Instituto Juan March
de Estudios e Investigaciones

121 CENTRO DE REUNIONES
INTERNACIONALES SOBRE BIOLOGÍA

Workshop on

Structural Genomics and Bioinformatics

Organized by

B. Honig, B. Rost and A. Valencia

C. M. Dobson
T. Gaasterland
L. Holm
B. Honig
D. T. Jones
M. Kanehisa
C. D. Lima
M. Linial
A. McDermott
A. G. Murzin
C. Orengo
M. Orozco
M. C. Peitsch
B. Rost
A. Sali
C. Sander
M. J. Sippl
J. M. Thornton
A. Valencia

Instituto Juan March
de Estudios e Investigaciones

121 CENTRO DE REUNIONES
INTERNACIONALES SOBRE BIOLOGÍA

Workshop on

Structural Genomics and Bioinformatics

Organized by

B. Honig, B. Rost and A. Valencia

C. M. Dobson
T. Gaasterland
L. Holm
B. Honig
D. T. Jones
M. Kanehisa
C. D. Lima
M. Linial
A. McDermott

A. G. Murzin
C. Orengo
M. Orozco
M. C. Peitsch
B. Rost
A. Sali
C. Sander
M. J. Sippl
J. M. Thornton
A. Valencia

The lectures summarized in this publication were presented by their authors at a workshop held on the 12th through the 14th of March, 2001, at the Instituto Juan March.

Depósito legal: M-18.217/2001
Impresión: Ediciones Peninsular. Tomelloso, 27. 28026 Madrid.

INDEX

PAGE

INTRODUCTION: Alfonso Valencia 7

Session 1
Chair: Janet M. Thornton 11

Chris Sander: 16,000 targets for structural genomics 13

Christopher M. Dobson: Protein folding, evolution and disease 14

Michal Linial: Clustering the universe of protein sequences in light of structural and functional genomics 15

David T. Jones: Recognizing old and new folds in genomes 16

Short talk:
Seán I. O'Donoghue: Bioinformatics tools for structural genomics 17

Session 2
Chair: Manuel C. Peitsch 19

Christopher D. Lima: Approaching function through structure 21

Terry Gaasterland: Homology-based annotation of the *Drosophila melanogaster* genome 22

Alexey G. Murzin: Finding biologically important similarities between protein structures 23

Manfred J. Sippl: The role of protein structure in genomics 24

Session 3
Chair: Manfred J. Sippl 25

Minoru Kanehisa: Prediction of protein-protein interaction networks from genome information 27

Christine Orengo: What do structure classifications reveal about the universe of protein folds and protein evolution 28

Burkhard Rost: Do we aim at getting one structure per fold? 29

Andrej Sali: Comparative protein structure modeling of genes and genomes 30

PAGE

Short talk:

Federico Gago: From sequence to structure to function: the importance of overlapping binding sites for zinc finger-containing proteins on transcription regulation and possible modulation by anticancer ecteinascidins..... 31

Session 4
Chair: Chris Sander..... 33

Janet M. Thornton: From structure to function: functional diversity within protein superfamilies. 35

Liisa Holm: Classifying the universe of protein structures. 36

Alfonso Valencia: Bioinformatics approaches to the detection of protein interaction networks..... 38

Manuel C. Peitsch: The accuracy and use of protein models derived from comparative methods..... 40

Session 5
Chair: Minoru Kanehisa..... 43

Ann McDermott: Mechanistic structural and dynamics studies of proteins through solid state NMR. 45

Modesto Orozco: New challenges in drug design..... 46

Short talk:

José M. Carazo: The role of three-dimensional cryo-electron microscopy in structural genomics..... 47

Barry Honig: Combining bioinformatics biophysics to understand protein function..... 49

POSTERS..... 51

Federico Abascal: Clustering in sequence space. Identification of orthologous families. 53

Emil Alexov: pH dependence of the folding free energy – implication to the function of the proteins..... 54

Patrick Aloy: Automated structure-based prediction of functional sites in proteins - applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking..... 55

Gordana Apic: Domain combinations in archaeal, eubacterial and eukaryotic proteomes..... 56

PAGE

Francisco J. Blanco: NMR structural analysis of hypotetical proteins of mth0677 and mth0121 proteins from *Methanobacterium thermoautotrophicum*..... 57

Rosana Chehín: Structural predictions of *Escherichia coli* NDH-2..... 58

Xavier de la Cruz: Mapping disease-causing single amino acid polymorphisms to structure..... 59

Damien Devos: Practical limits of function prediction..... 60

Joaquín Dopazo: Finding protein families using an unsupervised neural network..... 61

Andrew Harrison: GRATH..... 62

Christophe Lambert: Homology modeling methodology improvement for low homologous protein sequences..... 63

Florencio Pazos: Similarity of phylogenetic trees as indicator of protein-protein interaction..... 64

Andreas Prlic: WILMA - a genome annotation database for the nematode *C. elegans*... 65

Javier Sancho: Protein cavities created by mutation: a precise structural prediction by molecular mechanics in the CHARMm1 force field..... 66

Dennis Vitkup: Completeness in structural genomics..... 67

LIST OF INVITED SPEAKERS..... 69

LIST OF PARTICIPANTS..... 71

INTRODUCTION

A. Valencia

Large-scale sequencing is filling up the catalogue of natural proteins at a breathtaking speed. Today, we have available not just a large number of sequences, but also glimpses of the genetic inventory of entire organisms. It is widely assumed that this will improve our understanding of cells, in particular, and of life, in general. This may appear like science-fiction, however, structural genomics – the marriage between protein structure determination and genomics – has begun, already. The meeting addressed the challenges for bioinformatics resulting from structural genomics in two ways: (1) How can bioinformatics help structural genomics initiatives? (2) What can bioinformatics profit from the flood of new structures?

Structure determination will be accelerated by, and will profit from genomics. Basing research and technical developments (such as drug design) on all three pillars (sequence, structure, and function) will constitute a major step towards a better understanding of life. Structure/function determination will benefit from genomics in two ways. (1) The mass of available sequences will facilitate quick determination of structure for most existing folds. (2) Sequences for entire organisms will help to unravel missing links in functional pathways, to explore alternative pathways, and to widen our understanding of principle mechanisms and of evolutionary cross-links.

Over the last two years the first serious proposals for carrying out structural genomics have been accepted for funding in various countries (USA, Japan, Germany, England and France). In particular, the National Institutes of Health (NIH, USA) has funded seven pilot projects for 2000-2005. The first conclusions from these pilot projects appear as follows. (1) The major problem is to set up the 'machinery' for large-scale protein expression, and purification. (2) The second major bottleneck are automatic crystallisation-robots, assignment time (for NMR), and accessible Synchrotron time for X-ray. (3) Target selection by bioinformatics is focused at (i) avoiding to duplicate structures for proteins similar to proteins of already determined structure, and at (ii) dissecting proteins into structural domains prior to knowing the structure to facilitate structure determination and to cover structure space. (4) The most important long-term challenge for bioinformatics is to develop methods that will use the wealth of experimental information produced.

The meeting have provided an overview of the status of the experimental techniques for high-throughput structure determination (X-ray crystallography, NMR, combinations of

the two with other techniques). We have addressed questions such as: What are the problems and bottlenecks in large-scale protein structure resolution?, Have we succeeded in solving structures on a large scale? What are the prospects?

As well as issues related with the development of theoretical methods associated to structural genomics, including protein sequence analysis (target selection; separation into families), protein structure prediction (completing the missing information) and prediction of protein function (finding missing links). The presentation have addressed practical issues related with the size of the problem (Can bioinformatics cope with the flood of data?, and issues related with the interpretation of the data, i.e. What are we going to learn from the huge amount of data produced?. Finally, one of the key points of common interest are the problems related with the integration of structural data with the other genomic information, e.g. functional experiments; expression data, sequencing, genome comparision, and variation data.

During the meeting it has also been possible to discuss key questions related with the goals of structural genomics and the implications for Bioinformatics, addressing questions such as: Will the structural genomics initiatives produce a significant gain in functional information? Will this translated into functional annotation? What will be the impact of the enterprise on the advance of biology/health?

A. Valencia

Session 1
Chair: Janet M. Thornton

16,000 targets for structural genomics

Chris Sander

MIT Center for Genome Research, Cambridge, MA, USA

Structural genomics has the goal of obtaining useful three-dimensional models of all proteins, by a combination of experimental structure determination and comparative model building. Currently, useful models of high quality are available only for a minor fraction of protein domains. We estimate the scope of structural genomics, i.e., the total effort requried to make models available for nearly all protein domains. We then evaluate different strategies for reaching optimal information return on effort. Taking into account the strong dependence of model quality on sequence similarity between template and target, we perform test calculations in the Pfam protein family database simulating varying scenarios of how to cover protein space by 3D models. For the protein space considered, the strategy which maximizes structural coverage requires about 7 times fewer structure determinations compared with the strategy where targets are selected at random. As most protein families provide a large number of alternative targets, broad coverage of sequence space can be achieved even with a relatively low success rate of protein structure determination. With a choice of reasonable model quality corresponding to 30% sequence identity and the goal of 90% coverage, we extrapolate from the fraction of residues in complete genomes which can be assigned to Pfam families and estimate the total effort of structural genomics. Using an optimized strategy for achieving completeness, it will take on the order of 16,000 carefully selected structure determinations to construct useful atomic models for the vast majority of all proteins. The actual number is likely to be higher, as practical implementation in a global effort is not likely to be fully optimized. Taking into account the current rate of structure determination and making plausible assumptions about advances in technology, such an effort can be accomplished within a decade, provided significant funding becomes available.

Co-authors of the work reported here are Dennis Vitkup (MIT), Eugene Melamud (CARB), and John Moult (CARB)
(Nature Structural Biology, 2001, in press).

Protein folding, evolution and disease

Christopher M. Dobson

Oxford Centre for Molecular Sciences, University of Oxford, New Chemistry Laboratory, South Parks Road, Oxford OX1 3QT, United Kingdom

Considerable progress has been made recently in understanding the fundamental principles that govern protein folding. Central to much of this progress has been the development of ideas as to the nature of the energy surface or landscape for a folding reaction. These ideas have arisen from a combination of theoretical analysis and experimental investigation (Dinner et al., *TIBS 25*, 331-339, 2000). Of particular importance in the latter has been the concerted application of a wide range of experimental techniques each able to describe aspects of the structural changes taking place during the folding process. NMR spectroscopy and protein engineering have both been key methods in this approach because of their ability to provide structural and dynamical information at the level of individual residues (Vendruscolo et al., *Nature 409*, 641-645, 2001). Some new approaches to utilising and combining these approaches will be described.

Recently, our research has also focussed on the question of what happens if proteins do not fold correctly, or if they subsequently find themselves in an environment where at least partial unfolding takes place. We have been investigating in particular the nature of protein fibrils of the type associated with amyloidogenic diseases. One system of particular interest to us has been c-type lysozyme. This protein has been for some time one of our model systems for studying fundamental aspects of folding. The discovery that clinical cases of amyloidosis are connected with single point mutations in the lysozyme gene has therefore enabled us to explore the molecular basis of this disease in a well-defined model system. This work has recently been extended by the discovery that many proteins not associated with clinical manifestations of disease can form amyloid fibrils in the laboratory under carefully chosen conditions (Chiti et al., *PNAS 96*, 3590-3594, 1999). This finding has enabled the nature of the structure and mechanism of formation of these fibrils to be explored in some detail (Dobson, *TIBS 24*, 329-332, 1999).

In addition to these experimental approaches, we have recently been investigating the molecular basis for the aggregation of proteins by the use of site-directed mutagenesis. The results indicate that the propensity of a denatured protein to aggregate can be altered substantially by judicious substitution of a small number of residues that can be chosen not to perturb the native protein stability or function (Villegas et al., *Protein Sci. 9*, 1700-1708 2000). In more recent work on acylphosphatase we have found that a few key regions of the protein sequence dominate the rate of aggregation of unfolded states of the protein, and that these regions are distinct from those that determine the folding behaviour (Chiti et al., in preparation). This talk will report some of these recent results from our laboratory and speculate on their possible significance for interpreting general characteristics of protein structures and the evolution of the sequences that encode the various folds.

"Clustering the universe of protein sequences in light of structural and functional genomics"

Michal Linial, Elon Portugaly and Yonatan Bilu

The number of proteins whose structure was solved at high resolution lags way behind the number of proteins sequenced (1). It is a major obstacle in studying protein structure to predict which proteins belong to new, currently unknown superfamilies or folds. In an attempt to address this problem we mapped all proteins with solved structures onto a graph of all known protein sequences provided by ProtoMap (2, 3). We wish to sort proteins according to their likeliness to belong to new superfamilies. We hypothesize that proteins within neighboring clusters tend to share common structural superfamilies or folds. If true, the likelihood to find new superfamilies (and folds) increases in clusters that are distal from other solved structures within the graph 4). On the basis of this hypothesis, we define an order relation between unsolved proteins according to their "distance" from solved structures in the graph. Based on that order relation we sort about 48,000 proteins (from the most likely to belong to new superfamilies to the most unlikely). Our list may be partitioned to three parts: the first contains ~35,000 proteins sharing a cluster with a known structure; the second contains ~6,500 proteins in clusters neighboring clusters containing known structures; the third contains the rest of the proteins – 6096, in 1274 clusters. Over 97% of third part proteins have no significant pairwise sequence similarity to any solved protein (E score worse than 0.1).

We test the quality of the order relation using datasets of solved structures that were not considered when the order was defined. The tests show that our order is significantly (P-value ~ 10-5) better than a random order. More interestingly, even when we ignore the first part, or when we consider only the third part, the order performs better then random (P-values: 0.002 and 0.15 respectively). For most test sets, an order derived from PSI-BLAST results performed worse than our method (P-values: 0.008 and 0.21). We tested an order derived from the refinement of our order using PSI-BLAST results. The last order usually performed better than either of the first two. Herein we present a method for selecting target proteins to be used in the Structural Genomics Project. We will also discuss the potential of sequence-based protein classifications for functional predictions using the yeast genome as a test case (5).

This study is supported by the Ministry of Science (Tashtiot), Internet2 and the Horwitz foundation.

