the opinion of this panel, other vital problems will only be addressable through the development of novel architectures, not by huge machines with very large theoretical peak power but limited efficiency, for the applications of interest. This is already at an advanced stage in the USA and Japan, and there is an extreme danger that Europe will be left behind. Examples of grand challenges we will face in the close future are:

- Simulations of biological systems that are thousands of times larger than those possible today. An example would be realistic cell membrane models, including drug permeation and binding.
- Simulations that are thousands of times more computationally complex than those possible today. An example would be simulations of biomolecules based purely on quantum mechanical principles, without the approximations that are currently needed for computational tractability.
- Simulations that cover timescales thousands of times longer than those possible today. An example would be studies of the dynamics of nucleic acids, of protein folding and of molecular motors, processes that take place on the timescale of seconds.

In the medium term, a significant effort is required to ensure that legacy software can be adapted to perform efficiently on Exascale resources. In the long term, it is likely that the most groundbreaking developments will come from radically new software and algorithms that have been designed from their inception with the parameters of Exascale resources in mind and written in next-generation parallel programming languages. There is no basic core kernel that covers all bio-simulation applications. Due to the diversity of programs and problem formulations, flexible general-purpose architectures are of fundamental importance to research and development in many areas of molecular simulation. However, this is not sufficient: commercial application-specific hardware for molecular simulation available in the US (Anton[15]) and soon in Japan (MDGRAPE4) is currently two orders of magnitude faster than any software running on general-purpose parallel systems, and this significant difference in performance is likely to persist, or even grow, in the future. Anton solves standard problems in molecular simulation (it does not allow multi-scale simulations). Furthermore, Anton (and in the near future MDGRAPE4) cannot be significantly upgraded which is not a major issue for the existing machines since such algorithms have been relatively stable over recent years. The major consequence is that every time Anton developers upgrade the hardware, they will need to re-program the software again. If Europe does not invest in this area, the consequence will be that the research will move to countries with the necessary technology.

Cutting-edge simulations can produce data at rates 10-100 times larger than the rate that was typical ten years ago, and complete data sets can currently reach up to 1000 times larger sizes (that is, on the multi-terabyte scale). This trend will continue. The management and analysis of these data sets may soon present a computational challenge that is as complex and vital as the data generation. The data-model will need to be adapted (or maybe radically rethought) for optimal performances in the Exascale environment. Progresses with the major scientific challenges will also require experimental data of high quality and consistency. The ongoing close collaboration between externalists and researchers in computer simulation has proven to be of immense value for Life Sciences. The future expansion will pose new demands for computation to keep the pace of experimental approaches.

[15] The D. E. Shaw Research company changed this year his policy of not selling the ANTON machine, which is now worldwide commercially available.

4.1.4.1 Multi-scale simulation

Structural genomics initiatives are beginning to encompass many of the important organisms, while proteomics initiatives increase our knowledge of the structural space of drug-targets. Massive sequencing projects (e.g. 25,000 cancer genomes), transcriptomics and functional genomics are deciphering the molecular mechanism of cellular action, and a variety of spectroscopic techniques are providing a picture of how tissues and organisms work (see Figure 5). The challenge of Multi-scale simulation is to integrate multiple simulations layers in different scales to reach a unified vision of living systems (from atoms to tissues). There is an obvious complexity in merging such techniques that will not have necessary the same hardware requirements (memory, disk space, processor, etc.). The Exascale systems will need to integrate this multi-scale scenario providing a simple user-interface in order to become a functional platform to support the Life Sciences and Health research project of the next years.

