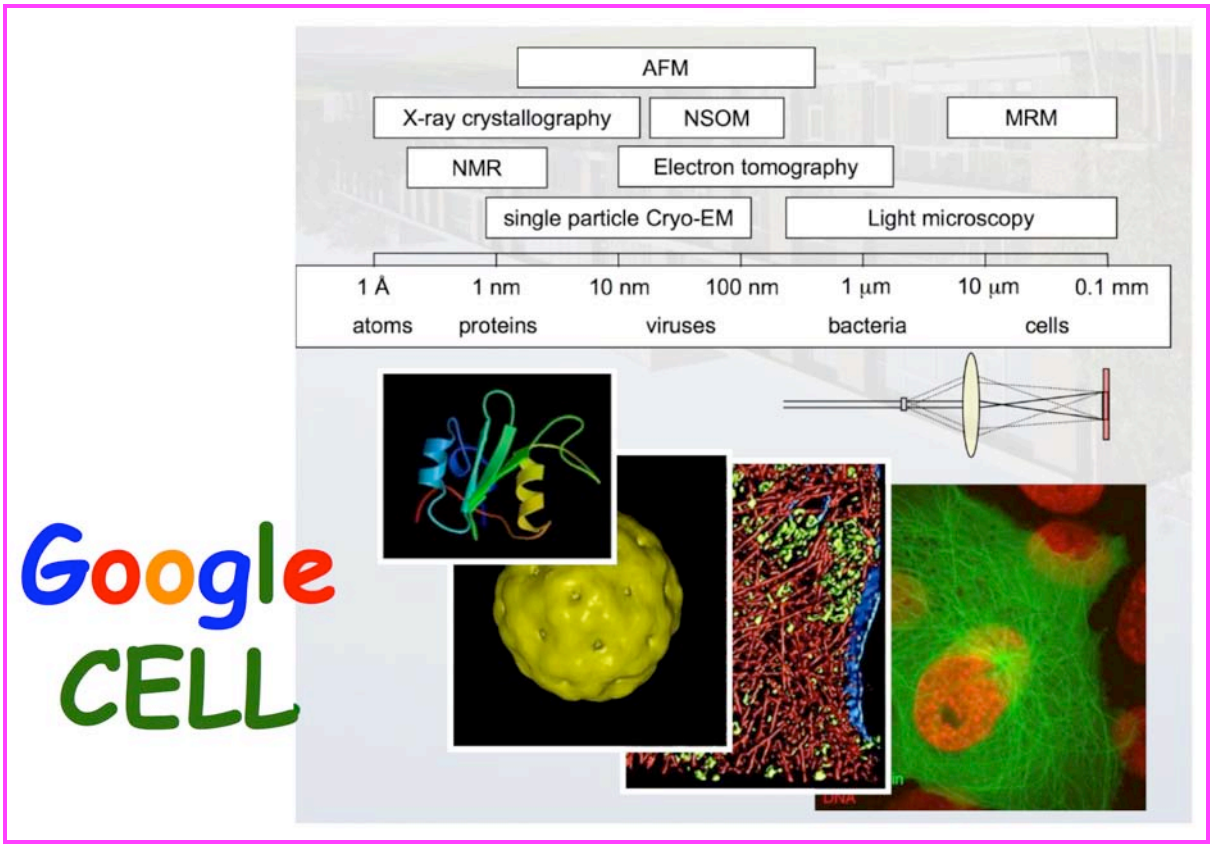
Forum for European Structural Proteomics

Assessment of Requirements for Bioinformatics in Structural Genomics and Structural Proteomics



A Roadmap for Strategic Future of European Directions in Structural Genomics and Structural Proteomics