References:

1. Linial, M. and Yona, G. (2000) Methodologies for target selection in structural genomics. Prog. Biophy. Mol. Biol. 73. 297-320.
2. Yona, G., Linial, N. and Linial M. (1999) ProtoMap – Automated classification of all proteins sequences: a hierarchy of protein families, and local maps of the protein space. Proteins 37, 360-378.
3. Yona, G., Linial, N. and Linial M. (2000) ProtoMap – A classification of all proteins sequences and hierarchy of protein families. Nucleic Acid Research 28. 49-55
4. Portugaly, E. and Linial, M. (2000) Estimating the probability of a protein to have a new fold – A statistical computational approach. Proc. Natl. Acad. Sci., USA 97. 5161-5166.
5. Bilu Y. and. Linial, M. (2001) On the Predictive Power of Sequence Similarity in Yeast. Proceedings RECOMB - Computational Molecular Biology (in press).

Recognizing old and new folds in genomes

David T. Jones

Institute for Cancer Genetics and Pharmacogenomics, Department of Biological Sciences, Brunel University, Uxbridge, Middlesex, U.K.

Protein fold recognition is a very effective means for predicting protein tertiary structure from sequence, as seen in the various CASP prediction experiments (e.g. Jones, 1999a). Despite this success, the better fold recognition methods often employ some degree of human expert intervention, which is clearly impractical if these methods are going to be applied to the annotation of uncharacterised genome sequences.

Here I will be discussing several automatic methods which have been developed for predicting protein structure on a genome-wide scale. These methods range from an extended version of a previously published fold recognition method for genome sequences, to a method which is capable of predicting entirely novel protein folds. In addition, I will also discuss various approaches which can be used to identify protein sequences which have novel folds.

References:

Jones, D.T., Tress, M., Bryson, K. & Hadley, C. (1999) Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure. Proteins. S3, 104-111.

Jones, D.T. (1999) GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. J. Mol. Biol. 287, 797-815.

Bioinformatics tools for structural genomics

Seán I. O'Donoghue and Joachim Meyer

LION Bioscience AG, Waldhoferstr. 98, Heidelberg 69123, Germany

The new science of structural genomics will require enormous resources, not only those associated with the actual structure determination, but also resources for integrating the diverse information (cloning, expression, crystallization trial, preliminary NMR spectra) within each structural genomic project, and between projects. Once structures become available, another challenge arises to analyze and integrate all this structural information; methods such as DALI that can analyze large numbers of structures and define simplified views in a simplified 'structure-space' will be important in maintaining the overview as the number of structures increases.

We will describe modifications and improvements we are making to our bioSCOUT and the SRS systems in anticipation of these requirements. bioSCOUT is the direct descendent of the GeneQuiz system; GeneQuiz was the first system to do whole-genome structure prediction; bioSCOUT, it's commercial descendent, was the first commercial system for such analysis. SRS is a widely used system for integrating diverse biological databases into a single meta-database.

In the case of SRS, we are extending its range of databases to include those directly relevant to structural analyses and structural genomics projects. Including its knowledge of fold libraries. In the case of bioSCOUT, we are improving the breadth and accuracy of its structural modeling methods, and adding further prediction methods that take advantage of 3D structures when they are known or can be modeled. We are also adding novel methods for integrating and presenting the large scale structural data that will be obtained by structural genomics.

Session 2
Chair: Manuel C. Peitsch

Approaching function through structure

Christopher D. Lima

Weill Medical College of Cornell University
1300 York Ave.
New York, NY 10021
lima@pinky.med.cornell.edu
http://pinky.med.cornell.edu
Phone: 212 746 6449
Fax: 212 746 4843

Structural genomics intends to use the intrinsic information content of multiple genome projects to prioritize a subset of conserved gene families with little known function for crystallographic and biochemical study.

Genomics provides an unbiased view of what nature considers important to preserve and maintain life. The importance of several gene families has gone undetected due to the failure of traditional biochemical or genetic screens to identify their function.

A primary goal of structural genomics will be to provide the science of structural biology with the same global understanding of three-dimensional biological space that sequence genomics has given to the linear content of the genomes. Utilization of structural information in the design and execution of further biochemical and functional screens will yield one of the first major dividends of the structural genomics effort.

The New York Structural Genomics Research Consortium (http://www.nysgrc.org/) intends to develop high-throughput methods for target selection, cloning, protein expression, protein purification, and x-ray crystallographic structure determination. The major goals of this project are to develop high-throughput methods for x-ray crystallographic structure determination as applied to proteins discovered through genomics. The specific aims of this proposal are 1) Identify target protein sequences for structural genomics, 2) Develop high-throughput *E. coli* expression of soluble target proteins, 3) Develop high-throughput production of target proteins, 4) Develop high-throughput biophysical characterization of target proteins, 5) Develop high-throughput crystallization of target proteins, 6) Develop high-throughput experimental strategies for MAD crystallography, 7) Develop high-throughput synchrotron data collection with target protein crystals, 8) Develop and use an Internet-based computational pipeline for protein crystallography, 9) Develop high-throughput molecular replacement tools for crystallography, 10) Develop high-throughput comparative modeling for structural genomics, 11) Develop efficient annotation and dissemination of protein structures and models, 12) Develop and Internet-based "web-book" for use by the NYSGRC.

Homology-based annotation of the *Drosophila melanogaster* genome

Shuba Gopal, Mark Schroeder, Ursula Pieper, Alexander Sczyrba, Gulriz Aytekin-Kurban, Stefan Bekiranov, J. Eduardo Fajardo, Eashwar Narayanan, Roberto Sanchez, Andrej Sali, Terry Gaasterland1

The approach to annotating a genome critically affects the number and quality of genes identified in the genome sequence. Genome annotation based on stringent gene identification is prone to underestimate the complement of genes encoded in a genome. In contrast, over-prediction of putative genes followed by exhaustive computational sequence, motif, and structural homology search will find rarely expressed, possibly unique, new genes at the risk of including biologically non-functional genes among predicted genes. A two-stage approach that combines the merits of stringent genome annotation with the benefits of over-prediction has been developed and applied to the genome of *Drosophila melanogaster*. After filtering 19,410 plausible genes, the search validated 1,042 genes beyond the 13,601 currently annotated genes [1]. The approach applies to genomes of all organisms, including human.

This talk will discuss the results of applying our annotation process to the Drosophila genome and the impact on strategies for structural genomics target selection.

Finding biologically important similarities between protein structures

Alexey G. Murzin

Centre for Protein Engineering, MRC Centre, Hills Road, Cambridge CB2 2QH, UK

Almost every new protein structure is similar to some extent to already known structure. The detection and classification of structural similarities facilitate greatly our understanding of protein structure, function and evolution. The structural classification of proteins (SCOP) database hierarchically organises all proteins of known structure according their structural similarity and far and near evolutionary relationships. Central to the SCOP hierarchy is the notion of that the majority of extant proteins come from, and, in principle, can be traced back to a relatively small number of ancestral proteins. All descendants from the same ancestor form a single protein superfamily and share the ancestral common fold. However, the members of the same superfamily may well have diverged beyond any recognisable sequence similarity, and they even may have evolved very different functions. These multifunctional superfamilies provide the most valuable insight into the evolution of genomes and proteomes.

In the common fold of a superfamily, there can be specific structural details identified that are called hereafter as the superfamily characters. These characters are conserved within superfamily presumably because of their structural and/or functional role. They can help to improve our understanding of natural history of proteins. The superfamily characters can be used for the detection of sequence families of unknown structure that may belong to a given superfamily that will allow the prediction of structures and functions for members of these families. The structural role of superfamily characters is not restricted to the stabilisation of protein folds, and there may be other functional roles than just being a scaffold for functionally important side chains in the active site. There are other possible structural and functional roles for the superfamily characters that will be discussed in my talk. New structural roles include the discrimination of alternative globular structures formed by the same secondary structure elements. New functional roles can be assigned to particular backbone conformations that contain no conserved sequences but contribute the mainchain groups to the active sites. The new roles for superfamily characters give new insights into the origin and evolution of protein function.

The role of protein structure in genomics

Manfred J. Sippl

University of Salzburg, Center of Applied Molecular Engineering, Jakob-Haringer-Str. 3, A-5020 Salzburg, Austria, Tel: +43(662) 8044-5797, Fax: +43(662) 454889,
Email: sippl@came.sbg.ac.at

The annotation of genes in terms of the three-dimensional structure, molecular function, and biological role of their protein products will occupy genomic research in the years to come. In essence, annotating genes and genomes, requires the creation of data bases, containing biologically relevant aspects of genes, gene procucts, and their interactions. To be useful these data bases need to be machine readable, compatible with bioinformatics tools, error free, and up to date.

In the annotation we have to distinguish two types of data. The first type is obtained from direct experimental observations, like X-ray and NMR structures, or the reactions catalyzed by enzymes. The second source of data is obtained from computational tools, like homology searches, or molecular models.

We develop an annotation scheme that uses protein domains obtained from X-ray and NMR experiments as primary data. Domains are defined in terms of structural/functional units. The classification is organized in hierarchical layers, that reflect function (family), homology (superfamily) and structure (fold type).

The annotation of sequence libraries and genomes links individual sequences to protein domains via computational tools, like sequence homology searches and fold recognition techniques. The domain classification serves as a reference to organize structural and functional features of sequences. A second big advantage is its impact on computational tools. For example, multiple alignments derived from structural families and superfamilies enhance sensitivity and coverage of sequence search tools, fold recognition searches in domain libraries are more accurate, and the interpretation of search results is more reliable. We discuss the construction of the domain classification and several applications.

Session 3
Chair: Manfred J. Sippl

Prediction of protein-protein interaction networks from genome information

Minoru Kanehisa

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan
kanehisa@kuicr.kyoto-u.ac.jp

The analysis of sequences and 3D structures of DNAs, RNAs, and proteins has been extremely useful for understanding their molecular functions. In general, however, the biological function of a cell or an organism is a result of many interacting molecules; it cannot be attributed to just a single molecule. Under the KEGG (http://www.genome.ad.jp/) project we have been organizing a reference database of known protein-protein interaction networks in living cells, including metabolic pathways, various regulatory pathways, and molecular complexes. At the same time we have been developing computational methods for knowledge-based prediction of protein-protein interaction networks and higher order functions from the complete genome sequences.

KEGG is a composite database of different data objects shown below:

Database	Data object	Content
PATHWAY	Network	Known pathways and complexes
GENES	Genome	Gene catalogs of individual organisms
EXPRESSION	Transcriptome	Microarray gene expression profiles
BRITE	Proteome / Protein universe	Protein-protein interactions and relations
LIGAND	Environment	Chemical compounds and reactions

All of these data objects are integrated with the concept of graphs. The network prediction involves a conversion from a set of genes in the genome to a network of gene products in the cell, which is considered as a conversion between two graphs: the genome graph with genes as nodes to the network graph with gene products as nodes. The prediction can be facilitated by additional experimental data on the transcriptome and the proteome, as well as computational results on the protein universe, which are again represented as graphs with either genes or proteins as nodes. I will report on an extension of our method to detect loose graph similarities termed correlated clusters (Ogata et al., NAR 28:4021-8, 2000) and its application to find empirical relationships among the genome, the protein universe, and the metabolic network.

References:

1. Ogata, H., Fujibuchi, W., Goto, S., and Kanehisa, M.; A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. Nucleic Acids Res. 28, 4021-4028 (2000).
2. Kanehisa, M.; Pathway databases and higher order function. Adv. Protein Chem. 54, 381-408 (2000).

What do structure classifications reveal about the universe of protein folds and protein evolution

Christine Orengo, Andrew Harrison, Frances Pearl, James Bray, David Lee, Daniel Buchan, Janet Thornton

Since the 1970s the structural databases have grown exponentially with ~19,000 protein chains now determined. At UCL we have developed a structural classification (CATH) which groups proteins according to their (C)lass, (A)rchitecture, (T)opology or fold and (H)omologous superfamily. Various automatic protocols have been implemented for recognising similarities at each level in the classification. CATH currently contains ~28,000 domains grouped into ~1200 superfamilies and ~650 fold groups. Population statistics and variability data have been compiled for different levels in the classification.

In order to speed up the classification to keep pace with structure genomics initiatives, we have recently developed a protocol for recognising folds within multidomain proteins. This uses graph theory to identify corresponding structural cliques in proteins, based on similarities in their secondary structure orientations and midpoint separations. The method can be used to aid domain boundary assignment, which is one of the major bottlenecks in structure classification.

We have also been developing protocols to expand functional annotation within CATH superfamilies by identifying related gene sequences in Genbank and displaying any associated functional data in the CATH dictionary of homologous superfamilies (Bray et al). Several methods have been optimised and validated including Hidden Markov Models (Sonnhammer et al., Karplus et al.) and 1D-Profiles such as PSI-BLAST (Altschul et al.). We have used these methods to integrate ~300,000 domain sequences into CATH. The algorithm DomainFinder (Pearl et al.) is used to identify domain boundaries within the gene relatives.

Structural annotations for gene regions within each completed genome are also provided in the CATH Gene-3D web resource together with statistics for fold distributions within the genomes. This resource can be used to identify gene regions for which no structural data currently exists and is being used to facilitate target selection for structure genomics.

Analysis of functional properties in these expanded CATH superfamilies has shown that below 40% sequence identity function is poorly conserved for single domain proteins and below 60% identity for multidomain proteins suggesting that at low sequence identities more experimental structures will be required to predict protein function. Hidden Markov models have been built at different levels within the CATH database to improve identification of functional subgroups, where possible.

References:

A Rapid Classification Protocol for the CATH Domain Database to Support Structure Genomics. F.E.Pearl, A. Shepherd, G.Reeves, D. Buchan, J.Bray, D.Lee, I.Sillitoe, A.Todd, A. Harrison, J.M.Thornton, C.A.Orengo *Nucl. Acids Res.* (2001) 29:223-7.

The CATH Dictionary of Homologous Superfamilies (DHS): a consensus approach for identifying distant structural homologues. Bray JE, Todd AE, Pearl FM, Thornton JM, Orengo CA.Protein Eng. (2000) 13:153-65.

From structure to function: approaches and limitations. Thornton JM, Todd AE, Milburn D, Borkakoti N, Orengo CA. Nat Struct Biol. (2000) 7 Suppl:991-4.

Do we aim at getting one structure per fold?

Burkhard Rost

One representative of each 'structural family' instead of one representative per fold! One goal of structural genomics is to determine one high-resolution structure for each existing protein fold (1). Do we know the protein sequence for a representative of each fold family, already? Assuming an affirmative answer, can bioinformatics systematically identify one representative for each unknown fold from sequence? The answer is definitively negative. The vast majority of similar folds have less than 10% sequence identity, i.e. lie in the midnight zone of sequence alignments (2, 3). Even advanced sequence searches (4) cannot reach into this zone (5, 6). Thus, there is no way to identify exactly one representative of all folds represented in current sequence databases. However, bioinformatics can identify a set of 'structural families' such that all members of that family are likely to have a similar fold. Currently, we know about 10,000 'structural families'. For about 2300 of these we have high-resolution structures. (Interestingly, almost 20% of these sequence-unique structures (400 of 2300) have been determined over the last ten months.) Hence, we will have to solve structures for about 7,700 protein families. If the current rate of determining sequence-unique structures continues, we may have structures for all families in 16 years. Large-scale initiatives in structural genomics may half this number.

