CROSS DISCIPLINARY GENOMICS

Up-and-coming Advances in Genome Sciences

Abstract book

SESSION 1: COMPUTATIONAL GENOMICS

EUGENE V. KOONIN – *NCBI, Bethesda, USA*

Reassessment and generalization of the key concepts of molecular evolution in the post genomic era

Comparative genomics, and in particular phylogenomic study of microbial genomes, yields a wealth of data that clash with the accepted fundamental models of evolutionary biology and call for new, more general models. Here I will describe three such models, each pertaining to a key concept in evolutionary biology.

1. The notion of a stable genome is replaced by the model of a dynamic supergenome, given the compelling evidence that evolution of microbes is a web-type process dominated by horizontal gene transfer. The supergenome is the entire gene pool that is accessible to a given species or another group of organisms. Reconstruction of microbial evolution indicates that genomes exchange genes with the supergenome at variable but typically extremely high rates, with several gene losses and gains per a nucleotide substitution per gene.
2. The concept of the Tree of Life (TOL) that is typically represented by the phylogenetic tree of 16S ribosomal RNA is replaced by the "Forest of Life" which consists of phylogenetic trees for each individual genes that generally have different topologies due to extensive horizontal gene exchange. However, comparative analysis of the tree topologies in the Forest of Life reveal significant coherence that is well reflected in the topology of nearly universal trees, primarily those for genes encoding translation system components. Thus, the TOL model can be re-conceptualized as a Statistical Tree of Life (STOL) which refkects the central trend in the Forest of Life.
3. Phylogenomic analysis also leads to the replacement of the Molecular Clock model with the more general model of Universal Pacemaker of genome evolution. Unlike Molecular Clock, which postulates approximate conservation of gene-specific evolutionary rates throughout the history of genomes, the Universal Pacemaker only requires conservation of the relative rates of gene evolution remain constant across long evolutionary spans. Thus, under this model, synchronous, genome-wide change of evolutionary rates is a global feature of genome evolution. Comprehensive phylogenomic analysis shows that the Universal Pacemaker model provides for a

substantially better fit of individual gene tree topologies to the species tree than the traditional Molecular Clock model.

I submit that evolutionary biology evolves under the same paradigm as physics whereby new, general models replace the models of the previous generation but embed them as "low energy approximations."

FRANCESCA CICCARELLI – *King's College London, UK*

From cancer genomics to targeted therapy

In my lectures I will discuss on the recent advances in on our understanding of cancer genetics and evolution. I will start by reviewing the accumulating evidence of cancer heterogeneity in terms of acquired genetic mutations and genomic rearrangements. I will then describe the impact of these novel results on modelling cancer evolution and will describe how this can help in identifying cancer-specific targets to be used in therapy.

ROB FINN – *EMBL-EBI, Cambridge, UK*

Big Data: Computational approaches to real-time analysis of protein sequences

The advancement of DNA sequencing technologies has meant that large-scale sequencing projects are routine in most academic institutes. This has led to exponential growth of comprehensive sequence databases such as UniProtKB that has millions of entries. It has also led to the application of DNA sequencing to new areas of science, such as rapid clinical and environmental diagnostics.

So how do we know when we have found something new and interesting? Only a tiny fraction of sequences are ever going to be experimentally characterized. We have developed indispensable fast, powerful and reliable computational methods for inferring, or "annotating", sequence function. Protein and non-coding RNA family databases, such as Pfam and Rfam, try and group and classify sets of sequences using probabilistic models, and offer a solution to the 'big data' problem. Advances in algorithms (HMMER and Infernal) and computational technologies means that it is possible to search large sequence collections in minutes and even seconds.

While these approaches have allowed us to keep pace with the growth of UniProtKB, this database is dwarfed by the number sequences that are being submitted to the EBI metagenomics portal. This year alone, some 7 billion proteins have been analysed. We have truly entered the realm of 'big data'. How we scale, store and visualise this data is the next major challenge, so that we can provide insights into the ecology of complex microbial communities.

CHRISTINE ORENGO – *University College London, UK*

A Structural Perspective on the Evolution of Protein Functions

Powerful tools for comparing protein structures and protein sequences have allowed us to analyse proteins from more than 7000 completed genomes and identify 2700 evolutionary domain superfamilies. These superfamilies cover nearly 70% of domains from all kingdoms of life and are captured in our resource (CATH-Gene3D). More detailed phylogenetic analyses of the highly populated superfamilies, accounting for nearly two thirds of all known domains, identified some particularly promiscuous superfamilies that can be traced back to the last universal common ancestor (LUCA) and in which relatives can diverge considerably to acquire modified structures and functions. Some structural frameworks seem particularly suited to supporting diverse residue arrangements in the active sites, and considerable structural variations on the surfaces of the domains. We also find a surprising number of examples of convergent evolution within a superfamily where very different catalytic machineries are associated with similar enzymatic chemistries, showing that these scaffolds enable multiple routes to the same function. Phylogenetic analyses of protein families can also yield insights into evolution of novel chemistries or substrate specificities and functional analyses can be combined with thermodynamic analyses to reveal the energetic considerations associated with functional divergence.

