MRes Bioinformatics and Systems Biology

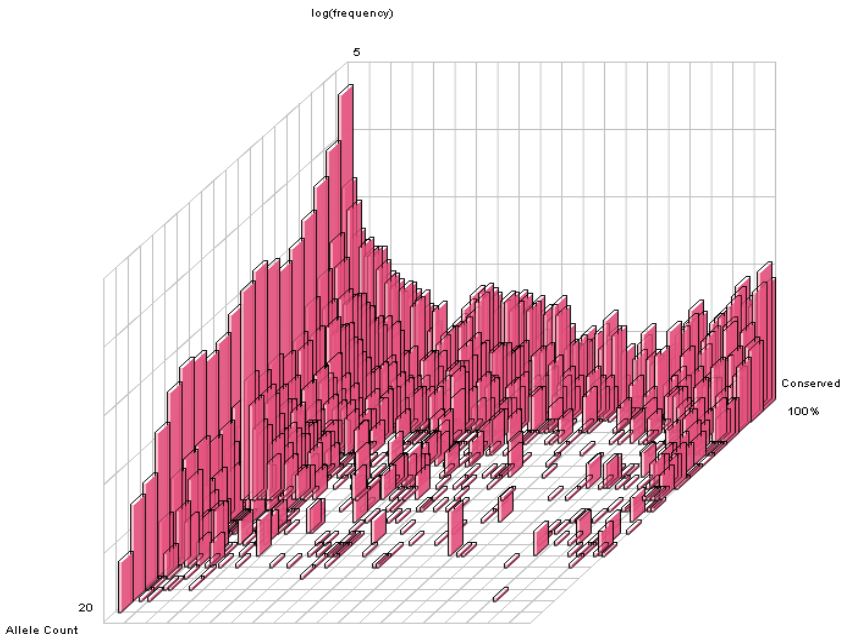
Project list Nov 2008

The evolution of the influenza virus: a Bayesian approach

Supervisors: Adrian Shepherd and David Moss (a.shepherd and d.moss@mail.cryst.bbk.ac.uk)

The influenza virus is a constantly evolving pathogen and each year new vaccines have to be developed that are based on the current viral strains that are circulating in the population. There is also the possibility that re-assortment of genetic material can take place in infected animals such as birds and pigs, resulting in new viral strains that may reach the human population causing a pandemic before effective vaccines can be developed.

Here at Birkbeck we have constructed a database of 40034 polypeptide sequences for influenza virus isolates and examining the mutations that affect transmissibility and lethality. We will then go on to predict the likelihood of mutations that might lead to seriously pathogenic strains. This information, together with information on the distribution of T-cell epitopes would inform vaccine design



The figure above shows the results of analysing 40034 polypeptide sequences derived from 7954 influenza-A strains. The vertical bars show the distribution of 14 million predicted MHC Class I binding peptides in bins, classified according to sequence conservation across strains versus number of MHC alleles predicted to bind.

In this project we would investigate the time course of viral evolution by investigating the mutations caused by genetic drift in the fast mutating proteins, hemagglutinin and neurominidase. We would construct statistical models (using, for example, a Bayesian approach) from mutation matrices to show the probability of a given observed protein sequence mutating into another observed sequence. These models would be tested by their ability to predict the appearance of mutated sequences, using the sequence data in our database that is taken from dated isolates.

This project would require some knowledge of probability theory. Some knowledge of the programming language R would be beneficial.

Prediction of regulatory regions in mammalian genomes

Supervisor: Alona Sosinsky (a.sosinsky@mail.cryst.bbk.ac.uk)

Functional annotation of the non-protein coding DNA of eukaryotic genomes is a challenging problem in modern computational biology. This project includes the development of new method for prediction of regulatory regions in mammalian genomes based on phylogenetic footprinting of orthologous sequences. New method will be validated using a set of regulatory regions based on experimental data from the ENCODE Project Consortium and the Eukaryotic Promoters Database. In combination with other approaches the new method will be used for prediction of new target genes for SF-1 transcription factor in adrenal gland in order to identify new functions for this transcription regulator.