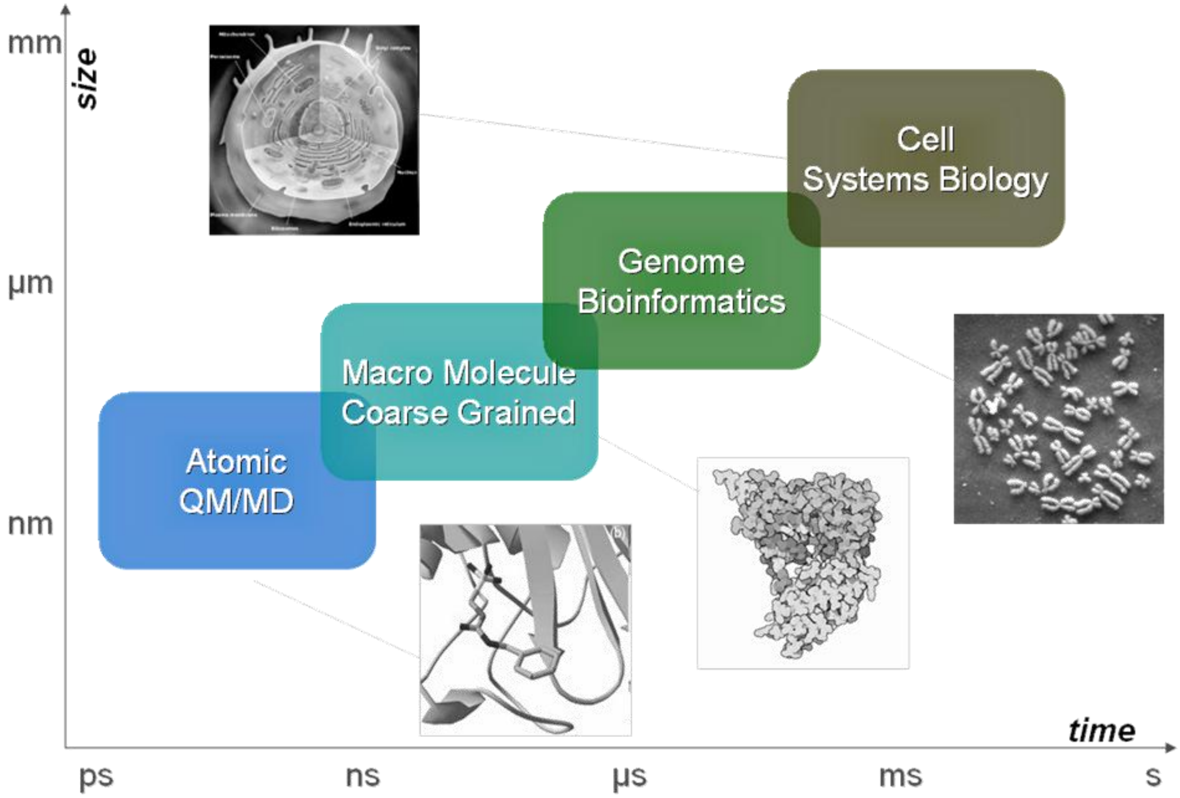

Figure 5: Multi-scale Simulation in Life Sciences

4.2 Needs of education and training

There is a clear asymmetry in education and training needs. Life scientists and clinicians must learn how to use the best of breed *in silico* science. In the other hand programmers need to better understand end user’s needs in more detail. This requires significant efforts to develop methods that are able to address pressing Life Science problems with the help of Exascale computing. There is a critical need to train the computing Life Science community in the special demands of parallel computing (programming, performance optimization, etc.), and to prepare them for using HPC in combination with systems and integrative biology. Unfortunately, there are very few places in Europe providing education in bioinformatics and computational biology. Priorities in training are:

- **Method development**: Many present-day techniques are for fundamental reasons not scalable towards the envisioned access architectures. For many challenging problems, codes addressing various aspects of the problem must be combined in an interdependent set of simulations/calculations. Method development in this direction is still in its infancy and will probably become the most significant obstacle towards the exploitation of Exascale computing for Life Science applications.

- **Memory management**: Memory can become a major bottleneck for Life Sciences Exascale computing challenges. A better understanding of how to optimize this resource will reduce future costs in updating codes.
- **Integration of optimized libraries in scripts**: Genomics and Systems Biology are fast-evolving disciplines and the driving force for code development will be non-expert programmers. It is important that bio-programmers understand which regions of the code (Exascale-demanding) should be developed (eventually by specialists) using efficient programming languages. They also need to understand how to link these libraries in high-level/scripting languages as Java, Perl or Python.
- **Data storage**: as data access (memory/disk) is much more costly than flops, training should focus on how to efficiently store information, including compression methods and database storage models.
- **Data integration and analysis**. Tools such as Hadoop and MapReduce can expedite searches through the large, irregular data sets that characterize some life sciences problems. These tools can be effective for retrieving and moving through huge volumes of complex data, but they do not allow researchers to take the next step and pose intelligent questions. A related issue is that these tools may be fine for working with a few terabytes of scientific data, but become cumbersome to use when data sets cross the 100-terabyte threshold. Effective tools for scientific data integration and analysis on this scale are largely lacking today
- **Code parallelization**: The trend of the hardware industry is to increase flop power by multiplying the number of cores. However, this trend has resulted in unbalanced architectures. This implies enormous efforts in code parallelization. In the coming years it will be important to train researchers in, at least, standard parallel programming models (MPI, OpenMP).
- **Benchmarking** Support benchmarking performance to evaluate hardware and software alternatives: Software tools such as BSC-Tools[16] are able to identify inefficient blocks of code, or bottlenecks, in applications. Such performance tools will provide developers witha powerful aid towards reaching Exascale challenges. Another vital aspect of benchmarking is the verification of the scientific quality of the results by setting standard protocols for comparison and by developing meta-servers that can combine multiple approaches. The diversity of areas makes this issue computationally complex.
- **Computational methods training:** Overall, the groups involved in Exascale challenges for Life Science will require expertise in code parallelization, applied mathematics, mathematical modelling, statistics, biology, biochemistry, biophysics, data analysis, data visualization and biological simulation. Therefore, we will need to focus on training in computational methods for those coming from a biological, as opposed to a physical sciences, background. In parallel, computational biologists need to learn how to design software and build friendly interfaces for experimentalists.

