Teaching medical students to use supercomputers – a personal reflection

Andrea Townsend-Nicholson[1*]

[1]Research Department of Structural & Molecular Biology, Division of Biosciences, University College London, London, WC1E 6BT, United Kingdom
**Author for correspondence.*

Abstract

At the "Kick Off" meeting for CompBioMed (compbiomed.eu), which was first funded in October 2016, I had no idea that one single sentence ("I wish I could teach this to medical students") would lead to a dedicated programme of work to engage the clinicians and biomedical researchers of the future with supercomputing. This programme of work which, within the CompBiomed Centre of Excellence, we have been calling "the CompBioMed Education and Training Programme", is a holistic endeavour that has been developed by and continues to be delivered with the expertise and support from experimental researchers, computer scientists, clinicians, HPC centres and industrial partners within or associated with CompBioMed. The original description of the initial educational approach to training has previously been published [1]. In this chapter, I describe the refinements to the programme and its delivery, emphasising the highs and lows of delivering this programme over the past six years. I conclude with suggestions for feasible measures that I believe will help overcome the barriers and challenges we have encountered in bringing a community of users with little familiarity of computing beyond the desktop to the petascale and beyond.

Key words: high-performance computing, university education, medical student, undergraduate, molecular biosciences, experimental-computational workflow, metagenomics, next-generation sequencing, computational biology, computational biomedicine

Introduction

In 2016, we had started the programme by designing training in computational biomedicine with two different modalities: 1) as part of a credit bearing unit within a taught undergraduate degree programme; and, 2) as an extracurricular training course of shorter duration. As part of a taught programme of study, we focused on medicine because we wanted to engage clinicians with computational predictions that were sufficiently accurate that they could use them to inform their clinical practice. We included the molecular biosciences (biochemistry and molecular biology) to encourage the translation of basic research into clinical applications and to facilitate personalised medicine initiatives. Also, as Head of Teaching, I was responsible for the design and delivery of the Molecular Biosciences degree programme and could ensure the provision of appropriate domain-level knowledge to facilitate the integration of relevant computational methods into the curriculum. The extracurricular training course was something that we ran as part of the PRACE Advanced Training Courses (PATC) delivered at the Barcelona Supercomputing Center as a Winter School in Computational Biomedicine [2]. The trainees on the PATC course were primarily MSc students of varying backgrounds, mostly computer science but with some from biological disciplines.

The taught programmes of study were first delivered at UCL to both medical and molecular biosciences students. Medical student course delivery was through a course called a "student selected component (SSC)". These are mandatory courses within the curriculum that provide medical students with the opportunity to learn about specific areas in medicine that are of interest to them. Molecular Biosciences course delivery was through a third year Specialist Research Project. Both the SSC and the Specialist Research Project allowed students to conduct a research project in metagenomics based on the Human Microbiome Project [3]. These courses are consistently well-received by the students taking them and the feedback is used to tune the courses to make engagement with supercomputing easier for those unfamiliar with the technology.

Course Developments

It quickly became clear, in our first year of implementation, that although students were able to upload images, audio and video to a number of different social media platforms, the vast majority of them had no experience with the command line, with connecting to a remote system or with running job scripts on a machine. To address this in all subsequent years, we have ensured that the taught content of the medical students' SSC includes an introduction to the command line and to high performance computing (HPC) before beginning the metagenomic analyses. For the Molecular Biosciences students, this introductory material is provided in their first year of study. Currently, our undergraduate students in Molecular Biosciences engage with HPC in their first year, learning to connect to a remote system and to transfer files to and from this system. They learn how to submit jobs to a scheduler and to retrieve and visualise their results.

Our next great challenge came with the COVID-19 pandemic. We had been running metagenomics courses that integrated experimental and computational work. In 2020, we (and the rest of the planet) found ourselves unable to ensure that the experimental component was able to be achieved during lockdown. However, this provided an opportunity to obtain data from public repositories for analysis. Between 2017 and 2020, we had noted that a number of students were finding metagenomic sequence data to include in their analyses that was beyond that which they had generated themselves experimentally and we took the opportunity to provide next generation sequence datasets and links to data repositories during the pandemic. This worked exceptionally well for the course and has had the added benefit of helping to expand our students' interest in data science. It also has allowed us to run our metagenomics courses online as a purely computational

course with no experimental component. Although this was not what we had originally envisaged, this computational-only offering has actually helped us to expand our teaching delivery beyond UCL and into other universities with the establishment of new courses.

Course Delivery

Location Location Location

It takes a great deal of effort to establish a course or module within a university curriculum but once it is running, it is likely to be within the curriculum for a period of time that can be measured in years, if not decades. In the summer of 2020, we ported our medical student SSC from UCL to the University of Sheffield. This wasn't an obvious 'lift and shift' because SSCs have different formats in different medical schools. The UCL SSC block format is a course with 3 hours per week for 8 weeks in total and it takes place in years 1, 2 and 6 of study; students complete either one double block or two single block SSCs in each of these years. The Sheffield SSC format is a research SSC that comprises five weeks of dedicated study with an academic supervisor and takes place in year 2. Despite these differences, we have successfully run the metagenomics SSC at both UCL and Sheffield since 2020 (see Table 1).