Previous projects

A comparison of plant and human metabolites using QSAR and artificial neural networks

Supervisors: Irilenia Nobeli (Kings College London) and Adrian Shepherd (a.shepherd@mail.cryst.bbk.ac.uk)

Cognate (native) small molecules are becoming an increasingly important topic in bioinformatics research. On the one hand, there is considerable interest in the properties and activities of small molecule inhibitors, as their role is central to the development of drugs for therapeutic purposes. On the other hand, there is a growing interest in mapping the metabolomes of organisms as part of the emerging field of systems biology[1-4].

Most of our knowledge on the structure and properties of small molecules relates to the metabolomes of bacteria and humans. However, the small molecules of

plants are an important and neglected topic, as they have the potential to provide a rich pool of diverse, unexploited leads for the pharmaceutical industry. Hence a study of the known plant metabolite structures and their properties is of fundamental interest to drug design. A comparison to human metabolites is also essential, as an overlap in the space defined by plant and human structures and properties would alert us to potential interference from plant inhibitors with the normal metabolic pathways of humans (which would, in general, be undesirable).

The student on this project will undertake a comparison of the virtual spaces of human and plant metabolites defined by the values of Quantitative Structure Activity Relationship (QSAR) descriptors. Karakoc *et al.*[4] showed recently that human metabolites occupy compact and distinct clusters in QSAR space that do not overlap with the clusters defined by conventional drugs, antimicrobials or bacterial metabolites. They also showed that artificial neural networks (ANNs) used to derive binary classifiers based on the QSAR descriptors provide the most accurate predictions in distinguishing between the different classes of small molecules. We believe that the same approach will be effective when comparing plant and human metabolites.

This project will utilise structural data for human metabolites available from the Human Metabolome Database (www.hmdb.ca/). Structures of plant metabolites can be obtained from a variety of sources, most notably the AraCyc database (http://www.arabidopsis.org/tools/aracyc) and the KEGG Ligand database (www.genome.ad.jp/kegg/ligand.html). Molecular descriptors can be calculated with software available at King's College London. Dr Shepherd will provide software for designing and training ANNs.

References

Hattori M, Okuno Y, Goto S, Kanehisa M. 2003. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. J Am Chem Soc 125:11853-11865.

Nobeli I, Ponstingl H, Krissinel EB, Thornton JM. 2003. A structure-based anatomy of the E.coli metabolome. J Mol Biol 334:697-719.

Nobeli I, Spriggs RV, George RA, Thornton JM. 2005. A ligand-centric analysis of the diversity and evolution of protein-ligand relationships in E.coli. J Mol Biol 347:415-436.

Karakoc E, Sahinalp SC, Cherkasov A. 2006. Comparative QSAR- and Fragments Distribution Analysis of Drugs, Druglikes, Metabolic Substances, and Antimicrobial Compounds. J. Chem. Inf. Model. 46:2167-2182.

A display tool for integrated annotations

Supervisor: Andrew C.R. Martin, UCL Biochemistry
(martin@biochemistry.ucl.ac.uk)

Recent work in my group as led to the development of 'APAT' (Automated Protein

Annotation Tool), a system designed for taking a protein sequence and running it through a number of annotation/prediction servers either locally or remotely over the web. The system wraps the output from each tool in XML and a display program is used to convert the XML to HTML (including tables and graphs) for output. The current display tool displays the output of each tool separately, but in a consistent format. Some forms of annotation/prediction, particularly those which apply to individual amino acids (e.g. secondary structure prediction, membrane-spanning regions, post-translational modification sites), could usefully be displayed as a set of annotations on a single view of the sequence. This project will develop a new display tool to fulfill this requirement. The project will make heavy use of Perl and the Perl XML::DOM module as well as HTML and CSS for output.

