

3D Structure Databases - Uses for Biological Problem solving

The course will teach the basic principles aspects of 3D database technology and the associated tools for data analysis to bioscientists wishing to understand the wealth of structure information available. The course is aimed at PhD students and postdocs to give them a familiarity with how structure data can be used in their own projects.

Databases for 3D structural data for proteins and nucleic acids, together with the associated access tools have matured into a major tool for molecular biology. The course is intended to cover the background to relational databases and the computational aspects of characterizing structure of biological macromolecules

The importance of databases in biological research has been stressed in the recent Nature technology feature by Buckingham [1]. In the United States, the National Science Foundation (NSF) has announced a new initiative, 'Biological Databases and Informatics Program Announcement' [2], with the belief that future advances in the biological sciences will depend both upon the creation of new knowledge and upon effective management of proliferating information. Further general background can be found in references 3 and 4.

1. S. Buckingham Data's future shock (2004) *Nature* **428**, 774-777
2. Biological Databases and Informatics Program Announcement NSF 02-058
<http://www.nsf.gov/pubs/2002/nsf02058/nsf02058.html>
3. Michael Y. Galperin (2004) The Molecular Biology Database Collection: 2004 update. *Nucleic Acids Research*. **32**, Database issue D3-D22
4. Andrej Sali, Robert Glaeser, Thomas Earnest & Wolfgang Baumeister (2003) From words to literature in structural proteomics *NATURE*, **422**, 216-225

Provisional Timetable

Lecturers:

Professor Janet Thornton
Dr Roman Laskowski
Dr Kim Henrick
Dr Phil McNeil
Dr Sameer Velankar
Mr Dimitris Dimitropoulos
Dr Jaime Prilusky
Dr Loredana Lo Conte
Professor Bob Spence
Dr Tom Oldfield
Dr. Philip E. Bourne
Ms. Kyle Burkhardt

Dr Sue Jones
Dr Hannes Ponstingl
Dr Eugene Krissinel
Dr Thomas Oldfield
Mr Adel Golovin
Dr Gerard Kleywegt
Dr Helen Berman
Dr Christine Orengo
Dr James Milner-White
Dr John Westbrook
Dr Robert Finn

Monday 20th September

9:00-9:40 Structure analysis *Professor Janet Thornton (EBI)*

1. Todd A.E, Orengo C.A, Thornton J.M. (2002) Plasticity of enzyme active sites. *Trends Biochem Sci.* **27** 419-26.

2. Steward RE, MacArthur MW, Laskowski RA, Thornton JM. (2003) Molecular basis of inherited diseases: a structural perspective. *Trends Genet.* **19**, 505-13.
3. Sanishvili R, Yakunin AF, Laskowski RA, Skarina T, Evdokimova E, Doherty-Kirby A, Lajoie GA, Thornton JM, Arrowsmith CH, Savchenko A, Joachimiak A, Edwards AM. (2003) Integrating structure, bioinformatics, and enzymology to discover function: BioH, a new carboxylesterase from *Escherichia coli*. *J Biol Chem.* **278**, 26039-45.

9:40-10:20 An overview of the RCSB Protein Data Bank *Dr. Helen M. Berman*
RCSB Protein Data Bank Rutgers, The State University of New Jersey

A description of the resources for data deposition, validation and query offered by the RCSB PDB will be given.

1. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235-242.
2. Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J.D. and Zardecki, C. (2002) The Protein Data Bank. *Acta Crystallogr D* **58**, 899-907.
3. John Westbrook, Zukang Feng, Li Chen, Huanwang Yang and Helen M. Berman (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Research*, **31**, 489-491
4. Bhat,T.N., Bourne,P., Feng,Z., Gilliland,G., Jain,S., Ravichandran,V., Schneider,B., Schneider,K., Thanki,N., Weissig,H. et al. (2001) The PDB data uniformity project. *Nucleic Acids Res.*, **29**, 214-218.
5. Westbrook,J., Feng,Z., Jain,S., Bhat,T.N., Thanki,N., Ravichandran,V., Gilliland,G.L., Bluhm,W., Weissig,H., Greer,D.S. et al. (2002) The Protein Data Bank: unifying the archive. *Nucleic Acids Res.*, **30**, 245-248.

