available template for modelling protein-protein "docking" interactions.

The Kbdock project is run as a collaboration between the Capsid and Orpailleur teams at the Loria/Inria research center in Nancy. It is funded and supported by Inria, the CNRS, and the University of Lorraine, as well as specific ANR ("Agence Nationale pour la Recherche") grants. The Kbdock program is available through its online interface. It may also be queried programmatically by expert users in order to execute complex or specialised queries. Recent developments to Kbdock make use of a novel protein structure alignment algorithm called "Kpax" that we have developed [3]. This allows queries in Kbdock to span structural neighbours of the retrieved DDIs, thus allowing Kbdock to search over more distant regions of protein structure space and to propose protein docking templates that cannot be found using conventional sequence-based or structure-based comparison techniques.

We are currently working to link KBdock's structural domain binding site classification with the widely used ExPASy Enzyme Classification scheme. In order to achieve this, we are developing efficient data-mining approaches to process the millions of sequence-function associations that are now available in large molecular biology databases, such as Swiss-Prot and TrEMBL, which together build the UniProt Knowledgebase at the European Bioinformatics Institute.

Links:
http://kbdock.loria.fr
http://kpax.loria.fr, http://hex.loria.fr
Protein Data Bank:
http://www.rcsb.org/pdb/home/home.do
Pfam domain classification:
http://pfam.xfam.org/
UniProt Knowledgebase:
http://www.ebi.ac.uk/uniprot
ExPASy: http://enzyme.expasy.org

References:
[1] A. W. Ghoorah et al.:"Spatial clustering of protein binding sites for template based protein docking", Bioinformatics, 2011 Oct 15; 27(20):2820-7.
[2] A. W. Ghoorah et al.: "A structure-based classification and analysis of protein domain family binding sites and their interactions", Biology (Basel), 2015 Apr 9; 4(2):327-43.
[3] D. W. Ritchie et al: "Fast protein structure alignment using Gaussian overlap scoring of backbone peptide fragment similarity", Bioinformatics. 2012 Dec 15; 28(24):3274-81.

Please contact:
David Ritchie
Inria, France
E-mail dave.ritchie@inria.fr

The Source of the Data Flood: Sequencing Technologies

by Alberto Magi, Nadia Pisanti and Lorenzo Tattini

Where does this huge amount of data come from? What are the costs of producing it? The answers to these questions lie in the impressive development of sequencing technologies, which have opened up many research opportunities and challenges, some of which are described in this issue. DNA sequencing is the process of "reading" a DNA fragment (referred to as a "read") and determining the exact order of DNA bases (the four possible nucleotides, that are Adenine, Guanine, Cytosine, and Thymine) that compose a given DNA strand. Research in biology and medicine has been revolutionised and accelerated by the advances of DNA and even RNA sequencing biotechnologies.

The sequencing of the first human genome was completed in 2003 (though a draft was already complete by 2001). This achievement cost 100 million US dollars and took 13 years. Now, only 12 years later, the cost for the equivalent process has dropped to just over $1,000 and takes just a few days. As a consequence, sequencing the human exome (the coding part of the genome), for example, or the whole genome, has become common practice in medicine, and genomic sequencing of many other species has paved the way to new challenges and research fields in the life sciences.

The first breakthrough in the history of DNA sequencing was the development of Frederick Sanger's method using chain-termination with dideoxy nucleotides in 1977, which earned Sanger his second Nobel Prize in 1980. Sanger sequencing (SSeq) is a sequencing-by-synthesis technique since it requires DNA polymerase enzymes to produce the observable output by means of nucleotide radiolabelling or fluorescent dyes.

From around 2005-2006, second-generation sequencing (SGS) produced a dramatic cost reduction, and from that point on, we have seen a growing diffusion of sequencing machines that have revolutionised clinical research and practice, as well as molecular biology investigations in genomics [1]. The higher error rate (compared with SSeq) is balanced out by the impressive throughput of SGS platforms. Though they still rely on sequence-by-synthesis (using chemi- or photo-luminescence), these platforms are based on various implementations of cyclic-array sequencing. SGS methods (in particular Illumina and Roche) are widely used for the investigation of the whole spectrum of genomic variants from single nucleotide variants (SNVs) to structural variants (SVs). Other implementations of SGS include SOLiD and Complete Genomics' nanoball sequencing. Notably, Ion Torrent was the first platform based on non-optical (i.e., electrochemical) methods