The current version of APAT may be downloaded from
http://www.bioinf.org.uk/apat/

Excellent Perl programming skills are essential.

Deevi, S. Vishnu V. and Martin, Andrew C. R. (2006) An extensible automated protein annotation tool: standardizing input and output using validated XML, *Bioinformatics*, **22**,291-296.

A critical assessment of NMR structure determination methods using simulated data

Supervisor: Mark Williams (m.williams@mail.cryst.bbk.ac.uk)

Nuclear magnetic resonance spectroscopy provides a picture of proteins having variable conformation and undergoing motions, both local and global, on timescales from picoseconds to days. Much of this conformational variability is either functionally necessary or at least modulates function though its effect on thermodynamics or kinetics. Models for the structure of proteins determined by NMR are thus represented in the PDB by families of structures with a variety of conformations. However, the range of conformations of the protein within the family is influence by many factors; variation in the real protein structure, uncertainty in the interpretation of experimental data, lack of data or as a consequence of errors in the analysis and refinement procedures. Many methods have now been developed for the calculation of families of structures from spectral data and 'improvements' appear regularly. However, there is presently no way in which to readily compare the reliability and accuracy of structures produced by these methods nor to measure how well the resulting family of structural models represents the actual behaviour of the protein. To address these problems, this project will create simulated spectra from naturalistic models of protein structure in solution in such a way that the origins of all features of the spectra are known. The ability of different structure determination methodologies to account for the spectral features and the accuracy with they can reproduce the original models can then be critically assessed.
Taking advantage of recent increases in compute power of the Birkbeck and UCL high performance compute clusters, molecular and essential dynamics simulation methods (GROMACS) will be used to generate families of structures of several proteins, relaxation matrix methods (MARDIGRAS/YARM/FIRM) will then be used to

create artificial experimental data from these structures. This data will be made available as a web-based resource. A variety of semi-automated analysis and structure determination (e.g. CCPN/ARIA/CANDID/FastNMR) programs will be applied to the data to derive families of structures. These families will be compared to the simulated family using standard structural analysis software and a variety of statistical measures. By and large the project could be carried out using existing software, but there will be some essential scripting (probably using python) in order to exchange data between programs and to further automate some procedures. The programming aspect could be expanded to include implementation of additional methods into the standard CCPN analysis software.

Schneider *et al*. (1999). Influence of internal dynamics on accuracy of protein NMR structures: derivation of realistic model distance data from a long molecular dynamics trajectory. J. Mol. Biol. 285, 727-740.

Linge JP et al (2003). Refinement of protein structures in explicit solvent. Proteins: Struct. Funct. Gen. 50, 496-506.

Lindorff-Larsen *et al*. (2005). Simultaneous determination of protein structure and dynamics. Nature 433, 128-132.

Nilges *et al*. (2006) Error distribution derived NOE distance restraints. Proteins: Structure, Function, and Bioinformatics 64, 652-664.

A comparison of molecular dynamics and continuum models of the TATA-Binding Protein:DNA interaction

Supervisor: Mark Williams (m.williams@mail.cryst.bbk.ac.uk)

The functioning of cells is governed by the organised assembly and disassembly of molecular complexes. These complexes form in the crowded environment of the cell, which is full of other proteins, miscellaneous small molecules and salts. During the formation of a complex its surroundings are also rearranged, and this rearrangement may have a significant effect on the equilibrium population of the complex. A great challenge for computational biophysics is to develop a way of sufficiently accurately simulating such large systems of interacting molecules. The computational effort required to carry out a simulation increases significantly with the number of simulated molecules. Consequently, in order to deal with larger systems, simplifications must be made to the way in which the molecules and their interactions are represented. How much simplification can be made without losing essential features of the behaviour of the molecules? Indeed, which features are essential? This project seeks to compare different types of model of a protein-DNA interaction in order to begin to answer these questions.