SESSION 2: NUCLEAR ARCHITECTURE OF CHROMOSOMES

LUCA GIORGETTI – *Institut Curie, Paris, France*

Structural and transcriptional fluctuations at the X inactivation center

Transcriptional regulation occurs in the context of dynamic chromosome architecture. Recently a new level of chromosome organization, Topologically Associating Domains (TADs), was uncovered by chromosome-confirmation-capture (3C) techniques. To explore TAD structure and function, we developed a polymer model that can extract the full repertoire of chromatin conformations within TADs from population-based 3C data. The model predicts the degree to which chromosomal contacts vary between cells. It also identifies interactions within single TADs that stabilize boundaries between adjacent TADs. Combining the model's predictions with high-resolution DNA FISH and quantitative RNA FISH for TADs within the X-inactivation center (Xic), we dissect the relationship between transcription activity of a promoter and its spatial proximity to cis-regulatory elements in single cells. We demonstrate that contacts between potential regulatory elements occur in the context of fluctuating structures rather than stable loops and propose that such fluctuations can contribute to asymmetric expression in the Xic at the onset of X inactivation.

ROMAIN KOSZUL – *Institut Pasteur, Paris, France*

Addressing genomic and metagenomic limitations with chromosomes third dimension

Assembling genomes from the short reads generated by most high-throughput DNA sequencing technologies remains a computationally expensive and error-prone task. Because the experiments needed to close gaps in draft assemblies are costly and time consuming, published genomes are usually left unfinished. In particular, repeated or duplicated regions are challenging to resolve, impairing the study of genome rearrangements. Here, we show that genome-wide chromosome conformation capture (3C) data can be used to overcome these limitations, and present a computational approach rooted in polymer physics that determines the most likely genome structure using chromosomal contact data. The methodology has been validated on yeast, the industrial fungus /T. reesei /and human and can reach a resolution down to 2 restriction fragments. The method has been further extended to the genomic analysis of microbial populations in their natural environment, which remains limited by the difficulty to assemble full genomes of individual species. Using controlled mixes of bacterial and yeast species, we developed meta3C, a high-throughput chromosome conformation capture experiment that allows characterizing individual genomes within a metagenome and their average chromosome organization. Not only can meta3C libraries be used on species with sequenced genomes, but the reads can also be used for the /de novo /assembly, scaffolding and 3D characterization of unknown genomes. This first meta3C study highlights the remarkable diversity of microorganisms chromosome organization, while providing an elegant and integrated approach to metagenomic analysis.

SESSION 3: SYNTHETIC BIOLOGY – GENOME ENGINEERING

YAAKOV BENENSON – *ETH, Zurich, Switzerland*

Molecular computing meets synthetic biology

Source: Unconventional Computation and Natural Computation. 13th International Conference. Proceedings: LNCS 8553. Publisher: Springer, Cham, Switzerland

One of the motivations behind computing with molecules is to "computerize" living systems, for example to prevent disease or control artificial tissues. Biology, however, is already very good at computing - the human brain being one example. Even on a single cell level information is constantly being processed, and the development of a functional organism from a single fertilized cells is controlled by an ingenious if only partially understood program encoded in DNA. Does this mean that the efforts to "write" new molecular programs are redundant? Not at all - natural programs have taken three billion years to evolve and, despite their beauty, are very difficult to alter in any way.

In my view the optimal approach is to balance the engineering principles inspired by computer science and engineering such as universal models, reprogrammability, modularity, etc., with the harsh reality of cell and organismal biology. The simple fact is that we do not know yet, even at the theory level, whether it is possible to perform reliable information processing in actual living cells as opposed to idealized "well-mixed reactors". Despite these limitations, the field of molecular computing in cells, or biological computing, has made significant steps forward with new design principles, new architectures, and new exciting experimental results. These developments also inform basic biological research.

DAVID BIKARD – *Institut Pasteur, Paris, France*

Studying and fighting bacteria with the help of CRISPR

CRISPR loci and the associated Cas genes are the adaptive immune system of archeae and bacteria. The Cas9 protein, a RNA-guided nuclease from the S. pyogenes CRISPR system, can be used as a tool for genome engineering. The approach relies on the ability to re-program CRISPR specificity through a small RNA, which directs Cas9 to cleave a target genomic locus and selects for the recombination of a homologous template containing a desired mutation. This method allows the easy engineering of any genomic loci in a diversity of bacteria. The catalytic site of Cas9 can also be mutated, giving dCas9 (dead Cas9), and repurposing it as a RNA-guided DNA binding protein. This binding is strong enough to block transcription and can thus be used to easily knock-down any gene in the cell. Another area of focus is the development of sequence-specific antimicrobials using CRISPR. When directed to cleave the host cell genome, the Cas9 nuclease leads to cell death. Phagemids can be used as vectors to deliver self-targeting CRISPR systems to bacterial populations. The CRISPR can be programmed to kill bacteria carrying antibiotic resistance or virulence genes, leaving the rest of the microbiota intact.

