2.5 Bioinformatics 67

2.5 Bioinformatics 67

ESSENTIALS Bioinformatics may be defined as ‘conceptualizing biology in terms of molecules and applying “informatics techniques” (e.g. applied mathematics, computer science and statistics) to understand and organize the information associated with these molecules, on a large scale’. Clinical bioinformatics may be defined as ‘the clinical application of bioinformatics-associated sciences and technologies to under- stand molecular mechanisms and potential therapies for human diseases’. To achieve these aims: (1) data must be curated to facilitate stand- ardized access to existing information and to allow the submission of new entries to data sets; (2) analysis tools should be developed drawing upon both computational and biological/clinical expertise; (3) all analyses must be interpreted in a biologically/clinically mean- ingful manner. Introduction If clinical bioinformatics is to deliver the integration of molecular and clinical data and thereby translate research knowledge into ef- fective ‘personalized’ medicine, then two broad constituencies need to be supported. Clinicians at the point of care need to understand and integrate, perhaps via decision support mechanisms, entities such as genotype/phenotype correlations, biomarker discovery, and pharmacogenomics; while researchers require accurate, struc- tured and (ideally) coded clinical data, as well as biological reference data sets. Ever accelerating technological advances and precipitous falls in the cost of high-throughput technologies (e.g. whole genome sequencing, expression profiling, high resolution image processing and others) means that there is a veritable deluge of available data. Accompanying falls in the cost of the substantial computational power and associated data storage needed also mean that there is opportunity for meaningful analysis. Ever faster turnaround times (currently measured in hours) mean that it is now feasible to intro- duce next-generation sequencing (NGS) into workflows directly contributing to patient care. NGS technologies allow the identification of single nucleo- tide polymorphisms, point mutations, and insertions or deletions (indels) as well as larger structural changes such as translocations, rearrangements, inversions, duplications, and copy number vari- ations. When investigating somatic mutations e.g. in tumours, comparison with germline samples facilitates variant detection. In rare diseases the comparison within a given trio, proband and both parents, serves a similar process. The need for defined metrics to inform strict criteria-based quality assurance is crucial if a clinical bioinformatics pipeline is to be setup. NGS has additional capabilities to investigate cellular proper- ties over and above the determination of genomic sequence alone. Epigenomics deals with the chemical modifications of nucleic acids (e.g. 5′ methylation, and the consequent effect on gene expres- sion). NGS offers the potential to identify changes across the entire genome by capturing epigenetic information from multiple genes simultaneously. Given that for some tumours epigenetic status re- flects the overall prognosis, such analyses may provide substantially enhanced prognostic information. However, simply aggregating

patient-specific clinical data with genetic, expression, or other data will not automatically lead to better clinical outcomes. Clinical data is often unstructured and incomplete, being spread across multiple paper and electronic sys- tems, hence clinically meaningful semantic vocabulary standards are needed (see next). At a cellular level it is clear that biology is not solely dependent on the genome sequence alone, and in 2012 it was estimated that the biological function of approximately half of all human genes remained unknown. Projects such as the Encyclopaedia of DNA Elements (ENCODE), currently building a comprehensive list of the functional elements in the human genome, and the Kyoto Encyclopaedia of Genes and Genomes (KEGG), which supports machine executed models of system-level biological pathways, are important. Without these the ability to in- terpret analyses and to draw relevant biomedical patient-specific conclusions will be severely impeded. 2.5 Bioinformatics Afzal Chaudhry