[16] CEPBA-Tools Team@BSC Home. http://www.bsc.es/plantillaF.php?cat id=52.

5. Building Exascale (2012-2020)

The priorities set out by the experts include new techniques for (i) data management and large storage systems (increase of shared memory capacity) (ii) interactive supercomputing, (iii) data analysis and visualization, (iv) multi-level simulation and (v) training. As Life Sciences and Health is a very heterogeneous field it will be necessary to have several application-oriented initiatives developed in parallel, although they can share similar agendas.

5.1 Funding of e-communities

The experts agree that many challenging Life Sciences problems cannot be addressed with present-day simulation methodologies. This problem goes beyond the adaptation of existing software to new computational platforms and involves a general lack of scalability as well as missing concepts of multi-scale, multi-model interactions that are required to efficiently exploit Exascale computing platforms. Such hurdles can best be overcome by nucleating communities of scientists, from Life Science research, bioinformatics, and computer science, who will work together to address the problems and to develop innovative solutions. Such communities can be fostered by programs such as the E-science/E-infrastructure schemes implemented in the present ICT program of FP7. A vigorous expansion of such activities is required in order to generate methods and implementations capable of correctly exploiting the new computational resources for Life Science applications. It must be acknowledged that Life Science research rewards applications and method development only in the context of successful applications. In order to generate a sustainable and effective set of codes for Life science applications it is important to nucleate and consolidate the scientific community at the European scale. The formation of broad communities targeting Exascale method development would be a tremendous benefit for R&D efforts in Europe, because it would enable the transfer of such technologies to the European end-user, generating a competitive advantage over other regions.

The experts of the panel are eager to apply the model of USA co-design centers focused on Exascale physics applications, such as the Centre for Exascale Simulation of Advanced Reactors (CESAR), the Co-Design for Exascale Research in Fusion (CERF), the Flash High-Energy Density Physics Co-Design Center, or the Combustion Exascale Co-Design Center. The idea is to create a center with academic and industrial participation, with computational and experimental interfaces, oriented to the Life Sciences and Health field. PRACE-Tier 0 centers will be a valuable resource for building centers that could focus on computational challenges for Tissue Simulation, Molecular Dynamics, Cell Simulation, Genome Sequencing and Personalized Medicine. This would provide the opportunity to capitalize on the Exascale for the strategic sector of Health and to synchronize the timeline of grand challenges, such as the Human Brain Project (which targets completion in 2023, if appropriate computational facilities are available; see Figure 3). Without European funding in Exascale initiatives, Europe will lose the unique and strategic advantage it now has to reach a dominant position in simulations in Life Sciences and medicine.

5.2 Exascale Skeleton: Pillar Applications in Life Sciences

While most of the software in use today “could” be used in the Exascale, most of the software that “should” be used has not yet been developed. On the other hand, there are software packages available today whose “functionality” (but not necessarily the code itself) needs to be ported to Exascale platforms. These applications will currently not run efficiently on Exascale computers without enormous efforts in method development. Concerns exist in the panel over the fact that most current algorithms cannot scale up to using 10^5 or 10^6 slow processors. Rather, there is a need to completely reconsider which parallelization approaches should be used for Exascale Pillar Applications (see below), and then to adapt software to these new approaches. Analyzing the challenges facing the Life Sciences, we have defined ten "Pillar" applications that should cover the full range of Life Sciences HPC needs (see Figure 6):

- **Quantum Chemistry** The current capability of first-principles quantum chemistry is used to study neurotransmitters, helical peptides and DNA complexes. Quantum chemistry calculations are precise but expensive. Exascale should make feasible calculations that are unthinkable today. The vital applications are Dalton[17], GAMESS[18], Gaussian[19] and CPMD[20].
- **Chemical Informatics** It is becoming unfeasible to fully explore and predict 1D, 2D and 3D chemical properties of small molecules with databases of tens of millions of compounds. Drug discovery based on small molecules will need to deal with the increasing size of databases (up to 1 billion entries today). Vital applications such as OpenEye[21] should be ready for Exascale systems within 2020.
- **Stochastic Models and Biostatistics** Stochastic methods will be applied to model complex biological systems, to simulate large coarse-grained systems, to sample conformational space for molecular docking, or to predict secondary structure of RNA. Personalized medicine is based on the so-called Single Nucleotide Polymorphism (SNP) association studies to identify mutations as bio-markers for genes that predispose individuals to diseases. The existing multi-SNP methods are only capable of handling 10 to 100 SNPs. Exascale systems should provide methods which could handle much higher dimensionality
- **Sequence Analysis** With the increased amount of data generated in laboratories around the world, basic protein and DNA sequence-based calculations are becoming a significant bottleneck in research. For example in phylogeny (reconstruction of ancient proteins), right now, Bayesian approaches cannot be applied to more than 200 sequences (200 base pair long), and new methods will increase the complexity. Vital applications of this are BWA[22], BLAST/BLASTMPI[23], CLUSTALW[24], HMMER[25] and MrBayes[26].
- **Molecular Modelling** Molecular Modelling is a key discipline for rational drug design. Computational tools in this area allow scientists to model pharmaceutical target structures and calculate protein-drug docking energy. Molecular Modelling represents one of the main Exascale challenges. Vital applications include Gromacs[27], AMBER[28], NAMD[29], Autodock[30], Glide[31], Dock[32], Flexx[33], FTDock[34], LigandFit[35] and ROSETTA[36].

