



BBSRC REVIEW OF THE COMPUTATIONAL REQUIREMENTS OF THE BIOLOGICAL SCIENCES

FINAL VERSION

EXECUTIVE SUMMARY

1. BBSRC has recognised the need to develop a fuller picture of the computational requirements of the biosciences to underpin future strategy development and investments in infrastructures and resource provision for quantitative and integrative biology. The BBSRC Tools and Resources Strategy Panel agreed that a review be undertaken by an independent expert group to examine past and current usage of computational resources provided at the national, regional and local level, as well as provide a forward-look on the computational requirements of the biosciences.
2. There is a broad range of computational architectures available to the UK bioscience community, from desktop PCs to local clusters, and local and national high performance computing (HPC) facilities. The review group observed that there was a broad range of local HPC architectures available to the biosciences within UK higher education institutions, although these were primarily used by the engineering and physical sciences community. The group concluded that the current national HPC facility was currently only suitable for those research areas which were numerically intensive and were amenable to data parallelisation e.g. biomolecular simulation. It did not lend itself easily to research areas which required parallelisable processing capacity e.g. bioinformatics. The needs of areas such as systems biology, computational ecology and image analysis/reconstruction will be met by a range of developments in both hardware and software provision.
3. In reporting on the current and future requirements of the biosciences, four major bottlenecks in the provision of computational infrastructures for the biosciences were identified. These were:
 - data issues, including data storage infrastructures, management and curation;
 - awareness about, and access to, the national HPC facilities and support;
 - the limited computational skills of traditionally-trained bioscientists and the lack of appropriate training directed towards experimental biologists wishing to develop computational skills;
 - the lack of suitable software tools currently available on HPC platforms.
4. The review group acknowledged the roles of both BBSRC and the organisations involved in the management of national HPC within the UK in encouraging usage of the national facilities by the biosciences.

Conclusions

5. Conclusion 1
Use of the national HPC facility is a niche activity in the biosciences. There is uncertainty over whether the previously anticipated expansion of usage will be realised. There is a need to bridge the gap between biosciences and HPC at the national level.
6. Conclusion 2
Local 'high-end' HPC facilities are not widely used in the biosciences. This mirrors the situation with the national facility. There is a need to bridge the gap between biosciences and HPC at the local level.
7. Conclusion 3
The national facility is used mainly for biomolecular simulations. The local 'high end' HPC facilities show a greater range of bioscience applications. Efforts to broaden the use of the national facility in the biosciences have met with mixed success and there is an apparent lack of sustained engagement. Most bioscience users of the national facility participate in software development, working to increase the utility of the service rather than being a customer.
8. Conclusion 4
The national HPC facility is useful for applications that are numerically intensive and require data parallel approaches. This will only cover a subset of the computationally intensive biological research challenges.

9. Conclusion 5

There are major bottlenecks that limit the widespread use of computational approaches in the biosciences (data, skills and training). Furthermore, there is a gulf between most bioscience users and the national HPC service with bottlenecks in skills and training, awareness and access, and software.

10. Conclusion 6

There is a growing demand for computation in the biological sciences and a range of architectures (and associated software) will be required to tackle the challenges ahead. It is important that bioscience utilises all appropriate infrastructures. Collapsing into the lower levels of the Branscombe pyramid could be to the detriment of scientific advancement in some key areas.

Recommendations

11. Recommendation 1

BBSRC should work with partners to ensure that national HPC services support applications that are of value to significant numbers of users within the life sciences community, where the users are interested in analysing data of interest and not necessarily engaged with computational methods development.

12. Recommendation 2

BBSRC should work with other partners to develop a programme of activities to increase awareness of, access to, and use of HECToR by the bioscience community. This will require additional BBSRC investment but, overall, should increase the benefits arising from the Council's original financial commitment.

13. Recommendation 3

BBSRC should continue to be involved in future national HPC services. However, there would need to be a significant increase in bioscience users for BBSRC to consider any increase above the current financial commitment.

14. Recommendation 4

It is essential that cutting edge bioscience research is underpinned by a range of appropriate computational hardware and software. BBSRC should:

- support the development and long-term sustainability of appropriate software tools for the biosciences through their incorporation into appropriate funding mechanisms;
- revisit the decision to cease a dedicated equipment funding stream.

GENERAL CONTEXT

15. The co-ordination of High Performance Computing (HPC) activities for UK academic research is the joint responsibility of the Research Councils. The Research Councils are responsible for setting the strategic direction of HPC developments, planning the procurement and location of HPC services, establishing access arrangements and promoting widespread use. The Engineering and Physical Sciences Research Council (EPSRC) manages the cross-Council High Performance Computing Programme on behalf of, and in consultation with, the other Research Councils.
16. BBSRC is a minor partner in the cross-Council HPC programme, contributing approximately 5% of the budget of the current national HPC service. As such, any BBSRC strategic review of HPC in the biosciences sits within the broader landscape of the cross-Research Council HPC Programme.
17. The High End Computing Strategy Committee (HSC) advises the cross-Council HPC Programme on the strategy for HPC. HSC advises on procurements, access to facilities and the nature of support for High End Computing. HSC receives advice from a Technology Watch Panel (TWP) on many aspects of the technology requirements for internationally-competitive research in the UK using HPC. This includes developments in both hardware and software tools. HSC also receives advice from the Applications Panel which is responsible for monitoring and advising HSC on the development of High End Computing applications, and in particular for producing an Applications Roadmap for the UK. It works closely in collaboration with the TWP. BBSRC nominates user representatives to the HSC and the TWP.
18. Recent high-level strategic outputs (reports and activities), arising from the cross-Council HPC Programme, that have contributed to the development of UK academic HPC include:
 - Publication by EPSRC of an International Review of HPC in the UK (2005)¹;
 - A Strategic Framework for High End Computing² (2006) and the associated publication Challenges in High End Computing³ (2006);
 - Commencement of the process for identifying the needs for, and starting the procurement of, the next UK national HPC resource (to replace HPCx and overlap with HECToR). An initial case was submitted to the Large Facilities Capital Fund prioritisation exercise, and EPSRC is currently developing a 10-year plan for HPC provision;
 - The recent initiation of a preparatory phase project, funded under the European Strategy Forum on Research Infrastructures (ESFRI) call, for a possible European HPC procurement: Partnership for Advanced Computing (PrACE). This project will run for two years from January 2008, and EPSRC is acting as managing agent for all the Research Councils. There is no explicit commitment to multilateral funding at this stage.
19. The current national HPC service, High End Computing Terascale Resource (HECToR), represents a next-generation academic HPC resource for the UK and a significant investment for the Research Councils. EPSRC managed the procurement for HECToR on behalf of the other Research Council partners and continues to act as the managing agent for the service. The University of Edinburgh HPCx Ltd are contracting and directing the HECToR hardware, providing the accommodation and system management, as well as helpdesk, website and administration. NAG Ltd. provides the computational science and engineering (CSE) support for users of the service.
20. The current HECToR management structure is comprised of three groups, a strategic management board (H-SMB) of partner organisations, a scientific advisory committee (H-SAC) and a CSE performance review group (CSE-PRG). Community academic representation is made via the H-SAC.

