Computational Approaches to the Analysis of Biomolecular Sequences, Structures and Their Functions and Applications to Biotechnology and Clinical Data Studies (23 - 27 Mar 2020)

Organizing Committee

Co-chairs
Igor N. Berezovsky (Bioinformatics Institute, A*STAR)
Lars Nordenskiöld (Nanyan Technological University)
Frank Eisenhaber (Bioinformatics Institute, A*STAR)

Members
Zhang Louxin (National University of Singapore)
Choi Kwok Pui (National University of Singapore)
Day 1. Monday, March 23, 2020
Workshop and Tutorial on Sequence-Based Analysis of Biomolecules and Molecular Modelling for Therapeutics

Morning Session 9:30 – 12:15
Chair: Frank Eisenhaber

9:50 – 10:20
Michael Galperin, NCBI/NIH, USA
Talk “Functional annotation of microbial genomes, one COG at the time”

ABSTRACT
Functional annotation of microbial genomes, one COG at a time
Michael Y. Galperin NCBI, NLM, National Institutes of Health, Bethesda, Maryland, USA

Genome sequencing projects continue to flood public databases with sequences of deduced proteins, only a small fraction of which has been ever studied experimentally or could be studied in detail any time soon. The only feasible way to assign functions to these proteins is to predict them through computational analysis. The Clusters of Orthologous Genes (COG) database, first created in 1997, has been a popular tool for functional annotation of bacterial, archaeal, and eukaryotic genomes. This talk will discuss the key reason for the success of COGs, including 1) its reliance on complete microbial genomes, which allows unequivocal assignment of orthologs and paralogs; 2) family-based approach, which used the functions of the characterized members of the protein family (COG) to assign function to the entire family and describe the range of the potential functions when there were more than one, 3) careful manual curation of the COG names and protein contents, and 4) a convenient classification of COGs into sensible functional groups. We will use COG annotations to discuss the problems of assigning protein function and the common errors that complicate this process.

References
   https://academic.oup.com/nar/article/43/D1/D261/2439462
ABSTRACT
Using COG phylogenetic patterns in comparative genome analysis: a tutorial
Michael Y. Galperin NCBI, NLM, National Institutes of Health, Bethesda, Maryland, USA

Protein domain databases, such as CDD, Pfam and InterPro, allow an easy assignment of previously uncharacterized proteins to the known protein (super)families. The COG database, while smaller than these, has an advantage of allowing the user to see which members of the given protein family (COG) are not encoded in a given genome. This tutorial will use this feature to 1) distinguish orthologs and paralogs in microbial genomes, 2) identify errors in completed genome sequences, and 3) compare phyletic profiles of individual protein chains of multisubunit enzymes and of the enzymes belonging to the same pathway.

References
ABSTRACT
We first give an introduction of some general principles, techniques and checks in protein sequence analysis. Next, we tackle an interesting example of sequence and function analysis, the lonely guy enzyme, an important plant cytokinin processing enzyme. We then illustrate the function of cytokinin looking at some systems biological effects of this hormonal system. Currently, we are experiencing a big data wave permitting large-scale sequence analysis of mRNA data. To illustrate this, we show how signatures for therapeutic responders in leukemia (B-precursor ALL) are identified. On a more molecular level, different immune responses of human monocyte-derived dendritic cells and macrophages against A. fumigatus infection can be identified and modelled. We finally summarize our key insights into some general rules for bioinformatical protein function and sequence analysis as well as indicate some important limitations of our knowledge of the protein universe.
ABSTRACT

Chromatin is a complex structure composed of densely compacted DNA and a myriad of molecular factors. Biological processes within chromatin remain highly organized despite spatial crowding of the system, which is believed to be due to the organization of chromatin into community structures. Capturing and understanding these community structures is thus the first step in studying how chromatin structure relates to its function and regulation. To achieve this, we modelled data from high-throughput chromosome conformation capture (Hi-C) using a Markov State Model defined on the interaction network. Various aspects of Transition Path Theory were used to develop a concrete definition of structural partitions and their interactions in chromatin, which enabled us to detect the structural hierarchy in chromatin organization, and characterize whole-genome structure for further analysis.
ABSTRACT

How bacteria see the world
Michael Y. Galperin NCBI, NLM, National Institutes of Health, Bethesda, Maryland, USA