Bioinformatics crucial to avoid the traps of sequence space, i.e. non-globular structures. Should structural genomics initiatives target to determine one representative for each of the 8,000 families? Obviously, one important objective in prioritising the list of targets is to avoid non-globular proteins. One important challenge for bioinformatics is to determine these proteins. We already know that about 20-30% of all proteins contain transmembrane helices (7). About 30% of all human proteins contain regions of more than 70 residues without any apparent signal for regular secondary structure (8). About 4% of all eukaryotic proteins contain long coiled-coil regions. Obviously, such regions should be excluded from structural genomics. However, should we exclude the respective proteins? How do we best determine which targets to go for first?

Systematic concept for target selection: driven by aspects of function. The first round of targets will obviously focus on large families to maximise the direct impact of the determined structure on biology. However, this concept will cover less than 20% of sequence space. How to prioritise the remaining majority of structural families? We suggest exploring functional criteria. For example, select proteins interacting to many others. Group proteins according to sub-cellular localization, target extra-cellular (< 20% of proteome), nuclear proteins, or long globular fragments of helical membrane proteins.

References:

1. Rost, B. (1998) *Structure* **6,** 259-263.
2. Rost, B. (1997) *Folding & Design* **2,** S19-S24.
3. Yang, A. S. & Honig, B. (2000) *J. Mol. Biol.* **301,** 679-689.
4. Altschul, S., Madden, T., Shaffer, A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. (1997) *Nucl. AcidsRes.***25,** 3389-3402.
5. Rost, B. (1999) *Prot. Engin.* **12,** 85-94.
6. Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. & Chothia, C. (1998) *J. Mol. Biol.* **284,** 1201-1210.
7. Liu, J. & Rost, B. (2001) *Prot. Sci.*, in submission.
8. Liu, J., Tan, H. & Rost, B. (2001) *Structures*, in preparation.

Comparative protein structure modeling of genes and genomes

Andrej Sali

The Rockefeller University, New York, USA
email sali@rockefeller.edu

Structural genomics aims to determine or accurately predict 3D structure of most proteins (1). This aim will be achieved by a focused, large-scale determination of protein structures by X-ray crystallography and NMR spectroscopy, combined efficiently with accurate protein structure prediction. Comparative protein structure modeling will be discussed in this context. To allow large-scale modeling, we automated fold assignment, sequence-structure alignment, comparative model building, and model evaluation (2). These steps were implemented mostly in our program Modeller, which is available on the Web at http://guitar.rockefeller.edu/.

The modeling pipeline has been applied to all of the approximately 400,000 protein sequences in the TrEMBL database, resulting in 3D models for segments of approximately 200,000 proteins. These models are stored in the ModBase database (3), accessible over the web at http://pipe.rockefeller.edu/modbase. Several examples of how comparative modeling can be useful in the biological analysis of individual proteins as well as whole genomes will be described.

References:

1. R. Sanchez, U. Pieper, F. Melo, N. Eswar, M.A. Marti-Renom, M.S. Madhusudhan, N. Mirkovic, and A. Sali. Protein structure modeling for structural genomics. Nat. Struct. Biol. 7, 986-990, 2000.

2. R. Sanchez and A. Sali. Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. Proc. Natl. Acad. Sci. USA 95, 13597-13602, 1998.

3. R. Sanchez, U. Pieper, N. Mirkovic, P. I. W. de Bakker, E. Wittenstein and A. Sali. ModBase, a database of annotated comparative protein structure models. Nucl. Acids Res. 28, 250-253, 2000.

From sequence to structure to function: the importance of overlapping binding sites for zinc finger-containing proteins on transcription regulation and possible modulation by anticancer ecteinas cidins

Raquel García-Nieto & Federico Gago

Departamento de Farmacología, Universidad de Alcalá, E-28871 Madrid, Spain

The mdr1 gene encodes a highly conserved 180-kDa membrane P-glycoprotein (P-gp) that mediates the efflux of different xenobiotics from the cytoplasm and is found overexpressed in cells exhibiting a multidrug resistance (MDR) phenotype. The proximal promoter of the mdr1 gene does not contain a TATA box but contains several regulatory regions, including an inverted CCAAT box at -79 to -75, and a downstream G-rich site (-46 to -61) [1] that contains overlapping sites to which both Sp1 (GGGGCGTGG) [2] and the general transcription factor EGR-1 (GCGTGGGCT) [3] can bind specifically. This dual Sp1/EGR-1 site is required, in addition to the inverted CCAAT box, to mediate acute induction of mdr1 gene expression by trichostatin A [4]. Activation by 12-O-tetradecanoylphorbol-13-acetate has been shown to be mediated by EGR-1 [3] and can be inhibited by the Wilms' tumor suppressor, WT1 [5]. Binding of these zinc finger-containing transcription factors to the major groove of specific G/C-rich DNA sequences is based on a relatively simple set of contacts that resembles a well-studied protein-DNA recognition code [6]. Activation of mdr1 expression can be effectively inhibited by nanomolar concentrations of antitumour ecteinascidin ET743 [7,8]. This marine compound has been shown to bind covalently to the DNA minor groove and to induce bending of the DNA toward the major groove [9]. The striking structural similarity in DNA architecture between a DNA-EGR-1 complex and a DNA-ET743 complex led us to propose that ET743 might be selectively recognizing a DNA stretch that is already preorganized for binding upon association with a specific zinc finger-containing protein, very likely the transcription factor Sp1 [10]. Facilitation of drug binding by a protein-induced DNA deformation could account for the unusually high potency and the remarkable effects on gene transcription of ecteinascidin-like molecules. We now redefine the consensus sequence for Sp1, propose a molecular model for the DNA binding domain of this ubiquituous transcription factor, and probe the suitability of EGR-1, Sp1 and WT-1 binding sites as putative target sites for ET743.

References:

1. Madden MJ, Morrow CS, Nakagawa M, Goldsmith ME, Fairchild CR & Cowan KH. (1993). Identification of 5' and 3' sequences involved in the regulation of transcription of the human mdr1 gene in vivo. J. Biol. Chem. 268, 8290-8297. 2. Cornwell MM & Smith DE. (1993). Sp1 activates the MDR1 promoter through one of two distinct G-rich regions that modulate promoter activity. J. Biol. Chem. 268, 19505-19511. 3. McCoy C, Smith DE & Cornwell MM. (1995). 12-O-tetradecanoylphorbol-13-acetate activation of the MDR1 promoter is mediated by EGR1. Mol. Cell. Biol. 15, 6100-6108. 4. Hu Z, Jin S & Scotto KW. (2000). Transcriptional activation of the MDR1 gene by UV irradiation. Role of NF-Y and Sp1. J. Biol. Chem. 275, 2979-2985. 5. McCoy C, McGee SB & Cornwell MM. (1999). The Wilms' tumor suppressor, WT1, inhibits 12-O-tetradecanoylphorbol-13-acetate activation of the multidrug resistance-1 promoter. Cell Growth Differ. 10, 377-386. 6. Choo Y & Klug A. (1997). Physical basis of a protein-DNA recognition code. Curr. Opin. Struct. Biol., 7, 117-125. 7. Minuzzo, M.; Marchini, S.; Broggini, M.; Faircloth, G.; D'Incalci, M. & Mantovani, R. (2000). Interference of transcriptional activation by the antineoplastic drug ecteinascidin-743. Proc. Natl. Acad. Sci. USA, 97, 6780-6784. 8. Jin, S.; Gorfajn, B.; Faircloth, G. & Scotto, K. W. (2000). Ecteinascidin 743, a transcription-targeted chemotherapeutic that inhibits MDR1 activation. Proc. Natl. Acad. Sci. USA, 97, 6775-6779. 9. García-Nieto, R.; Manzanares, I.; Cuevas, C. & Gago, F. (2000). Bending of DNA upon binding of ecteinascidin 743 and phthalascidin 650 studied by unrestrained molecular dynamics simulations J. Am. Chem. Soc. 122, 7172-7182. 10. García-Nieto, R.; Manzanares, I.; Cuevas, C. & Gago, F. (2000). Increased DNA binding specificity for antitumor ecteinascidin ET743 through protein-DNA interactions? J. Med. Chem. 43, 4367-4369.

Session 4
Chair: Chris Sander

From structure to function: functional diversity within protein superfamilies

Janet M. Thornton [(1)] Annabel Todd [(2)] & Christine Orengo [(2)]

(1) University College and Birkbeck College, Biochemistry and Molecular Biology Department, Gower Street, London WC1E 6BT, UK
(2) University College London, Biochemistry and Molecular Biology Department, Gower Street, London WC1E 6BT, UK

Structural Genomics promises a flood of structures for proteins whose functions are unknown. Currently the major route to infer function is through sequence or structural similarity, providing evidence of an evolutionary relationship from which some functional relationship is often predicted. If this approach is to be successful, it is necessary to understand the conservation and diversification of function within homologous families during evolution.

The recent growth in protein databases has revealed the functional diversity of many protein superfamilies. We have assessed the functional variation of homologous enzyme superfamilies containing two or more enzymes, as defined by the CATH [1] protein structure classification, by way of the Enzyme Commission (EC) scheme [2]. Combining sequence and structure information to identify relatives, the majority of superfamilies display variation in enzyme function, with 25% of superfamilies in the PDB having members of different enzyme types. We determined the extent of functional similarity at different levels of sequence identity for 486,000 homologous pairs (enzyme/enzyme and enzyme/non-enzyme), with structural and sequence relatives included. For single- and multi-domain proteins, variation in EC number is very rare above 40% sequence identity, and above 30%, the first three digits may be predicted with an accuracy of almost 90%. For more distantly related proteins sharing less than 30% sequence identity, functional variation is significant, and below this threshold, structural data are essential for understanding the molecular basis of observed functional differences. To explore the mechanisms for generating functional diversity during evolution, we have studied in detail 31 diverse structural enzyme superfamilies for which structural data are available. A large number of variations and peculiarities are observed, at the atomic level through to gross structural rearrangements. Almost all superfamilies exhibit functional diversity generated by local sequence variation and domain shuffling. Commonly, substrate specificity is diverse across a superfamily, whilst the reaction chemistry is maintained. In many superfamilies, the position of catalytic residues may vary despite playing equivalent functional roles in related proteins. The implications of functional diversity within supefamilies for the structural genomics projects are discussed. More detailed information on these superfamilies is available at http://www.biochem.ucl.ac.uk/bsm/FAM-EC/.

References:

1. Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. & Thornton. J.M., (1997). CATH - A Hierarchic Classification of Protein Domain Structures. *Structure*, 5, 1093-1108.
2. Todd, A.E., Orengo, C.A. & Thornton, J.M. (2000). Evolution of Function in Protein Superfamilies. *Journal of Molecular Biology*, (Accepted)

Classifying the universe of protein structures

Liisa Holm

Structural Genomics Group, EMBL-EBI, Cambridge CB10 1SD, UK

The rapid growth in the number of experimentally determined three-dimensional protein structures has sharpened the need for comprehensive and up-to-date surveys of known structures. The generally accepted classification scheme for protein structures has four hierarchical levels (our terms in parenthesis) which correspond to

1. supersecondary structural motifs (attractors in fold space)
2. the topology of globular domains (fold types)
3. remote homologues (functional families)
4. homologues with sequence identity above 25% (sequence families).

We have developed computational procedures to derive such a classification fully automatically. I will particularly report on recent developments in class assignment and identifying the homologue-to-analogue transition.

Classic work on protein structure classification has made it clear that a structural survey is best carried out at the level of domains. Our method for automated domain identification from protein structure atomic coordinates is based on quantitative measures of compactness and recurrence [1].

The central concept is a map of fold space, which is derived from all-against-all structure comparison using distance matrix alignment (Dali program). We have identified five densely populated attractor regions in fold space [2]. Each fold space region is represented by an archetype and a shortest-path criterion is used to assign structures to attractors.

Fold types are defined as clusters of structural neighbours in fold space. The radius of the clusters has been chosen empirically and groups together structures which have topological similarity [3].

Structural similarity accompanied by similarity of function makes common evolutionary descent a plausible inference. We have developed a procedure to identify the analogue to homologue transition computationally [4]. Functional families are clusters in fold space where all pairs have a high average neural network prediction for being homologous. The neural network weighs evidence coming from: overlapping sequence neighbours, clusters of identically conserved functional residues, E.C. numbers, Swissprot keywords. SCOP superfamilies are recovered at high accuracy and good coverage.

The Dali Domain Dictionary (http://www.ebi.ac.uk/dali/domain) provides a map of the currently known regions of the protein structure universe and unifies distantly related protein families using a uniform set of criteria. It is useful for the analysis of folding principles and for maximizing the information return from experimental structure determination.

References:
[1] Holm L, Sander C (1998) Dictionary of recurrent domains in protein structures. Proteins, 33, 88-96.
[2] Holm L, Sander C (1996) Mapping the protein universe. Science, 273, 595-603.
[3] Holm L, Sander C (1993) Protein structure comparison by alignment of distance matrices. J Mol Biol 233, 123-138.
[4] Dietmann S, Park J, Notredame C, Heger A, Lappe M, Holm L (2001) A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary v.3. Nucl Acids Res 29, 55-57.

Bioinformatics approaches to the detection of protein interaction networks

Alfonso Valencia

Protein Design Group. CNB-CSIC. Cantoblanco, Madrid 28049. Spain
Tfn. +34-1-585 45 70, Fax. +34-1-585 45 06, email: valencia@cnb.uam.es
http://www.pdg.cnb.uam.es

The considerable amount of information available about individual protein components in the form of genome sequences (0), protein structures, and functional genomics (gene expression patterns) demands further work for its integration. The reconstruction of protein interaction networks is the obvious next step in this direction.

I will present three complementary computational efforts for the study of protein-protein interactions.

The first approach is based on the study of the patterns of variation in multiple sequence alignments. We have previously demonstrated that the weak signal left by evolution in the form of correlated mutations is enough to single out the right inter-domain docking solution amongst many wrong alternatives (1). These predictions have been tested in different experimental systems (2-3). The extension of this method to the detection of interacting partners makes possible the prediction of large collections of interactions at genomic scale (4).

The second approach is based on the application of information extraction techniques (5-6) to the retrieval of protein interactions directly from the scientific literature (Medline abstracts, 7). The results obtained in different complex biological systems are very informative even if number of critical areas still require further development.