[17] http://dirac.chem.sdu.dk/daltonprogram.org/

[18] http://www.msg.ameslab.gov/gamess/

[19] http://www.gaussian.com/

[20] http://www.cpmd.org/

[21] http://www.eyesopen.com/

[22] http://bio-bwa.sourceforge.net/

[23] blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download

[24] http://www.ebi.ac.uk/Tools/msa/clustalw2/

[25] http://hmmer.janelia.org/

[26] mrbayes.sourceforge.net/

[27] www.gromacs.org/

[28] http:// ambermd.org/

[29] www.ks.uiuc.edu/Research/namd/

[30] autodock.scripps.edu/

[31] www.schrodinger.com/

[32] dock.compbio.ucsf.edu/

[33] www.biosolveit.de/FlexX/

[34] www.sbg.bio.ic.ac.uk/docking/ftdock.html

[35] www.accelrys.com

- **Network Medicine** In the recent years it has become apparent that many common disorders such as cancer, cardiovascular diseases and mental diseases are often caused by multiple molecular abnormalities. As mathematical systems theory shows, the scale and complexity of the solution should match the scale and complexity of the problem. Network medicine has multiple potential biological and clinical applications. For example, the understanding of the effects of cellular interconnectedness may offer better targets for drug development, more accurate biomarkers to monitor diseases and better disease classification. Exascale computing will be the necessary infrastructure to move from a static to a dynamic understanding of biological pathways and protein interaction networks (the human interactome connects 25,000 protein-coding genes and ~1,000 metabolites).
- **Cell Simulation** It is estimated that eukaryotic cells contain about 10,000 different proteins (with close to 1,000,000 copies per protein). Whole cell and sub-cellular simulations (e.g. membranes) will require huge computational resources and efficient coupled multi-scale simulation applications. This is a main task area of the Brain Simulation project
- **Tissue Modelling** As described in previous sections, tissue simulations (like heart, respiratory system and brain) are going to be key issues for animal substitution in drug testing. Future medicine will be based on virtual patient models and this should increase both drug safety and efficacy.