Signal transduction systems allow the bacterial cells to sense a variety of environmental parameters, such as the temperature, pH, salinity, osmotic pressure, presence of nutrients and poisons, and to respond to the changes in these parameters by adjusting their behavior and metabolism. In this talk, I will discuss 1) the key advantages of two- (or multi-) component signal transduction over one-component regulatory systems; 2) unusual properties of archaeal two-component signalling; 3) diversity of the bacterial sensory systems, and 4) the recent progress in understanding the mechanisms and evolution of c-di-GMP-mediated signaling. In addition, I will show how the ability to enumerate all environmental sensing systems encoded in a given bacterium provides a better understanding of its behavior and offers certain clues in the search for novel antibacterial drug targets and/or an insight into the ways to suppress production of virulence factors altogether.

References


ABSTRACT

Identification of novel domains of microbial signal transduction systems: a tutorial
Michael Y. Galperin NCBI, NLM, National Institutes of Health, Bethesda, Maryland, USA

Many signal transduction proteins have unrecognized ligand-binding and/or signal-transducing domains. This tutorial will show how such proteins could be identified in the protein databases and how the new domains could be identified and characterized. This would include preparing a sequence alignment, identification of the most conserved residues using the sequence logo, secondary structure prediction, identification of the conserved domain architectures, prediction of the likely function and generation of the submission to one of the popular domain databases.

References
Day 2. Tuesday, March 24, 2020
Workshop and Tutorial on the Physics and Evolution of Protein Structure and Function

Morning Session 9:30 – 12:00
Chair: Igor N. Berezovsky

9:30 – 10:30
Gustavo Caetano-Anolles,
University of Illinois at Urbana-Champaign, USA
Talk “Phylogenomic and the evolutionary rise of hierarchy and community structure in biological systems”

ABSTRACT
Gustavo Caetano-Anollés
Department of Crop Sciences and C.R. Woese Institute for Genomic Biology,
University of Illinois, Urbana, IL 61801, USA. E-mail: gca@illinois.edu

Phylogenomics and the evolutionary rise of hierarchy and community structure in biological systems

Networks describe how parts associate with each other to form integrated systems, which are often structured by modularity and hierarchy. Here we explain the structure of biological networks with a biphasic (bow-tie) theory of module emergence. In a first phase, parts are at first weakly linked and associate variously. As they diversify, they compete with each other and are often selected for performance. The emerging interactions constrain their structure and associations. This causes parts to self-organize into modules with tight linkage. In a second phase, variants of the modules diversify and become new parts for a new generative cycle of higher-level organization. The paradigm predicts the rise of hierarchical modularity in evolving networks at different timescales (nanosecond to billions of years) and complexity levels, which we confirm with phylogenomic data.
11:00 – 12:00
Zhen Wah Tan, Bioinformatics Institute, A*STAR, Singapore
Tutorial “Modeling allostery: global protein structural changes in response to local perturbations”

ABSTRACT
Characterizing the global structural effects of allosteric binding is a challenging task, often requiring extensive molecular dynamics simulations to reveal the change in the structural rigidity of different regions. To enable rapid screening of multiple potential allosteric sites, we have developed a structure-based statistical mechanical model that allows for quick estimates of changes in structural dynamics of ligand-bound or mutated proteins. In this talk, we will introduce two web resources we have developed: First, the AlloSigMA server (http://allosigma.bii.a-star.edu.sg) allows users to visualize and explore model predictions for user-defined protein structures, and screen for new --- or latent --- allosteric sites. Second, the AlloMAPS database (http://allomaps.bii.a-star.edu.sg) contains pre-computed data for a collection of (i) classical allosteric proteins, (ii) proteins associated with pathological SNPs, and (iii) protein chains representing a wide variety of structural folds.
13:30 – 14:00
Gustavo Caetano-Anolles, University of Illinois at Urbana-Champaign, USA
Tutorial “‘Memory’ describing evolutionary rationale”

ABSTRACT
‘Memory’ describing evolutionary rationale
Understanding how biological molecules evolve requires untangling their structural organization with phylogenomic approaches. Here, we will summarize current challenges in a ‘Memory’ segment that describes the evolutionary rationale.
14:00 – 14:15
Melvin Yin, Bioinformatics Institute, A*STAR, Singapore

Selected short talk “Deriving structure profiles corresponding to functional loops in proteins”