The third approach is based on the application of clustering techniques (self organizing maps, 8) and information retrieval methods (9) to the analysis of expression array data. The results obtained in different systems lead to the discovery of new functional links between up to now unrelated genes (10).

The challenge for the future is the integration of the information contained in the interaction networks with the proteomic and structural genomic data.

The work in protein interaction prediction was carried out by F. Pazos, the tools for information retrieval were developed C. Blaschke, the application of SOM to expression arrays was developed by J. Herrero and J. Dopazo (CNIO). The combined application of clustering and IE to expression arrays was mainly the work of J.C. Oliveros-Collazos.

References:

0.- Iliopoulos, I., Tsoka, S., Andrade, M.A., Janssen, P., Audit, B., Tramontano, A., Valencia, A., Leroy, C., Sander C., Ouzounis, C.A. (2000) Genome sequences and great expectations. Genome Biology 2: 0001.1-0001.3
1.- Pazos F, Helmer-Citterich M, Ausiello G, Valencia A (1997) Correlated mutations contain information about protein-protein interaction. J. Mol. Biol. 272, 1-13
2.- Gdssler C S, Buchberger T, Laufen M P, Mayer H, Schrvder A, Valencia A, Bukau D (1998) Mutations in DnaK chaperone affecting interaction with the DnaJ co-chaperone. Proc. Natl. Acad. Sci. USA. 95: 15229-15234.

3.- Azuma Y, Renault L, García-Ranea JA, Valencia A, Nishimoto T, Wittinghofer A (1999) Model of the Ran-RCC1 Interaction using Biochemical and Docking Experiments. J. Mol. Biol. 289:1119-1130.
4.- Pazos F, Valencia A (2001) In silico two-hybrid system for the selection of physically interacting proteins pairs. Submitted.
5.- Andrade M A, Valencia A (1997) Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. Development of a prototype system. ISMB 5, 25-32
6.- Andrade M A, Valencia A (1998) Automatic extraction of keywords from scientif text: Application to the knowledge domain of protein families. Bioinformatics 14, 600-607
7.- Blaschke C, Andrade MA, Ouzounis C, Valencia A (1999) Automatic extraction of biological information from scientific text: protein-protein interactions. ISMB 7: 60-67.
8.- Herrero J, Valencia A, Dopazo J (2000) A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioionformatics, in the press.
9.- Blaschke C, Oliveros- Collazos JC, Valencia A (2000) Mining functional information associated to expression arrays. Functional and Integratetive Genomics, in the press.
10.- Oliveros-Collazos JC, Blaschke C, Javier Herrero J, Dopazo J, Valencia A (2000) Assessing the functional information associated to expression profile similarities. Genome Informatics Series, in the press.

The accuracy and use of protein models derived from comparative methods

Manuel C. Peitsch, Torsten Schwede, Alexander Diemand and Nicolas Guex

GlaxoSmithKline. and Swiss Institute of Bioinformatics
WTC I, 10, route de l'aéroport, 1215 Genève

Comparative protein modelling methods have now been around for some time. It is clear that the correctness and accuracy of the models has made some progress, but the methods are still far from being able to duplicate the experimental control structures. What are the causes of these structural differences? and to what end can protein models derived from comparative methods be used?

Assessing the quality of a model

The quality of a model is determined by two distinct criteria, which will determine its applicability. First, the correctness of a model is dictated by the quality of the sequence alignment used to guide the modelling process. If the sequence alignment is wrong in some regions, then the spatial arrangement of the residues in this portion of the model will be incorrect. The first edition of the community-wide experiment known as Critical Assessment of protein Structure Prediction (CASP) already underscored that most severe modelling errors can be traced back to sequence alignment mistakes. This remains, despite many efforts to address this issue, the main weakness of comparative protein modelling. Second, the accuracy of a model is essentially limited by the deviation of the used template structure(s) relative to the experimental control structure. This limitation is inherent to the methods used, since models result from an extrapolation. As a consequence, the core Cα atoms of protein models which share 35 to 50% sequence identity with their templates, will generally deviate by 1.5 to 1.0 Å from their experimental counter parts as do experimentally elucidated structures. One should however not overlook the contributions of the templates to the model accuracy. The templates, which are obtained through experimental approaches, are subject to structural variations not only caused by experimental errors and differences in data collection conditions - such as the temperature, but also because of different crystal lattice contacts and the presence or absence of ligands. Furthermore, X-ray crystallography and NMR generally yield 3D structures with an even broader rmsd spread. It is thus crucial to know the experimental conditions under which the modelling templates were collected as this has a direct impact on the accuracy of the derived models and thereby on their potential use.

About the use of protein models

Protein model obtained with comparative modelling methods can be classified into three broad categories: i) models which are based on incorrect alignments between target and template sequences. Such alignment errors, which generally reside in the inaccurate positioning of insertions and deletions, are caused by the weaknesses of the alignment algorithms and can often not be resolved in the absence of a control experimental structure. It is however often possible to correct such errors by producing several models based on alignment variants and by selecting the most "sensible" solution. Nevertheless, it turns out that such models are often useful as the errors are not located in the area of interest, such as within a well conserved active site. ii) Models based on correct alignments are of course much

better, but their accuracy can still be medium to low as the templates used during the modelling process have a medium to low sequence similarity with the target sequence. Such models, as the ones described above, are however very useful tools for the rational mutagenesis experiment design. They can however not be of great assistance during detailed ligand binding studies. iii) The last category of models comprises all those which were build based on templates which share a high degree of sequence identity (> 70%) with the target. Such models have proven useful during drug design projects and allowed the taking of key decisions in compound optimisation and chemical synthesis. For instance, models of several species variants of a given enzyme can guide the design of more specific non-natural inhibitors.

However, nothing is absolute and there are numerous occasions in which models falling in any of the above categories, could either not be used at all or in contrast proved to be more useful and correct as first assessed.

In our experience, several applications of medium-accuracy models have proven successful. These can be classified into three categories:

Interpreting the impact of mutations on protein function. Potential link to diseases:

One of the first uses one can make of a model structure is to interpret the impact a mutation can have on the overall function of a protein. Although the development of objective scoring functions has begun only recently, "visual inspection" associated with a good knowledge of the rules underlying protein structure has proven useful in defining the broad reasons for mutant malfunction. With the upcoming high throughput production of single nucleotide polymorphisms (SNP), objective scoring functions will be crucial to make maximum use of the information. Indeed, a sizeable proportion of the SNPs will alter the translated protein sequences, and thus interpreting the potential functional effects of these mutants will be crucial to elucidate the molecular basis of human diseases.

Prioritisation of residues to mutate to determine protein function:

The discovery of gene function will require a sustained experimental effort, which includes the creation of molecular mutants. The prioritisation of residues to mutate will be greatly optimised by considering the 3-D structure of the target protein.

Providing hints for protein function:

This is probably the broadest and least defined spectrum of potential applications for 3-D models. The common feature of these applications is that models can be used to formulate a hypothesis around a protein, which can then be tested in experimental settings. It is well known that low, yet significant, degrees of sequence similarity are often not sufficient to attribute a function to a protein. In such cases, protein modelling can provide useful insights and help determine or confirm a potential functional assignment. Furthermore, one can use models to create hypothesis around potential enzymatic activities and possible ligand binding functions.

Session 5
Chair: Minoru Kanehisa

Mechanistic structural and dynamics studies of proteins through solid state NMR

Ann McDermott, Tatyana Igumenova, Chad Rienstra and Sharon Rozovsky

Columbia University, Department of Chemistry

Membrane proteins are highly under-represented in our knowledge of protein structure and function; not only the structural biology, but also the enzymology, the thermodynamics, conformational dynamics and computational studies of these systems holds many puzzles for the future. We have carried out solid state NMR studies of protein motion for various enzymes, utilizing site-specific isotopic labels. For example, for Triose Phosphate Isomerase and cytochrome P-450 we have investigated aspects of the conformational dynamics under turnover conditions, on the timescale of the chemical catalytic events; our data offer insight into the anomalous isotope effects previous reported, and into the relation between sequence, dynamics and function. We are currently establishing and rehearsing solid state NMR methods for characterizing uniformly labeled proteins, particularly in terms of their conformational dynamics and structure. Our immediate targets include rather abundant and small membrane proteins involved in bioenergetics: LH1 and the c subunit of ATP synthetase. To illustrate and rehearse the methods that will be involved, we have assigned the multidimensional carbon and nitrogen NMR spectra of microcrystalline BPTI and ubiquitin, in collaboration with Montelione (Rutgers) and Zilm (Yale) and Wand (U. Penn). Furthermore, we have also demonstrated methods for detecting tertiary contacts based on long-range distance (ca. 4.5 6A) measurements. In conjunction with torsion measurements these will be used to attempt to determine the structures of all of these targets.

New challenges in drug design

Modesto Orozco. Departament de Bioquímica i Biologia Molecular.Facultat de Química. Universitat de Barcelona. Martí i Franquès 1. Barcelona 08028. Spain

The progress in genomic is generating a large amount of data that makes it necessary to modify the process of drug design development. There are new challenges in drug design, among others the development of: i) methods for target prediction, ii) new theoretical methods for structural prediction, iii) strategies for massive virtual screening, iv) more accurate methods for lead optimization, iv) techniques for detection of cross-interactions and secondary effects, and v) methods for the development of individual-oriented therapies.

Within this general issue, I will summarize the work of our group in the development of new strategies for the detection of cross-interactions, and secondary effects, presenting the different levels of accuracy necessary for the prediction of the therapeutic effect of drugs. I will also summarize the work that our group is performing in the development of 3-D databases of pathologic mutations, which can be used to gain insight into the development of individual-oriented therapies.

The role of three-dimensional cryo-electron microscopy in structural genomics

Carazo, J.M.(1), Barcena, M.(1), Chagoyen, M.(1), Jimenez-Lozano, N.(1), Fuller, S.(2), Radermacher, M. (3) and Barton G (4)

(1) National Center for Biotechnology-CSIC, Campus Universidad Autonoma, 28049 Madrid
(2) Division of Structural Biology, Wellcome Trust Center for Human Genetics, University of Oxford, Rossevelt Drive, OXford OX3 7BN, UK
(3) MPI for BioPhysics, Franfurt. (4) EMBL-European Bioinformatics Institute, Genome Campus, Hinxton, Cambs CB10 1SD, UK

This contribution analyes the possible role of the methodological approach commonly known as "Three-dimensional cryo electron microscopy (cryo-EM)" within the context of Structural Genomics. Indeed, cryo-EM complements other experimental techniques (namely NMR and X-ray crystallography) in the quest to help understanding biological functions through the structural characterization of macromolecules.

In brief, cryo-EM aims at solving the structure of biological macromolecules by combining thousands of transmission electron microscopy images of the specimen under study. The approach is quantitative, being one form of inverse problem found in the general area of "Tomography" (and formally identical to the case of Medical Tomography). In contrast with other techniques, cryo-EM allows the study of very large macromolecular complexes.

In a few cases cryo-EM has been able to contribute with the structure of a number of rather complicated proteins at atomic resolution by electron crystallography (i.e, bacteriorhodopsin, the light-harvesting complex, tubulin, AQP-1Š). Still, the bulk of results points at resolutions in the range between circa 0.5 to 2 nm. With resolutions still improving, 'single-particle' analyses as well as studies on icosahedral viruses are already depicting secondary structure. Moreover, cryo-EM can be combined in several ways with X-ray diffraction to enhance the resolution of cryo-EM and the applicability of crystallography. With these considerations in mind, we put forward the point of the critical role of 3D-EM in the Structural Genomics endeavor in the following areas:

1. Structural analysis at atomic resolution of a relatively small number of specially difficult-to-solve proteins (for instance, membrane proteins)

2. Structural analysis at a resolution of between circa 0.5 to 2nm of large macromolecules. A number of key applications for these medium resolution data are now possible:

   a. Study of conformational changes:3D-EM can provide medium-low resolution information of a given macromolecule with the aim of studying its polymorphism under different conditions such as pH, ionic strength, cofactors... The many works on ribosome structure and function by cryo-EM, to name just the case of one biological specimen, provides plenty of success stories on this matter.

   b. Studies of macromolecules too large to be solved by NMR or too flexible to be easily crystallized. Conceptually, the approach is like solving a puzzle in which the pieces are provided at atomic resolution by NMR or X-ray diffraction, but the pattern

only emerges when the information from cryo-EM comes into play. Extreme examples are the on-going works in the tomographic reconstructions of whole cells.

Accessibility and standardization of 3D-EM data has always been an issue, but the on-going database efforts such as the IIMS project ("Integrating Information on Macromolecular Structure, Coordinated by the European BioInformatics Institute, EBI) aims at providing an homogeneous environment in which cryo-EM data could be easily interfaced with X-rays and NMR data.

In essence, and in a somehow provocatively framed sentence open to myriad of comments, 3D-EM offers a path to "understand the cell at atomic resolution", rather than a set of pieces in search of function. (Bohm, 2000; Baumeister and Steven, 2000).

References:

* Bohm J, Frangakis AS, Hegerl R, Nickell S, Typke D, Baumeister W. Toward detecting and identifying macromolecules in a cellular context: Template matching applied to electron tomograms. Proc Natl Acad Sci U S A. 2000 Nov 21.

* Baumeister W, Steven AC, Macromolecular electron microscopy in the era of structural genomics. Trends Biochem Sci 2000 Dec 1;25(12):624-631

Combining bioinformatics biophysics to understand protein function

Barry Honig

Howard Hughes Medical Institute, Department of Biochemistry and Molecular Biophysics, Columbia University, 630 W. 168 St., New York, NY 10032

The increasing numbers of proteins whose three-dimensional structures have been determined will have major impact on the ability to exploit genomic data. Sequence alignments will become more meaningful, protein structure prediction will become more accurate, and the prediction of protein function will become increasingly refined and precise. Such developments will require that sequence, structure and biophysical information be fully integrated and correlated with biological data in as much detail as possible. We have been developing a series of computational tools with the goal of detecting relationships among amino acid sequence, protein structure and protein function. Some of these tools and their application to understanding the diverse biological functions of different protein families, and of specificity differences within families, will be described.

