BioSample Databases at NCBI and EBI

July 7, 2010

NCBI and EBI are in the process of establishing new databases that will contain information about *biological samples* used in experiments, such as sequencing, genotyping, gene expression, proteomics or metabolomics profiling assays. In this letter we will refer to these databases as BioSample Database at NCBI (BioSDn) and BioSample Database at EBI (BioSDe). Each sample in each of these databases will be issued a stable accession number. The data will be exchanged between BioSDn and BioSDe; all sample information in one of these databases will also be held in the other database. The space of the accession numbers will be shared between the two databases. It will be possible to reference the samples in BioSDn and BioSDe from other databases run in NCBI or EBI respectively.

The *biosamples* can be either actual physical samples, such as, blood samples, or sources of such samples, such as cell lines, animal strains or (anonymised) human individuals. The information about human subjects and access to it will be compliant with all relevant ethical requirements. Discussing these requirements is not the purpose of this document; below we will concentrate mostly on issues related to non-human samples, or samples from human cell lines where broad consent has been given and no ethical issues arise.

Information about the sample will include:

- Species;
- The material sampled, e.g., – organs, tissues, cell type;
- Phenotypic information – including disease states and clinical information about the individual;

We refer to all such data as *sample data* or *sample information*. Samples may have relationships between them, for instance a particular sample can be derived from a particular cell line or individual already described and accessioned in the databases.

A particular set of biosamples submitted to BioSDn or BioSDe directly may be referenced subsequently from many experiments. We will refer to this set of samples as *reference biosamples*. Example of these may be some commonly used cell lines or mouse strains. It is our intention to pre-populate BioSDn and BioSDe with information about such commonly reused samples or sample sources, to issue them accession numbers and make it easy to reference these from other databases at EBI and NCBI (such as GenBank, ENA, SRA, ERA, ArrayExpress, GEO, PRIDE) without the need to resubmit the sample information with every new experiment or assay.

NCBI and EBI have started working toward developing the data exchange format as well as submission formats and tools. The early reference sample information providers will give important use cases that will guide us in developing the format as well as common sample attributes and vocabularies for describing these. Since the data will be exchanged between BioSDn and BioSDe, the sample descriptions will have to be submitted only to one of these databases.

The BioSample Databases will be searchable. For instance the user will be able to find all samples in the database having certain attributes or described using certain keywords. The sample descriptions will always include information about their providers.