10:20-11:00 Crystals, Symmetry and Protein Assemblies *Dr Kim Henrick (EBI)*

1. Henrick,K. and Thornton,J.M. (1998) PQS: a protein quaternary structure file server. *Trends Biochem. Sci.*, **23**, 358-361.

11:00-11:30 coffee break

11:30-12:10 Protein-DNA Interactions: analysis and prediction *Dr Sue Jones*

The 3D structures of over 700 proteins bound to DNA molecules have been determined. These proteins have diverse structural folds, and achieve binding and recognition of DNA in many different ways. This lecture will give an overview of the prominent characteristics of DNA-binding proteins, and explain how common physicochemical properties and conserved structural motifs can be used in a predictive manner to identify novel DNA-binding proteins.

1. Jones S. & Thornton J.M. (2004) Searching for functional sites in protein structures. *Current Opinion in Chemical Biology.* 8, p3-7.
2. Jones S, Shanahan H, Berman H.M. & Thornton J.M. (2003) Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic Acids Research* 31, p7189-7198.
3. Jones S, Barker J, Nobeli I & Thornton JM. (2003): Using structural motifs to identify proteins with DNA binding function. *Nucleic Acids Research.* 31, p2811-2823.

4. **Jones S.** & Thornton J.M. (2003) Protein-DNA interactions: the story so far and a new method for prediction. *Comparative and Functional Genomics*. 4, p428-431.
5. **Jones S.**, van Heyningen P, Berman HM & Thornton JM. (1999) Protein-DNA interactions: a structural analysis. *Journal of Molecular Biology* 287, p877-896.

Additional Notes on Protein-Protein Interactions: classification, analysis and prediction

Interactions between proteins are fundamental to many diverse biological processes including signal transduction, enzyme inhibition and cell adhesion. These interactions can be classified as 'obligate' or 'non-obligate'. Obligate interactions form the basis of the quaternary structure of multimeric proteins, and non-obligate interactions occur between proteins that exist independently as well as in complexes. This lecture will give an overview of the characteristics of protein-protein complexes from 3D structures, and explain how these features vary dependant upon the class of the complex. A method for the prediction of protein interaction sites will then be described based on the analysis of patches on the protein surface.

- **Jones S** & Thornton JM. (1999) Protein domain interfaces: characterisation and comparison with oligomeric protein interfaces. *Protein Engineering* 13, p77-82.
- **Jones S** & Thornton JM. (1997): Prediction of protein-protein interaction sites using patch analysis. *Journal of Molecular Biology* 272, p133-143.
- **Jones S** & Thornton JM. (1997): Analysis of protein-protein interaction sites using surface patches. *Journal of Molecular Biology* 272, p121-132.
- **Jones S** & Thornton JM. (1996): Principles of protein-protein interactions derived from structural studies. *Proceedings of the National Academy of Science (USA)* 93, p13-20.

12:10-12:50 Using the 'Thornton-Group' WWW Database Services for Structural Research *Dr Roman Laskowski EBI*

1. R.A. Laskowski, J.D. Watson and J.M. Thornton (2003). From Protein structure to biochemical function. *J. Struct. Funct Genomics*, **4**, 167-177.
2. Laskowski RA. (2001) PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res.* **29** 221-2.
3. Luscombe N M, Laskowski R A, Westhead D R, Milburn D, Jones S, Karmirantzou M, Thornton J M (1998). New tools and resources for analysing protein structures and their interactions. *Acta Cryst.*, **D54**, 1132-1138.
4. Craig T. Porter, Gail J. Bartlett, and Janet M. Thornton (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucl. Acids. Res.* **32**: D129-D133.
5. G J Bartlett, C T Porter, N Borkakoti & J M Thornton, *Journal of Molecular Biology* (2002) **324**, 105-121.

12:50-14:00 lunch

14:00-14:40 Quaternary Structure Inference of Proteins from their Crystals *Dr Hannes Ponstingl EBI*

Protein-Protein Interactions: The basic principles which determine the strength and geometry of protein-protein complexes by Patch Analysis and other methods.

1. H. Ponstingl, T. Kabir & J. M. Thornton (2003) Automatic inference of protein quaternary structure from crystals, *J. Appl. Cryst.* **36**, 1116-1122.
2. H. Ponstingl, K. Henrick & J. M. Thornton (2000) Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins* **41**, 47-57.