BIOSCIENCE CONTEXT

21. The Report 'Challenges in high end computing'³ (2006) identified a series of challenges for the non-medical life sciences:
 - Systems biology: to position Europe in the next four years to host the first '*in silico*' cell;
 - Chromatin dynamics: study of nucleosome dynamics;
 - Large scale protein dynamics: large conformational changes in proteins;
 - Protein association and aggregation: *in silico* protein complex formation and simulation of crowded, extracellular protein environments;
 - Supramolecular systems; systematic analysis of protein 'machines' such as ribosomes and polymerases.
22. In identifying these challenges, the report set out new targets for the biomolecular simulations community, a research area that has obtained most benefit from HPC resources to date, together with additional challenges outside of the established HPC bioscience areas, picking out systems biology in particular.
23. Systems biology and the *in silico* cell ambition lie at the heart of BBSRC's 10-year vision and associated strategic and delivery plans. BBSRC is therefore committed to enabling the use of HPC in bioscience areas that have not previously engaged with HPC to a significant degree, particularly in areas of high strategic importance.
24. At a BBSRC Tools and Resources Strategy Panel meeting, informed by a paper written by Professor Mark Sansom (Oxford), the Panel endorsed the BBSRC commitment to the HECToR project, and considered the level of investment (£3.3M/ equates to 5% of total) to be appropriate. This commitment represented a significant increase over BBSRC investment in the previous national high performance computer (HPCx), which equated to approximately 1.5% of project costs. The increased commitment reflected an anticipated expansion in use into new areas such as systems biology.
25. The Panel further recognised the need to develop a fuller picture of the computational requirements of the biosciences and agreed that a strategic review should be undertaken. This would provide a measure of HECToR uptake by the bioscience community at an early stage of operation and provide an evidence base to inform future BBSRC investment in next generation HPC. The review also aimed to provide some insight into computing architectures and their provision in the biosciences.
26. The BBSRC Tools and Resources Strategy Panel agreed that the review should focus on a number of biological research areas identified as having significant computational requirements. These were systems biology, biomolecular simulations, structural biology, bioimaging, mathematical modelling (e.g. of agricultural systems) and bioinformatics.

REVIEW PROCESS

27. The terms of reference for the review were agreed by the Tools and Resources Strategy Panel:
 - To describe the current computational infrastructure provision in the biosciences (HPC and non-HPC) and key areas of application, and to identify any bottlenecks in this provision;
 - To develop a view of the computational needs of UK bioscience research over the next five to ten years;
 - To provide advice to the Tools and Resources Strategy Panel on the positioning of BBSRC interest in the support and development of appropriate computational infrastructure such as the next generation HPC.

28. An expert group, chaired by Professor Steve Homans (University of Leeds and member of the Tools and Resources Strategy Panel), was established to undertake the review. The membership spanned bioinformatics and computer science, and a range of application areas including biomolecular simulations, structural biology, systems biology and bioimaging. The Review Group membership is at **Annex 2**. A representative from the EPSRC attended as an observer.
29. The Review Group met three times during the review process: in August 2008 to examine the background evidence and identify informational needs; in December 2008 to analyse the portfolio of HPC funded grants and agree a community questionnaire; and in February 2009 to consider the questionnaire responses and identify the review conclusions and recommendations. The final meeting also included a presentation from NAG Ltd. The review report was drafted by BBSRC officials (A. Collis) with input from the expert group members and chair. The final version was approved by the chair.
30. The Review Group was provided with a range of evidence including:
- A list of BBSRC funded research grants using HPC (i.e. Class I access). This covered HECToR and the two previous national academic HPC facilities (CSAR, HPCx). **Annex 3a**.
 - A list of current EPSRC-funded research projects, using HPC, that were relevant to the biosciences (i.e. Class I access). **Annex 3b**.
 - Bioscience-oriented projects supported via HECToR Class II access. **Annex 3c**. [Note: Class II access covers (a) pump-priming activity (Class IIa) aimed at introducing new-users to the national service whose previous experience may have been on university based or mid-range systems; and (b) compute time in support of distributed computational science and engineering applications aimed at researchers who do not currently have access to the HECToR system.]
 - A list of BBSRC-funded projects using university or regional HPC facilities. **Annex 3d**.
 - Details of related e-infrastructure projects funded by BBSRC's Research Equipment Initiative (**Annex 3e**);
 - A summary and the responses to a community questionnaire targeted at bioscience research areas considered to have significant computational requirements. The questionnaire is at **Annex 4**. A quantitative summary of the responses is at **Annex 5**.

FRAMEWORK FOR DISCUSSION

31. The US Branscomb Pyramid⁴ describes the hierarchy of provision of computational architectures. It shows computational resources split into four requirements, each level building upon the previous. It implies a small number of national 'leadership' facilities, underpinned by a computational infrastructure based around national/regional facilities, clusters and workstations. The Review Panel considered the applicability of the Branscomb Pyramid to UK computational architectures, noting that it had informed the construction of the community questionnaire. It was recognised that the cluster level and the national/regional facilities level might blur, especially around parallel processing architectures. The responses to the questionnaire confirmed that this was indeed the case, but distinct observations on the two tiers could still be made.
32. The driver for the review was HPC usage in the biosciences, but placed within the broader context of bioscience computational requirements. In order to maintain this focus, a number of areas were excluded from the review process. These were:
- The Research Councils e-Science programme. This will shortly be examined by the RCUK International Review of e-Science;
 - Computational services for large groups of researchers beyond those provided by HPC.
33. The review process did not yield any examples of HPC usage in other disciplines that could be used to inform thinking in the biosciences. Furthermore, only very limited information was obtained on

international HPC provision (mainly relating to hardware capacity). Industrial partnership was rarely mentioned, but where this was the case it was recognised that working closely with the hardware supplier was beneficial in pushing forward research applications.

TERM OF REFERENCE 1: TO DESCRIBE THE CURRENT COMPUTATIONAL INFRASTRUCTURE PROVISION IN THE BIOSCIENCES (HPC AND NON-HPC) AND KEY AREAS OF APPLICATION AND TO IDENTIFY ANY BOTTLENECKS IN THIS PROVISION.

Current Computational Infrastructures

National HPC Facilities

34. The current national HPC facility, HECToR, is described in detail at **Annex 1**. In order to examine the extent to which national HPC facilities have been used in bioscience research, the Review Panel examined the portfolio of BBSRC awards from 2003 onwards (**Annex 3**), together with funded projects of relevance to the biosciences from the EPSRC HPC programme that were identified as 'current' on the grants database in November 2008. Details of unfunded BBSRC applications were also provided.
35. The key features of the portfolio are set out below:
- Since 2003 BBSRC has supported eight projects. There are no new BBSRC funded bioscience projects on HECToR (class I access). The two 'live' bioscience projects were transferred from HPCx. There is currently one new Class II project relevant to the biosciences on HECToR, another was initiated on HPCx but was not transferred across. BBSRC has received seven research grant applications requesting Class I access to HECToR. One was unsuccessful and the remainder are undergoing peer review.
 - There are five EPSRC funded projects of relevance to the life sciences, four are relevant to medicine. The grants database indicates that from 1 April 2009 only one project will be on-going.
 - Twelve BBSRC grant applications requested time on HPCx, of which five were funded. Four were supported by the IBM outreach funding stream (computer time only), of which only three were completed. A further project was funded through responsive mode. No other Class I projects were funded in 2004, 2005 and 2006. Three Class II bioscience projects were supported on HPCx and two of these were completed.
 - Institutions involved in projects using national HPC facilities are: Oxford, Southampton, UCL, Manchester, Bristol, JIC, Edinburgh and Leeds.
 - The University of Oxford (Professor Mark Sansom) is the most frequent user of national HPC facilities with three projects since 2003. No other user has accessed the facility more than once *via* BBSRC funding mechanisms.
36. The portfolio analysis indicated that the national service is not widely used in the biosciences. This was confirmed by the questionnaire responses with ten out of fifty-seven responses stating that they had used HECToR or a previous national service.
37. It is clear from the analysis that the use of the national HPC facility is currently a niche activity in the biosciences involving only a handful of researchers. There was little evidence to indicate that the demand for access would expand as it had remained largely unchanged over 5 years. BBSRC's £3.3M (5%) commitment to HECToR was a significant increase upon previous investments (HPCx 1.5%). The analysis raised uncertainty over whether this anticipated expansion of usage of the national HPC facility will be realised unless direct action is taken to engage and retain new bioscience users. Class I access has not delivered a significant number of new users and the decline in responsive mode success rate could cause further restriction. The lack of Class II bioscience projects is worrying as this is the principal route for the engagement of new users.

