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 Commettiee

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 – 12:30
  Chair: Frank Eisenhaber

  9:30 – 10:30

  Frank Eisenhaber, Bioinformatics Institute, A*STAR, Singapore

  Welcome by Organisers *and*

  Talk "About the limited potential, yet instrumental role of computational biology and the target identification bottleneck due to the decline in biomolecular mechanism discovery after 2000"

  ABSTRACT
  Computational biology is all about studying biomolecular sequences, structures and biological images as well as composite data types such as clinical data and the understanding of mechanisms that lead determine the phenotype for a given genome. The discovery of sequence homology was the prime reason for pushing bioinformatics into the center stage as it connects various areas of life science research via common biomolecular mechanisms. Although there is still no coherent biological theory, nevertheless, computational biology became the main driver for hypothesis generation in life sciences.

  It is generally believed that full human genome sequencing was a watershed event in human history that boosted biomedical research, biomolecular mechanism discovery and life science applications. At the same time, researchers in the field of genome annotation see that there is a persisting, substantial body of functionally insufficiently or completely not characterized genes (for example, ~10,000 protein-coding in the human genome) despite the availability of full genome sequences. A survey of the biomedical literature shows that the number of reported new protein functions had been steadily growing until 2000 but the trend reversed to a dramatic decline thereafter (1,2) when, at the same time, the annual amount of new life science publications doubled between 2000 and 2017.

  This reduction in the supply of newly characterized pathways has profound implications for the drug development pipelines in industry as well as for research. The example of the SUGCT function discovery (3) shows the significance that many uncharacterized genes will have in aging, metabolic diseases and their complications.

  1) Darkness in the human gene and protein function space: Widely modest or absent illumination by the life science literature and the trend for fewer protein function discoveries since 2000. Sinha S, Eisenhaber B, Jensen LJ, Kalbuaji B,

Eisenhaber F. Proteomics. 2018 Nov;18(21-22):e1800093. doi: 10.1002/pmic.201800093

2) A decade after the first full human genome sequencing: when will we understand our own genome? Eisenhaber F. J Bioinform Comput Biol. 2012 Oct;10(5):1271001. doi: 10.1142/S0219720012710011

3) Knockout of the non-essential gene SUGCT creates diet-linked, age-related microbiome disbalance with a diabetes-like metabolic syndrome phenotype. Niska-Blakie J, Gopinathan L, Low KN, Kien YL, Goh CMF, Caldez MJ, Pfeiffenberger E, Jones OS, Ong CB, Kurochkin IV, Coppola V, Tessarollo L, Choi H, Kanagasundaram Y, Eisenhaber F, Maurer-Stroh S, Kaldis P. Cell Mol Life Sci. 2019 Nov 13. doi: 10.1007/s00018-019-03359-z

11:00 – 11:45

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.

11:45 – 12:00
Chinh Trah To Su, Bioinformatics Institute, A*STAR, Singapore
Selected short talk "A holistic view of proteins: impact on antibody engineering and drug discovery"

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.

Afternoon Session 13:30 – 16:00
Chair: Frank Eisenhaber

13:30 – 14:00
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.

14:00 – 14:30

Swati Sinha, Bioinformatics Institute, A*STAR, Singapore

Selected short talk "Discovery of an Antifungal Compound BII-Rafflesfungin: A Computational Perspective"

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 [HMHDA-L-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 – 16:00

Marek Mutwil, School of Biological Sciences, Nanyang Technological University, Singapore
Talk "Evolutionary lessons learned from transcriptome data"

ABSTRACT
Marek Mutwil, SBS, NTU, Singapore
Evolutionary lessons learned from transcriptome data

To understand the evolution of various organs and metabolic pathways, it is necessary to understand the functions of the underlying gene products. Classical genomic approaches based solely on gene sequences are useful, but have shortcomings, as they cannot readily reveal which genes work together in a pathway, i.e., form a functional gene module. Consequently, to study the evolution of new traits, we need to integrate the classical genomic approaches with predicted functional gene modules, which can be identified by studying, for example, gene expression, protein-protein interactions or co-expression networks.

I discuss how a variety of transcriptomic approaches can be used to answer evolutionary questions that cannot be easily addressed with genomics. More specifically, I will present how transcriptomes can be used to identify essential genes, discover metabolic pathways and study the evolution of new cell wall types and organs. I will also describe the existence of a gene expression program that is conserved over 1,500,000,000 years of evolution.

- **Day 2. Tuesday, March 24, 2020**
  Workshop and Tutorial on the Protein and Chromatin Structure and Function

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

  9:30 – 10:30

  Konstantin Pervushin, School of Biological Sciences, Nanyang Technological University, Singapore
  Talk "Abeta chaperones in Alzheimer's disease: friends or foes?"

  ABSTRACT

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.