ABSTRACT
Functional segments in proteins are highly conserved across function-based families in both sequence and structure. We use sequence alignment and motif search techniques to automatically identify such segments in labelled proteins, and consolidate them into generalised structure profiles. These are manually verified to correspond to known functional loops. From the profiles, we build descriptors that contain the necessary structural information to replicate the environment that lead to the original protein performing the corresponding chemical reaction. These can be used in identification of current or previous functions in unlabelled proteins through a structure match scoring function, or in de novo protein design.
ABSTRACT
Phomafungin is a recently reported broad spectrum antifungal compound but its biosynthetic pathway is unknown. We combed publicly available Phoma genomes but failed to find any putative biosynthetic gene cluster that could account for its biosynthesis. Therefore, we sequenced the genome of one of our Phoma strains (F3723) previously identified as having antifungal activity in a high-throughput screen. We found a biosynthetic gene cluster that was predicted to synthesize a cyclic lipodepsipeptide that differs in the amino acid composition compared to Phomafungin. Antifungal activity guided isolation yielded a new compound, BII-Rafflesfungin, the structure of which was determined. We describe the NRPS-t1PKS cluster ‘BIIRfg’ compatible with the synthesis of the cyclic lipodepsipeptide BII-Rafflesfungin [HMHDAL-Ala-L-Glu-L-Asn-L-Ser-L-Ser-D-Ser-D-allo-Thr-Gly]. We propose a mechanism for BII-Rafflesfungin biosynthesis, which involves the formation of the lipid part by BIIRfg_PKS followed by activation and transfer of the lipid chain by a predicted AMP-ligase on to the first PCP domain of the BIIRfg_NRPS gene.
15:00 – 15:30
Gustavo Caetano-Anolles, University of Illinois at Urbana-Champaign, USA
Tutorial “Structure’ defining molecular complexities”

ABSTRACT
Structure’ defining molecular complexities
Understanding how biological molecules evolve requires untangling their structural organization with phylogenomic approaches. Here, we will summarize current challenges in a ‘Structure’ segment that defines molecular complexities.
ABSTRACT
The structure, memory and linkage of molecular systems

Understanding how biological molecules evolve requires untangling their structural organization with phylogenomic approaches. Here, we will summarize current challenges in a ‘Linkage’ segment that discusses approaches to network science necessary for understanding biological systems.
Day 3. Wednesday, March 25, 2020
Workshop and Tutorial on Theoretical Models of Biomolecules

Morning Session 9:30 – 12:00
Chair: Igor N. Berezovsky

9:30 – 10:30
Marianne Rooman, Université Libre de Bruxelles, Belgium
Talk “Structure-based computational approaches for large-scale predictions of protein biophysical properties. Application to protein design, natural evolution and disease phenotypes”. Part 1

11:00 – 12:00
Marianne Rooman, Université Libre de Bruxelles, Belgium
Talk “Structure-based computational approaches for large-scale predictions of protein biophysical properties. Application to protein design, natural evolution and disease phenotypes”. Part 2

ABSTRACT
Biophysical properties of proteins, such as their stability, thermal resistance, solubility and binding affinities, are of key importance for them to function properly and to avoid aggregation and disease issues. To gain insights into these matters, we developed a physics-based software suite to predict the modifications of these properties upon amino acid mutations. These machine learning predictors use experimental or modelled protein structures as input and are extremely fast, allowing large scale screening of all possible amino acid substitutions in huge sets of proteins. Important ingredients are physics-based model structures and statistical potentials that have the advantage of reproducing the relative contributions of amino acid interactions to the folding free energy in various protein environments and at different temperatures. These tools have a series of challenging applications:

• The rational design of proteins with modified properties aiming to optimize a wide series of biotechnological processes is a straightforward application. The predictors can be used to guide and restrict the huge, time-consuming, search in sequence space and propose sets of mutants that have increased heat resistance, stability, solubility and remain active in non-physiological conditions.

• Another challenging problem is the classification of human genetic variants into disease-causing and benign and the understanding of their effect at the molecular and interactome scales. Our computational tools allow the fast identification of variants at the genome scale, which are deleterious because they affect protein stability, protein-protein interactions, function, or DNA charge properties.