Conclusion 1: Use of the national HPC facility is a niche activity in the biosciences. There is uncertainty over whether the previously anticipated expansion of usage will be realised. There is a need to bridge the gap between biosciences and HPC at the national level.

Local HPC Facilities

38. The portfolio analysis showed that between 2003 and 2007 BBSRC funded eight projects that utilised local HPC facilities (i.e. at a regional or university level). This represented six users and there was only one user from this group who had also used the national facility. Questionnaire responses indicated that a total of ten researchers had used local HPC and an even greater number (27 respondents) considered that they had access to local HPC facilities.
39. Respondents described the local HPC facilities available to them and, in doing so, set out a broad range of compute hardware, extending from a three node Linux cluster, to a 600 core cluster to an IBM Blue Gene. Beowulf (Linux) clusters were commonly cited, as well as distributed computing resources (OXGRID, CAMGRID, NWGRID). This broad interpretation of 'local HPC' made it difficult to draw firm conclusions on this group of facilities, but by focusing on two distinct classes of hardware (i.e. the 'high end' facilities of 1000+ core/CPU and much smaller clusters of 150 core / CPU or below) the Review Group was able to make some observations on availability, areas of usage, how the facility meets the needs of bioscientists and whether the resource acted as a development platform for the national facilities.
40. Eight respondents indicated that their institution had one or more 'high end' facilities used for a range of numerically intensive tasks including large finite element analysis, molecular dynamics, Monte Carlo methods and quantum mechanical calculations. The primary users of these systems were in the physical sciences and engineering. However, one respondent commented on applications in a 'broad range of scientific computing tasks in the biological domain', and another commented on increasing use in simulating tissue/organism mechanics and function. Two respondents recognised that these high end systems have to achieve a consensus configuration and are not necessarily ideal for bioscientists due to the high activation barrier to optimising code and to implementing effective parallelisation. The same two respondents stated that the systems act, to some extent, as a development platform for national facilities, setting up and optimising calculations.
41. The smaller clusters (i.e. 150 core / CPU or below) described by respondents as 'local HPC', demonstrated much wider availability and a greater diversity of biological applications. Activities mentioned included proteomics (Mascot), genomics (GCG, JEMBOSS, SRS, BLAST), high throughput sequencing (MAQ), mathematics and statistics (R, MatLab), simulations (NAMD), hosting bioinformatics server capacity and hosting Solexa platforms. Other biological applications identified as using 'local HPC' included bioinformatics (several applications), cyro-EM image reconstruction, molecular simulation, protein crystallography, NMR, molecular/granular dynamics, heart modelling, high microarray probe design and sequence comparisons. None of the smaller clusters were identified as providing a development platform for national facilities.
42. The Review Group observed that most high-end local HPC shows a similarity with the national facility being used for numerically intensive tasks and aimed principally at the physical sciences and engineering community. The smaller computer clusters have wide spread use in the bioscience research areas identified as 'computationally intensive'. This 'falling away' of bioscience applications on the 'high-end' hardware could be due to: a lack of use in biology of the methodologies suited for HPC; HPC offering an inappropriate architecture for certain biological applications; and the high activation barrier for moving bioscience research onto HPC.
43. The Review Group considered that users preferred to utilise local resources due to ease of access, a perception of greater control, and greater collaboration with software developers and computer support.

Furthermore, the accumulation of data on local facilities could require significant level of efforts and resource to replicate on national HPC.

44. The portfolio analysis of projects using HPC (national service and local facility) may represent an underestimate of total usage in the biosciences. The portfolio searches only identified a grant as using HPC if usage was mentioned in the title or abstract. However, the questionnaire responses on the national service were broadly in line with the questionnaire responses, thus it is unlikely that there were a large pool of additional researchers that remain unidentified.

Conclusion 2: Local ‘high end’ HPC facilities are not widely used in the biosciences. This mirrors the situation with the national facility. There is a need to bridge the gap between bioscience and HPC at the local level.

Group Clusters

45. As indicated in paragraph 28, ‘low-end HPC’ clusters of medium size servers and workstations, linked together with high speed local area networks, were considered as being the work-horses of modern bioscience. **Annex 6** indicates that they were used by 26 out of 38 respondents with widespread application across biomolecular sciences, systems biology and bioinformatics. Well established computational biology groups tended to mix multi CPU servers in different configurations, and clusters were being used in applications where some parallelisation or batch processing of experiments was required (e.g. sequence comparison, cell simulations). The use of clusters to tackle bioscience applications had expanded in recent years, and there was a recognition that this increasing use of computer clusters in bioscience would inevitably lead some researchers to move their research onto national facilities, in order to guarantee that appropriate compute power was available to them.

PCs and Workstations

46. PCs and workstations, using a number of operating systems, were identified as being crucial for the majority of bioscience research, both in the management of research (writing papers, general data analysis) and within the laboratory. 35 out of the 38 respondents in **Appendix 5** used PCs or workstations. Desktop computers were widely used both by computational biologists for method development and by experimental biologists for data analysis. Examples of how desktop computers were used within an experimental context included being used for graphics and imaging, as well as within protein crystallography, which is well supported with software provided by the Computational Collaborative Project (CCP)⁴.

Key Bioscience Research Areas

47. The Review Group recognised that the computational bioscience community was small but growing, and that it had a diversity of requirements. The research areas that were making use of HPC were considered, followed by an examination of the computational architecture needs of a number of different bioscience areas. It was recognised that general HPC needs may take some time to emerge in new areas of application (e.g. systems biology). Research areas that are more closely allied to research in physics and computational chemistry were in a better position to make use of existing HPC architectures.

Research areas using HPC

48. The portfolio analysis of BBSRC-funded grants identified the scientific areas that had used national HPC facilities (Class I). Of the eight projects, five were biomolecular simulations (including mechanistic enzymology, ion channels and membranes), one was systems biology (gene networks) and one was virtual tissue engineering. The Review Panel noted that the IBM/BBSRC outreach initiative, that aimed to provide HPCx compute time to bioscience users, had reinforced traditional user areas (i.e.

biomolecular simulations) rather than broaden access. However, examination of the quarterly reports for HPCx indicated that the overall outreach activity had prompted some work on retina modelling, cardiac simulations and genome-wide searches for RNA localisation. These new avenues did not appear to generate any Class I or Class II projects.

49. Since 2003, four Class II projects have been supported on HPCx and HECToR. Two projects were on biomolecular simulations (biomolecular computational chemistry and quantum chemistry studies of rusticyanin protein crystals) and two projects were on bioinformatics (parallelisation of Matlab codes and checkpoint start for R jobs). The project on Matlab was not progressed on HPCX and did not transfer to HECToR. None of these projects led to access being secured through the Class I route.
50. Examination of the six BBSRC-funded projects using local HPC showed three in systems biology, two in biomechanics and one in genomics. Despite the small numbers involved, there appeared to be a distinct difference between the Class I bioscience projects on the national facility and the equivalent of Class I access using local facilities.
51. The Review Group considered that most of the bioscience users of HPC were also software developers working in computational biology. While this group is important, it was recognised that biologists will only start to make large-scale use of HPC when services were provided that could be readily used by them to run models and analyse data (i.e. the user and developer are different people).