What has been particularly notable about the transfer is that although primarily a computational project, a domain/subject specialist is clearly required for effective teaching delivery of this SSC. From 2020, we have needed to have an academic in each institution collaborating in the delivery of the module: for this particular SSC, the domain specialist is at UCL (where the SSC was originally developed) and the computational expert is at Sheffield. The successful establishment of computational biomedicine courses will be greatly facilitated by identifying educators possessing both domain and computational knowledge. We have, from 2021, expanded to include a taught course at the University Pompeu Fabra (UPF; Barcelona, Spain) in our CompBioMed curriculum. In this case, the course organiser is both domain and computationally expert and no co-delivery of content is required. We are currently developing a Sheffield SSC project with significant computational biomedicine content and will port it to UCL to see what training and support are required to run it as a standalone SSC module there. We are also working with our Core and Associate Partners in CompBioMed to find training modalities that will allow us to expand our SSC training to medical schools across Europe, particularly in the EU13 (member states admitted to the EU since 2004) and in HPC-under-represented countries (see Figure 1).

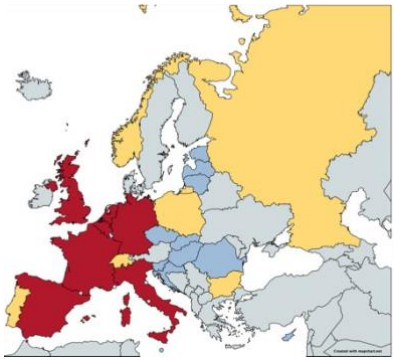

Figure 1 Expanding CompBioMed SSC Delivery Across Europe

The geographic location of countries with CompBioMed Core Partners (red), Associate Partners (yellow) and EU13 countries with medical schools (medium blue) facilitates opportunities for teaching computational biomedicine in medical schools across Europe (figure from A. Marzo © 2021).

HPC Resource

We had originally been using the training allocation of the CompBioMed grant to provide access to compute for our students. This was provided by a number of HPC centres, including the Edinburgh Parallel Computing Centre [4] in the UK and SURF [5] in the Netherlands. However, the pandemic also provided us with the opportunity to explore different methods of course delivery. Specifically, we looked at cloud-based methods, whereas we had previously been using federated HPC resources. In the summer of 2020, together with our colleagues in Alces Flight and one of our undergraduate students, we built nUCLeus – a proof of concept cloud HPC education environment [6] that we were able to showcase at SC20. We have reverted to federated resources at present, but we can reactivate nUCLeus whenever needed and we envisage using it for training delivery in the future. We have also explored the use of Google Colaboratory to analyse open datasets with cloud computing platforms [7].

Academic Year	Medical Student Course	University (Year)	Medicine students	Delivery
2017-2018	SSC	UCL (Year 1)	20	Face to Face
	SSC	UCL (Year 2)	20	
2018-2019	SSC	UCL (Year 1)	20	Face to Face
2019-2020	SSC	UCL (Year 1)	20	Face to Face
2020-2021	SSC	UCL (Year 1)	20	Online
	SSC	USFD (Year 2)	20	
2021-2022	SSC	UCL (Year 1)	8	Face to Face
	SSC	UCL (Year 2)	30	Face to Face
	SSC	USFD (Year 2)	17	Hybrid
	Inter-Departmental	UPF (Year 5)	0/24	Face to Face
2022-2023*	SSC	UCL (Year 1)	12	Face to Face
	SSC	UCL (Year 2)	60	Face to Face
	SSC	USFD (Year 2)	12	Hybrid
	Inter-Departmental	UPF (Year 5)	tba	Face to Face
		TOTAL	259*	

Table 1 Medical Students on HPC

The location and delivery mode of medical student SSC modules delivered from 2017 as part of the CompBioMed Education and Training Programme. The UPF course is multidisciplinary and integrates students from medicine, biology, biomedical engineering and data science degrees.

Challenges and Barriers

There are two major challenges that we have collectively faced in embedding computational biomedicine training in the educational curriculum of medical and biomedical undergraduate degrees. The first is access to HPC resources and the second is funding to support the teaching programme.

In 2016, a mere 10% of the UK's national supercomputers were being used for research in the life and medical sciences and only 0.2% of this was being used for medical applications [1]. This access was for research and none of it was being used for teaching at an undergraduate level. Life and medical sciences are primarily experimental/clinical rather than computational and are an underrepresented demographic in HPC. They are sufficiently underrepresented that for some time now we have been running a 'train the trainers' programme, upskilling the digital skills of academics as well as students to enable them to make use of computing resource beyond their laptop or local server. If you are not familiar with a system, you are not going to be able to use it and our goal is to ensure that everyone in this domain has access to the training they need to develop the skills they require to integrate HPC into their professional practice.