• The question if and how natural evolution acts on DNA and protein sequences to ensure mutational robustness and evolvability has also been tackled through proteome-scale applications of our predictors. We demonstrated that both the standard genetic code and the codon usage is
partly optimized to limit the destabilizing effects of random mutations, and even more, to ensure translation accuracy.
9:30 – 10:30  
Lanyuan Lu, Nanyang Technological University, Singapore  
Talk “Constructing protein structure ensemble using solution experimental data”

ABSTRACT
Developing algorithms for identifying structure ensembles of protein systems on the basis of small angle X-ray scattering or nuclear magnetic resonance is an emerging research field due to the recent experimental technological advances. The determination of protein structure ensemble from those solution experimental data is usually a highly ill-posed problem and suffers from the over-fitting issue. We systematically investigated the problem using combined molecular simulation and experimental approach, and it was discovered that a well-defined structure ensemble based on low-resolution scattering data can only be obtained from a few low energy structures. A similar conclusion was also achieved using paramagnetic relaxation enhancement data and computer simulation. Results on proteins containing multiple domains and flexible links, such as the non-structural proteins of dengue virus, will be discussed.
ABSTRACT
Infectious diseases due to bacterial and viral pathogens are a significant threat to human health. The interaction and assembly of such pathogens with host cells is critical to infection, the immunological response, and therapeutic interventions. We use multiscale simulations and integrative modelling approaches to investigate these processes, from atomic resolution to highly simplified coarse-grained levels. Here, I will give an overview of this research, with a particular focus on: (1) flavivirus dynamics during its infective life cycle; and (2) mammalian immune receptor responses upon exposure to bacterial pathogens.

For the former, we especially study dengue virus, responsible for more than 400 million infections per year. In serious cases, the virus can cause dengue hemorrhagic fever and shock syndrome, often facilitated by the host via antibody-dependent enhancement. In order to fill existing gaps in our molecular understanding of the viral life cycle, we have been integrating structural, biophysical, and genomic experiments with multiscale modelling, towards the “virtual dengue virus”. We have reconstructed and refined the complete dengue envelope in near-atomic resolution, under different states of maturation. These virtual viruses serve as a platform for following mechanistic stages along the life cycle, such as endosomal membrane remodeling and fusion, and enabled us to identify how virus particles interact with antibodies to modulate or even facilitate infectivity, with important consequences for vaccine development.

For the latter, we are interested in the mechanisms by which endotoxic molecules released from bacterial pathogens over-stimulate immune receptors of the innate host defense system, which can lead to sepsis, a condition that kills millions of people each year. To better understand this process, we have developed computational models to trace in atomic and molecular detail the cascades and intermediates associated with transfer of endotoxins from bacterial surfaces to host immune receptors. We have also leveraged these models, to establish previously undisclosed modes of action of anti-endotoxic peptides that occur naturally during wound healing. Collectively, this work has helped to unravel key determinants governing interactions between bacterial pathogens and the host, and should help towards the search for novel therapeutics to tackle ongoing issues of antimicrobial resistance.
Afternoon Session 13:30 – 16:15
Chair: Lars Nordenskiöld

13:30 – 14:30
Alexander Lyubartsev, Stockholm University, Sweden
Talk “Multiscale modeling of macromolecular systems by structure-based coarse-graining”

ABSTRACT
Molecular simulations of many phenomena related to biomolecular systems, soft matter and nanomaterials requires consideration of length scales above 10 nm and time scales longer than 1 mks, which necessitates the use of coarse-grained (low resolution) models, when each site of the model represents a group of atoms, and the solvent is often omitted. While many of coarse-grained models used in different studies in recent years relay on empirically parametrized interaction potentials, the systematic structure-based coarse-graining approach is based on determination of coarse-grained potentials from atomistic (high resolution) simulations.