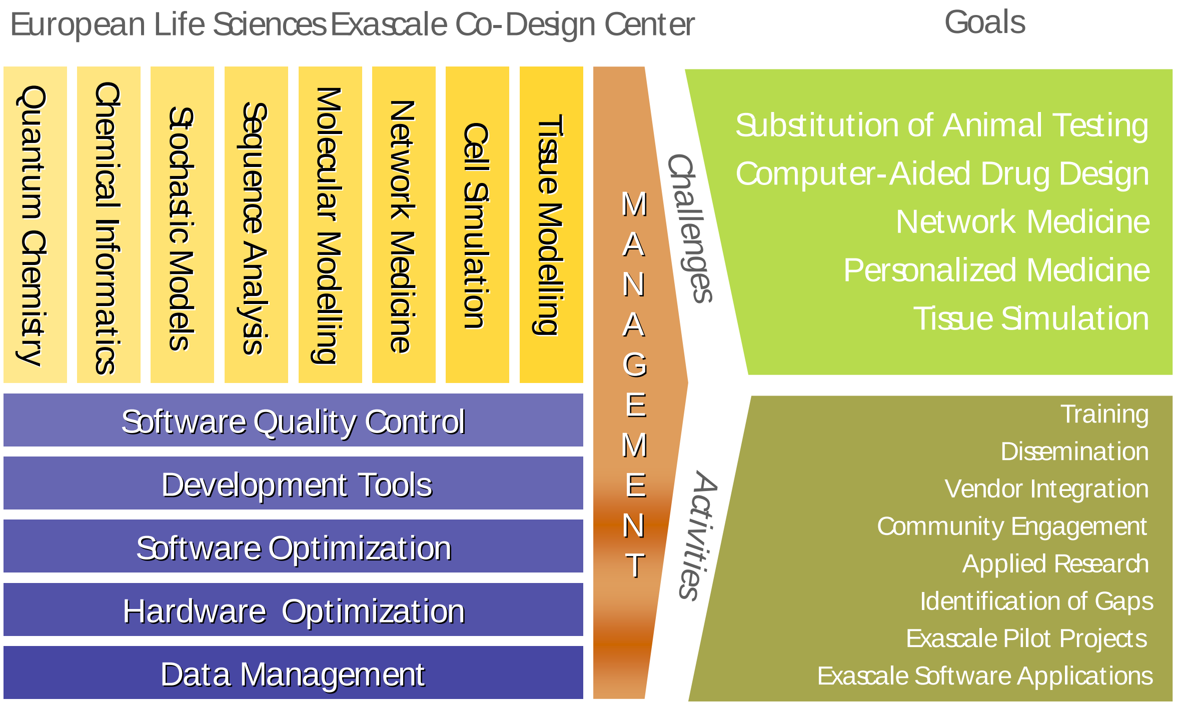

Figure 6. Proposed schema of Exascale co-design center. The center is organized in 14 teams: 8 scientific pillars, 5 cross-cutting units and 1 management team. The main goal of the center is to reach the Exascale in 2020

[36] http://www.rosettacommons.org/

5.3 Transverse issues

Transverse issues are technical areas that cut across the 10 scientific pillar applications. We identify five main transverse issues for the development of Exascale computing in the Life Sciences (see Figure 6)

- Software Quality Control. This covers application testing and scalability benchmarking. It will not be possible to reach Exascale performance if we do not coordinate best practices among the different application pillars.
- Development Tools. The Life Science community is not inherently expert at developing code. Programming in Java, Perl or Python has been widely adopted for ease of use reasons. Unfortunately, this is often coupled to poor performances and poor scalability. This section will target the development of new and efficient development tools for non-experts.
- Software Optimization. This area should analyse code run-time performances to identify bottlenecks and improve calculation times in the context of future Exascale computers. Some parts of the codes developed in high-level languages should be moved to more efficient alternatives. Key applications will need HPC programming expertise to improve parallelization, to scale to multiple cores, and to be adapted to GPU-based systems, FPGAs, and very probably also to heterogeneous architectures.
- Hardware Optimization. It has already been demonstrated that optimization of Molecular Dynamics based on developing specific purpose hardware is much more competitive than strategies purely based on software optimization. This area will study the use of hardware-specific solutions to unlock bottlenecks in the Life Sciences.
- Data Management. Data management represents one of the main issues in Life Science that affects all application pillars. This area should deal with confidentiality (source to computing data links, administration, security protocols and back up systems) and increasing data size (dynamic data access, data compression etc.).

5.4 Timeline

To implement the Life Sciences and Health Exascale computing applications the experts propose a timeline to build an Exascale center for Life Sciences (see table 3). The center will require the combined expertise of vendors, hardware architects, system software developers, Life Science researchers and computer scientists working together to make informed decisions about features in the design of the hardware, software and underlying algorithms. The aim is to organize a center with distributed nodes in Europe, with the Barcelona Supercomputing Center the suggested host for the headquarter-node. The kick-off team (2012) will be composed of 20 members organized in units addressing Management duties, the Scientific Pillars and the Transverse issues. The mission of the kick-off team is to design a full Field Work Proposal and, by 2013, to approve a deployment plan for the center and fix the goals of the pilot projects (to be completed before 2018). Pilot projects have tasks, defined cooperatively by the center, by the scientific community and by industry partners that will target high-impact areas of Life Sciences: Substitution of Animal Testing, Computer-Aided Drug Design, Network Medicine, Personalized Medicine and Tissue Simulation.

The kick-off team will need to grow gradually in order to cover the full range of Pillar Applications (8 in total, see Figure 6) and to tackle every one of the transverse issues (5 in total). The center should start to be fully operational before 2014, hosting in a first phase of typically 3 to 8 research members per group/unit, and approximately 75 employees. Before 2016, the human resources should almost double in size and groups will be typically composed of 5 to 12 researchers. Working in collaboration with hardware and software vendors, the center should reach the goals of Exascale pilot projects by 2018. After this milestone, the co-design center should work to gradually provide Exascale computing to the Life Sciences and Health community by 2020, as well as to strengthen the role of staff dedicated to training and dissemination activities.

Table 3. Time line Milestones

Year	Milestones
2012	Kick-off team and Field Work Proposal (approx. 20 employees)
2013	Approval of Development plan and Pilot Projects
2014	First deployment phase (approx. 75 employees)
2016	Second deployment phase (approx. 135 employees)
2018	Completion of Exascale Pilot Projects
2020	Full deployment of Exascale computing for Life Sciences (approx 180 employees)

5.5 Provisional Costs

The calculated cost (see table 4) to launch and support the Life Sciences Exascale co-design center is €153M (total cost for 2012-2020). This cost does not include hardware and low-layer software development. The center will not develop future hardware architectures, but will ensure that they are well suited to Life Science applications. Human resources are given in Man * Year.