14:40-15:20 Validation of protein structures ... Or: just because it was published in Nature, doesn't mean it's true ! *Dr. Gerard Kleywegt Uppsala Sweden*

With the explosive growth of the number of experimentally determined macromolecular structures, "structural awareness" is becoming an important aspect of many disciplines, ranging from medicinal chemistry to cell biology. This means that many scientists who have not been specifically trained in the area want to make use of structural information in order to explain the molecular basis of their own research findings, to plan new experiments, to design novel ligands, substrates or inhibitors, etc. What these users of structural information often are unaware of is that there are limitations to and uncertainties in the experimentally determined structures. A number of protein structures have been published (often in prestigious journals) that turned out to be partly or entirely incorrect. A few examples will be given, and simple ways in which non-experts can assess the overall reliability of structures will be discussed. However, as technology improves, such gross errors are less and less likely to occur. On the other hand, mistakes in the details of the structures are much easier to make, and concomitantly more difficult to detect. It is often in these details, however, that the value of a structure lies, since they reveal the molecular basis of interactions. This is particularly true for non-macromolecular entities (ligands, inhibitors, substrate-analogues, sugars, ions, etc.). Some of the pitfalls and limitations of the use of structural information will be discussed, with a view to structure-based design. In the practical, some of the basics of protein structure validation will be reviewed, and the use of various databases (PDB, PDBsum, PDBREPORT and EDS) to assess the quality of deposited protein structures will be explained.

References:

1. G J Kleywegt, "Validation of protein crystal structures" (Topical review), *Acta Crystallographica*, **D56**, 249-265 (2000).
2. AM Davis, SJ Teague & GJ Kleywegt, "Applications and limitations of X-ray crystallographic data in structure-based ligand and drug design", *Angewandte Chemie International Edition*, **42**, 2718-2736 (2003)
3. EDS Viewer for structures and electron density maps – interpretation of validation criteria

15:20-16:00 3D databases and data warehouse technology *Dr Phil McNeil (EBI)*

16:00-16:30 coffee break

16:30-17:10 Clustering of 3D structures and representative sets *Dr Thomas Oldfield EBI*

1. Li,W., Jaroszewski,L. and Godzik,A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282-283.

17:10-17:50 Sequences and 3D structures (Integration of 3D data and sequence databases) *Dr Sameer Velankar EBI*

The "Structure integration with function, taxonomy and sequence (SIFTS) initiative" aims to work towards the integration of various bioinformatics resources. One of the

major obstacles to the improved integration of structural databases such as MSD <<http://www.ebi.ac.uk/msd/>> and sequence databases like UniProt <<http://www.ebi.uniprot.org/index.shtml>>, which are primary archival databases for structure and sequence data, is the absence of up to date and well-maintained mapping between corresponding entries. We have worked closely with the UniProt group at the EBI to clean up the taxonomy and sequence cross-reference information in the MSD and UniProt databases. The project was started in the year 2001 and has resulted in creating a robust mechanisms for exchanging data between the two primary data resources. This has dramatically improved the quality of annotation in both databases and is aiding the continuing improvements of legacy data. In the longer term this project will allow for not only the better and closer integration of derived-data resources but will continue to improve the quality of all data in the primary resources. This information is vital for the reliable integration of the sequence family databases such as Pfam <<http://www.sanger.ac.uk/Software/Pfam/>> and Interpro <<http://www.ebi.ac.uk/interpro/>> with the structure-oriented databases of SCOP <<http://scop.mrc-lmb.cam.ac.uk/scop/>> and CATH <<http://www.biochem.ucl.ac.uk/bsm/cath/>>. This information has been made available to the eFamily group <<http://www.efamily.org.uk/>> and now forms the basis of the regular interchange of information between the member databases (MSD, Uniprot, Pfam, Interpro, SCOP and CATH).

Tuesday 21st September

Participants split into 2 groups.