Conclusion 3: The national facility is used mainly for biomolecular simulations. The local 'high end' HPC facilities show a greater range of bioscience applications. Efforts to broaden the use of the national facility in the biosciences have met with mixed success and there is an apparent lack of sustained engagement. Most bioscience users of the national facility participate in software development, working to increase the utility of the service rather than being a customer.

Architectural Requirements

52. The Review Group examined the questionnaire responses to determine the architectural requirements of areas identified as computationally intensive. In the following, **data parallelism** refers to the architectures in which several processors perform the same task on different items of data. **Process parallelism** is associated with architectures that carry out a task by distributing processes across computing nodes. Sufficient information was provided on bioinformatics, systems biology and biomolecular simulations:
 - Bioinformatics: Databases, data management and large scale data analysis were considered to require few CPUs, but large amounts of physical memory (e.g. 256GB RAM per node), fast input/output (I/O) (i.e. the communication time between the computer and the outside world) and good connectivity (i.e. network bandwidth to nodes on which analyses may be taking place). It was generally considered that many bioinformatics jobs, including sequence analysis and comparisons, were **process** parallelisable with Beowulf clusters and workstation farms being a popular choice of architecture. It was also noted that a reason for the popularity of these architectures was their broad applicability and the often manageable amount of investment (time and money) required to adapt codes developed for single CPU machines to these systems.
 - Systems Biology: The learning of models that explain experimental observations utilises algorithms that may not be readily process parallelisable. Machine learning and finite element analysis (e.g. used to handle large datasets or complex models) requires fast processors and interconnects with a lot of RAM, allowing data to be accessed in order. Monte Carlo or sensitivity analysis, where large numbers of repeats are required, can be undertaken on large numbers of slower computers. Cellular simulations are expected to require massively process parallel

approaches utilising HPC. The range of mathematical approaches under the broad systems biology banner will require a number of different architectures.

- Biomolecular simulations: These require massively parallel calculations. The architectures required are generally **data** parallel systems suited to numerically intensive tasks that, on the whole, are more relevant to the physical sciences and where the software tools have been developed to use these architectures. Large scale simulations cannot be efficiently run on clusters as they require shared memory parallel computers. Predictive biomolecular modelling is now becoming possible using HPC. Protein crystallography does not require HPC.

Conclusion 4: The national HPC facility is useful for applications that are numerically intensive and require data parallel approaches. This will cover only a subset of the computationally intensive biological research challenges.

Bottlenecks in Computational Provision

53. The Review Group considered the evidence on national HPC facilities (see paragraphs 21-24), other computational infrastructures and the questionnaire responses to identify a series of bottlenecks in computational provision. The major bottlenecks are set out in this section, whilst the Review Group's considerations on how to overcome the bottlenecks are covered under Term of Reference 2.

Awareness and access

54. There is a lack of awareness of the national HPC facility in the biosciences. Some questionnaire respondents had not heard of the facility, whilst others were unclear about its relevance to their research, and other respondents did not know how to access the facility. It was noted that the HECToR website may be off-putting to biologists as it appears to assume that a potential user knows exactly what they want. Access to expert knowledge about what the facility could deliver for the biosciences is required. At the same time, HPC support services (e.g. NAG Ltd) appeared to be under-utilised and underappreciated in biology. This pointed to a gulf between the needs and skill level in most parts of the biosciences and what the HPC service expects to provide to potential users. There was a concern that the mechanisms to obtaining access to national facilities were burdensome, particularly for those researchers just starting to use HPC. There was a general opinion that access to national facilities should be made easier, in order for researchers to undertake some preliminary trials of code portability/utility and make full use of the facility.

Data

55. Data capacity was considered to be a major bottleneck. Data storage infrastructures (e.g. for databases or file archives) were a particular concern, with increasingly large volumes of data (primary and derived) being generated by computationally-intensive (molecular dynamics) and data-rich (e.g. functional genomics, metabolomics, imaging and next generation sequencing) projects. The related issues of data management and curation (at the research group level and the infrastructures) were also identified as important.

Software

56. The lack of suitable codes on the national facility is limiting its use in the biosciences; one either brings compatible code to HECToR for running a particular calculation, or brings the calculation to run on an existing code. The widespread use of high level programme languages (e.g. R, Matlab, Python and Perl) in the biosciences creates an additional barrier to the uptake of HPC and it was noted that the only systems biology application on the national HPC facility had been rewritten in C to remove the dependency on Matlab. The Review Group recognised that the software currently available for biomolecular simulations may require refinement. Other bioscience areas (e.g. whole organism,

landscape) will potentially require a host of different codes. Porting and optimising codes were considered to be non-trivial tasks that required partnership between biologists and software engineers.

Skills/Training

57. Training was another major bottleneck and gaps in skills were identified across the board. The Review Group considered that there was a general lack of computational training aimed at bioscientists at both the under- and postgraduate levels. This meant that training was required at the post-doctoral level but this was also in short supply. Furthermore, there was a need for courses to re-position, moving away from standard bioinformatics applications and towards incorporating mathematical and computational approaches. The Review Group also identified a severe lack of the specialised staff required to develop and port codes and implement biological requirements. HPC programming skills were considered to be very scarce in the bioscience community and institutional computing staff were often overloaded with general duties; few were dedicated to meeting the software and hardware needs of the biosciences.

Conclusion 5: There are major bottlenecks that limit the widespread use of computational approaches in the biosciences (data, skills and training). Furthermore, there is a gulf between most bioscience users and the national HPC service with bottlenecks in skills and training, awareness and access, and software.

TERM OF REFERENCE 2: TO DEVELOP A VIEW OF THE COMPUTATIONAL NEEDS OF UK BIOSCIENCE RESEARCH OVER THE NEXT 5-10 YEARS

Getting the most out of current HPC provision

58. The Review Group noted that, in the context of UK HPC, the biosciences were a minor player, with apparently fragmented needs. As such, the benefits arising from the HECToR service had, to date, been limited. Members considered that in the short-to-medium term there was a need to tackle the identified bottlenecks that were limiting the use of high-end HPC in the biosciences, and the national facility in particular. A programme of demystification, community collaboration and a building up of expertise is necessary for researchers to embrace the scope and applicability of HPC in the biosciences. Positioned appropriately, the programme of activities would encourage the use of both local and national HPC facilities. The programme may require a small amount of additional investment but it would help ensure that a better return is made on BBSRC's financial commitment to HECToR.

59. To demystify HPC a number of potential exercises to raise awareness were identified. These included:

- Outreach activities (e.g. road shows, conference presentations, seminars, workshops), to provide a general overview of information on what HPC resources are available, what classes of problems they are designed for, how to access them and the training and support available.
- Show-casing of successful projects that have used HPC and yielded high impact publications. These projects should be in areas of importance to biologists, set out the type of problems being solved and the type of architecture and software required for the successful computational approach.
- Facilitation in the form of a dedicated bioscience consultancy and advice service, to work with researchers to identify potential projects and to provide follow-up support and problem-specific advice.

60. Bearing in mind the architecture of the national facility, the bioscience community should work with software developers to identify a list of methods applicable to groups of biologists that, in principle, could operate on the national HPC. The Review Group considered that significant work is required to make sure that appropriate, reliable and efficient codes are available on HECToR. The Review Group

queried whether the 'bottom-up' Class II access was the appropriate route to take this forward. A more 'top-down' approach of lobbying the HECToR Strategic Management Board and the Scientific Advisory Committee was considered necessary.