Cutting edge bioinformatic tools, together with reliable, well maintained and state-of-the-art databases, are essential to the SP/SG endeavour. Fortunately, such capabilities are largely generated and maintained by the EBI, which is an outstation of EMBL, by the Swiss Institute for Bioinformatics, the Wellcome Trust Sanger Institute and numerous bioinformatics resources within universities and research institutes throughout Europe. Many of these bodies have now been brought together under the umbrella of ELIXIR, the European Life Sciences Infrastructure for Biological Information (http://www.elixir-europe.org) which will be funded in the framework of the ESFRI roadmap for biomedical sciences, and will have a mandate to ensure **long-term** funding for the EBI, which until now has not been the case, as well as integrating its capabilities with those of bioinformatics bodies across Europe. It is thus anticipated that many of the requirements of the SP/SG endeavour, and of INSTRUCT in particular, as well as of individual structural biology groups, will be taken care of by ELIXIR. In the following, therefore, we will only address briefly, specific aspects of the SG/SP endeavour which may need to be prioritized within the framework of INSTRUCT in particular, and of the SG/SP consortia in general, or coordinated with ELIXIR and other bioinformatics bodies.

It may be anticipated that the sizes of the various databases will increase dramatically. Consequently, efforts will have to be invested in ensuring that interplay between them will be transparent to all users, not just to bioinformatics. The maintenance of large databases is also an enormous and extremely complex task. This can be seen in the case of the DNA sequence database, which is maintained as a very successful 3-way collaboration, via an international agreement, between EBI, Genbank (US) and DDBJ (Japan). To quote from the ELIXIR proposal (http://www.elixir-europe.org):

> The size of this database has grown exponentially since its inception. From the 1980's to mid 1990's, the database doubled in size every 23 months, and from the mid 1990's the rate of growth had increased such that the database doubled every 16 months. Recently a sister database has been established for unassembled DNA sequence (The Trace Repository) to handle the growth in data collected in this form, also as an international collaboration with the US. This database already contains 1 billion records and is doubling every 11 months. At 22 Terabytes the database is already one of the largest single scientific database in the world. It can be predicted with confidence that, within a decade, bioinformatics requirements will be comparable with the computational requirements of the physical sciences. New technologies are likely to drop sequencing costs by several orders of magnitude in the next decade. This will drive the collection of sequence data on a new scale as it will become cheap enough to be used as a routine screen in human genetics research. While it is anticipated that in future these databases will grow even faster, the current doubling rate of every 11 months is already faster than the 'Moore's law' growth of both computer disk storage and CPU speeds. The effect of such high growth rates is that computer infrastructure to support such resources must grow physically even if it is being upgraded continuously. Although support for the EMBL database has been sufficient to keep up with storing the raw data

submitted to it, it has not been sufficient to keep up with the work required to optimally organize, annotate or provide services based on this data. The setting up of a European Trace Repository for raw sequence data was only possible at all through the cost being underwritten by the Wellcome Trust at the Sanger Institute. By contrast the US partner of these international collaborative data resources, NCBI (National Centre for Biotechnology Information), has been substantially better funded (appearing as a line item in the congressional budget) and has since the 1980's dominated world bioinformatics data services, with well supported services such as Blast, Entrez and PubMed.

One may anticipate a significant expansion of the current electron microscopy (EM) database in Europe, i.e. EMDB (http://www.ebi.ac.uk/msd-srv/docs/emdb). EMDB is the first outcome of collaboration between the European Network of Excellence, 3D-EM (http://www.3dem-noe.org), and the NIH-funded partnership for a Unified Data Resource for CryoEM (http://emdatabank.org) recently established between the EBI, the Research Collaboratory for Structural Bioinformatics (RCSB) at Rutgers, and the National Center for Macromolecular Imaging at Baylor College of Medicine.

Computational biology will be a key element in integrating the multi-layered structural information acquired, ranging from the detailed atomic level information acquired by crystallography and NMR through SAXS and neutron scattering through to electron and light microscopy. Specifically, it will be necessary to significantly improve tools for processing images obtained by the various microscopic techniques, as well as to emphasize the kind of tools being developed by, for example, Andrew Sali (Alber et al. 2007a; Alber et al. 2007b; Robinson et al. 2007), for fitting individual proteins and protein complexes into larger structures visualized by EM tomography. There will also be a requirement to develop tools that will integrate structural and dynamic aspects, so as to achieve the vision of Dino Moras recently proclaimed at the FESP Workshop at the SPINE Annual Conference in Montecatini, Italy, Aug 2005 'watching molecules dancing in the cell'. Such integration will be crucial to implementing the recommendations of FESP concerning a trajectory in which SG/SP will approach the solution of increasingly complex structures. INSTRUCT should provide a framework in which both the core centres and associated groups will work to towards this objective.