The interaction between TATA binding protein and its cognate DNA sequence is essential for the initiation of transcription in eukaryotes and archea. The TATA-binding protein from the thermophilic and halophilic organism *Pyroccocus woesei* is one of the best studied protein-DNA interactions both structurally and thermodynamically. Both the effects of mutation and of changes in the environment (salt, temperature) have been extensively investigated. It is thus an excellent system for the evaluation of the capability of different levels of model to

explain structure/thermodynamics/function relationships. It is envisaged that standard software will be used to model wild-type and mutant complexes in variety of solution conditions using GROMACS (MD) and APBS (Poisson-Boltzmann continuum). The results from these methods will be compared to each other and to the experimental data. A suitably prepared student could extend the project to encompass other simulation methods, such as Brownian Dynamics.

Ladbury & Williams (2004). The extended interface: measuring non-local effects in biomolecular interactions. Current Opinion in Structural Biology 14, 562-569.
Bergqvist S, Williams MA, O'Brien R and Ladbury JE (2002). Reversal of protein halophilicity by limited mutation strategy. Structure 10, 629-637.

A Combined Bioinformatics and Structural Biology Analysis of Bardet-Biedl Syndrome

Supervisors: John Sgouros (j.sgouros@mail.cryst.bbk.ac.uk), Phil Beales (Institute of Child Health), and Helen Saibil

Bardet-Biedl syndrome (BBS, OMIM 209900) is a genetically heterogeneous disorder affecting several organs and systems. Genetic studies have revealed mutations in at least 12 genes, including novel, and apparently vertebrate-specific, members of the type II chaperonin superfamily. Analysis of the pathways involved, including comparative studies in *Caenorhabditis elegans*, suggests a ciliary defect implicating the centrosome and basal body. Among other symptoms, BBS patients develop obesity, type 2 diabetes and progressive retinal degeneration, thus making the syndrome an interesting candidate for understanding the molecular biology of the above diseases.

The aim of the proposed MRes project is to improve the understanding of the molecular defects in BBS using a combined bioinformatics and structural biology approach.

The objectives of the project are:

- Computational detection of potential regulatory elements and/or RNA genes in BBS loci by comparative sequence analysis of syntenic regions in human and other vertebrates and including functional data from the ENCODE project.
- Establishment of the biochemical pathways involved by analysing a) expression profiles of BBS genes in datasets from NCBI GEO and EBI ArrayExpress and b) protein interaction maps of BBS genes and their orthologues.
- Analysis of single nucleotide polymorphisms (SNPs) in BBS genes, especially in a population-specific context, using data from HapMap.

References

Beales P.L. (2005) Lifting the lid on Pandora's box: the Bardet-Biedl syndrome. *Curr. Opin. Genet. Dev.* **15:** 315-323

Badano J.L., Mitsuma N., Beales P.L., Katsanis N. (2006) The Ciliopathies: An Emerging Class of Human Genetic Disorders. *Annu. Rev. Genomics Hum. Genet.* **7:** 125-148

Stoetzel C. et al. (2006) Identification of a Novel BBS Gene (BBS12) Highlights the Major Role of a Vertebrate-Specific Branch of Chaperonin-Related Proteins in Bardet-Biedl Syndrome. *Am. J. Hum. Genet.* in press.