61. In light of the current limited bioscience user base it was considered that demonstrator projects would need to be supported to provide a portfolio of successful case studies. This could involve facilitated interaction between individual biologists and NAG Ltd to identify proof-of-concept projects in application areas where the use of HPC might not yet be obvious (e.g. ecology, biochemical pathways, imaging) followed by incorporation and running of the relevant codes. This would be similar to the HPCx IBM Outreach Call but would need to take account of the need to retain bioscience users over the longer-term.
62. Community building was also identified as important. The Review Group considered that there is a need to create a vibrant and productive community to share, on a regular basis, knowledge and expertise of direct relevance to current and emerging scientific challenges and HPC. This could lead to a faster evolution of computational approaches to science, due to the shortcuts achieved by sharing ideas and code, and the creation of broadly common protocols / pipelines. This would help achieve consensus requirements and overcome the apparently fragmented nature of bioscience requirements.
63. Many of the questionnaire respondents considered training as crucial to increasing the understanding and use of HPC and identified a host of different components for incorporation into courses and longer term education activities:
 - Training modules covering parallel computing to include computational architectures, parallelisation of algorithms, and potential applications for biologists, computational biologists and programmers working in the biosciences, as well as associated details on different hardware technologies and when to use them.
 - Increasing awareness of the problems e.g. increased dataset size, increasing processing time and how simple serial approaches may not scale. Understanding potential solutions and how they can benefit and further enable science.
 - A focus on local IT teams (e.g. in BBSRC sponsored Institutes) to ensure they are briefed on HPC infrastructure and services.
 - Targeting courses at a range of different levels from postgraduates to PIs. Training early career researchers would help fill the gaps in skilled researchers with expertise in HPC.
64. It was noted that the current training activities offered by NAG Ltd are directed at programmers. Training for biologists needs to be tailored to two bioscience groups with different needs – computational biologists who develop methods and experimental biologists who may benefit from the use of a high performance platform (but who do not want to contribute directly to the development of the applications). Supporting training directed at both of these groups would boost the size of the potential user community for HECToR. Exposing NAG Ltd to the level of HPC understanding in the bioscience community together with the views of biologists would be beneficial to both parties.

Recommendation 1: BBSRC should work with partners to ensure that national HPC services support applications that are of value to significant numbers of users within the life sciences community, where the users are interested in analysing data of interest and not necessarily engaged with computational methods development.

Recommendation 2: BBSRC should work with other partners to develop a programme of activities to increase awareness of, access to, and use of HECToR by the bioscience community. This will require some additional BBSRC investment but, overall, should increase the benefits arising from the Council's original financial commitment.

Computational needs over the next 5-10 years

65. The Review Group considered that it was difficult to predict medium- to long-term requirements but recognised that bioscientists will continue to need access to a range of computing infrastructures. Awareness and training will remain important to ensure that the discipline utilises all levels of the Branscomb Pyramid and is not collapsed into just PC and local cluster usage to the detriment of scientific advancement. The national HPC facility is only one of several types of infrastructure, including the Grid, that is required in the biosciences and new developments such as Cloud or utility computing may have an increasing role to play.
66. Based on the questionnaire responses, the Review Group considered a number of the computationally intensive areas, highlighting potential developments that will raise both hardware and software challenges. They also identified two important emerging research areas that raise computational issues.

Molecular Dynamics and Simulations

67. Biomolecular modelling methods are getting to a stage of being truly predictive. The ability to model molecular systems will greatly improve the development of integrated, complex models at greater length scales. While there are still significant hurdles to overcome, the time when biomodelling is as predictive as fluid dynamics modelling for aerodynamics is now in sight. Thus biomolecular modelling and simulation will move from being a research area for computational biologists to a tool used by the wider community.

Modelling of Biological Systems

68. There has been a large investment in systems biology which relies on simulations of biological systems. These, and future, systems-level analyses will require the integration of multiple data sources (functional and genomic, experimental and computational). High-end computing has an identifiable role in facilitating complex computational analysis, simulations and parameter optimisation for large-scale models. Future analyses are likely to entail larger and more complex simulations of larger and more complex models.
69. The scaling-up of models from the molecular to population and ecosystem levels will require significant computational power if it is to be achievable in the next decade. Modelling the properties of populations of organisms and their biotic and abiotic interactions is often carried out using 'agent-based' modelling, currently on desktop PCs. However, it is recognised that with increasingly complex and larger-scale ecosystem models at an increasingly finer-grain of description, the level of computational power required is raised.

Emerging Areas

70. The Review Group identified two new emerging areas that will require compute power and possibly HPC infrastructures in the future. These were computational ecology and biological image analysis and reconstruction:
- Computational ecology: There is an increased awareness of the potential use of modelling and simulation in ecology and environmental sciences, such as to explore climate change scenarios and life cycle analysis of landscapes under different use (e.g. biofuel crops or arable crops).
 - Image analysis and reconstruction: This was identified as a new 'data-rich' area where there is a growing need for computer power. Future developments will involve running mathematical models on images, combined with high-throughput and high content approaches, as the science of visualisation and analysis comes through to application in the biosciences. There is currently only a

single example of HPC use in imaging (in EM) and very little software being developed to meet forthcoming needs.

Conclusion 6: There is a growing demand for computation in the biological sciences and a range of architectures (and associated software) will be required to tackle the challenges ahead. It is important that bioscience utilises all appropriate infrastructures. Collapsing into the lower levels of the Branscomb Pyramid could be to the detriment of scientific advancement in some key areas.

TERM OF REFERENCE 3: TO PROVIDE ADVICE TO THE TOOLS AND RESOURCES STRATEGY PANEL ON THE POSITIONING OF BBSRC INTEREST IN THE SUPPORT AND DEVELOPMENT OF APPROPRIATE COMPUTATIONAL INFRASTRUCTURE SUCH AS THE NEXT GENERATION HPC.

Project X and a 10 year forward plan for HPC procurement

71. It is typical for academic HPC investments to be staggered such that the facilities' project lives overlap. As such, groundwork has now started in anticipation of the next generation HPC procurement, to supersede HPCx and to overlap and partially supersede HECToR. This potential, next generation, national HPC facility procurement has been labelled 'Project X' and EPSRC is currently developing a 10-year plan for HPC provision, incorporating planning for Project X. This includes identification of a long-term service provider and regular injections of capital rather than developing each next generation HPC from scratch. EPSRC Council approved development of a full business case based on a scientific case made in Spring 2008 and a proposed approach has been noted by the Research Councils UK Research and Development Group.
72. The Review Group considered that Project X must incorporate the requirements of the biosciences and be informed by the evidence presented in this report. It is important that the project includes representation from bioscience communities above and beyond biomolecular simulations and that these needs are reflected in the final configuration. Based on current usage, there is no case for increasing BBSRC's financial commitment in future national facilities beyond current provision. However, changes in usage should be closely monitored and underpin any final decision. The Review Group considered that BBSRC should take a similar position in any UK discussions on the European Partnership for Advanced Computing (PrACE).

Recommendation 3: BBSRC should continue to be involved in future national HPC services. However, there would need to be a significant increase in bioscience users for BBSRC to consider any increase above the current financial commitment.

The Hartree Centre

73. In 2008, the Department for Innovation, Universities and Skills announced a £50M Large Facilities Capital Fund investment, through the Science and Technology Facilities Council (STFC), to fund a computational resource as part of the Gateway initiatives. The Hartree Centre is intended to be a multidisciplinary resource that tackles a number of computational challenges in a variety of disciplines. The Centre intends to complement the national HPC facility by adopting mission-led research grand challenges, including some in the life sciences. Through consultation with the research community areas such as the virtual cell, plant cell walls and the brain have been identified as potential topics.
74. The Review Group noted the approach proposed by the Hartree Centre and recognised that this could lead to the submission of research grant applications to BBSRC. It was considered that proposals should be focused in areas identified as strategically important by BBSRC and where there were potentially a large group of bioscience beneficiaries. Furthermore, it was recognised that partnership

and collaboration between STFC researchers and experimental bioscientists would be essential for the development and subsequent utility of these proposed multi-scale simulations.