Today's low cost of sequencing allows any reasonably funded lab to sequence multiple genomes, thus raising new clinic and research issues. Sequencing several individuals of the same species, for example, is now common for personalised medicine and for under-

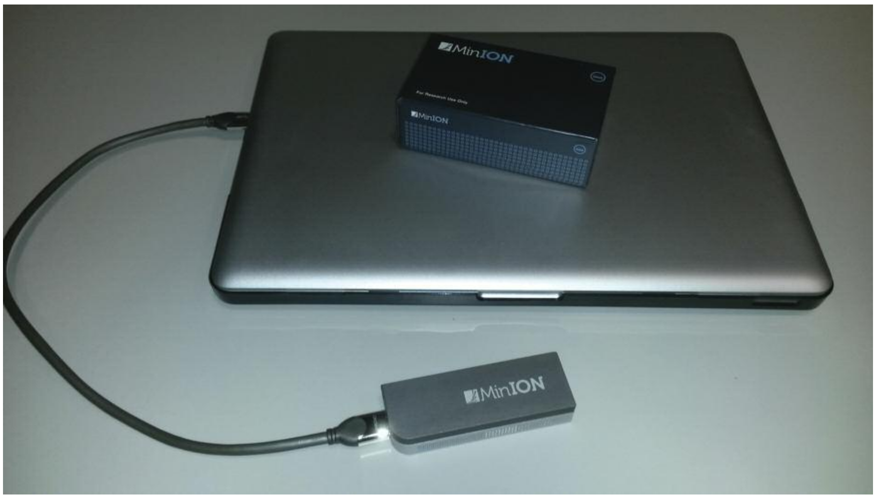

Figure 1: USB-sized Nanopore MinION.

standing how we differ genetically from each other. Also, RNA-Seq – that is sequencing genes that are transcribed for proteins synthesis – allows us to take a (possibly comparative) picture of which genes are expressed under certain conditions. Moreover, sequencing several strands of the same species allows us to investigate intra-species mutations that involve the mobile element of the genomes. Finally, "metagenomics" studies the microbiology of genetic material that is recovered from non-cultivated environments (e.g., soil, gut, sea-depths) and sequenced.

A new revolution is currently underway with "third-generation sequencing" (TGS) techniques [2]. These platforms are based on single molecule real time sequencing, a single DNA molecule sequencing approach. While PacBio platforms exploit an optical detection method, Oxford Nanopore Technologies are based on ionic current measurements [3]. Both platforms show high error rates, though the length of the reads produced is up to thousands of base-pairs. When we have a robust (i.e., low error rates) technology capable of such long reads – which will probably be soon – we will certainly witness new challenges in genomics.

The machine costs and sizes vary considerably, as does the throughput (that is, the amount sequenced base-pairs per hour), even within the same generation. In general, both price and size grow with the throughput. Machines range from the huge and expensive Illumina HiSeq (as big as a closet) to the smaller (desktop-sized) Illumina MiSeq and Ion Torrent, and the even smaller USB-sized Nanopore MinION, shown in Figure 1, passing through the desktop sized ones. Other performance parameters can instead be grouped according to generation: Table 1 reports the cost per base of sequencing, the length of fragments that can be output, and the error rate for each technology generation.

Method	Cost per Base ($/Mbp)	Read Length	Error Rate
SSEQ	400 (up to 2007)	300-1000	10^{-5}-10^{-2}
SGS	0.015 (2015)	$O(10^2)$	10^{-2}
TGS	0.5 (PacBio), 0.65 (Nanopore)	$O(10^3)$	10^{-1}

References:

[1] T. P. Niedringhaus et al.: "Landscape of next-generation sequencing technologies", Analytical chemistry 83.12 (2011): 4327-4341.

[2] C. S. Pareek, R. Smoczynski, A. Tretyn: "Sequencing technologies and genome sequencing", Journal of applied genetics 52.4, 2011, 413-435.

[3] G. F. Schneider, C. Dekker: "DNA sequencing with nanopores", Nature biotechnology 30.4, 2012, 326-328.

Please contact:
Nadia Pisanti
University of Pisa and Inria
E-mail: pisanti@di.unipi.it