68 SECTION 2 Background to medicine Components needed for clinical bioinformatics Data storage Flexible, extensible data warehousing is essential to accommodate the volume of clinical and biomedical information. The ability to support multiple data sources containing heterogeneous data sets is crucial, and data structures must also be able to accommodate sparse data sets as it is unlikely that any individual clinical record will contain information on all possible concepts. Typical warehouse designs are built upon a ‘dimensional fact’ model. Here, a fact is a concept relevant to decision-making (e.g. an observation made at a specific point in time such as a blood pressure measurement), while a dimension describes some attri- bute of that fact (e.g. the blood pressure was measured with the patient supine). Use of a common set of semantic terms to support data aggregation/interoperability Control of the metadata dictionary describing all of the facts in the warehouse is essential. For example, for laboratory tests, normal ranges, and assay types may change over time and/or may vary from one laboratory to another. It is impossible then to meaning- fully aggregate data over time or from multiple laboratories un- less the results (facts) are interpreted using the relevant metadata (dimensions). Examples of such metadata include the struc- tured vocabulary/ontologies listed in Table 2.5.1. Hierarchical ontological terminologies reflecting clinical meaning such as SNOMED CT are preferred over more epidemiological orien- tated classification systems such as ICD-10. Dimensions should ideally be described using elements from a defined archetype rep- resented in a definition set such as the openEHR reference model. When using natural language processing technology to extract structured standardized data from unstructured information, the extraction should be ‘supervised’ by the metadata dictionaries to allow data from text-based records to be amalgamated with that from a coded record. Meaningful analytical tools There are multiple data sets and tools to support the analysis of bio- informatics data (Tables 2.5.2 and 2.5.3). Typically, these focus on assessing similarities between molecular sequences based on align- ment. Sequence-based data sets significantly outnumber structural-based data sets because of the relative ease by which sequence data can be obtained. Additional analyses, often using specialized soft- ware, are needed over and above simple alignment analyses to detect clinically relevant structural genomic alterations. Protein orientated databases are often categorized as either pri- mary, detailing the linear amino-acid sequence, or secondary, containing derived information. Secondary-based analyses may consider motifs or electrostatic interactions that are contiguous in three-dimensional space but not in the linear sequence. Some macromolecular three-dimensional structure databases contain a hierarchical taxonomy to help identify evolutionary relationships. Ultimately the most value is seen by combining data from multiple sources—clinical, sequence, structure, expression, and function (or as many of these as exist). This may not always be straightforward due to variations in nomenclature and data formats,

although web- based gateways supporting the traversal of multiple databases are becoming more effective. Examples of clinical/research areas benefitting from clinical bioinformatics strategies See Table 2.5.4. Oncology research Oncology research has tended to focus on single gene and single pathway analysis. However, NGS offers both multiple simultaneous analyses and extremely high sequence coverage thus substantially increasing sensitivity. International consortia such as the Cancer Genome Atlas are sequencing thousands of cancers to generate data sets across different cancer subtypes. Computational theories Table 2.5.1 Clinical classifications/terminologies/structured vocabularies available in the United Kingdom Name (acronym) Full name Clinical related entities described URL (accessed November 2018) ICD-10 International Classification of Disease 10 Diagnoses https://www.who.int/classifications/icd/icdonlineversions/en/ SNOMED CT Systematized Nomenclature of Medicine— Clinical Terms Symptoms, signs, diagnoses http://www.ihtsdo.org/snomed-ct dm+d NHS Dictionary of Medicines and Devices Medication http://dmd.medicines.org.uk/ NLMC National Laboratory Medicine Catalogue Laboratory investigations https://nlmc.x-labsystems.co.uk/ LOINC Logical Observation Identifiers Names and Codes Laboratory investigations https://loinc.org/ NICIP National Interim Clinical Imaging Procedure code set Radiological investigations https://digital.nhs.uk/services/terminology-and- classifications/national-interim-clinical-imaging-procedure-nicip-code-set OPCS 4.7 Office of Population Censuses and Surveys Classification of Interventions and Procedures version 4.7 Procedures https://isd.digital.nhs.uk/trud3/user/guest/group/0/ pack/10 HPO Human Phenotype Ontology Phenotypic abnormalities encountered in human disease https://hpo.jax.org/app/

 including pathway network analysis and graph theory can be used to model tumour-related regulatory networks and interactions, allowing complex interactions to be understood. The predictive power of multigene biomarker panels, now potentially scaled into the many thousands of genes analysed simultaneously as opposed to just tens of genes, can be profoundly enhanced (e.g. in one study a panel of 2300 genes could discriminate adenocarcinoma of the lung from healthy tissue with 100% accuracy). Pharmacogenomics As the genetic/molecular basis of the metabolism and mechanism of action of drugs becomes increasingly understood, so therapy can be individually tailored to some degree. Recent examples include trastuzumab for HER2-positive breast cancer and imatinib for chronic myeloid leukaemia and conditions associated with tyrosine kinase- based mutations. NGS can identify somatic variants which help to direct therapy (e.g. as resistant tumour clones emerge—melanoma with the BRAF mutation V600E is susceptible to vemurafenib while the p61 BRAF V600E variant is not). In haematology the potential to stratify an individual's sensitivity to warfarin (VKORC1 and CYP2C9 gene polymorphisms) will help to guide appropriate dosing and avoid potentially life-threatening events. Infectious diseases The far smaller size of viral and bacterial genomes makes it possible to sequence the genome of infecting pathogens. In the case of the 2009 H1N1 influenza pandemic, bioinformatics tools were able to describe within a few hours of the first identification of a novel muta- tion a possible mechanistic explanation by which it was able to mani- fest such a severe phenotype. Computational analysis of genome sequence and protein structures can help in identifying likely drug susceptibility (e.g. the enterohaemorrhagic O104:H4 E. coli outbreak in Germany in 2011), while individual infecting strains can be typed and traced over both time and geographical distribution so sup- porting more appropriate and economical public health strategies. Digital imaging Even among the most experienced histopathologists there can be considerable interobserver variation in certain conditions. Objective algorithms to identify tumour grading and to