Software development and sustainability

75. Software was identified as a major bottleneck in accessing HPC; the lack of software simply excluded certain bioscience research communities from using HECToR. This absence of widely available, biologist-friendly software is not restricted to HPC and in bioimaging it is seen as the single most important area in need of attention. Software development platforms are needed to provide a focus for the research community in developing software, harnessing any existing distributed efforts and avoiding duplication. The platforms can also undertake training and promote the suite of tools to the user as well as undertake a library role. The Collaborative Computational Projects, in particular CCP4, are seen as flagships in software development. The Review Panel considered that BBSRC should consider adopting this model in areas such as systems biology and bioimaging. The provision of a stable software base might be hoped to open the door to a community of bioscientists who are able to make use of HPC to analyse their data without having themselves developed the underlying algorithms.

Hardware Sustainability

76. The Review Group noted that SRIFII and SRIFIII were the most commonly cited funding mechanism for local HPC facilities (**Annex 3e**). Four more respondents said their HEI HPC was university funded, and two more cited a mixture of SRIF and university funds. The presence of university resources was patchy and without long-term guarantee. In the absence of SRIF and other equipment funding streams (e.g. REI) and the introduction of Full Economic Costing, the Review Group considered it might become increasingly difficult to sustain and procure local facilities in the future. The Review Group considered that it is essential that cutting-edge bioscience research is underpinned by a range of appropriate computational infrastructures. The need for these infrastructures is expected to increase in the years ahead. BBSRC should revisit the decision to cease a dedicated equipment funding stream and HEI's need to plan effectively to ensure sustainable support through full economic costing.

Recommendation 4: It is essential that cutting edge bioscience research is underpinned by a range of appropriate computational hardware and software. BBSRC should:

- **Support the development and long-term sustainability of appropriate software tools for the biosciences through their incorporation into appropriate funding mechanism**
- **Revisit the decision to cease a dedicated equipment funding stream.**

CONCLUSIONS AND RECOMMENDATIONS

Conclusion 1: Use of the national HPC facility is a niche activity in the biosciences. There is uncertainty over whether the previously anticipated expansion of usage will be realised. There is a need to bridge the gap between biosciences and HPC at the national level.

Conclusion 2: Local 'high end' HPC facilities are not widely used in the biosciences. This mirrors the situation with the national facility. There is a need to bridge the gap between bioscience and HPC at the local and facility level.

Conclusion 3: The national facility is used mainly for biomolecular simulations. The local 'high end' HPC facilities show a greater range of bioscience applications. Efforts to broaden the use of the national facility in the biosciences have met with mixed success and there is an apparent lack of sustained engagement. Most bioscience users of the national facility participate in software development, working to increase the utility of the service rather than being a customer.

Conclusion 4: The national HPC facility is useful for applications that are numerically intensive and require data parallel approaches. This will cover only a subset of the computationally intensive biological research challenges.

Conclusion 5: There are major bottlenecks that limit the widespread use of computational approaches in the biosciences (data, skills and training). Furthermore, there is a gulf between most bioscience users and the national HPC service with bottlenecks in skills and training, awareness and access, and software.

Conclusion 6: There is a growing demand for computation in the biological sciences and a range of architectures (and associated software) will be required to tackle the challenges ahead. It is important that bioscience utilises all appropriate infrastructures. Collapsing into the lower levels of the Branscomb Pyramid could be to the detriment of scientific advancement in some key areas.

Recommendation 1: BBSRC should work with partners to ensure that national HPC services support applications that are of value to significant numbers of users within the life sciences community, where the users are interested in analysing data of interest and not necessarily engaged with computational methods development.

Recommendation 2: BBSRC should work with other partners to develop a programme of activities to increase awareness of, access to, and use of HECToR by the bioscience community. This will require some additional BBSRC investment but, overall, should increase the benefits arising from the Council's original financial commitment.

Recommendation 3: BBSRC should continue to be involved in future national HPC services. However, there would need to be a significant increase in bioscience users for BBSRC to consider any increase above the current financial commitment.

Recommendation 4: It is essential that cutting edge bioscience research is underpinned by a range of appropriate computational hardware and software. BBSRC should: a) Support the development and long-term sustainability of appropriate software tools for the biosciences through their incorporation into appropriate funding mechanisms b) Revisit the decision to cease a dedicated equipment funding stream

References

¹ International Review of High Performance Computing in the UK (2005)

www.epsrc.ac.uk/ResearchFunding/FacilitiesAndServices/HighPerformanceComputing/InternationalReview/IntRevReport.htm

² Strategic Framework for High End Computing (2006)

www.epsrc.ac.uk/ResearchFunding/FacilitiesAndServices/HighPerformanceComputing/HPCStrategy/2006StrategicFramework.htm

³ Challenges in High End Computing

www.epsrc.ac.uk/ResearchFunding/FacilitiesAndServices/HighPerformanceComputing/HPCStrategy/Challenges.htm

⁴ US Branscomb Pyramid

This is described in International Review of High Performance Computing in the UK (2005). See ¹ above.

HIGH END COMPUTING TERA SCALE RESOURCE (HECToR)

The System

1. HECToR (High End Computing Terascale Resource) is the current UK national supercomputer. The system is a Cray XT4, with 60 XT4 cabinets, comprising 5664 compute nodes with 2.8 GHz Operton processors, and 6GB RAM per node, and a total of 11,328 cores. At a Phase 1 peak performance of 63TFlops¹, this is approximately four times the performance of the previous resource (HPCx), at 15TFlops. A series of scheduled upgrades will increase this performance. The system is scheduled to run from 2007-2013.
2. As it stands, HECToR is currently a batch-driven system, so for e.g. visualisation software, interactivity would be a problem. Future software upgrades might allow part of the machine to be reserved for individuals that might alleviate this potential problem.

Governance

3. EPSRC managed the procurement for HECToR on behalf of the other Research Council partners (in full consultation) and continues to act as managing agents for the service. BBSRC's contribution to the procurement was £3.3 million, or 5% of the total allocation. UoE HPCX Ltd. are contracting and directing the HECToR hardware including the accommodation and system management, as well as the helpdesk, website and administration. NAG Ltd. provides the computational science and engineering support for the users of the service. Cray Inc. provides and maintains the HECToR hardware and systems software. STFC staff at Daresbury Laboratory are involved in the provision of the service.
4. The current management structure is comprised of three groups, a strategic management board (H-SMB) of partner organisations, a scientific advisory committee (H-SAC) and the CSE performance Review Group (CSE-PWG). Community academic representation is made via the H-SAC. The BBSRC representative is Dr Charles Laughton, University of Nottingham.

HECToR Access

5. There are two mechanisms through which researchers can access time on HECToR. The first is by applying for compute time through a normal responsive mode grant (Class I access). The access time is assigned a nominal value although as the contribution to the national HPC has already been paid, this does not include a cash contribution. The time is claimed in the form of Allocation Units (AU's). AU's are a measure of time on the instrument only, and do not correlate directly to other resources e.g. disk space.
6. The second mechanism is Class II access. This can be considered to be 'pre-peer review' access, and is intended to allow researchers the opportunity to investigate the potential of running their codes on a national high performance computing resource. Class II access is generally too limited in duration to allow a research project to be completed. Class II access can be further separated into:
 - Class IIa access. This is intended to allow existing users access to HECToR to support a full application through responsive mode. It is also aimed at new users whose previous experience may have been on university based and mid-range pieces of equipment. The limit for this type

¹ FLOPS (floating point operation per second) are the standard unit of operation used to measure compute performance, particularly relating to floating point operations

of access is currently 100,000 AUs for BBSRC users. Applications of this type are contingent on a positive technical assessment from the HECToR service.