Here a multiscale modeling approach based on the inverse Monte Carlo method is presented, in which radial distribution functions (RDF) and distributions of internal degrees of freedom of molecular structure, obtained in high-resolution atomistic simulations, are used to reconstruct effective potentials which reproduce the same structural properties within low-resolution coarse-grained model. The statistical-mechanical equations expressing canonical properties such as RDFs in terms of potential parameters can be inverted and solved numerically according to the iterative Newton scheme. The approach is illustrated on several examples of varying complexity: ionic solution; ionic liquids, coarse-grained lipid model, coarse-grained DNA model and nucleosome core particles. It is demonstrated further how effective potentials, derived exclusively from atomistic simulations, can be used to model such phenomena as lipid self-assembly, formations of vesicles and other ordered structures at varying lipid composition and concentration of different components, modeling of long DNA fragments at varying ionic conditions. A software package MagiC implementing the inverse Monte Carlo method for computation of effective potentials for coarse grained models of arbitrary structure from atomistic trajectories is presented.
ABSTRACT
Viruses are metastable macromolecular assemblies that undergo reversible conformational rearrangements in solution referred to as ‘breathing’ that are critical for sensing host-specific environmental cues, receptor interactions and for initiating host entry. Host specific environmental cues including temperature, osmolyte and pH trigger large changes in viral assembly leading to formation of disassembly intermediates and eventual disassembly with release of genomic material. Amide hydrogen deuterium exchange mass spectrometry (HDXMS) is a powerful tool for measuring hydrogen bonding propensities and solvent accessibility of proteins and protein complexes. This makes it especially suitable for probing the breathing dynamics, measurement of quaternary contacts of viral particles in solution as well as to probe the effects of host-specific conditions and for mapping epitopes and paratopes of neutralizing antibodies. In this talk, I will present a short overview of mass spectrometry applications in structural biology for mapping protein assemblies and interfaces with lipids and nucleic acids with dengue and zika viral dynamics as the focus. Our HDXMS results show that each dengue/zika serotype/strain display a distinct breathing dynamics profile with an entirely different expansion response to temperature. At 37 °C, DENV serotype 2 strains show temperature-specific changes with the biggest change at the Envelope E-intradimeric interface. However, at 40 °C, a different set of temperature-specific loci were detected in strains of DENV serotype 1. These unique serotype dynamic profiles underscore the importance of protein quaternary contacts, packing of lipid bilayer and RNA genome in each viral particle. HDXMS has also enabled detailed maps of epitope and paratopes for 2 (1 stabilizing and 1 destabilizing) antibody-dengue virion complexes. I will conclude by describing how HDXMS can reveal the role of genomic RNA in viral stability and dynamics in a model plant RNA virus, Turnip Crinkle Virus (TCV).
ABSTRACT
Antibodies are classified into five major immunoglobulin (Ig) isotypes and they can have different numbers of antigen binding sites. For example, IgG has two antigen binding sites, whereas IgM has up to twelve. While the former is the isotype of choice for most currently available therapeutic antibodies, the latter has superior complement-activating and cell-agglutinating abilities. Utilizing IgM is therefore an attractive option for future immunotherapy. But is having more antigen binding sites always better? In this study, we performed an integrative multiscale modelling and simulation of two well-known monoclonal antibodies used to treat human epidermal growth factor receptor 2 (HER2)-positive breast cancer: Pertuzumab and Trastuzumab. We found that only Pertuzumab IgM can utilize all of its twelve antigen binding sites to bind to HER2 extracellular domain, while similar binding in Trastuzumab IgM is hindered by steric clashes. This is subsequently validated by cell count assay showing that Pertuzumab IgM is more efficient at inhibiting the proliferation of HER2-overexpressing cells compared to its IgG counterpart and to Trastuzumab IgM. Our study highlights the importance of understanding the molecular details of antigen-antibody interaction for the design and isotype selection of therapeutic antibodies.
Day 5. Friday, March 27, 2020
Biomedical Modelling, Informatics, and Clinical Data Studies

Morning Session 9:30 – 12:00
Chair: Frank Eisenhaber

9:30 – 10:15
Samuel Gan, Bioinformatics Institute, A*STAR, Singapore
Selected short talk “Antibody engineering, Scientific Phone Apps, Viral research, and Device prototyping in APD Lab”