FARREN ISAACS – *Yale University, New Haven, USA*

Programming Genomes to Expand Life's Functional Repertoire

A defining cellular engineering challenge is the development of high-throughput and automated methodologies for precise manipulation of genomes. To address these challenges, we develop methods for versatile genome modification and evolution of cells. Multiplex automated genome engineering (MAGE) simultaneously targets many locations on the chromosome for modification in a single cell or across a population of cells, thus producing combinatorial genomic diversity. Conjugative assembly genome engineering (CAGE) facilitates the large-scale assembly of many modified genomes. Our methods treat the chromosome as both an editable and evolvable template and are capable of fundamentally re-engineering genomes from the nucleotide to the megabase scale. I will present one application of MAGE to generate combinatorial genomic variants from a complex pool of synthetic DNA to diversify target genes in order to optimize biosynthetic pathways. Then, I will also describe the integration of MAGE and CAGE to engineer a Genomically Recoded Organism (GRO), replacing all 321 UAG stop codons with the synonymous UAA stop codon in E. coli. This GRO exhibited improved properties for incorporation of nonstandard amino acids that expand the chemical diversity of proteins in vivo. The GRO also exhibited increased resistance to T7 bacteriophage, demonstrating that new genetic codes could enable increased viral resistance. This work increases the toolbox for genomic and cellular engineering with the goal of expanding the functional repertoire of organisms.

SESSION 4: SINGLE CELL GENOMICS

VALENTINA PROSERPIO – *EBI, Cambridge, UK*

Multi-state modelling of T cell differentiation reveals three discrete cell states with increasing rates of cell division

Cell differentiation requires changes in gene expression, and is frequently accompanied by changes in cell cycle status. We use T-helper cell as a model to ask how cell cycle is linked to differentiation, and whether this process is gradual or switch-like. We determine gene expression profiles across consecutive generations of differentiated and undifferentiated cells in Th2 polarized cells. Analysis of the transcriptome allowed us to predict the existence of three cell states during the differentiation from naïve cells, representing Early (E), Activated (A) and Th2-like cells (T). We validate this 3-state "EAT model" at the single cell level by single cell qPCR. Next, flow-cytometry data allowed us to construct a Markov branching model to extract the cellular rates of death, division and differentiation. Our multi-scale modelling on diverse data types shows that Th2 differentiation occurs through two switch-like transitions across three states, and reveals both their cellular phenotype and molecular characteristics.

GAËL YVERT – *ENS, Lyon, France*

Particle Genetics: mapping single-cell Probabilistic Trait Loci of the genome

We are exploring a novel angle of genetics by studying how genotypes shape the statistical properties of single-cell phenotypic traits. Using yeast as an experimental model system, we found that natural genetic backgrounds confer different statistical properties of single-cell molecular and cellular traits. For example, some wild yeast strains display elevated cell-cell trait variability as compared to other strains. We identified genomic loci causing probabilistic changes and we propose to call them single-cell Probabilistic Trait Loci. Identifying such loci extends the usual QTL or GWAS scan approach by providing a probabilistic framework where cellular individualities are considered. For example, one locus we identified is a cis-eQTL (variant in one gene affecting its own expression) that also acts non-deterministically in trans on the expression of another gene. In multicellular organisms, this non-deterministic approach may be particularly appropriate to study macroscopic phenotypes that emerge from rare cells, such as cancer, as it may directly interrogate hidden roots of incomplete penetrance and disease predisposition.

JOHN MARIONI – *EMBL-EBI, Cambridge, UK*

Computational challenges in single-cell transcriptomics

Recent technical developments have enabled the transcriptomes of hundreds of cells to be assayed in an unbiased manner. These approaches have enabled heterogeneity in gene expression levels across populations of cells to be characterized as well as facilitating the identification of new, and potentially physiologically relevant, sub-populations of cells.

However, to fully exploit such data and to answer these questions, it is necessary to develop robust computational methods that take account of both technical noise and underlying, potentially confounding, variables such as the cell cycle.

In this presentation I will begin by briefly describing how we used spike-ins to quantify technical noise in single-cell RNA-seq data, thus facilitating identification of genes with more variation in expression levels across cells than expected by chance. Subsequently, I will discuss a computational approach that uses latent variable models to account for potentially confounding factors such as the cell cycle before applying it to study the differentiation of Th2 cells. I will show that accounting for cell-to-cell correlations due to the cell cycle allows identification of otherwise obscured sub-populations of cells that correspond to different stages along the path to fully differentiated Th2 cells.