- Class IIb access. This access class is aimed at researchers wishing to apply for dCSE support from NAG along with their application. It is envisaged that this route will be used to support short-term activities such as code development. Applications of this type are contingent on confirmation of support from dCSE.

HECToR Academic Community Liason

7. A number of community liaison activities are run by the HECToR partners:

- A continuing series of HECToR user meetings are run, to introduce the service to the users, discuss issues arising and detail the application process.
- A large portfolio of UK-wide training courses (ongoing list available on the HECToR website) are available, some of which will be aimed at the new or uninitiated user. These are free to HECToR users and to BBSRC-, EPSRC- and NERC-funded users.
- NAG Ltd provides CSE support for HECToR. The CSE Service exists to help the user community to make the best use of the HECToR machine by providing training, web-based resources, and assistance with porting, optimisation and tuning of software. In addition, a number of other training and support activities are provided. The CSE Service will be provides a variety of training courses covering topics ranging from good software development practises to optimising and tuning code for the Cray XT4. The CSE team is responsible for providing a Technical Assessment of any grant applications requesting time on HECToR. Support for new and existing users are provided, including assisting in porting code and providing dedicated project training. In addition, projects or consortia using HECToR and funded by EPSRC, NERC or BBSRC, may apply for dedicated PDRA-level support from the CSE team to assist them with developing, porting and tuning their software. This could involve seconding a NAG employee to the project, sub-contracting to a third party with the necessary expertise, "buying-out" an existing member of the project team or hiring a person to work on the project for an agreed period. Any training required will be provided by the CSE team.

Codes

8. A number of third party application codes are already available and running on HECToR. A number of these codes are relevant to the biosciences, specifically for the running of molecular dynamics simulations. They are GROMACS, LAMMPS, NAMD, AMBER and CHARMM.

REVIEW OF THE COMPUTATIONAL REQUIREMENTS OF THE BIOSCIENCES**REVIEW GROUP**

The Review Group membership is shown in **Table 1**. The Group was chaired by Professor Steve Homans, University of Leeds and consisted of representatives from UK academia, BBSRC-sponsored institutes and European institutes. Members marked as * had previously worked in industry.

Table1

Name	Institution
Prof. Steve Homans (Chair)	University of Leeds
Dr. Charles Laughton	Univ. Nottingham
Prof. Mark Sansom	University of Oxford
Prof. Norman Paton	University of Manchester
Prof. Christopher Rawlings*	Rothamsted Research Institute
Dr Rachel Errington	Cardiff University
Dr Penny Rashbass	University of Sheffield
Dr John Overington*	Inpharmatica (now European Bioinformatics Institute)

RESEARCH GRANT FUNDING INVOLVING HPC

Annex 3a

A list of BBSRC funded research grants using HPC (i.e. Class I access). This covered HECToR and the two previous national academic HPC facilities (CSAR, HPCx).

CSAR

Project	Principal Investigator	Institution	Year	Collaborators	Collaborating Institution	Funding Stream	Value of award (£)
DFT calculations for ion channels and transport proteins	Sansom, M	University of Oxford	2003	Domene, C	University of Oxford	Responsive Mode	33,570
IntBioSim: an integrated approach to multi-level biomolecular simulations ²	Sansom, M	University of Oxford	2003	J. Essex	Southampton	Bioinformatics and E-Science Programme II	916,081
				P. Coveney	UCL		
				D. Gavaghan	Oxford		
				McKeever	Oxford		
				J. Gurd	Manchester		
A. Mulholland	Bristol						

HPCx

Project	Principal Investigator	Institution	Year	Collaborators	Collaborating Institution	Funding Stream	Value of award (£)
Towards a virtual outer membrane (vOM)	Sansom, M	University of Oxford	2003			IBM Outreach	IBM Outreach programme Included compute time only
Modelling enzyme catalysis	Mulholland, A	University of Bristol	2003			IBM Outreach	
Virtual forced evolution of catalytic transition metal complexes	Durrant, M	John Innes Centre	2003	Pickett, C	University of East Anglia	IBM Outreach	
Life sciences proposal for software development on the HPCx	Dicks, J	John Innes Centre	2003			IBM Outreach	
IntBioSim: an integrated approach to multi-level biomolecular simulations	Sansom, M	University of Oxford	2003	See table 1 above			
Multiple light input signals to the gene network of the circadian clock	Millar, AJ	University of Edinburgh	2007			BBSRC responsive mode,	783,502

² Note that this project also requested HPCx resources, which were subsequently migrated to HECToR – hence this project appears as a user on three of the national HPC services.

HECToR

Project	Principal Investigator	Institution	Funding Stream	Collaborators	Collaborating Institution	Value of award (£)
Multiple Light Inputs to the Gene Network of the Circadian Clock	A. Millar	University of Edinburgh	See Table 2 above			
IntBioSim: an integrated approach to multi-level biomolecular simulation	M. Sansom	University of Oxford	See Table 1 above			

Annex 3b

A list of EPSRC funded projects relevant to the biosciences supported by the High Performance Computing Programme identified as current in November 2008 (i.e. Class I access).

Project	Principal Investigator	Institution	Collaborators	Collaborating Institution	Value of Award (£)	Start Date	End Date
High Performance Computing for Simulation of Radioprobng, a Novel Method for Studying Nucleic Acid Structure and Recognition	Dr CA Laughton	Nottingham	H. Nikjoo	MRC Radioation and Genome Stability Unit	Not Available	1/6/04	30/11/07
GENIUS: Grid Enabled Neurosurgical Imaging Using Simulation	Professor PV Coveney	University College London	R. Blake	STFC	157,003	1/10/07	31/12/08
			J. Brooke	Manchester			
			S. Booth	Edinburgh			
			S. Pickles	Manchester			
			S. Brew	UCL			
High throughput electrophysiological, electromagnetic & electromechanical cardiac virtual tissue engineering	Professor A Holden	Leeds	R. Clayton	Sheffield	92,311	1/10/05	30/9/09
HPC Software for Medical Imaging	Dr D Atkinson	University College London	J. Schnabel	Oxford	173,695	1/10/07	31/3/09
HPC Software for Modelling Chemical Diffusion through Skin Membranes	Professor PK Jimack	Leeds	C. Goodyer	Leeds	121,928	1/10/07	31/3/09

Annex 3c Bioscience-oriented projects supported via HECTOR Class II access.

Project	Principal Investigator	Institution	Allocation
Establishing a checkpoint restart for R jobs	P Ghazal	University of Edinburgh	100K Au

Annex 3d A list of BBSRC funded grants projects using university or regional HPC facilities.