An important area of structural biology, which should not be neglected is homology modeling. This is an area that has benefited enormously from the large number of new protein families and folds whose 3D structures have been determined in the structural genomics initiatives world-wide. But it is not, perhaps, sufficiently appreciated that there are already a few cases in which this technique has proved of utility both in drug design and in engineering proteins with modified specificity or novel functions (Okumoto et al. 2005; Becker et al. 2006; Dooley et al. 2006; Lager et al. 2006; Röthlisberger et al. 2008). Obviously, the larger the database at the disposal of the scientist, the better will be the quality of the homology models generated, whether of native proteins, engineered proteins or of drug-protein complexes. Thus, although protein structures arising out of structural genomics projects have not yet led to a drug in clinical use, the situation might well change in the not to distant future. Indeed, the quality of the homology models being generated is improving continuously (Moult 2008), and even the so-called 'boring' proteins and 'low-hanging fruit', such as those generated in the initial stages of the PSI and in Japan, can significantly

strengthen the data basis at the disposal of the homology modeler. Although we do not recommend additional funding projects just to increase the number of folds or structures, as this is happening serendipitously, we would strongly recommend funding for carefully chosen projects involved in development of new algorithms for protein prediction or design of new protein functions.

While the number of sequenced genomes continues to grow, experimentally verified functional annotation of whole genomes remains patchy. SG projects are yielding many protein structures of unknown function. Experimental investigation of the function of such proteins is costly and time-consuming. Consequently, effort should be invested in improving computational methods for prediction of protein function. An increasing number of methods for predicting protein function from sequence or on the basis of structural data are becoming available, but biologists making use of such tools should take care to be aware of their strengths and weaknesses (Lee et al. 2007).

Currently, the bulk of function assignments for newly sequenced genomes are performed by copying annotations from similar protein sequences. It is now well recognized that these procedures can give rise to annotation errors, and more seriously, to chains of mis-annotation (Gilks et al. 2005). It is thus desirable that such procedures are progressively replaced by less error-prone methods (Brown et al. 2007) and that present annotations are subject to a process of remediation.

One possible strategy to make structural data more readily accessible to the scientific community can involve the creation of databases (possibly in the Proteopedia format) dedicated to specific subsets of the structural biological knowledge, such as membrane proteins or metalloproteins. Although this strategy must be limited to avoid excessive proliferation and useless fragmentation of information, scientists may better retrieve the desired data by recurring to resources designed for their particular area of research, especially since (as mentioned above) the size of all-inclusive databases is continuously increasing.

At a more mundane level, it will be necessary to invest in databases collating such information as purification protocols, expression vectors and gene constructs, so as to streamline the protein production pipeline and eliminate redundancy, see FESP document on "Research Infrastructures in Structural Proteomics: Assessment of Needs for Protein Production in Structural Genomics / Structural Proteomics Projects" (http://www.ec-fesp.org/FESP/report.html). Implementation of a standardized trans-European LIMS system would be a worthy objective, since it could greatly facilitate transfer and exchange of information.

One of the most severe criticisms made in a recent assessment of the NIGMS PSI (http://www.nigms.nih.gov/News/Reports/PSIAssessmentPanel2007.htm) was that dissemination of the data generated by the PSI had been poor. "PSI efforts to facilitate the use of structures and materials by the broad scientific community have remained *ad hoc* and low throughput while the structural pipeline has moved successfully to high throughput. The use of structural information is best driven by scientists who wish to understand biological

mechanisms, therefore deposition of structures in the PDB does not reach the key audience. This failing is critical because the long-term value of PSI-generated structures will depend on the value of the information being generated." This criticism also holds true for data generated by structural biologists in general, and attempts to rectify this situation, *viz.* to make structural data readily accessible to biologists other than structural biologists, should be given a very high priority. Some promising steps have already been made in this direction, including the developments of:

- Kinemage (Richardson and Richardson 1992)
- ProteinExplorer (http://www.umass.edu/microbio/chime/pe_beta/pe/protexpl)
  FirstGlance (http://molvis.sdsc.edu/fgij)
- iSee (Abagyan et al. 2006)
- PDBSUM (Laskowski 2007)
- TOPSAN (http://www.topsan.org/TOPSAN)
- Preparing Enhanced Figures in IUCr Journals (Einspahr and Guss 2008)
- Proteopedia (http://www.proteopedia.org)

References:

Abagyan, R., Lee, W.H., Raush, E., Budagyan, L., Totrov, M., Sundstrom, M., and Marsden, B.D. 2006. Disseminating structural genomics data to the public: from a data dump to an animated story. *TIBS* **31:** 76-78.