Morning session Group 1

Tutorials (15 min intro/demo – 45 minute tutorial)

- **Searching 3D protein structures for conserved DNA-binding motifs:** *Dr Sue Jones*
- **Using the ‘Thornton-Group’ WWW Database Services for Structural Research** *Dr Roman Laskowski EBI (2hours)*
CATRES, Catalytic Site Atlas (CSA), NetFunc, EC->PDB, Pita - Protein InTerfaces and Assemblies, Receptor Structure and Function, Protein Side-Chain Interactions, Practical: Structural Genomics
- **Pfam and MEROPS** *Robert Finn Sanger*

Morning session Group 2

Lectures (40 minutes – intro/demos)

- **Protein Structure Classification and Genome Annotation: Technologies and Insights from the CATH Database.** *Professor Christine Orengo UCL London UK*
- **SCOP: Structural Classification of Proteins.** *Dr Loredana Lo Conte (LMB Cambridge UK)*
- **SSM fold characterization** *Dr Eugene Krissinel EBI*

Afternoon session Group 1

Lectures (40 minutes - intro/demos)

- **Protein Structure Classification and Genome Annotation: Technologies and Insights from the CATH Database.** *Professor Christine Orengo UCL London UK*
- **SCOP: Structural Classification of Proteins.** *Dr Loredana Lo Conte (LMB Cambridge UK)*
- **SSM fold characterization** *Dr Eugene Krissinel EBI*

Afternoon session Group 2

Tutorials (15 min intro/demo – 45 minute tutorial)

- **Searching 3D protein structures for conserved DNA-binding motifs:** *Dr Sue Jones*
- **Using the ‘Thornton-Group’ WWW Database Services for Structural Research** *Dr Roman Laskowski EBI (2hours)*
CATRES, Catalytic Site Atlas (CSA), NetFunc, EC->PDB, Pita - Protein InTerfaces and Assemblies, Receptor Structure and Function, Protein Side-Chain Interactions, Practical: Structural Genomics
- **Pfam and MEROPS** *Robert Finn Sanger*

[**Motif and protein structures** *Dr. Gerard Kleywegt Uppsala Sweden*]

Protein Structure Classification and Genome Annotation: Technologies and Insights from the CATH Database. *Professor Christine Orengo UCL London UK*

CATH is a novel hierarchical classification of protein domain structures, which clusters proteins at four major levels, Class(C), Architecture(A), Topology(T) and Homologous superfamily (H). Class, derived from secondary structure content, is

assigned for more than 90% of protein structures automatically. Architecture, which describes the gross orientation of secondary structures, independent of connectivities, is currently assigned manually. The topology level clusters structures according to their topological connections and numbers of secondary structures. The homologous superfamilies cluster proteins with highly similar structures and functions. The assignments of structures to topology families and homologous superfamilies are made by sequence and structure comparisons.

I will illustrate concepts behind protein structure comparison and classification using the CATH database. I will also present methods for providing structural annotations for genome sequences.

1. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. and Thornton, J. M. (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093-1108.
2. Pearl, F.M.G, Lee, D., Bray, J.E, Sillitoe, I., Todd, A.E., Harrison, A.P., Thornton, J.M. and Orengo, C.A. (2000) *Assigning genomic sequences to CATH* Nucleic Acids Research. **28**. 277-282

SCOP: Structural Classification of Proteins. Dr Loredana Lo Conte (LMB Cambridge UK)

Nearly all proteins have structural similarities with other proteins and, in some of these cases, share a common evolutionary origin. The SCOP database, created by manual inspection and abetted by a battery of automated methods, aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known. As such, it provides a broad survey of all known protein folds, detailed information about the close relatives of any particular protein, and a framework for future research and classification.

I will take you behind the scenes. All we learned so far. What is missing.

1. Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536-540.
2. Andreeva A., Howorth D., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucl. Acid Res.* **32**, D226-D229.
3. Lo Conte L., Brenner S. E., Hubbard T.J.P., Chothia C., Murzin A. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acid Res.* **30**, 264-267.

Pfam and MEROPS Robert Finn Sanger

Pfam is a database of two parts, the first is the curated part of Pfam containing over 7459 protein families. To give Pfam a more comprehensive coverage of known proteins we automatically generate a supplement called Pfam-B. This contains a large number of small families taken from the PRODOM database that do not overlap with Pfam-A. Although of lower quality Pfam-B families can be useful when no Pfam-A families are found.