Afternoon Session 13:30 – 16:00
Chair: Igor N. Berezovsky

13:30 – 14:30

Amartya Sanyal, School of Biological Sciences, Nanyang Technological University, Singapore

Talk "3D genome: from organization to function"

ABSTRACT

All genomic processes take place in the context of chromatin, the organization of genomic DNA with histones and hundreds of other proteins and RNAs inside the nucleus. The higher order folding of chromatin controls transcription and other nuclear processes, and is important for cellular identity and cell state transitions. At the beginning of the talk, I will provide a brief overview about the organization of higher-order structures of chromatin and their detection using chromosome conformation capture (3C)-based methods. Next, I will narrate our latest work on cancer drug resistance that involves interplay between genetic and epigenetic mechanisms, leading to transcriptional reprogramming. At the end, I will explain a simplified low-input 3C-seq protocol that we developed to interrogate chromatin organization in patient samples.

15:00 – 16:00

Zhen Wah Tan, Bioinformatics Institute, A*STAR, Singapore

Tutorial “Capturing communities in diffusive chromatin polymers via Markov State Modeling and Transition Path Theory”

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.

Day 4. Thursday, March 26, 2020
Coarse-Graining in Molecular Simulations and Reconstruction of Biomolecules and their Assemblies

Morning Session 9:30 – 12:00
Chair: Lars Nordenskiöld

9:30 – 10:30
Lanyuan Lu, Nanyang Technlogical 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.

11:00 – 12:00

Peter J. Bond, Bioinformatics Institute, A*STAR, Singapore

Talk "Multiscale modelling approaches to large systems in biology: understanding infectious diseases and the host response"

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.

15:00 – 16:00

Ganesh Srinivasan Anand, Department of Biological Sciences, National University of Singapore, Singapore
Talk "Whole Viral Particle Metastability and Dynamics by HDX Mass Spectrometry"

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).

16:00 – 16:15

Firdaus Samsudin, Bioinformatics Institute, A*STAR, Singapore

Selected short talk, "To bind or not to bind: IgM versus IgG in Pertuzumab and Trastuzumab"

Firdaus Samsudin[a], Joshua Yi Yeo[a], Samuel Ken-En Gan[a], and Peter J. Bond[a,b]
[a]Bioinformatics Institute, A*STAR, Singapore
[b]National University of Singapore, Singapore

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 – 10:30

Gerhard Grüber, School of Biological Sciences, Nanyang Technological University, Singapore
Talk "Amino acid sequence insertions or deletions inside Mycobacterium tuberculosis's respiratory complexes provide fitness for the pathogen and new pathways for TB-drug development"

ABSTRACT

The causative agent of Tuberculosis (TB) *Mycobacterium tuberculosis* (*Mtb*) encounters unfavourable environmental conditions in the lungs, including nutrient limitation, low oxygen tensions and/or low/high pH values. These harsh conditions in the host triggers *Mtb* to enter a dormant state in which the pathogen does not replicate and uses host-derived fatty acids instead of carbohydrates as an energy source. Independent to the energy source, the bacterium's energy currency ATP is generated by oxidative phosphorylation, in which the F_1F_O-ATP synthase uses the proton motive force generated by the respiratory complexes, including two terminal oxidases: a proton-pumping cytochrome $bc_1{:}aa_3$ oxidase supercomplex and a non-proton pumping, less energetically efficient, cytochrome *bd* oxidase. The later reveals regions and amino acid residues unique to mycobacteria, opening a door for TB-specific compound design.

A special feature of the mycobacterial F-ATP synthase is its inability to establish a significant proton gradient during ATP hydrolysis, and its latent ATPase activity, to prevent energy waste and to control the membrane potential. Recently, we have identified unique epitopes of mycobacterial F_1F_O-ATP synthase subunits absent in their prokaryotic or mitochondrial counterparts which contribute to the regulation of the low ATPase activity. Most recent structural insights into individual subunits, the F_1 domain or the entire mycobacterial enzyme added to the understanding of mechanisms, regulation and differences of the mycobacterial F_1F_O-ATP synthase compared to other bacterial and eukaryotic engines. Novel insights providing the basis for the design of new compounds targeting this engine and even novel regimens for multidrug resistant TB will be presented.

11:00 – 12:00

Lim Soon Wong, National University of Singapore, Singapore
Talk "A genome assembly quality assessment measure that integrates contiguity, completeness, and correctness"

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.

Afternoon Session 13:30 – 16:15
Chair: Frank Eisenhaber

13:30 – 14:30

Mahsa Paknezhad, Bioinformatics Institute, A*STAR, Singapore
*Tutorial "*Reinforcement learning and its biomedical applications"

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

Liang 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.

16:00 – 16:15

Rohan Williams, Singapore Centre for Environmental Life Sciences Engineering (SCELSE), National University of Singapore, Singapore

Selected short talk "Sequence similarity networks: a working model for defining the 'known-unknowns' of gene function in microbiomes"

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).