Alber, F., Dokudovskaya, S., Veenhoff, L.M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B.T., Rout, M.P., and Sali, A. 2007a. Determining the architectures of macromolecular assemblies. *Nature* **450:** 683-694.

Alber, F., Dokudovskaya, S., Veenhoff, L.M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B.T., Sali, A., and Rout, M.P. 2007b. The molecular architecture of the nuclear pore complex. *Nature* **450:** 695-701.

Becker, O.M., Dhanoa, D.S., Marantz, Y., Chen, D., Shacham, S., Cheruku, S., Heifetz, A., Mohanty, P., Fichman, M., Sharadendu, A., Nudelman, R., Kauffman, M., and Noiman, S. 2006. An integrated in silico 3D model-driven discovery of a novel, potent, and selective amidosulfonamide 5-HT1A agonist (PRX-00023) for the treatment of anxiety and depression. *J. Med. Chem.* **49:** 3116-3135.

Brown, D.P., Krishnamurthy, N., and Sjolander, K. 2007. Automated protein subfamily identification and classification. *PLoS Comput. Biol.* **3:** e160.

Dooley, A.J., Shindo, N., Taggart, B., Park, J.G., and Pang, Y.P. 2006. From genome to drug lead: identification of a small-molecule inhibitor of the SARS virus. *Bioorg. Med. Chem. Lett.* **16:** 830-833.

Einspahr, H., and Guss, M. 2008. A new service for preparing enhanced figures in IUCr journals. *Acta Crystallographica Section F* **64:** 154-155.

Gilks, W.R., Audit, B., de Angelis, D., Tsoka, S., and Ouzounis, C.A. 2005. Percolation of annotation errors through hierarchically structured protein sequence databases. *Math. Biosci.* **193:** 223-234.

Lager, I., Looger, L.L., Hilpert, M., Lalonde, S., and Frommer, W.B. 2006. Conversion of a putative Agrobacterium sugar-binding protein into a FRET sensor with high selectivity for sucrose. *J. Biol. Chem.* **281:** 30875-30883.

Laskowski, R.A. 2007. Enhancing the functional annotation of PDB structures in PDBsum using key figures extracted from the literature. *Bioinformatics* **23:** 1824-1827.

Lee, D., Redfern, O., and Orengo, C. 2007. Predicting protein function from sequence and structure. *Nat. Rev. Mol. Cell. Biol.* **8:** 995-1005.

Moult, J. 2008. Comparative Modeling in Structural Genomics. In *Structural Proteomics and its Impact on the Life Sciences*. (eds. J.L. Sussman, and I. Silman), pp. (in press). World Scientific Publishing Company.

Okumoto, S., Looger, L.L., Micheva, K.D., Reimer, R.J., Smith, S.J., and Frommer, W.B. 2005. Detection of glutamate release from neurons by genetically encoded surface-displayed FRET nanosensors. *Proc. Natl. Acad. Sci. USA* **102:** 8740-8745.

Richardson, D.C., and Richardson, J.S. 1992. The kinemage: A tool for scientific communication. *Protein Sci.* **1:** 3-9.

Robinson, C.V., Sali, A., and Baumeister, W. 2007. The molecular sociology of the cell. *Nature* **450:** 973-982.

Röthlisberger, D., Khersonsky, O., Wollacott, A.M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J.L., Althoff, E.A., Zanghellini, A., Dym, O., Albeck, S., Houk, K.N., Tawfik, D.S., and Baker, D. 2008. Novel Kemp elimination catalysts by computational enzyme design. *Nature***:** (in press).