Function and specificity are often coded on the protein surface and are not necessarily evident from sequence and structural patterns. An interesting example is provided by the C2 domain family, whose role is to target proteins to different locations in the cell, often in a calcium dependent fashion. C2 domains that target different sub-cellular locations are quite similar in sequence and structure. As will be described, their specific function can only be understood when the physical chemical properties of the protein surface are considered in some detail. Our analysis of C2 domains, which is based in part on homology modeling, allows us to provide a more precise annotation of function than is possible with traditional methods. Another example of our approach is provided by SH2 domains that also mediate subcellular targeting. Based on a novel multiple structure alignment procedure we are to improve the detection of remote homologs and to obtain optimal sequence alignments. Our approach to predicting specificity is based on a quantitative analysis of the biophysical properties of binding sites of known structure: For each SH2/peptide complex we identify the residues that provide the most significant energetic contributions to binding and then map these onto sequence space. This allows us to identify residues that are potential important for binding specificity. Our predictions are in excellent agreement with experiment.

Our approach combines sequence, structure and biophysical methods in novel ways and will prove useful in the analysis of many other protein families. This type of integrated analysis can provide general insights that could not be obtained from standard bioinformatics techniques nor from the examination of individual proteins of known structure.

POSTERS

Clustering in sequence space. Identification of orthologous families

Federico Abascal and Alfonso Valencia

Understanding the way in which sequence similarity and functional similarity are related will be a key step for the prediction of function from the sequence information. In this sense sequence similarity have to be seen in the context of the relations between sequences forming families of related sequences in the same or different organisms. The evolutionary relationships between sequences are commonly described with a terminology imported from the field of paleontology and evolutionary studies. The basic, but difficult, concept of homology has been qualified in terms of ortologous homology and paralogous homology, perhaps for first time by (Fitch, 1970). Simplifying, orthologues are those genes that in different species have evolved from an common ancestral gene by speciation, whereas paralogues are those whose similarity is due to a duplication event inside a given genome. Orthologous genes tend to conserve function, while paralogues can acquire new functionalities.

We have tested a way of representing the sequence space as a graph and applied a clustering algorithm, based on the *Normalized Cut Algorithm* (Shi and Malik, 1997), to that space to find which groups of sequences are more strongly connected between them than to others. This analysis has been done in the genomic context, which is the only context that allows the identification of orthology relationships. For the validation of the results we have compared with COGs database (Tatusov *et al.*, 1997). Since we obtained many small clusters in the results, we devised a way of joining with neighbor clusters based on their distance and on the genome representation inside them that improved the results, making them much more similar to COGs. The reason why the clustering itself produced small clusters is that in the sequence context we used there were not enough sequences to define the limits of families and subfamilies, that were confused with phylogenetic biases.

References:

Fitch WM. (1970) Distinguishing homologous from analogous proteins. *Syst Zool.* 19(2):99-113.

Shi J, Malik J. (1997) Normalized cuts and image segmentation. *Proc. Of the IEEE Conf. On Comp. Vision and Pattern Recognition* 731-737.

Tatusov RL, Koonin EV, Lipman DJ. (1997) A Genomic Perspective on Protein Families. *Science* 278: 631-636.

pH dependence of the folding free energy – implication to the function of the proteins

E. Alexov and B. Honig*

Department of Biochemistry and Biophysics, Columbia University,
630W 168 Street, New York, NY 10032

The pH dependence of the free energy of folding was computed for a representative set of proteins from all SCOP classes. Electrostatic free energies were calculated with a method based on the Poisson Boltzmann equation that also accounts for conformational changes that accompany changes in ionization state. For most proteins studied, the electrostatic free energy has a "flat" region, i.e. a pH interval where the free energy is almost constant. In the majority of cases this region is centered at physiological pH. Incorrectly folded proteins, specifically the EMBL misfolded set, do not exhibit this pH independent behavior suggesting that native proteins have been specifically "designed" to be stable with respect to fluctuations in pH unless these are of functional importance.

Automated structure-based prediction of functional sites in proteins - Applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking

Patrick Aloy , Enrique Querol , Francesc X. Aviles & Michael J.E. Sternberg

A major problem in genome annotation is whether it is valid to transfer the function from a characterised protein to a homologue of unknown activity. This poster shows that one can employ a strategy that uses a structure-based prediction of protein functional sites to assess the reliability of functional inheritance. The first step was to automate and benchmark the evolutionary trace approach (Lichtarge, et al. J. Mol. Biol. 257, 342-358, 1996). From a multiple sequence alignment one identifies invariant polar and charged residues which are then mapped onto the protein structure. The predicted functional site is formed from spatial clusters of these invariant residues. For 79% of 86 proteins examined, the method will yield information about the observed functional site. The following strategy is proposed to assess the validity of inheriting the function from protein A to protein B. First predict the functional site for protein A without B and its close homologues. Then include B and its homologues and one infers that the proteins have related functions if at least one functional site remains. This procedure was tested on 18 pairs proteins with unrelated function and 70 sets of proteins with related function. The accuracy of prediction that two proteins have related functions was 94% compared to a random of 79% This automated method could be linked to schemes for genome annotation. Finally, we examined the use of functional site prediction in protein-protein and protein-DNA docking. The use of predicted functional sites is shown to filter putative docked complexes with a similar discrimination to that obtained by manually including biological information about active sites or DNA binding residues.

Domain combinations in archaeal, eubacterial and eukaryotic proteomes

Gordana Apic

Domains are the building blocks of all globular proteins, and are units of compact three-dimensional structure (Murzin et al., 1995; Orengo et al., 1997) as well as evolutionary units (Riley & Labedan, 1997). There is a limited repertoire of domain families (Chothia, 1992; Wolf et al., 2000), so that these domain families are duplicated and combined in different ways to form the set of proteins in a genome. Proteins are gene products, and at the level of genes, duplication, recombination and fusion are the processes that produce new genes. We attempt to gain an overview of these processes by studying the structural domains in the proteins of seven genomes from three different phylogenetic groups. The domain and superfamily definitions in the Structural Classification of Proteins Database (Murzin et al., 1995) are used, so that we can view all pairs of adjacent domains in genome sequences in terms of their superfamily combinations. We find 624 out of the 764 superfamilies in SCOP in these genomes, and the 624 families occur in 585 pairwise combinations. Most families are observed in combination with one or two other families, while a few families are very versatile in their combinatorial behaviour. This type of pattern can be described by a scale-free network. We also study the N-to-C terminal orientation of domain pairs and domain repeats. Finally, we compare the set of the domain combinations in the genomes to those in PDB, and discuss the implications for structural genomics.

References:

Chothia, C. (1992) One thousand families for molecular biologist. Nature, 357, 543-544.

Murzin, A., Brenner, S.E., Hubbard, T. & Chothia,C. (1995) SCOP: A structural classification of protains databasefor investigation of sequences and structures. J.Mol.Biol., 247, 536-540.

Orengo, C.A.,Michie, A.D., Jones.D.T., Swindells, M.B. & Thornton,J.M. (1997) CATH - a hierarchic classification of protein structure. Structure, 5, 1093-1098.

Riley, M. & Labedan, B. (1997) Protein evolution viewed through Escherichia coli protein sequnces : introducing the notion of a structural segment of homology, the module. J.Mol.Biol., 268, 857-868

Wolf, Y.I., Grishin, N.V. & Koonin. E.V. (2000) Estimating the number of protein folds and familes from complete genome data. J.Mol.Biol., 299, 897-905.

NMR structural analysis of hypotetical proteins of mth0677 and mth0121 proteins from *Methanobacterium thermoautotrophicum*

Francisco J. Blanco and Manuel. Rico

Instituto de Estructura de la Materia, CSIC Serrano 119, 28006 Madrid, Spain
paco@malika.iem.csic.es

The proteome of the archeon Methanobacterium Thermoautotrophicum is the subject of a comprehensive effort to test the feasibility of structural proteomics1. Two hypothetical sequences (mth0677 and mth0121) display solution NMR spectra with a dispersion of signals indicative of folded conformations (Arrowsmith et al., personnal communication). The structure of these proteins will be modeled from the amino acid sequence using threading and fold recognition methods2,3 with special attention to evaluate the possibility of the proteins adopting new folds. The solution structures will be solved by standard multinuclear and multidimensional NMR techniques. It is expected that structural based functional characterization will provide information about the functional role of these proteins3.

References

1.- Christendat et al., (2000) Nature Struct. Biol. 7, 903-909.
2.- Olmea et al., (1999) J. Mol. Biol. 295, 1221-1239.
3.- Sánchez et al., (2000) Nature Struct. Biol. 8S, 986-990.
4.- Thornton et al., (2000) Nature Struct. Biol. 8S, 991-994.

Structural predictions of *Escherichia coli* NDH-2

Rosana Chehín[1], Viviana A. Rapisarda[1], Ricardo N. Farías[1], Eddy M. Massa[1] & Javier De Las Rivas[2]

[1]INSIBIO(CONICET-UNT) - Tucumán - Argentina. *E-mail: rosana@unt.edu.ar*
[2]IRNA (CSIC) -Salamanca - España. *E-mail: jrivas@gugu.usal.es*

NADH dehydrogenase-2 (NDH-2) is a flavoprotein of the *E. coli* respiratory chain consisting of a single polypeptide species of MW 47,200 encoded by *ndh* gene. Our previous studies showed that NDH-2 promotes the electron transfer from NADH to Cu (II), which is reduced to Cu (I). The NDH-2 cupric reductase activity depends on the presence either FAD or quinone. The enzyme was not able to reduce Fe (III) to Fe (II) and thus, it appears to be a copper-specific metal ion reductase. We have strong evidences that NDH-2 contains one Cu (I) mole per polypeptide mole.

No crystal structure is currently available from NDH-2 from *E .coli.* In the present work, we use bioinformatics to analyse the secondary and tertiary structure of this protein. Multiple alignments with all the found homologous proteins were constructed and used as input for secondary structure and solvent accessibility predictions. The presence of specific patterns and motifs was also search. Our computational predictions indicate that NDH-2 has 4 structural and functional domains. At the N-terminus there is a two-fold repeated secondary structure of βαβαββββββ pattern. The sequence alignment between these two regions show 28% of identity and 48% of similarity, suggesting that they have a common ancestor. The first domain presents a FAD binding motif (domain I), and the second one a NAD^+ binding motif (domain II). At the C-terminus, two transmembrane helixes were detected suggesting a region of anchorage to the membrane (domain IV). Between domain II and IV there is a region which contains 3 of the 4 cysteines present in NDH-2. This region exhibits a high degree of amino acid identity with a Zn finger motif, suggesting a probable copper binding site (domain III). Seven threading methods were used to recognize the fold of each domain and to produce a three-dimensional (3D) model of the protein. The threading results clearly indicate that NDH-2 belongs to the α/β class and has a *3 layer* βαβ *sandwich* architecture (according to CATH classification). A remote 3D model of the protein was obtained with acceptable quality scores. The results presented in this work are very useful in the understanding of biochemical and functional properties of NDH-2 from *E. coli.*

Mapping disease-causing single amino acid polymorphisms to structure

Carles Ferrer, Modesto Orozco, Xavier de la Cruz

Departament de Bioquimica, Facultat de Quimica, Universitat de Barcelona
Marti i Franques, 1 08028 Barcelona, Spain

It is well known that variations in the consensus sequence of a protein can cause dramatic alterations in its function, leading to disease. Our work is currently focused in describing, in structural terms, those single aminoacid polimorphisms that are responsible for human diseases. To this end we analysed a set of human proteins for which disease-associated polymorphisms are known. Every pathological variant was described in terms of secondary structure, solvent accessibility, location in surface cavities and degree of conservation in multiple sequence alignments. In addition, we analyze changes in some physico-chemical properties associated with the mutations. We studied free energy variations derived from aminoacid partition coeficients, and variations in secondary structure propensities. In this poster we discuss the results of the previous analysis and suggest some general characteristics of pathological mutations.

Practical limits of function prediction

Devos & Valencia

The widening gap between sequences and functions has lead to the practice of assigning a potential function to an uncharacterised protein based on sequence similarity with other proteins of experimentaly investigated function. Even if the reliability of those homology based functional assignments is not well characterized, it represent common practises in whole genomes functional assignments. We propose here a systematic approach to the study of the margins of error in homology based functional prediction by analysing the conservation of the functional annotations in a large set of structural alignments. In particular, we analyze five aspects of protein function, commonly used in genome annotation, namely: i) PDB header line, ii) Enzymatic function classification: DE code, the standard definition of the chemical nature of the enzymatic function; iii) Functional annotations in the form of keywords, describing the biochemical function such as the interactions with compounds, cofactors, substrates, regulators and other cellular components; iv) Classes of cellular function, capturing the main types of cellular activities in which proteins participate, e.g. "carbon compound metabolism" or "DNA biosynthesis"; and v) Conservation of the type of amino acid in the binding site, related with the binding activity of the protein, and in many cases, the specificity of binding different substrates and cofactors. The screening of the full range of sequence functional similarities allows us to present an initial picture of the relation between sequence and functional similarity, and in particular, to derive a theoretical error rate for homology-based functionnal assignments (1). With those data, we estimate the theoretical error rates of predicted functions in different genomes. Indeed, it is particularly interesting to think in the consequences of this study for whole genome annotations carried out by automatic systems (2) and to compare the expected level of error with the different values published by different groups of expert annotators (3, 4, 5).

References:
1.Devos and Valencia. PROTEINS 41, 1 98-107 2.Andrade et all.. Bioinformatics 1999; 15:391-412 3.Brenner. Trens Genet. 1999; 15: 132-133 4.Galperin and Koonin. In Silico Biol. 1998; 1:0007 5.Ouzounis et all.. Mol. Microbiol. 1996; 20: 985-900

Finding protein families using an unsupervised neural network

Javier Herrero and Joaquín Dopazo

Bioinformatics Unit, CNIO. 20220 Majadahonda, Spain

The idea of clustering proteins into families in a way as unsupervised as possible is not new. Within the context of the projects of automatic annotation and given the growing rate experienced by molecular databases, the availability of an efficient procedure for clustering big amounts of data it is now more necessary than ever. Since the classical work by Ferran and Ferrara (1991) there have been different proposals based on different criteria, including neural networks (see Wu 1997 for a revision) and similarity (Yona et al, 2000). Here we present an approach based on a different type of neural network, the Self Organising Tree Algorithm (SOTA) (Dopazo and Carazo, 1997). This new version of SOTA combines multifurcation and bifurcation. The rationale of the new topology for the network is as follows: the model that accounts for the relationships behind all the proteins one may want to study must be strictly bifurcating but, due to the fact that the information has been overrun by the noise along the evolution, it is impossible to recover the phylogenetic tree linking all the proteins. The approach taken here implies two steps. Firstly, the system grows as a multifurcating tree, which is an approach similar to the one taken by using self organising maps (Ferran and Ferrara, 1991; Andrade et al., 1997) although in this case is free of the restriction of the number of clusters. Once the system has information enough to detect a signal significantly higher than the random noise, then the network grows as a bifurcating tree. In this way, families are defined automatically based on a pure information contents criterion, and subfamilies are defined following an evolutionary model. The growth of the system can be stopped at the desired level of heterogeneity. In this modification, SOTA uses dipeptide frequencies and other coding schemes (see Wang et al., 1998) instead of the aligned sequences.

