GCBR Forum — Meeting 8

11 November 2024, 16:00-17:30 GMT

Meeting Summary

The eighth virtual meeting of the Global Core Biodata Resource (GCBR) Forum was held on 11 November 2024. It was attended by 41 representatives from 37 GCBRs, together with members of the GBC Secretariat. The meeting was chaired by GBC Executive Director and Forum co-chair, Guy Cochrane. The key items discussed are summarised briefly below.

Forum Discussion: AI & LLMs - Challenges and Opportunities for Biodata Resources

Forum members had a wide-ranging discussion on the potential application of artificial intelligence (AI) and large language models (LLMs) to enhance the work of biodata resources.

Several promising applications to AI tools were identified. This included recent work at EMBL-EBI to investigate the use of LLMs to help optimise helpdesk functionality. A ChatBot developed for the PRIDE database to assist in responding to user queries and in navigating help documentation has shown positive results, and EMBL-EBI is exploring the scope to extend this approach across its suite of database resources. The Reactome team has also been exploring the potential to apply ChatGPT to augment various knowledge extraction tasks, with retrieval augmented generation (RAG) showing promising results.

Examples of other promising applications for AI and LLMs included:

- EnzChemRED (Enzyme Chemistry Relation Extraction Dataset) — a training and benchmarking dataset developed at the SIB Swiss Institute of Bioinformatics to support the development of Natural Language Processing (NLP) methods to assist in enzyme curation.
- Tools such as Pfam-N, which was developed by InterPro to apply deep learning to expand protein family coverage
- Tools for literature classification and named entity recognition
- Tools to assist in software development, including the development of guide code, and website optimisation
- Potential applications in the arena of pathogen genomics to assist in prioritisation of data processing in line with public health needs (as assessed by data usage).

Whilst acknowledging the huge potential of AI approaches in several areas and the speed at which these technologies were developing, Forum Members emphasised the critical need for human validation of all AI outputs and clarity on the provenance of information. It was also the case that, for extraction of knowledge from literature, no AI-based approach is as yet close to the quality and reliability of human curations and none of the knowledge bases represented would currently consider incorporating AI-based extractions in their resources.

Other key conclusions were that:

- There is strong potential for biodata resources to collaborate in the development of benchmarks for AI tools, as doing this in isolation would be inefficient and prohibitively costly for small resources.
- There is a need for resources to consider the potential payback for AI approaches in relation to the cost of development, and how to evaluate this meaningfully.
- Whilst it is right that funders are encouraging biodata resources to utilise AI approaches where these can add value, it was important to recognise that this represented additional work for resources at the current time - particularly given the critical need for human validation of AI outputs. There may well ultimately be a significant payback but this will be down the track and the extent of the benefit is uncertain at the current time.

Update on GCBR Review Process and Future Selection Rounds

The Secretariat provided an update to the Forum on plans for the first periodic reviews of existing GCBRs and future GCBR selection rounds.

Forum Business

Forum members were updated on key ongoing GBC activities, notably:

- Development of the GBC White Paper on Sustainability, which was nearing finalisation and would be published in early 2025.
- The GCBR of the Week campaign through which GBC each week publishes a profile of one of the 52 GCBRs to celebrate their work and importance of sustainability.
- The GBC Open Letter Campaign, which will launch in November 2024 to highlight the critical importance of biodata resources and their long-term funding and sustainability.

Next Forum Meeting

It was noted that the next virtual meeting of the Forum would take place in February 2025.