Provisional costs are given in K€ with a yearly flat rate of 100k€/FTE

Table 4. Provisional costs for Life Sciences Exascale co-design center

Resources	**Human Resources by year → 2015**	**Integrated (4 years) Provisional Costs 2012 → 2015**	**Human Resources by year 2016 → 2020**	**Integrated (5 years) Provisional Cost 2016 → 2020**
Management Team	10	4 000	40	20 000
Pillar Research Groups	50	20 000	150	75 000
Trans-verse Units	10	4 000	60	30 000
Total resources	70		250	
TOTAL in k€		**28 000**		**125 000**

6. European strengths and weaknesses

6.1 Sources of competitiveness for Europe (Strengths)

6.1.1 ELIXIR

The Life Science community is well organized in a network structure at the national (for example, the Spanish National Institute of Bioinformatics) and European levels (ELIXIR). This facilitates information flow and the acceptance by the community of new software or innovative IT architectures. The European Strategy Forum on Research Infrastructures (ESFRI) was created in 2002. The Council of the European Union mandated ESFRI to create a European Research Infrastructures Roadmap which was first published in 2006 and identified 10 Research Infrastructures (RIs) in Biological and Medical Sciences (BMS) which will be essential in fulfilling Europe’s potential to meet the Grand Challenges. ELIXIR represents an opportunity to develop an engagement strategy to ensure that the Genomics and Life Science community take full advantage of Exascale HPC facilities. Other relevant initiatives are the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI), the European Advanced Translational Research Infrastructure in Medicine (EATRIS), European Clinical Research Infrastructures Network (ECRIN) and the European Research Infrastructure on Highly Pathogenic Agents (ERINHA)

6.1.2 EMBL-EBI

The European Bioinformatics Institute (EBI) is part of the European Molecular Biology Laboratory (EMBL) and also supported by the Welcome Trust. The EBI hosts the major, core biomolecular resources of Europe – collecting, archiving and distributing data throughout Europe and beyond (at June 2010 there were on average 4.0 million requests per day). EMBL-EBI coordinates the preparatory phase of the ELIXIR.

6.1.3 Participation in research consortiums such as ENCODE, ICGC and iHEC

Another source of competitiveness for Europe is the participation in large scale genomics projects, in which massive sequencing is a fundamental component. An example of this is ENCODE, an NIH lead project addressing the detailed mapping of all the genetic control elements in the human genome and other species. There is also the iHEC, an international consortium, including a leading European consortium, which proposes the collaboration in the mapping of all the elements related with epigenetic control in hundreds of samples representing developmental stages and diseases. In the case of the International Cancer Genome Consortium (ICGC), the association of research initiatives from Australia, Canada, China, France, the UK, Germany, India, Italy, Japan, Spain, USA and two European Union consortia. ICGC runs in parallel and competes with the USA based Cancer Genome Atlas, lead by the NIH National Cancer Institute. The consortium will sequence the genome of 2,500 patients of cancer, what with the current technology is far less of a challenge than the real tour-de-force that the analysis and integration of this information represents for the supercomputer centers in Europe and elsewhere.

6.1.4 Human Brain Project

The European Blue Brain project is currently the only project in the world working on biologically detailed simulations of the whole brain using high performance computing. The “Human Brain Project” (part of the FET Flagship Program) will continue and expand Europe’s strategic advantage in this domain. The Human Brain Project will systematically collect and analyse data on the brain, derive organizing principles and build brain models with as much biological detail as technically possible. As brain science and medicine advance, the models will evolve to incorporate ever-richer data and knowledge. Building such models represents an enormous challenge that will shape the future of supercomputing, and provide the technologies we need to create realistic simulations of life processes. Combined with high-level mathematical theories of brain function, they will make it possible to build a new class of brain-like hardware devices and computer architectures. As the project models larger and larger numbers of neurons in ever-greater detail, it will need more and more computing power. Consequently, one of its key activities will be the development and management of a large-scale supercomputing infrastructure. This will be developed at the Julich Supercomputing Center. The Human Brain Project will use this computing power to completely change current methods of working in neuroscience and medicine. Simulation will provide clinical researchers with a powerful new tool, offering them a new way of understanding brain diseases and helping to design and test new drugs, more effective than any drug we have today. Many other Life Science and medical projects (cancer, heart, diabetes, kidney, bone, blood, etc.) will potentially benefit from the advances made in the brain project and place Europe in the leading position in simulation science applied to Life Sciences and medicine. The main partners are thirteen of the most important research institutions in Germany, the UK, France, Spain, Switzerland, Sweden, Israel, Austria and Belgium. Overall the project brings together more than a hundred organizations covering disciplines as different as neuroscience, genetics, applied mathematics, computer science, robotics and the social sciences.