References:

Andrade, MA, Casari, G. Sander, C. Valencia, A. (1997) Classification of protein families and detection of the determinant residues with an improved self-organzing map. Biol Cybern 76:441-450. Dopazo, J., Carazo, J.M. (1997) Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. J. Mol. Evol 44:226-233. Ferran E.A. and Ferrara P. (1991) Topological maps of protein sequences. Biol Cybern 65:451-458. Wang, HC, Dopazo, J., de la Fraga, LG, Zhu, YP Carazo, JM (1998) Self-organising tree-growing network for the classification of protein sequences. Prot. Sci. 7:1-10. Wu, CH (1997) Artificial neural networks for molecular sequence analysis. Comput. Chem 21:237-256. Yona, G., Linial, N., Linial, M (2000) ProtoMap: automatic classification of protein sequences and hyerarchy of protein families. Nucl. Acids Res. 28:49-55.

GRATH

Andrew Harrison, Frances Pearl, Tim Slidel, Janet Thornton and Christine Orengo

We have produced an algorithm, GRATH, that is able to provide rapid and accurate fold assignments for any novel protein. GRATH transforms secondary structures into vectors and turns the fold of a protein into a graph of the vectors. By comparing two graphs and looking for cliques, GRATH is able to find the largest amount of structural overlap between two proteins. GRATH extends the work of Grindley et al. in several ways. Firstly, our graphs include more information about the geometrical relationships between secondary structure. Also, we have tested empirically its fold assignments against the CATH database and so have been able to determine its reliability. We find that GRATH predicts the correct fold with an accuracy of 90%. GRATH is particularly accurate for large domains (> 8 secondary structure within the domain).

Homology modeling methodology improvement for low homologous protein sequences

C. Lambert , N. Léonard, X. De Bolle and E. Depiereux

Unité de Recherche en Biologie Moléculaire, Facultés Universitaires Notre-Dame de la Paix, rue de Bruxelles 61, 5000 Namur, Belgium

The aim of our work is to propose a reliable methodology for homology modeling, especially when the protein of interest shares a low percentage of identities (20-30%) with the template.

Our strategy consists in the blind modeling of PDB's proteins for which the closest structure has an identity rate between 20 and 30%. Similar sequences are fetched (PSI-BLAST[1]) in a non-redundant sequence databank. Then, as far as possible, two sets of sequences are built. The first one contains all the best hits above a given similarity cutoff (E value). The second one contains a subset of the sequences, after dropping too redundant ones. This method aims at creating different conditions to run multiple alignment programs and extracting different consensus and in order to raise the confidence of the sequence-structure alignment.

The two sets are then submitted to five alignment programs: ClustalW[7], Dialign2[5], Match-Box[3], Multalin[2] and PRRP [4]. A pairwise alignment between the target and template sequences is extracted from each multiple alignment and the final sequence-structure alignment is obtained from the consensus between all the pairwise alignments. A tri-dimensional model is built using MODELLER[6] on this final alignment. As a control, another model is built from the rough sequence-structure alignment provided by PSI-BLAST[1], and compared with the model obtained using our methodology.
The last steps of our scheme is the a priori assessment of each model using statistical methods (Procheck, Whatcheck, Verify 3D) and its final validation by the comparison with the crystallographic structure. The global RMS between the model and the real structure and the length of boxes closer than 1.0Å RMS are taken into account to evaluate the model quality.

References:
1. Altschul SF, et al. (1997). Nucleid Acids Research 25(17): 3389-3402 2. Corpet F (1988) Nucl. Acids Res. 16:10881-10890. 3. Depiereux E, et al. (1997). Comput. Appl. Biosci. 13(3): 249-256. 4. Gotoh O (1996) J. Mol. Biol. 264:823-838 5. Morgenstern, B. (1999). Bioinformatics 15(3): 211-8. 6. Sali A and Blundell TL (1993). Journal of Molecular Biology 234(3): 779-815. 7. Thompson JD, et al. (1994). Nucleic Acid Research 22(22): 4673-4680.

Similarity of phylogenetic trees as indicator of protein-protein interaction

Florencio Pazos and Alfonso Valencia

Most cellular functions involve protein-protein interactions. The deciphering of the large set of interacting pairs of proteins in an organism has lead to the development of high throughput experimental techniques, such as yeast two hybrid screening. Only recently some computational approaches have emerged trying to determine the possible interactions using information related with the protein sequences. For systems such as ligands and receptors it has been previouly proposed a corelation between their phylogenetic trees based on the notion of co-evolution, a principle that seems to be aplicable to the few isolated cases studied so far. Here we test in statistical terms if the amount of information that can be retrieved from the study of the similarity between evolutionary trees is enough for detecting interacting proteins among a large collection of possible partners.

WILMA - a genome annotation database for the nematode *C. elegans*

Andreas Prlic

Genome annotation involves two major steps: The first is to identify the location of the coding regions on the genome ("gene structural annotation"), the second is to infer biologically relevant information ("functional annotation").

Here, we focus on functional annotation of the approximately 19.000 predicted protein sequences of the nematode *C. elegans*. We applied a collection of different protein sequence analysis tools in an automated way. These tools include sequence searches, sequence patterns, transmembrane predictions, signal peptide predictions, protein structure predictions among other methods. Additional *C. elegans* information available in public databases was also collected, e.g. cDNA data, exon/intron borders or names of gene loci.

This vast amount of biological information has been stored in a relational database, WILMA. An easy to use web interface allows access to all results as well as an integrated view of the data belonging to a certain *C. elegans* protein sequence. In addition, SQL commands can be used to investigate relationships between different types of data and derive statistics on a genome wide scale in short time.

Protein cavities created by mutation: a precise structural prediction by molecular mechanics in the CHARMm1 force field

Machicado C., Bueno, M., & Sancho J.

Department of Biochemistry and Molecular and Celular Biology, University of Zaragoza
50009-Zaragoza (Spain)

Protein cavities might be useful for a number of purposes. If the structure of a protein is known, they can be easily created by site directed mutagenesis, and their size and shape can, in principle be predetermined, provided they do not collapse. X-ray studies on T4 lysozyme mutants2 indicate, however, than cavity collapses are not infrequent events, which certainly complicates the design.

As a tool to assist the efficient design of protein cavities we have sought to develop a minimisation strategy that can predict with precision the fate of cavities created by mutagenesis. To that end, we have performed energy minimisation of cavity forming-mutants of T4 lysozyme by both constrained and unconstrained pathways. In the constrained path, often recommended for energy minimisation, hydrogens were relaxed first, then side chains, and finally the whole molecule. Despite the logic of the approach, we find the unconstrained path fit the experimental structures better than the constrained one. We have also assessed if an initial steepest descents step, performed before the conjugate gradient step, improves the minimisation. Our data indicate the steepest descents step is unnecessary. To determine the reliability of the all-atom atom approximation, we also performed all of the above minimisations with united atom models. This approximation gave structures with similar, only slightly higher, RMS deviations than the all-atom model, and a 60-70% savings in computer time.

Based on these results, we propose for energy minimisation of protein cavities the use of an unconstrained path with conjugate gradients and an all-atom atom representation. Using this procedure we have minimised the structure of twelve virtual mutations performed, one by one, on the structure of native T4 lysozyme. The minimised structures closely fit the crystal structures of the corresponding mutants (0.3-0.6 Å root-mean-square deviation in the position of atoms within 6 Å of the newly introduced side chain). It seems thus that the structure of protein cavities generated by mutagenesis can be confidently simulated regardless of the cavity tendency to collapse.

References:

(1) CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations, J. Comp. Chem. 4, 187-217 (1983), by B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus.

(2) Generation of ligand binding sites in T4 lysozyme by deficiency-creating substitutions. Baldwin E, Baase WA, Zhang Xj, Feher V, Matthews BW. J Mol Biol. 1998 277:467-85.

Completeness in structural genomics

Dennis Vitkup[1,2], Eugene Melamud[3], John Moult[3], Chris Sander[1,4]

[1]MIT Center for Genome Research, One Kendall Square, Building 300, Cambridge, MA, 02139, USA
[2]Department of Chemistry and Chemical Biology, Harvard University Cambridge, Massachusetts 02138, USA
[3]Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, 9600 Gudelsky Drive, Rockville, MD, 20850
[4]Millennium Pharmaceuticals, 640 Memorial Drive, Cambridge, MA, 02139, USA

Structural genomics has the goal of obtaining useful three-dimensional models of all proteins, by a combination of experimental structure determination and comparative model building. Currently, useful models of high quality are available only for a minor fraction of protein domains. We estimate the scope of structural genomics, i.e., the total effort required to make models available for nearly all protein domains. We then evaluate different strategies for reaching optimal information return on effort. Taking into account the strong dependence of model quality on sequence similarity between template and target, we perform test calculations in the Pfam protein family database simulating varying scenarios of how to cover protein space by 3D models. For the protein space considered, the strategy which maximizes structural coverage requires about 7 times fewer structure determinations compared with the strategy in which targets are selected at random. As most protein families provide a large number of alternative targets, broad coverage of can be achieved even with a relatively low success rate of protein structure determination. With a choice of reasonable model quality corresponding to 30% sequence identity and the goal of 90% coverage, we extrapolate from the fraction of residues in complete genomes which can be assigned to Pfam families and estimate the total effort of structural genomics. Using an optimized strategy for achieving completeness, it would take on the order of 16,000 carefully selected structure determinations to construct useful atomic models for the vast majority of all proteins. In practice, a variety of approaches to target selection is likely to increase the estimated total effort by a factor of three. This provides a strong incentive for global coordination of target selection.

LIST OF INVITED SPEAKERS

Christopher M. Dobson	Oxford Centre for Molecular Sciences, University of Oxford, New Chemistry Laboratory. South Parks Road, Oxford OX1 3QT (UK). Tel.: 44 1865 275 916. Fax: 44 1865 275 921. E-mail: chris.dobson@chem.oxford.ac.uk
Terry Gaasterland	Laboratory of Computational Genomics. The Rockefeller University. 1230 York Avenue, New York, NY. 10021-6399 (USA). Tel.: 1 212 327 77 55. Fax: 1 212 327 7765. E-mail: gaasterland@rockefeller.edu
Liisa Holm	Structural Genomics Group. EMBL-EBI, Cambridge CB10 1SD (UK). Tel.: 44 1223 494 454. Fax: 44 1223 494 468. E-mail: holm@ebi.ac.uk
Barry Honig	HHMI, Dept. of Biochemistry and Molecular Biophysics, Columbia University. 630 W. 168 St., New York, NY. 10032 (USA). Fax: 1 212 305 6926. E-mail: bh6@columbia.edu
David T. Jones	Institute for Cancer Genetics and Pharmacogenomics, Dept. of Biological Sciences, Brunel University. Uxbridge, Middlesex UB8 3PH (UK). Tel.: 44 1895 81 62 41. Fax: 44 1895 27 43 48. E-mail: David.Jones@brunel.ac.uk
Minoru Kanehisa	Bioinformatics Center, Institute for Chemical Research, Kyoto University. Uji, Kyoto 611-0011 (Japan). Tel.: 81 774 38 32 70. Fax: 81 774 38 32 69. E-mail: kanehisa@kuicr.kyoto-u.ac.jp
Christopher D. Lima	Weill Medical College of Cornell University. 1300 York Ave., New York, NY. 10021 (USA). Tel.: 1 212 746 6449. Fax: 1 212 746 4843. E-mail: lima@pinky.med.cornell.edu
Michal Linial	Dept. of Biological Chemistry. Life Science Institute, Givat Ram Campus, The Hebrew University, Jerusalem 91904 (Israel). Tel.: 972 2 658 54 25. Fax: 972 2 658 64 48. E-mail: michall@mail.ls.huji.ac.il
Ann McDermott	Columbia University, Dept. of Chemistry. 3000 Broadway, New York, NY. 10027 (USA). Fax: 1 212 932 12 89. E-mail: aem5@columbia.edu
Alexey G. Murzin	Centre for Protein Engineering, MRC Centre. Hills Road, Cambridge CB2 2QH (UK). Tel.: 44 1223 402 132. Fax: 44 1223 402 140. E-mail: agm@mrc-lmb.cam.ac.uk

Christine Orengo	Dept. of Biochemistry and Molecular Biology. University College. Gower Street, London WC1E 6BT (UK). Fax: 44 207 380 71 93. E-mail: orengo@biochemistry.ucl.ac.uk
Modesto Orozco	Departament de Bioquímica i Biologia Molecular. Facultat de Química. Universitat de Barcelona. Martí i Franquès 1, Barcelona 08028 (Spain). Tel.: 34 93 402 17 19. Fax: 34 93 402 12 19. E-mail: modesto@luz.bq.ub.es
Manuel C. Peitsch	GlaxoSmithKline and Swiss Institute of Bioinformatics. WTC I, 10, route de l'aéroport, 1215 Genève (Switzerland). Tel.: 41 22 799 43 01. Fax: 41 22 799 43 10. E-mail: manuel.peitsch@isb-sib.ch
Burkhard Rost	CUBIC, Columbia University. 630 West, 168 Street, New York, NY. 10032 (USA). Tel.: 1 212 305 3773. Fax: 1 212 305 7932. E-mail: rost@columbia.edu
Andrej Sali	The Rockefeller University. 1230 York Avenue, New York, NY. 10021-6399 (USA). Tel.: 1 212 327 7550. Fax: 1 212 327 7540. E-mail: sali@rockefeller.edu
Chris Sander	MIT Center for Genome Research, Cambridge, MA. (USA). Tel.: 1 617 252 1900. E-mail: sander@genome.wi.mit.edu
Manfred J. Sippl	University of Salzburg, Center of Applied Molecular Engineering. Jakob-Haringer-Str. 3, 5020 Salzburg (Austria). Tel.: 43 662 8044 5797. Fax: 43 662 45 48 89. E-mail: sippl@came.sbg.ac.at
Janet M. Thornton	Univ. College and Birkbeck College, Biochemistry and Molecular Biology Dept. Gower Street, London WC1E 6BT (UK). Tel.: 44 20 7679 7048. Fax: 44 20 7679 7193. E-mail: thornton@biochemistry.ucl.ac.uk
Alfonso Valencia	Protein Design Group. CNB-CSIC. Cantoblanco. Madrid 28049 (Spain). Tel.: 34 91 585 45 70. Fax: 34 91 585 45 06. E-mail: valencia@cnb.uam.es