1. The Pfam Protein Families Database Alex Bateman, Lachlan Coin, Richard Durbin, Robert D. Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon, Erik L. L. Sonnhammer, David J. Stud holme, Corin Yeats and Sean R. Eddy (2004) *Nucleic Acids Research Database Issue* **32**, D138-D141

2. Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K, Bairoch A. (2002). The PROSITE database, its status in 2002 *Nucleic Acids Res.* **30**:235-238
3. Julie D. Thompson, Frédéric Plewniak, Raymond Ripp, Jean-Claude Thierry and Olivier Poch (2001) Towards a Reliable Objective Function for Multiple Sequence Alignments. *J.Mol.Biol.* **314**, 937-951
4. Servant F, Bru C, Carrère S, Courcelle E, Gouzy J, Peyruc D, Kahn D (2002) ProDom: Automated clustering of homologous domains. *Briefings in Bioinformatics.* **3**, 246-251

SSM fold characterization *Dr Eugene Krissinel EBI*

SSM is a powerful interactive research tool for secondary structure matching that allows for comparing protein structures in 3D. The service provides for (i) pairwise comparison and 3D alignment of protein structures, (ii) multiple comparison and 3D alignment of protein structures, (iii) examination of a protein structure for similarity with the whole PDB or SCOP archives, (iv) best Ca-alignment of compared structures and (v) the ability to download and visualization of best-superposed structures using RasMol or RasTop. The results are linked to other services including OCA, SCOP, GeneCensus, FSSP, 3Dee, CATH, PDBsum, SwissProt and ProtoMap. SSM is recognized as a valuable tool in protein research, an aid to study protein function via structural similarities (used in drug design), protein conformations, choice of model for structure solution in X-ray experiments and many others. The development has also extensive integration with other MSD services.

1. E. Krissinel and K. Henrick, Protein structure comparison in 3D based on secondary structure matching (SSM) followed by Ca alignment, scored by a new structural similarity function. In: Andreas J. Kungl & Penelope J. Kungl (Eds.), *Proceedings of the 5th International Conference on Molecular Structural Biology*, Vienna, September 3-7, 2003, p.88.
2. E. Krissinel and K. Henrick, Common subgraph isomorphism detection by backtracking search. (2004) *Software: Practice and Experience*, **34**, 591-607.

Motif and protein structures *Dr. Gerard Kleywegt Uppsala Sweden*

Motif recognition (essentially, function-from-structure)
 SPASM server Motif recognition in nucleic acids structures (SPANNA)
 Motif recognition in proteins (SPASM)

1. Madsen, D. and Kleywegt, G.J. (2002). Interactive motif and fold recognition in protein structures. *J. Appl. Cryst.* **35**, 137-139.

WEDS 22nd September

Participants split into 2 groups.

Morning session Group 1

Tutorials (15 min intro/demo –45 minute tutorial)

- **Validation of protein structures ... Or: just because it was published in Nature, doesn't mean it's true !** *Dr. Gerard Kleywegt Uppsala Sweden (90min)*
- **Learning about structures using the RCSB PDB** *Dr. Philip E. Bourne Professor of Pharmacology University of California (60min)*
- **Evaluation of structure quality tutorial, using RCSB tools** *Ms. Kyle Burkhardt RCSB Protein Data Bank Rutgers, The State University of New Jersey (60min)*

Morning session Group 2

Lectures (7x30mins - intro/demo MSD tools)

- **MSDchem, MSDlite, MSDpro, MSDsite**
- **MSDmySQL , MSDmine**
- **Advanced AstexViewer integrated into 3D PDB searches**
Generic search systems for the search database has been written and made a public service as <http://www.ebi.ac.uk/msd-srv/msdlite> and <http://www.ebi.ac.uk/msd-srv/msdpro>. The system is written using java servlets and uses XML extensively for configuration of the database/search system interactions, the description of the user interface, and the return of results. The system is designed to translate user input from a series of values into an SQL query, which can then be executed on the database. The architecture of the server-side of the search system allows a high degree of flexibility and extending the range of searches SQL statements that can be created automatically). The system is highly configurable and can be moved easily to other databases simply by modifying XML dictionaries, which describe the database.