6.1.5 FET Flagship in Personalized Medicine

The aim of the IT Future of Medicine (ITFoM) project is to develop models of human pathways, tissues, and ultimately of the whole human, to create a “virtual patient” that will enable physicians to identify prevention schedules and treatments adapted to each person. Innovation in Information and Communications Technology (ICT) and computing has been primarily driven by the requirements of physics and commercial applications such as entertainment. But the growing requirements of individualised medicine are likely to overcome those of all other ICT development fields in the next future. In the first five years, the project is expected to establish integrated molecular/anatomical prototype models of man and develop IT techniques to individualise these models based on high throughput data sources. In the following five years, the project aims to develop an infrastructure for model-based individualized medicine and to interact with relevant stakeholders, governments, healthcare and insurance systems to implement this approach throughout the healthcare system. ITFoM brings together world-leading research groups from Europe and beyond. The Project will also involve industry representatives such as IBM, Intel, XEROX, Roche, Illumina, Life Technologies and Agilent.

6.1.6 PRACE

PRACE links the largest supercomputer centers in Europe and aims to provide a competitive supercomputer framework for European researchers. PRACE also facilitates the transfer of information between centers at both administrator and researcher levels and fuels research in computer science. In summary, PRACE defines an optimum framework to power Exascale initiatives.

6.1.7 Pharmaceutical Industry

Health is an area of science where Europe is a world leader. Europe is the major worldwide pharmaceutical drug-producing region (currently dominating its principal competitors, the USA and Japan). Pharmaceutical industry is a driving force for the economy and R&D activity in Europe. Small molecule design, target identification and protein binding predictions are key issues for computer-based drug design. These methods will increase the success rate for finding drug candidates, improving the competitiveness of our pharmaceutical industry. Some pharmaceutical companies have shown a high interest to approach GRID/cloud computing solutions (such as those proposed by Amazon or by the PRACE supercomputer network). As an example, the Pharma HPC Forum (sponsored by Intel) was formed in late 2007 by the main pharmaceutical companies (including AstraZeneca, Bristol-Myers Squibb, Eli Lilly, GlaxoSmithKline, Johnson & Johnson, Merck, Novartis, Pfizer, Sanofi-Aventis, Schering-Plough, and Wyeth). One of the main issues for these companies with respect to using external computing services has been the lack of security protocols to preserve industrial secrets.

6.1.8 Computational Biology and Bioinformatics Software

Europe is a major provider of software for biocomputing, especially in the molecular simulation arena (CPMD, GROMACS and DALTON among others). It also has a leading position in developing data mining tools for bioinformatics. European research laboratories are also leaders in the development of models to predict drug activity. Europe has a clear opportunity to dominate the development of Life Science codes for Exascale systems.

6.2 Europe Weaknesses

Many Grand Challenge problems cannot be solved by merely increasing size or resolution. They often require orders-of-magnitude improvements in scaling through radical hardware and software development. Competitors (notably the USA) with privileged access to specialized resources (e.g. Anton in the molecular dynamics world) will be a hard to match. The fact that the main hardware makers (Intel, NDIVIA, etc.) are mostly USA based “might” make Europe realise that they need their own chip makers.

In addition, there is currently no coherent unified data layer available for bio-data. There are too many small and disparate repositories with sub-optimal data management standards. This is intrinsic to the system where end users are looking for fast and easy data access, while developers struggle with file format standards and data integration. Europe has copies of many vital databases, leading some of them (e.g. ENSEMBLE, PFAM-Domain) but many others were created, and are maintained, outside Europe (Genebank, Pubchem, etc.)

6.3 Potential collaborations outside Europe

6.3.1 D E Shaw Research

The company D.E. Shaw Research organisation conducts basic scientific research in the field of computational biochemistry under the direct scientific leadership of David Shaw. The laboratory is involved primarily in the design of novel algorithms and machine architectures for high-speed molecular dynamics (MD) simulations of proteins and other biological macromolecules and the application of such simulations to basic scientific research in structural biology, biochemistry, and computer-aided drug design. The D.E. Shaw Research Lab (DESRES) has developed the Anton computer (an application-specific computer for molecular dynamics simulations). Anton is estimated to be 3-5 years ahead of any other supercomputer system, with the estimated price of approximately 15 M$. The experts recommend using the Exascale project to approach D.E. Shaw and try to obtain access to Anton computers within European facilities.

6.3.2 AMBER developer team

AMBER is a set of molecular mechanical force fields for the simulations of biomolecules (which are of the public domain, and are used in a variety of simulation projects) and a package of molecular simulation programs. AMBER is developed in an active collaboration between David Case at Rutgers University, Tom Cheatham at the University of Utah, Tom Darden at NIEHS (now at OpenEye), Ken Merz and Adrian Roitberg at the University of Florida, Carlos Simmerling at SUNY-Stony Brook, Ray Luo at UC Irvine, Ross C Walker at the U California San Diego and many others. AMBER developers have considerable experience in porting Molecular Dynamics to HPC systems and their package is widely used by Computational Biology community. Developers of the AMBER team taught a workshop co-sponsored by the BSC and preliminary discussions about cooperation towards the Exascale were maintained.