LIST OF PARTICIPANTS

Federico Abascal	Protein Design Group. Centro Nacional de Biotecnología. Campus Universidad Autónoma. Cantoblanco, 28049 Madrid (Spain). Tel.: 34 91 585 46 69. Fax: 34 91 585 45 06. E-mail: fabascal@cnb.uam.es
Emil Alexov	Department of Biochemistry and Biophysics. Columbia University. 630W 168 Street, New York, NY. 10032 (USA). Tel.: 212 305 02 65. Fax: 212 305 69 26. E-mail: ea388@columbia.edu
Patrick Aloy	European Molecular Biology Laboratory. Meyerhofstrasse, 1, 69117 Heidelberg (Germany). Tel.: 49 6221 387 306. Fax: 49 6221 387 519. E-mail: Patrick.Aloy@EMBL-Heidelberg.de
Gordana Apic	Lab. of Molecular Biology. MRC/University of Cambridge, Cambridge CB2 2QH (UK). Tel.: 44 1223 402479. Fax: 44 1223 213556. E-mail: apic@mrc-lmb. cam.ac.uk
Montserrat Bárcena	National Center for Biotechnology-CSIC. Campus Univ. Autónoma, 28049 Madrid (Spain). Tel.: 34 91 585 45 10. Fax: 34 91 585 45 06. E-mail: montse@cnb.uam.es
Francisco J. Blanco	Instituto de Estructura de la Materia, CSIC. Serrano 119. 28006 Madrid (Spain). Tel.: 34 91 561 68 00. Fax: 34 91 564 55 57. E-mail: paco@malika.iem.csic.es
Christine M. Blaumueller	EMBO Reports. Meyerhofstrasse 1, 69117 Heidelberg (Germany). Tel.: 49 6221 387 595. Fax: 49 6221 387 563. E-mail: Christine.Blaumueller@embo.org
José M. Carazo	National Center for Biotechnology-CSIC. Campus Universidad Autónoma, 28049 Madrid (Spain). Tel.: 34 91 585 45 43. Fax: 34 91 585 45 06. E-mail: carazo@cnb.uam.es
Mónica Chagoyen	National Center for Biotechnology-CSIC. Campus Univ. Autónoma, 28049 Madrid (Spain). Tel.: 34 91 585 45 10. Fax: 34 91 585 45 06. E-mail: monica@cnb.uam.es
Rosana Chehín	INSIBIO(CONICET-UNT). Chacabuco 461, 4000 Tucumán (Argentine). Tel.: 54 381 424 89 21. Fax: 54 381 250 686. E-mail: rosana@unt.edu.ar

Xiaojun Cheng	Memorial Sloan -Kettering Cancer Center. 1275 York Avenue, New York, NY. 10021 (USA). Tel.: 1 212 639 58 98. Fax: 1 212 717 36 27. E-mail: x-cheng@ski. mskcc.org
Xavier de la Cruz	Departament de Bioquimica, Facultat de Quimica. Universitat de Barcelona. Marti i Franques, 1, 08028 Barcelona (Spain). Tel.: 34 93 403 58 58. Fax: 34 93 402 12 19. E-mail: xavier@husky.bq.ub.es
Damien Devos	Protein Design Group. CNB-CSIC, Madrid 28049 (Spain). Tel.: 34 91 585 48 39. Fax: 34 91 585 45 06. E-mail: devos@ cnb.uam.es
Joaquín Dopazo	Bioinformatics Unit, CNIO, 20220 Majadahonda (Spain). Tel.: 34 91 509 70 69. Fax: 34 91 509 70 55. E-mail: jdopazo@cnio.es
Ernest Feytmans	Biochemistry Unit, FMS. The University of the West Indies. Eric Williams Medical Sciences Complex, Mt.Hope (Trinidad and Tobago). Tel.: 1 868 662 18 73. Fax: 1 868 662 18 73. E-mail: feytmans@yahoo.com
Federico Gago	Departamento de Farmacología. Universidad de Alcalá, 28871 Madrid (Spain). Tel.: 34 91 885 45 14. Fax: 34 91 885 45 91. E-mail: fgago@fisfer.alcala.es
Raquel García-Nieto	Departamento de Farmacología. Universidad de Alcalá, 28871 Alcalá de Henares (Spain). Tel.: 34 91 885 45 14. Fax: 34 91 885 45 91. E-mail: raquel.moreno@univ.uah.es
Paulino Gómez-Puertas	Protein Design Group. CNB-CSIC. Cantoblanco, Madrid 28049 (Spain). Tel.: 34 91 585 48 39. Fax: 34 91 585 45 06. E-mail: pagomez@cnb.uam.es
Andrew Harrison	Univ. College and Birkbeck College. University College London. Gower Street, London WC1E 6BT (UK). Tel.: 44 20 7 679 38 90. Fax: 44 207 679 71 93. E-mail: harry@ biochem.ucl.ac.uk
Jens B. Jensen	Display Systems Biotech A/S, Copenhagen (Denmark). Tel.: 45 70 222 202. Fax: 45 70 232 304. E-mail: jbj@displaysystems.com
Christophe Lambert	Unité de Recherche en Biologie Moléculaire. Facultés Universitaires Notre-Dame de la Paix. rue de Bruxelles 61, 5000 Namur (Belgium). Tel.: 32 81 72 4417. Fax: 32 81 72 4420. E-mail: Christophe.lambert@fundp.ac.be

Oscar Lao	Unitat de Biologia Evolutiva. Universitat Pompeu Fabra. Dr. Aiguader 80, 08003 Barcelona (Spain). Tel.: 34 93 542 28 39. Fax: 34 93 542 28 02. E-mail: oscar.lao@cexs.upf.es
Felipe A. Lombó	Universidad de Oviedo- IUOPA. Julián Clavería s/n, 33006 Oviedo (Spain). Tel.: 34 985 10 35 58. Fax: 34 985 10 31 48. E-mail: flb@sauron.quimica.uniovi.es
Debora Marks	Harvard Medical School. 250 Longwood Ave., Boston, MA. (USA). Tel.: 1 617 432 5132. Fax: 1 617 738 0516. E-mail: debbie@hms.harvard.edu
Joachim Meyer	LION Bioscience AG. Waldhoferstr. 98, Heidelberg 69027 (Germany). Tel.: 49 6221 4038 378. Fax: 49 6221 4038 201. E-mail: Joachim.Meyer@lionbioscience.com
Nebojsa Mirkovic	The Rockefeller University. 1230 York Avenue, New York, NY. 10021 (USA). Tel.: 1 212 327 72 06. Fax: 1 212 327 75 40. E-mail: mirkovn@rockvax.rockefeller.edu
Seán I. O'Donoghue	LION Bioscience AG. Waldhoferstr. 98, Heidelberg 69123 (Germany). Tel.: 49 6221 4038 363. Fax: 49 6221 4038 290. E-mail: odonoghue@lionbioscience.com
Florencio Pazos	Protein Design Group. CNB-CSIC, Madrid 28049 (Spain). Tel.: 34 91 585 46 69. Fax: 34 91 585 45 06. E-mail: pazos@gredos.cnb.uam.es
Manuel Pérez-Alonso	Univ. de Valencia, 46100 Burjasot, Valencia (Spain). Tel.: 34 96 398 31 79. Fax: 34 96 386 43 72. E-mail: alonsom@uv.es
Andreas Prlic	Center of Applied Molecular Engineering. University of Salzburg. Jakob Haringerstr. 3, 5020 Salzburg (Austria). Tel.: 43 662 804 457 98. Fax: 43 662 454 889. E-mail: andreas@came.sbg.ac.at
Javier Sancho	Department of Biochemistry and Molecular and Celular Biology. University of Zaragoza, 50009 Zaragoza (Spain). Tel.: 34 976 761 286. Fax: 34 976 762 123. E-mail: jsancho@posta.unizar.es
Dennis Vitkup	MIT Center for Genome Research. One Kendall Square, Building 300, Cambridge, MA.02139 (USA). Tel.: 1 617 495 4102. Fax: 1 617 496 4793. E-mail: vitkup@genome.wi.mit.edu
Daniel H. Wainstock	Cell and Molecular Cell. 1100 Massachusetts Avenue, Cambridge, MA. 02138 (USA). Tel.: 1 617 397 2825. Fax: 1 617 397 2810. E-mail: dwainstock@cell.com

Texts published in the
SERIE UNIVERSITARIA
by the
FUNDACIÓN JUAN MARCH
concerning workshops and courses organized within the Plan for International Meetings on Biology (1989-1991)

*: Out of stock.

*246 **Workshop on Tolerance: Mechanisms and Implications.**
Organizers: P. Marrack and C. Martínez-A.

*247 **Workshop on Pathogenesis-related Proteins in Plants.**
Organizers: V. Conejero and L. C. Van Loon.

*248 **Course on DNA - Protein Interaction.**
M. Beato.

*249 **Workshop on Molecular Diagnosis of Cancer.**
Organizers: M. Perucho and P. García Barreno.

*251 **Lecture Course on Approaches to Plant Development.**
Organizers: P. Puigdomènech and T. Nelson.

*252 **Curso Experimental de Electroforesis Bidimensional de Alta Resolución.**
Organizer: Juan F. Santarén.

253 **Workshop on Genome Expression and Pathogenesis of Plant RNA Viruses.**
Organizers: F. García-Arenal and P. Palukaitis.

254 **Advanced Course on Biochemistry and Genetics of Yeast.**
Organizers: C. Gancedo, J. M. Gancedo, M. A. Delgado and I. L. Calderón.

*255 **Workshop on the Reference Points in Evolution.**
Organizers: P. Alberch and G. A. Dover.

*256 **Workshop on Chromatin Structure and Gene Expression.**
Organizers: F. Azorín, M. Beato and A. A. Travers.

257 **Lecture Course on Polyamines as Modulators of Plant Development.**
Organizers: A. W. Galston and A. F. Tiburcio.

*258 **Workshop on Flower Development.**
Organizers: H. Saedler, J. P. Beltrán and J. Paz-Ares.

*259 **Workshop on Transcription and Replication of Negative Strand RNA Viruses.**
Organizers: D. Kolakofsky and J. Ortín.

*260 **Lecture Course on Molecular Biology of the Rhizobium-Legume Symbiosis.**
Organizer: T. Ruiz-Argüeso.

261 **Workshop on Regulation of Translation in Animal Virus-Infected Cells.**
Organizers: N. Sonenberg and L. Carrasco.

*263 **Lecture Course on the Polymerase Chain Reaction.**
Organizers: M. Perucho and E. Martínez-Salas.

*264 **Workshop on Yeast Transport and Energetics.**
Organizers: A. Rodríguez-Navarro and R. Lagunas.

265 **Workshop on Adhesion Receptors in the Immune System.**
Organizers: T. A. Springer and F. Sánchez-Madrid.

*266 **Workshop on Innovations in Proteases and Their Inhibitors: Fundamental and Applied Aspects.**
Organizer: F. X. Avilés.

267 **Workshop on Role of Glycosyl-Phosphatidylinositol in Cell Signalling.**
Organizers: J. M. Mato and J. Larner.

268 **Workshop on Salt Tolerance in Microorganisms and Plants: Physiological and Molecular Aspects.**
Organizers: R. Serrano and J. A. Pintor-Toro.

269 **Workshop on Neural Control of Movement in Vertebrates.**
Organizers: R. Baker and J. M. Delgado-García.

Texts published by the CENTRE FOR INTERNATIONAL MEETINGS ON BIOLOGY

1 **Workshop on What do Nociceptors Tell the Brain?**
Organizers: C. Belmonte and F. Cerveró.

*2 **Workshop on DNA Structure and Protein Recognition.**
Organizers: A. Klug and J. A. Subirana.

*3 **Lecture Course on Palaeobiology: Preparing for the Twenty-First Century.**
Organizers: F. Álvarez and S. Conway Morris.

*4 **Workshop on the Past and the Future of Zea Mays.**
Organizers: B. Burr, L. Herrera-Estrella and P. Puigdomènech.

*5 **Workshop on Structure of the Major Histocompatibility Complex.**
Organizers: A. Arnaiz-Villena and P. Parham.

*6 **Workshop on Behavioural Mechanisms in Evolutionary Perspective.**
Organizers: P. Bateson and M. Gomendio.

*7 **Workshop on Transcription Initiation in Prokaryotes**
Organizers: M. Salas and L. B. Rothman-Denes.

*8 **Workshop on the Diversity of the Immunoglobulin Superfamily.**
Organizers: A. N. Barclay and J. Vives.

9 **Workshop on Control of Gene Expression in Yeast.**
Organizers: C. Gancedo and J. M. Gancedo.

*10 **Workshop on Engineering Plants Against Pests and Pathogens.**
Organizers: G. Bruening, F. García-Olmedo and F. Ponz.

11 **Lecture Course on Conservation and Use of Genetic Resources.**
Organizers: N. Jouve and M. Pérez de la Vega.

12 **Workshop on Reverse Genetics of Negative Stranded RNA Viruses.**
Organizers: G. W. Wertz and J. A. Melero.

*13 **Workshop on Approaches to Plant Hormone Action**
Organizers: J. Carbonell and R. L. Jones.

*14 **Workshop on Frontiers of Alzheimer Disease.**
Organizers: B. Frangione and J. Ávila.

*15 **Workshop on Signal Transduction by Growth Factor Receptors with Tyrosine Kinase Activity.**
Organizers: J. M. Mato and A. Ullrich.

16 **Workshop on Intra- and Extra-Cellular Signalling in Hematopoiesis.**
Organizers: E. Donnall Thomas and A. Grañena.

*17 **Workshop on Cell Recognition During Neuronal Development.**
Organizers: C. S. Goodman and F. Jiménez.

18 **Workshop on Molecular Mechanisms of Macrophage Activation.**
Organizers: C. Nathan and A. Celada.

*19 **Workshop on Viral Evasion of Host Defense Mechanisms.**
Organizers: M. B. Mathews and M. Esteban.

*20 **Workshop on Genomic Fingerprinting.**
Organizers: M. McClelland and X. Estivill.

21 **Workshop on DNA-Drug Interactions.**
Organizers: K. R. Fox and J. Portugal.

*22 **Workshop on Molecular Bases of Ion Channel Function.**
Organizers: R. W. Aldrich and J. López-Barneo.

*23 **Workshop on Molecular Biology and Ecology of Gene Transfer and Propagation Promoted by Plasmids.**
Organizers: C. M. Thomas, E. M. H. Willington, M. Espinosa and R. Díaz Orejas.