search for other tissue- based measures of disease activity using level sets, fractal analysis, and machine learning can improve diagnosis. The adaptation of astronomical algorithms coupled with their application to large an- notated study cohorts is likely to provide a powerful set of analytical tools. In dermatology, texture analysis, neural network framework- based analyses, data mining of skin images and computer-based reconstruction of the skin surface have all been used to support re- search into reliable diagnostic strategies. Table 2.5.2 Publicly available databases of biological knowledge Domain Database URL (accessed November 2018) Nucleotide DDBJ https://www.ddbj.nig.ac.jp GenBank https://www.ncbi.nlm.nih.gov/genbank/ Genome COGs https://www.ncbi.nlm.nih.gov/COG/ Entrez genome https://www.ncbi.nlm.nih.gov/genome GeneCensus http://bioinfo.mbb.yale.edu/genecensus/ Protein—primary NRDB https://www.ncbi.nlm.nih.gov/protein OWL http://130.88.97.239/OWL/index.php SWISS-PROT https://www.uniprot.org/uniprot/?query=reviewed:yes Protein—secondary Pfam http://pfam.xfam.org/ PRINTS http://130.88.97.239/PRINTS/index.php PROSITE https://prosite.expasy.org/ Protein—macromolecular CATH http://www.cathdb.info/ PDBeFold http://www.ebi.ac.uk/msd-srv/ssm/ Protein Data Bank https://www.rcsb.org/ SCOP http://scop.mrc-lmb.cam.ac.uk/scop/index.html Functional/systems biology ENCODE https://genome.ucsc.edu/ENCODE/ KEGG https://www.genome.jp/kegg/ Vocabulary Gene Ontology http://geneontology.org/ Integrated systems/web gateways InterPro https://www.ebi.ac.uk/interpro/ Uniprot http://www.uniprot.org/ Table 2.5.3 Major research institutions providing access to a wide range of bioinformatics databases and analysis tools EMBL-European Bioinformatics Institute https://www.ebi.ac.uk/ NCBI (National Centre for Biotechnology Information) http://www.ncbi.nlm.nih.gov/ Wellcome Trust Sanger Institute https://www.sanger.ac.uk/science/tools

 Conclusions The need to deliver safe, timely, sustainable, and patient-centric care along with the need for evidence-based strategies means that any new technology first has to demonstrate clear translational research benefits before it can be adopted into routine practice. We now have the means to generate and analyse data sets that can lead to such breakthroughs, and this continues to develop at a breakneck pace. If the most is to be made of these new data sets and analyses then not only must appropriate clinical data be available for cor- relation, but there will also be a need for more structured training programmes and curricula to train clinicians in the analysis, interpretation, and use of such data. The UK Health Education England Genomics Education Programme Clinical Bioinformatics group reported in early 2015 and has recommended a series of steps and programmes for a range of healthcare staff to address the long-term goal of establishing a workforce fit for genomic medicine. The implementation of such recommendations is awaited. FURTHER READING Raza S (2014). Defining the role of a bioinformatician. http://www. phgfoundation.org/briefing/defining-the-role-of-a-bioinformatician Slade I, Burton H (2016). Preparing clinicians for genomic medicine. Postgraduate Medical Journal, 92, 369–71. Table 2.5.4 Examples of clinical/research areas benefitting from clinical bioinformatics strategies Clinical/research area Example DNA/RNA sequencing and expression profiling Comprehension of biomolecular pathways underlying malignant transformation Biomarker identification Improved classification, early diagnosis, prognostication, and tailoring of therapy Pharmacogenomics/proteomics Tailoring of therapy—likelihood of therapeutic benefit as well as likely propensity to side effects Pathogen genome/protein sequence/structure and function Description of putative mechanisms underpinning phenotypic manifestations Susceptibility to antimicrobial therapy Epidemiological analysis to identify environmental factors and support the relevant control mechanisms Digital imaging Machine

learning based improvements in cellular/tissue analysis and diagnosis Image analysis of three-dimensional geometric/structural anatomy using both visible (e.g. glaucoma), and nonvisible spectra (e.g. infrared analysis of meibomian gland morphology in dry eye syndromes)

Revision #1
Created 2026-01-22 16:41:02 UTC by Omar Ayman
Updated 2026-01-22 16:41:02 UTC by Omar Ayman