Symmetry, Assembly and Evolution of Multimeric Proteins

Supervisors: David Moss (d.moss@mail.cryst.bbk.ac.uk) and Christine Slingsby (c.slingsby@mail.cryst.bbk.ac.uk)

Many protein molecules associate as multimers that often display beautiful **symmetric structures**. The purpose of this project is to investigate the relation of this symmetry to protein stability, shape and mode of assembly. The work will start by analysing the PQS, a database maintained by the European Bioinformatics Institute that contains the atomic co-ordinates of protein multimers. From this database we shall extract a non-redundant set of multimers that will be used in the following investigations. We shall:

1. Classify the multimers in terms of their **point group symmetry** (point groups describe the type of symmetry exhibited)
2. Calculate the number and extent of the **interfaces** between the monomers in every multimer, and where different types of symmetry axis are involved, determine the symmetry relationship at each interface.
3. Assess how the shape of the asymmetric unit or polypeptide chain has an impact on the symmetry of the multimer
4. Investigate the **free energy** of formation of multimers in terms of both their symmetry and their buried surface area

This information will be used to examine **protein stability** from which we may also derive the **evolutionary** (and folding **pathways**) of complex assemblies. This project will appeal to students who are interested in protein structure and who wish to use programming skills to develop some simple software to analyse protein co-ordinates.

In silico **investigation of peptide presentation by HLA alleles in the autoimmune disease, multiple sclerosis**

Supervisors: David Moss (d.moss@mail.cryst.bbk.ac.uk) and Christine Slingsby (c.slingsby@mail.cryst.bbk.ac.uk)

Autoimmune diseases are associated with both genetic and environmental factors. **Multiple sclerosis** (MS) is one such disease that is characterised by progressive impairment of muscular function due to myelin damage in the central nervous system. Genetic susceptibility to MS is associated with particular HLA alleles.

These proteins present peptides to T-cells that are activated to secrete cytokines that lead to chronic inflammation and progressive demyelination of nerve cells. The object of this project is to establish computational protocols for the *in silico* **identification of autoreactive HLA-peptide complexes** that are implicated in MS. The peptides are derived from the protein α B-crystallin that is a myelin-associated protein that is upregulated in MS brains.

We shall use the molecular simulation package **NAMD** to calculate HLA-peptide dissociation constants in order to identify the peptides implicated in the disease. These calculations will use **thermodynamic integration** to yield free energy values and relative binding constants. These are very expensive calculations and will be carried out on our **IBM Blade Cluster** that currently has 67 dual Xeon 3.06GHz nodes interconnected by a 2Gb Ethernet. We are also collaborating with the London e-Science Centre to implement such calculations on a computational GRID.

The project will appeal to a student who is interested in algorithm development in the context of protein-ligand interactions in order to further the use of computational methods in the study of human disease.

References

Davies M N, Sansom, C E, Beazley C & Moss D S, A Novel Predictive Technique for the MHC Class II-peptide binding interaction, *Molecule Medicine*, (2003), **9**, 220-225.

Chou Y K, Burrows G G, LaTocha D, Wang C, Subramanian S, Bourdette, D N & Vandenbark, A A, CD4 T-cell epitopes of human α B-crystallin, *J Neuroscience Res*, (2004), **75**, 516-523.

Computational study of the dissociation of protein dimers

Supervisors: David Moss (d.moss@mail.cryst.bbk.ac.uk) and Adrian Shepherd (a.shepherd@mail.cryst.bbk.ac.uk)

Many X-ray studies show protein molecules with domain-domain interactions in the crystal. Such interactions are often such that **homodimers** are present in the crystal state. The question is whether these are dimers are just crystal artefacts, or are interactions that are significant in solution and hence possibly biologically relevant. An important step towards resolving this question would be the ability to calculate the **dissociation constants** of these putative protein complexes in solution. The object of this project is to use computer simulation to calculate dissociation free energies of protein dimers using the NAMD package and **thermodynamic integration**. The change of dissociation constants on mutating interface residues will be investigated and also the dimer-monomer equilibrium.

The required computations are very expensive and small dimers such as the **insulin dimer** will be chosen for initial studies. The results of the calculations will be compared with experimental dissociation constants. The work will be carried out

on our **IBM Blade Cluster** that currently has 67 dual Xeon 3.06GHz nodes interconnected by a 2Gb Ethernet. The project will appeal to a student who is interested in computer simulation and the molecular modelling of **protein-protein interactions**.