ABSTRACT
Many significant biomedical discoveries of old were made in the private property of famous scientists e.g. Leeuwenhoek and Archimedes. Today, discoveries are made in brightly-lit, hi-tech, ergonomic buildings that house research institutes. While such development is advantageous in many aspects, the spatial restriction of research into well-organized structures may delay and limit the spontaneity necessary for discoveries. The smartphone and peripheral mobile devices have the potential to not only increase the productivity and mobility of biomedical research, but also restore some freedom from spatial constraints. One possible way this can occur is the development of a mobile biomedical lab that allows researchers to carry out core research processes ‘on-the-go’ without being spatially restrained within a building or availability of equipment. This talk introduces the world of Scientific Phone Apps and device prototyping in APD Lab, with a sneak preview into the unique antibody engineering and viral research by the lab.
ABSTRACT
The reductionist approach is prevalent in biomedical science. However, increasing evidence now shows that biological systems cannot be simply considered as the sum of its parts. With experimental, technological and computational advances, we can now do more than view parts in isolation, thus we propose that an increasing holistic view (where a protein is investigated as much as a whole as possible) is now timely. To further advocate this, we investigate and discuss several studies and applications involving allostery, where distant protein regions can cross-talk to influence functionality. As a result, we believe that an increasing big picture approach holds great promise, particularly in the areas of antibody engineering and drug discovery in rational drug design.
ABSTRACT
We aim to identify the associations between key components on genome and epigenome levels, and to understand how the perturbation of these components can lead to disease. To accomplish this, we develop and apply machine learning, molecular modeling, molecular dynamics simulation methods, and use experimental data to guide us in designing new hybrid approaches. Such methods can provide a predictive power for the behavior of the system in response to disease and offer experimental leads.
ABSTRACT
The quality of a genome assembly is assessed by a variety of measures (e.g. NGA50, number of known core genes mapped, and number of translocation). These typically consider the contiguity, completeness, and correctness of the assembly separately and independently. It is often the case that one assembly is slightly superior to another assembly in some measures and is slightly worse in other measures. Thus one cannot easily tell which assembly is better. Indeed, even when one assembly is vastly superior to the other, these measures often suggest very small differences (e.g. 1-2%) between the assemblies. In this talk, I will describe a genome assembly quality assessment measure called PDR (pairwise distance reconstruction). The PDR is based on the average difference between the observed distance of two loci on an assembly and the actual distance on the reference genome. In other words, PDR determines how well the assembly can be used to inform on the distance between two arbitrary loci in the actually genome. It is easy to see that the more contiguous, complete, and correct an assembly is, the better it can inform on the distance between two arbitrary loci on the actual genome. Thus the PDR naturally generalizes and integrates these properties. As the PDR is averaged over all possible pairs of positions on a genome, it is impractical to compute it in a naïve manner for large genomes. Fortunately, an approximation to the PDR that is fast to compute accurately (error is less than E-10) is possible using piecewise integrals. I will show in this talk that this makes the PDR a very practical informative genome assembly quality assessment measure. This talk is based on the PhD thesis of my student Xie Luyu.
ABSTRACT
Reinforcement learning is a machine learning method that is gaining growing popularity in different areas of research. This talk will present an overview of multiple reinforcement learning algorithms. We will introduce Q-learning and policy gradient as two most widely known reinforcement learning methods. We will explain how these techniques are effectively utilized to address problems in biomedical data and image domain.
15:00 – 16:00  
Lian Kaicheng, Bioinformatics Institute, A*STAR, Singapore  
Tutorial “Computational imaging with deep learning: biomedical focus”

ABSTRACT  
Modern approaches to artificial intelligence using deep learning have achieved significant milestones in computer vision. Beyond image classification and analysis, AI techniques for imaging and image generation have also shown great promise. This talk will review the state of the art in generative AI for computational challenges such as super-resolution and style transfer, with a focus on biomedical applications and microscopy. Future opportunities and potential pitfalls will also be discussed.
ABSTRACT
It is well understood that a substantial proportion of predicted genes are classifiable as having no known function. Current approaches to defining gene function are highly dependent on comparative genomic methods and the use of reference databases such as KEGG. However, this kind of approach will by nature eliminate genes of unknown function from consideration, and this becomes particularly limiting when considering the functional repertoire of member species of complex microbial communities (microbiomes). Here I discuss some recent work that uses sequence similarity networks to provide a working model for organising functional genes irrespective of whether their biological function is currently understood or not. Briefly, protein sequence from genes predicted within metagenome assemblies are subjected to all-versus-all protein BLAST, and analysed as a network. I will discuss the validation of these networks against known functional databases (KEGG) and how we have applied this concept to the interpretation of gene catalogues from metagenome assemblies. I will illustrate this approach using metagenome assemblies obtained from bioreactor communities enriched for anaerobic ammonium oxidising bacteria (AnAOB), directed toward the ongoing search for the key genes underpinning the biochemistry of the anaerobic ammonium oxidisation reaction (ANAMMOX).