Afternoon session Group 1

Lectures (7x30mins - intro/demo MSD tools)

- **MSDchem, MSDlite, MSDpro, MSDsite**
- **MSDmySQL , MSDmine**
- **Advanced AstexViewer integrated into 3D PDB searches**

Afternoon session Group 2

Tutorials

- **Validation of protein structures ... Or: just because it was published in Nature, doesn't mean it's true !** *Dr. Gerard Kleywegt Uppsala Sweden (90min)*
- **Learning about structures using the RCSB PDB** *Dr. Philip E. Bourne Professor of Pharmacology University of California (60min)*
- **Evaluation of structure quality tutorial, using RCSB tools** *Ms. Kyle Burkhardt RCSB Protein Data Bank Rutgers, The State University of New Jersey (60min)*

Learning about structures using the RCSB PDB *Dr. Philip E. Bourne Professor of Pharmacology University of California,*

The RCSB PDB has been reengineered to include many new features and to integrate a variety of additional information related to macromolecular structure and function ranging from genomic information to disease states. A number of these features will be explored through several typical usage scenarios suited to novice users and more senior biologists.

1. Philip E. Bourne, Kenneth J. Address, Wolfgang F. Bluhm, Li Chen, Nita Deshpande, Zukang Feng, Ward Fleri, Rachel Green, Jeffrey C. Merino-Ott, Wayne Townsend-Merino, Helge Weissig, John Westbrook and Helen M. Berman (2004). The distribution and query systems of the RCSB Protein Data Bank *Nucleic Acids Research* **32**, Database issue D223-D225

Evaluation of structure quality tutorial, using RCSB tools *Ms. Kyle Burkhardt* *RCSB Protein Data Bank Rutgers, The State University of New Jersey*

In this one hour tutorial, users will learn how to evaluate the quality of a structure. Users will download a structure from the PDB and validate the structure using the RCSB developed online Validation Suite. Users will learn how to analyze the validation report as well as PROCHECK, NUCHECK, and SFCHECK results to determine structure quality.

MSDsite Service *Mr Adel Golovin EBI*

The research service, MSDsite has been developed to give access to 3D active site data. The three-dimensional environments of ligand binding sites have been derived from the parsing and loading of the PDB entries into a relational database. For each bound molecule the biological assembly of the quaternary structure has been used to determine all contact residues and a fast interactive search and retrieval system has been developed. Prosite pattern and short sequence search options are available together with a novel graphical query generator for inter-residue contacts.

Dimitris Dimitropoulos

MSDchem tutorial: Chemistry as the starting point of a search. Following the path from ligand chemistry to protein structure. This tutorial demonstrates in detail the searching capabilities of the MSD database and the MSDchem tool in identifying ligands using their basic chemical topology and chemical signature.

MSDmysql tutorial: Working with the bare MSD database in the popular mysql form using directly general purpose standard API's and programming languages. A way to use the MSD database infrastructure in the most flexible and powerfull way.

MSDmine tutorial: A web application for scientific discovery, data analysis and knowledge mining for the advanced researcher of the MSD database. From the simplest to most complex searches, that combine many different information entities, together with visualisation and cross-references. Online generation of charts and data drill and roll-up operations.

1. Golovin, T. J. Oldfield, J. G. Tate, S. Velankar, G. J. Barton, H. Boutselakis, D. Dimitropoulos, J. Fillon, A. Hussain, J. M. C. Ionides, M. John, P. A. Keller, E. Krissinel, P. McNeil, A. Naim, R. Newman, A. Pajon, J. Pineda, A. Rachedi, J. Copeland, A. Sitnov, S. Sobhany, A. Suarez-Uruena, G. J. Swaminathan, M. Tagari,

S. Tromm, W. Vranken and K. Henrick (2004) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Research*, **32**, Database issue D211-D216

2. [4] Boutselakis, H., Dimitropoulos, D., Fillon, J., Golovin, A., Henrick, K., Hussain, A., Ionides, J., John, M., Keller, P.A., Krissinel, E., McNeil, P., Naim, A., Newman, R., Oldfield, T., Pineda, J., Rachedi, A., Copeland, J., Sitnov, A., Sobhany, S., Suarez-Uruena, A., Swaminathan, J., Tagari, M., Tate, J., Tromm, S., Velankar, S. and Vranken, W. (2003) E-MSD: the European Bioinformatics Institute Macromolecular Structure Database *Nucleic Acids Research*, **31**, 458-462