Access to compute resource is another challenge in this domain. Life and medical science researchers typically have not had access to HPC systems to the same extent that other disciplines have had. This is slowly changing, but the combination of not knowing how to use something and the inability to access it to practice on is quite difficult to overcome. We have been very fortunate in having Tier 0, Tier 1 and Tier 2 HPC resource allocations made available to us for our education and training programme. Without this, we would not have been able to deliver our programme to almost 2000 students (see Table 2). Despite this large number, our focus on medical and Molecular Biosciences students has broadened significantly over the past year as students in other related fields ask whether they can take part in the courses we offer. We are keen to accommodate everyone interested in using HPC in our domain and are currently considering how to teach this content at scale, aiming to deliver locally to between 1000 and 2000 students per year.

Academic Year	Medicine students	Biosciences students	Extracurricular Course	Total number of students
2017-2018	40	89	40	169
2018-2019	20	104	38	162
2019-2020	20	108	31	159
2020-2021	40	86	35	161
2021-2022	55	370	40	465
2022-2023	84	472	16	572
	259	1229	200	1688

Table 2 Trainees in the CompBioMed Education and Training Programme
The number and origin of students trained from 2017 is shown. The increase in numbers from 2021 is due to embedding HPC training in all three years of the Molecular Biosciences degree programme.

Future Directions

It was not at all clear at the start of the programme, whether there would be the support needed to foster this kind of innovation. It turned out to be possible within both the grant and within the higher education system: engaging domain-specific practitioners with HPC is something that was very much supported by the grant and, as a result, the CompBioMed Education and Training Programme became an important part of our university teaching programme. This new knowledge has been well-received by our students, too, many of whom are completely taken with computational insights into the molecular apects of their subject. There is a significant demand for more access, more

courses, more training and I see students coming back again for new opportunities to engage with HPC. There are so many examples, but to highlight a few: a first year medical student who enjoyed the metagenomics SSC has gone on to do a third research project that involves molecular dynamics simulations of protein-protein interactions involved in neurological disorders; a first year medical student who chose to do an intercalated BSc degree in Mathematics, Computers and Medicine; and, several fourth year MSci project students have continued to work in the area of computational biomedicine as part of their PhD studies. The appetite for knowledge and experience in this area is significant and wonderful to see!

Having watched the programme grow to the point where we are easily teaching almost 600 students per year and knowing how we came to be able to do this, there are some things that I think we will need to consider going forward:

1. *There is an urgent need to support programmes that will create a culture of engagement with and integration of computational methodologies into professional practice for domains that are primarily experimental/clinical.*
2. *Local institutions should be encouraged to provide HPC for teaching or, at the very least, to implement schedulers that can be used to book a partition on their machine(s) for teaching events like classes and workshops, so that jobs can be run during the teaching session without getting stuck in the queue.*
3. *If local HPC systems do not operate under a 'free to use' model, there is the potential for inequality of access between students at institutions that make the resource freely available and those enrolled at institutions that charge to use it; for academics delivering HPC-based courses in the curriculum at institutions that charge to use local HPC resource, the teaching budget will need to include HPC access costs.*
4. *You don't need to deliver digital upskilling through dedicated taught courses. Key skills training, short workshops, week-long summer schools are all good ways of bringing this content to students. The advantage of doing it through a taught curriculum is that it is easier to keep track of what has been taught, to whom and when.*
5. *There will always be growth in the programme. We are now building more computationally-intensive workflows for teaching our medical and undergraduate students. These will require resource beyond that of a local cluster.*

Acknowledgements

I would like to specifically thank Carlos Teijeiro Barjas, Marco Verdicchio, Gavin Pringle, Andrew Narracott, Guillaume Hautbergue, Alberto Marzo, Oscar Camara Rey, Mariano Vazquez and Peter Coveney for their advice, support and contributions to the development of this novel education and training programme. I am grateful to the European Commission (grants 675451 and 823712) and to EPSRC (EP/X019446/1) for funding support.

References

1. Townsend-Nicholson, A (2020) Educating and engaging new communities of practice with high performance computing through the integration of teaching and research. Interface Focus 10:20200003. doi:10.1098/rsfs.2020.0003
2. https://www.bsc.es/education/training/other-training/online-short-course-hpc-based-computational-bio-medicine
3. https://www.hmpdacc.org
4. https://www.epcc.ed.ac.uk

5. https://www.surf.nl/en
6. Townsend-Nicholson, A, Gregory, D, Hoti, A, Merritt, C, Franks, S (2020) Demystifying the Dark Arts of HPC – Introducing biomedical researchers to supercomputers. https://sc20.supercomputing.org/proceedings/sotp/sotp_files/sotp104s2-file2.pdf
7. Poolman, TM, Townsend-Nicholson, A, Cain, A (2022) Teaching genomics to life science undergraduates using cloud computing platforms with open datasets. Biochem Mol Biol Educ 50 (5):446-449. doi: 10.1002/bmb.21646