*24 **Workshop on Deterioration, Stability and Regeneration of the Brain During Normal Aging.**
Organizers: P. D. Coleman, F. Mora and M. Nieto-Sampedro.

25 **Workshop on Genetic Recombination and Defective Interfering Particles in RNA Viruses.**
Organizers: J. J. Bujarski, S. Schlesinger and J. Romero.

26 **Workshop on Cellular Interactions in the Early Development of the Nervous System of Drosophila.**
Organizers: J. Modolell and P. Simpson.

*27 **Workshop on Ras, Differentiation and Development.**
Organizers: J. Downward, E. Santos and D. Martín-Zanca.

*28 **Workshop on Human and Experimental Skin Carcinogenesis.**
Organizers: A. J. P. Klein-Szanto and M. Quintanilla.

*29 **Workshop on the Biochemistry and Regulation of Programmed Cell Death.**
Organizers: J. A. Cidlowski, R. H. Horvitz, A. López-Rivas and C. Martínez-A.

*30 **Workshop on Resistance to Viral Infection.**
Organizers: L. Enjuanes and M. M. C. Lai.

31 **Workshop on Roles of Growth and Cell Survival Factors in Vertebrate Development.**
Organizers: M. C. Raff and F. de Pablo.

32 **Workshop on Chromatin Structure and Gene Expression.**
Organizers: F. Azorín, M. Beato and A. P. Wolffe.

*33 **Workshop on Molecular Mechanisms of Synaptic Function.**
Organizers: J. Lerma and P. H. Seeburg.

*34 **Workshop on Computational Approaches in the Analysis and Engineering of Proteins.**
Organizers: F. S. Avilés, M. Billeter and E. Querol.

35 **Workshop on Signal Transduction Pathways Essential for Yeast Morphogenesis and Cell Integrity.**
Organizers: M. Snyder and C. Nombela.

36 **Workshop on Flower Development.**
Organizers: E. Coen, Zs. Schwarz-Sommer and J. P. Beltrán.

*37 **Workshop on Cellular and Molecular Mechanism in Behaviour.**
Organizers: M. Heisenberg and A. Ferrús.

38 **Workshop on Immunodeficiencies of Genetic Origin.**
Organizers: A. Fischer and A. Arnaiz-Villena.

39 **Workshop on Molecular Basis for Biodegradation of Pollutants.**
Organizers: K. N. Timmis and J. L. Ramos.

*40 **Workshop on Nuclear Oncogenes and Transcription Factors in Hematopoietic Cells.**
Organizers: J. León and R. Eisenman.

*41 **Workshop on Three-Dimensional Structure of Biological Macromolecules.**
Organizers: T. L Blundell, M. Martínez-Ripoll, M. Rico and J. M. Mato.

42 **Workshop on Structure, Function and Controls in Microbial Division.**
Organizers: M. Vicente, L. Rothfield and J. A. Ayala.

*43 **Workshop on Molecular Biology and Pathophysiology of Nitric Oxide.**
Organizers: S. Lamas and T. Michel.

*44 **Workshop on Selective Gene Activation by Cell Type Specific Transcription Factors.**
Organizers: M. Karin, R. Di Lauro, P. Santisteban and J. L. Castrillo.

45 **Workshop on NK Cell Receptors and Recognition of the Major Histocompatibility Complex Antigens.**
Organizers: J. Strominger, L. Moretta and M. López-Botet.

46 **Workshop on Molecular Mechanisms Involved in Epithelial Cell Differentiation.**
Organizers: H. Beug, A. Zweibaum and F. X. Real.

47 **Workshop on Switching Transcription in Development.**
Organizers: B. Lewin, M. Beato and J. Modolell.

48 **Workshop on G-Proteins: Structural Features and Their Involvement in the Regulation of Cell Growth.**
Organizers: B. F. C. Clark and J. C. Lacal.

*49 **Workshop on Transcriptional Regulation at a Distance.**
Organizers: W. Schaffner, V. de Lorenzo and J. Pérez-Martín.

50 **Workshop on From Transcript to Protein: mRNA Processing, Transport and Translation.**
Organizers: I. W. Mattaj, J. Ortín and J. Valcárcel.

51 **Workshop on Mechanisms of Expression and Function of MHC Class II Molecules.**
Organizers: B. Mach and A. Celada.

52 **Workshop on Enzymology of DNA-Strand Transfer Mechanisms.**
Organizers: E. Lanka and F. de la Cruz.

53 **Workshop on Vascular Endothelium and Regulation of Leukocyte Traffic.**
Organizers: T. A. Springer and M. O. de Landázuri.

54 **Workshop on Cytokines in Infectious Diseases.**
Organizers: A. Sher, M. Fresno and L. Rivas.

55 **Workshop on Molecular Biology of Skin and Skin Diseases.**
Organizers: D. R. Roop and J. L. Jorcano.

56 **Workshop on Programmed Cell Death in the Developing Nervous System.**
Organizers: R. W. Oppenheim, E. M. Johnson and J. X. Comella.

57 **Workshop on NF-κB/IκB Proteins. Their Role in Cell Growth, Differentiation and Development.**
Organizers: R. Bravo and P. S. Lazo.

58 **Workshop on Chromosome Behaviour: The Structure and Function of Telomeres and Centromeres.**
Organizers: B. J. Trask, C. Tyler-Smith, F. Azorín and A. Villasante.

59 **Workshop on RNA Viral Quasispecies.**
Organizers: S. Wain-Hobson, E. Domingo and C. López Galíndez.

60 **Workshop on Abscisic Acid Signal Transduction in Plants.**
Organizers: R. S. Quatrano and M. Pagès.

61 **Workshop on Oxygen Regulation of Ion Channels and Gene Expression.**
Organizers: E. K. Weir and J. López-Barneo.

62 **1996 Annual Report**

63 **Workshop on TGF-β Signalling in Development and Cell Cycle Control.**
Organizers: J. Massagué and C. Bernabéu.

64 **Workshop on Novel Biocatalysts.**
Organizers: S. J. Benkovic and A. Ballesteros.

65 **Workshop on Signal Transduction in Neuronal Development and Recognition.**
Organizers: M. Barbacid and D. Pulido.

66 **Workshop on 100th Meeting: Biology at the Edge of the Next Century.**
Organizer: Centre for International Meetings on Biology, Madrid.

67 **Workshop on Membrane Fusion.**
Organizers: V. Malhotra and A. Velasco.

68 **Workshop on DNA Repair and Genome Instability.**
Organizers: T. Lindahl and C. Pueyo.

69 **Advanced course on Biochemistry and Molecular Biology of Non-Conventional Yeasts.**
Organizers: C. Gancedo, J. M. Siverio and J. M. Cregg.

70 **Workshop on Principles of Neural Integration.**
Organizers: C. D. Gilbert, G. Gasic and C. Acuña.

71 **Workshop on Programmed Gene Rearrangement: Site-Specific Recombination.**
Organizers: J. C. Alonso and N. D. F. Grindley.

72 **Workshop on Plant Morphogenesis.**
Organizers: M. Van Montagu and J. L. Micol.

73 **Workshop on Development and Evolution.**
Organizers: G. Morata and W. J. Gehring.

*74 **Workshop on Plant Viroids and Viroid-Like Satellite RNAs from Plants, Animals and Fungi.**
Organizers: R. Flores and H. L. Sänger.

75 **1997 Annual Report.**

76 **Workshop on Initiation of Replication in Prokaryotic Extrachromosomal Elements.**
Organizers: M. Espinosa, R. Díaz-Orejas, D. K. Chattoraj and E. G. H. Wagner.

77 **Workshop on Mechanisms Involved in Visual Perception.**
Organizers: J. Cudeiro and A. M. Sillito.

78 **Workshop on Notch/Lin-12 Signalling.**
Organizers: A. Martínez Arias, J. Modolell and S. Campuzano.

79 **Workshop on Membrane Protein Insertion, Folding and Dynamics.**
Organizers: J. L. R. Arrondo, F. M. Goñi, B. De Kruijff and B. A. Wallace.

80 **Workshop on Plasmodesmata and Transport of Plant Viruses and Plant Macromolecules.**
Organizers: F. García-Arenal, K. J. Oparka and P.Palukaitis.

81 **Workshop on Cellular Regulatory Mechanisms: Choices, Time and Space.**
Organizers: P. Nurse and S. Moreno.

82 **Workshop on Wiring the Brain: Mechanisms that Control the Generation of Neural Specificity.**
Organizers: C. S. Goodman and R. Gallego.

83 **Workshop on Bacterial Transcription Factors Involved in Global Regulation.**
Organizers: A. Ishihama, R. Kolter and M. Vicente.

84 **Workshop on Nitric Oxide: From Discovery to the Clinic.**
Organizers: S. Moncada and S. Lamas.

85 **Workshop on Chromatin and DNA Modification: Plant Gene Expression and Silencing.**
Organizers: T. C. Hall, A. P. Wolffe, R. J. Ferl and M. A. Vega-Palas.

86 **Workshop on Transcription Factors in Lymphocyte Development and Function.**
Organizers: J. M. Redondo, P. Matthias and S. Pettersson.

87 **Workshop on Novel Approaches to Study Plant Growth Factors.**
Organizers: J. Schell and A. F. Tiburcio.

88 **Workshop on Structure and Mechanisms of Ion Channels.**
Organizers: J. Lerma, N. Unwin and R. MacKinnon.

89 **Workshop on Protein Folding.**
Organizers: A. R. Fersht, M. Rico and L. Serrano.

90 **1998 Annual Report.**

91 **Workshop on Eukaryotic Antibiotic Peptides.**
Organizers: J. A. Hoffmann, F. García-Olmedo and L. Rivas.

92 **Workshop on Regulation of Protein Synthesis in Eukaryotes.**
Organizers: M. W. Hentze, N. Sonenberg and C. de Haro.

93 **Workshop on Cell Cycle Regulation and Cytoskeleton in Plants.**
Organizers: N.-H. Chua and C. Gutiérrez.

94 **Workshop on Mechanisms of Homologous Recombination and Genetic Rearrangements.**
Organizers: J. C. Alonso, J. Casadesús, S. Kowalczykowski and S. C. West.

95 **Workshop on Neutrophil Development and Function.**
Organizers: F. Mollinedo and L. A. Boxer.

96 **Workshop on Molecular Clocks.**
Organizers: P. Sassone-Corsi and J. R. Naranjo.

97 **Workshop on Molecular Nature of the Gastrula Organizing Center: 75 years after Spemann and Mangold.**
Organizers: E. M. De Robertis and J. Aréchaga.

98 **Workshop on Telomeres and Telomerase: Cancer, Aging and Genetic Instability.**
Organizer: M. A. Blasco.

99 **Workshop on Specificity in Ras and Rho-Mediated Signalling Events.**
Organizers: J. L. Bos, J. C. Lacal and A. Hall.

100 **Workshop on the Interface Between Transcription and DNA Repair, Recombination and Chromatin Remodelling.**
Organizers: A. Aguilera and J. H. J. Hoeijmakers.

101 **Workshop on Dynamics of the Plant Extracellular Matrix.**
Organizers: K. Roberts and P. Vera.

102 **Workshop on Helicases as Molecular Motors in Nucleic Acid Strand Separation.**
Organizers: E. Lanka and J. M. Carazo.

103 **Workshop on the Neural Mechanisms of Addiction.**
Organizers: R. C. Malenka, E. J. Nestler and F. Rodríguez de Fonseca.

104 **1999 Annual Report.**

105 **Workshop on the Molecules of Pain: Molecular Approaches to Pain Research.**
Organizers: F. Cervero and S. P. Hunt.

106 **Workshop on Control of Signalling by Protein Phosphorylation.**
Organizers: J. Schlessinger, G. Thomas, F. de Pablo and J. Moscat.

107 **Workshop on Biochemistry and Molecular Biology of Gibberellins.**
Organizers: P. Hedden and J. L. García-Martínez.

108 **Workshop on Integration of Transcriptional Regulation and Chromatin Structure.**
Organizers: J. T. Kadonaga, J. Ausió and E. Palacián.

109 **Workshop on Tumor Suppressor Networks.**
Organizers: J. Massagué and M. Serrano.

110 **Workshop on Regulated Exocytosis and the Vesicle Cycle.**
Organizers: R. D. Burgoyne and G. Álvarez de Toledo.

111 **Workshop on Dendrites.**
Organizers: R. Yuste and S. A. Siegelbaum.

112 **Workshop on the Myc Network: Regulation of Cell Proliferation, Differentiation and Death.**
Organizers: R. N. Eisenman and J. León.

113 **Workshop on Regulation of Messenger RNA Processing.**
Organizers: W. Keller, J. Ortín and J. Valcárcel.

114 **Workshop on Genetic Factors that Control Cell Birth, Cell Allocation and Migration in the Developing Forebrain.**
Organizers: P. Rakic, E. Soriano and A. Álvarez-Buylla.

115 **Workshop on Chaperonins: Structure and Function.**
Organizers: W. Baumeister, J. L. Carrascosa and J. M. Valpuesta.

116 **Workshop on Mechanisms of Cellular Vesicle and Viral Membrane Fusion.**
Organizers: J. J. Skehel and J. A. Melero.

117 **Workshop on Molecular Approaches to Tuberculosis.**
Organizers: B. Gicquel and C. Martín.

118 **2000 Annual Report.**

119 **Workshop on Pumps, Channels and Transporters: Structure and Function.**
Organizers: D. R. Madden, W. Kühlbrandt and R. Serrano.

120 **Workshop on Common Molecules in Development and Carcinogenesis.**
Organizers: M. Takeichi and M. A. Nieto.

---

*: Out of Stock.

The Centre for International Meetings on Biology
was created within the
Instituto Juan March de Estudios e Investigaciones,
a private foundation specialized in scientific activities
which complements the cultural work
of the *Fundación Juan March*.

The Centre endeavours to actively and
sistematically promote cooperation among Spanish
and foreign scientists working in the field of Biology,
through the organization of Workshops, Lecture
Courses, Seminars and Symposia.

From 1989 through 2000,
a total of 149 meetings,
all dealing with a wide range of
subjects of biological interest,
were organized within the
scope of the Centre.

Instituto Juan March de Estudios e Investigaciones
Castelló, 77 • 28006 Madrid (España)
Tel. 34 91 435 42 40 • Fax 34 91 576 34 20
http://www.march.es/biology

The lectures summarized in this publication were presented by their authors at a workshop held on the 12th through the 14th of March, 2001, at the Instituto Juan March.

All published articles are exact reproduction of author's text.

There is a limited edition of 450 copies of this volume, available free of charge.