Thursday

Morning session Group 1

- Tutorials – Using MSD Tools to solve Problems (60min)
- Tutorials Database Replication and use on your Desktop (60min)
MSDmysql and MSDmine
- Tutorials – Scripting languages + data mining *Dr Jaime Prilusky
Weizmann Israel (90min)*

Morning session Group 2

Lectures and Demos

- Visualisation data mining *Dr Thomas Oldfield EBI*
- What is visualisation and how are complex data Represented?
Professor Bob Spence (Imperial College London)
- Small Structural and Sequence Motifs Dr James Milner-White
University of Glasgow

Afternoon session Group 1

Lectures and Demos

- Visualisation data mining *Dr Thomas Oldfield EBI*
- What is visualisation and how are complex data Represented?
Professor Bob Spence (Imperial College London)
- Small Structural and Sequence Motifs Dr James Milner-White
University of Glasgow

Afternoon session Group 2

- Tutorials – Using MSD Tools to solve Problems (60min)
- Tutorials Database Replication and use on your Desktop (60min)
MSDmysql & MSDmine
- Tutorials – Scripting languages + data mining *Dr Jaime Prilusky
Weizmann Israel (90min)*

Scripting languages + data mining *Dr Jaime Prilusky Weizmann Israel*

Use of fast scripting languages (Perl, Python) for creating ad-hoc searching and analysis tools on top of existing databases

Visualisation

1. Tate, J.G., Moreland, J.L. and Bourne, P.E. (2001) Design and implementation of a collaborative molecular graphics environment. *J. Mol. Graph. Model.*, **19**, 280-287, 369-273.
2. Neshich, G., Togawa, R.C., Mancini, A.L., Kuser, P.R., Yamagishi, M.E., Pappas, G., Jr, Torres, W.V., Fonseca e Campos, T., Ferreira, L.L., Luna, F.M. et al. (2003) STING Millennium: a web-based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence. *Nucleic Acids Res.*, **31**, 3386-3392.
3. Bob Spence The Acquisition of Insight
<http://www.ee.ic.ac.uk/research/information/www/Bobs.html>

4. Watson, J. D. and Milner-White, E. J. 2002 The conformations of Polypeptide Chains where the main-chain parts of successive residues are enantiomeric. Their occurrence in Cation and Anion-binding regions of proteins. *Journal of Molecular Biology* **315**, 183-191
5. Watson, J. D. and Milner-White, E. J. 2002 A novel main-chain anion-binding site in proteins: The Nest. A particular combination of phi,psi values in successive residues gives rise to anion-binding sites that occur commonly and are found often at functionally important regions. *Journal of Molecular Biology* **315**, 171-182
6. Wan, W. Y. and Milner-White, E. J. 1999 A natural grouping of motifs with an aspartate or asparagine residue forming two hydrogen bonds to residues ahead in sequence: Their occurrence at alpha-helical N termini and in other situations. *Journal of Molecular Biology* **286**, 1633-1649
7. Furnas, G.W. Generalised fisheye views. (1986) Human Factors in Computing Systems. *CHI'86 Conference Proceedings*, Boston, April 13-17 pp 16-23.
8. George G. Robertson, Jock D. Mackinlay and Stuart K. Card. (1991) The Perspective wall: Detail and context smoothly integrated *CHI'91 Conference Proceedings*, pp 174-179.
9. Oldfield, T.J. Creating structure features by datamining the PDB to use as molecular replacement models. *Acta Cryst* **D57** 1421-1427.

Friday 24th September

9:00-9:40 Summary of 3D data and WWW Current scientific resources for 3D structure (Dr Kim Henrick EBI)

- **Clean data**
- **Data base architectures**

9:40-10:20 Now you have seen the services – a word about Conceptual basis for data analysis, problem solving and critical thinking – (Dr Tom Oldfield EBI)

10:20-11:00 Data Interchange/API's Data Integration problems / Query Interchange – Dr John Westbrook (RCSB)

11:00-11:30 coffee break

11:30-12:10 FeedBack – Wrapup – Chair Dr Helen Berman (RCSB)

12:10 Lunch and Depart