PI	Institution	Title	Facility used	Funding Stream	Collaborator + Institution	Value of award (£)	Start Date	End Date
Oliver, P	STFC	Supporting Bio-informatics Research on the National Grid Service (NGS)	(software developments for HPC)	e-Science development fund			1/10/2005	30/9/2007
Millar, AJ	University of Edinburgh	Centre for Systems biology at Edinburgh	IBM BlueGene Edinburgh	Integrative Systems Biology (systems centre)	Beggs, J, Edinburgh	8,579,303	1/5/2006	30/4/2011
Leigh-Brown, AJ	University of Edinburgh	Analysis of virulence determinants in full length H5N1 influenza genomes using computational modelling	IBM BlueGene Edinburgh	Combating Avian Influenza	61,012	369,349	1/10/2006	30/9/2009
O' Higgins, P	University of York	An investigation of tetrapod skull architecture using advanced computer modelling techniques	High Performance Cluster, Hull	Responsive Mode		20, 815	1/1/2007	31/12/2009
Millar, AJ	University of Edinburgh	Multiple Light Input Signals to the Gene Network of the Circadian Clock	IBM BlueGene Edinburgh	Responsive Mode			1/1/2007	31/12/2009
Evans, SE	University College London	An investigation of tetrapod skull architecture using advanced computer modelling techniques	High Performance Cluster, Hull	Responsive Mode			1/1/2007	31/12/2009
Fagan, M	University of Hull	An investigation of tetrapod skull architecture using advanced computer modelling techniques	High Performance Cluster, Hull	Responsive Mode	Curtis, N, University of Hull	321,504	1/1/2007	31/12/2009
O'Higgins, P	University of York	Mechanical Function of the Primate Craniofacial Skeleton	High Performance Cluster, Hull	Responsive Mode	243,254		1/5/2007	30/4/2010
Silver, R	University College London	An integrative study of neural coding in the vestibular cerebellum: from cellular physiology to models of network behaviour	Research Computing, UCL	ANR-BBSRC SysBio	Margrie, T, University College London	653,848	1/1/2008	31/12/2010

Annex 3e Details of related e-infrastructure projects, BBSRC Research Equipment Initiative

PI	Institution ³	Title	Year	Value of award (£)
Ranson, N	University of Leeds	Advance computing infrastructure for structural biology and bioinformatics at the University of Leeds	2003	197,958
Orengo, C	UCL	High performance computing infrastructure for structural biology and bioinformatics	2004	179,940
Laughton, C	University of Nottingham	A specialized computational resource for biomolecular simulation	2007	76,041

³ These applications include numerous co-applicants from the same institution.

COMPUTATIONAL REQUIREMENTS OF THE BIOLOGICAL SCIENCES

QUESTIONNAIRE FOR BIOSCIENCE RESEARCH COMMUNITY

Aim

The aim of this consultation is to gather evidence on the use of computational infrastructures, in particular high performance computing (HPC), in UK bioscience research. It focuses on research areas that have been identified as 'computationally intensive' – biomolecular sciences, systems biology, bioimaging and mathematical or computational biology. Although the questionnaire will be returned by e-mail, no personal information will be entered on the questionnaire, and only information from the questionnaire will be used for the purpose of collation. Typing can be done in the boxes only, please click on the left hand side of each box to start. Please return the questionnaire to the e-mail address at the end of the document.

General Information

1. Please give (up to) five keywords that describe your research area

2. What computational hardware do you use to conduct your research? Please provide a percentage estimate of the amount of usage of the various systems

National Supercomputers

Local Supercomputers

University, Department or Group Compute Clusters

Personal Computers or Workstations

3. Based on your experience, what computational bottlenecks exist in your area of research and how could they be overcome?

4. In your opinion, are different computational architectures required for different bioscience disciplines?

Yes

No

If YES, please provide an indication of what architectures might be required

5. Looking ahead, do you see an expanding role for HPC in the biosciences for:

Your Research

The Biosciences in General

Please expand on what those roles might be

Local HPC Facilities

6. Do you have access to and use HPC facilities via local computing at your HEI?
- If YES, please describe:
 - The facility
 - How it is funded and maintained and how is access obtained?
 - What you use it for?
 - How is it configured and managed to meet the needs of bioscientists?
 - Does it act as a development platform for national high-end facilities?

National HPC Facilities

7. Have you used HECToR or other national high performance computing infrastructures (e.g. HPCx, National Grid Service) for your research?

Yes

No

If YES: Please explain what you use the facility for.

If NO: Please explain why you do not use the facility (see bullets below and provide comments).

Local resources sufficient

Not enough information

Lack of availability of open source or commercial software/lack of skills or resources to develop software.

Mechanism of application too difficult or takes too long

Portability of data

Wrong type of computing architecture

Technology gap between your current work and HPC

Lack of Training

8. Do the national facilities meet the needs of the bioscience community effectively?

Yes

No

If NO please describe why not

9. Does BBSRC need to raise awareness to HECTOR amongst the bioscience community?

Yes

No

If YES, what form should this activity take?

10. Is there foundational work that needs to be undertaken to expand HPC usage in the biosciences – either with HECTOR or subsequent national facilities?

Yes

No

If so, please describe.

International HPC Facilities

11. How does UK HPC provision compare to that available in other countries? What could we usefully learn from international facilities?

Any Other Comments

12. BBSRC would be grateful to receive any other comments you may have that, whilst relevant to the review, are not prompted by the questions above.

For further information, contact:

Dr Michael Ball
Programme Manager
Tools and Resources Strategy Panel
BBSRC
michael.ball@bbsrc.ac.uk
01793 413282

QUESTIONNAIRE RESPONSES: A QUANTITATIVE SUMMARY

80 questionnaires were sent out by e-mail, with 59 usable replies. Three more responses indicated that the researcher no longer worked with HPC and so did not wish to complete the survey. Overall, a response rate of 77.5% was obtained.

27 respondents reported having access to HPC at a university level, with 12 reporting that they did not have access (69% with access, 21% with not).

37 respondents agreed that different computational architectures are required for different bioscience disciplines, ten disagreed

10 of the respondents had used HECToR or a previous national service, with the remaining 47 respondents not having used HPC.

10 of the respondents thought that the needs of the biosciences were met with existing HPC; 21 disagreed.

32 respondents commented that BBSRC could do more to raise awareness of HECToR amongst the community, and 6 believed that BBSRC should not.

23 respondents indicated that BBSRC should undertake further foundation work to encourage the usage of HPC within the biosciences. 5 believed that BBSRC should not.

INFORMATION ON BIOSCIENCE USAGE OF THE FOUR TIERS OF COMPUTATIONAL INFRASTRUCTURE

Each row represents a response (total=38) from an individual that completed the following questions from the questionnaire.

- Question 1: Please give (up to) five keywords that describe your research areas
- Question2: what computer hardware do you use to conduct your research? Please provide a percentage estimate of the amount of usage of the various systems.

Principal Scientific Area	National HPC	Local HPC	Compute Clusters	PCs or Workstations
Biomolecular Sciences				
Protein crystallography			60%	40%
Structural Biology				100%
Metalloproteins, structural biology				
Protein modelling				100%
Protein design, bioinformatics		5%	45%	50%
Analytical chemistry, metabolomics				100%
Biomolecular simulation, computational enzymology	10%		60%	30%
Computer simulations, molecular modelling	20%		70%	10%
Computational chemistry, molecular simulation		75%	15%	10%
Biopolymers, multiscale modelling				100%
Biophysics of protein folding, simulations		70%	30%	10%
Biophysics			80%	20%
Systems Biology				
Systems Biology		20%	30%	50%
Systems biology, neurone, simulation			30%	70%
Systems biology, cell signalling			1%	99%
Systems biology, modelling	5%	10%	25%	60%
Systems biology			80%	20%

Mathematical modelling of biological systems				
Mathematical modelling			30%	70%
Mathematical modelling				100%
Bioinformatics				
Bioinformatics		1%	60%	39%
Bioinformatics		50%	50%	
Bioinformatics, protein modelling			90%	10%
Bioinformatics, math modelling, plant systems biology			15%	85%
Computational genomics, bioinformatics			10%	90%
Gene expression				100%
Atlas, bioinformatics			50%	100%
Biodiversity informatics				100%
Animal sciences				
Morphometrics, simulation			60%	40%
Biomechanics	30%	30%	20%	20%
Palaeontology, anatomy			10%	90%
Computational physiology	10%	10%		80%
Other areas				
Immunology				100
Plant development, synthetic biology				100%
Computing support			100%	
Virology			30%	70%
EM, image processing			50%	50%
Food biochemistry				100%
Plant science – agriculture, ecology		5%	5%	90%
Total responses: 38	5	10	26	35