6.3.3 NAMD developer team

NAMD is a parallel molecular dynamics program for UNIX platforms designed for high-performance simulations in structural biology. Simulation of large molecules requires enormous computing power. One way to achieve such simulations is to use parallel computers. In recent years, distributed memory parallel computers have been offering cost-effective computational power. NAMD was designed to run efficiently on such parallel machines for simulating large molecules. The NAMD package includes Force Field Compatibility, Efficient Full Electrostatics Algorithms, Multiple Time Stepping, Interactive MD simulations and Load Balancing. NAMD was created at the University of Illinois at Urbana-Champaign. The principal investigators involved in its development are Klaus Schulten (Theoretical and Computational Biophysics Group), Laxmikant V. Kalé (Parallel Programming Laboratory) and Robert D. Skeel (Bionumerics Research Group, left Illinois in 2005).

6.3.4 NVIDIA

HPC architectures are moving to CPU-GPU hybrid architectures. Many bio-software show good performance when ported to GPU-based systems, but legacy code seems to be a major bottleneck. The CUDA Research Center Program (http://research.nvidia.com/content/cuda-research-centers) recognizes and fosters collaboration with research groups at universities and research institutes that are expanding the frontiers of massively parallel computing. Institutions identified as CUDA Research Centers are doing world-changing research by leveraging CUDA and NVIDIA GPUs.

The Barcelona Supercomputing Center has been named by NVIDIA as a CUDA Research Center since 2010. Through this initiative the bio-application Protein Energy Landscape Exploration (PELE) is going to be ported to GPUs.

6.3.5 The Beijing Genomics Institute (BGI)

In January 2010, BGI purchased 128 HiSeq 2000 sequencing systems, representing the largest single order for next-generation sequencing systems to date. The BGI was founded in Beijing in 1999 with the mission of supporting the development of science and technology, building strong research teams, and promoting the development of scientific partnership in genomics field. With the goal of excellence, high efficiency, and accuracy, BGI has successfully completed a large number of projects such as contributing 10% to the International Human HapMap Project, carrying out research to combat SARS and completing the sequencing of the deadly E. Coli strain that appeared in Germany. Other relevant NGS centers are the Genome Center of Washington University (that sequenced the 25% of Human Genome), the Baylor Human Genome Sequencing Center (HGSC), the MIT Broad Institute and the US Joint Genomics Institute (which has a specific access to the National Energy Research Scientific Computing HPC facilities).

6.4 Existing funded projects and funding agencies

PRACE and the Seventh Framework Programme (FP7) are the main funding sources so far for European Exascale initiatives. The FET Flagship call is a FP7 program to fund ambitious large-scale initiatives that aim to achieve a visionary goal. The Human Brain Project is a FET preparatory study led by Prof. Henry Markram (http://www.humanbrainproject.eu/). Under the same call, the IT Future of Medicine (ITFoM) is also in preparation. ITFoM is led by Prof. Hans Lehrach (Max Planck Institute for Mol. Genetics, Berlin) and aims at making general models of human pathways, tissues, diseases and ultimately of the human as a whole. This is the first time that the huge IT implications of worldwide, individualized patient care will be addressed in combination with genomics and medical requirements.

7. Conclusions

HPC computation in Life Sciences and Health is a large and diverse field. This diversity cannot be neglected when referring to our community. The panel of experts represents the main areas of Life Sciences: Genomics, Systems Biology, Molecular Simulation and Medical Simulation. Exascale computation will have an increasing impact in areas such as computer-aided drug design, computer modelling of animal testing, network medicine (network-based approach to diseases), personalized medicine and tissue simulation. Without an Exascale platform, the scientific community cannot deal with the grand challenges facing this field, including the simulation of human brain, the understanding of the blueprint of life, genomic-based medicine and multi-scale molecular simulations of increasingly large biological assembles. The panel wishes to stress its major concern, namely that Exascale problems in the Life Sciences will not be solved simply by Exaflop computing power. In fact, the panel agrees that Exascale should be taken to mean much more than Exaflops.

The experts also stress that many challenging Life Science problems cannot be addressed with present-day simulation methodologies. The experts of the panel are very open to apply the model of USA co-design centers focused on Exascale physics applications. The development of such a center within Europe and targeting Life Science problems has an estimated cost of 153 M € (2012-2020). While Europe is unlikely to be able to compete with USA, China and Japan in developing Exaflop computers, it can be a major provider of software for biocomputation, supporting its already strong Pharmaceutical industry. Europe has now the opportunity to further this goal by supporting the first Exascale co-design center for future medicine.