

Editorial

BIOINFORMATICS IN STRUCTURAL GENOMICS

The goals of structural genomics initiatives are to significantly expand the structural ‘coverage’ of sequence space. A number of specific objectives have been included within this broad definition. These include, the determination of the detailed three-dimensional structure for at least one representative of each protein fold that occurs in nature, the determination of enough structures so that all others can be built with homology models and the determination of enough structures so that functional information can be inferred from models of all others. Projects are being launched at different pace and with very different scopes in the USA, Japan and Europe (Table). The initial phase has shifted focus slightly away from ‘determining as many structures as possible’ to ‘finding large-scale semi-automatic solutions for the immediate technical bottlenecks’. These currently appear to be protein expression, purification and crystallisation.

The different initiatives have adopted various approaches to the problems of how to select the experimental targets, how to analyse the resulting structures, and how to optimally benefit from the structural information generated to learn about function. Obviously, many of the initial tasks of structural genomics may profit from exploring the potential of bioinformatics. However, the precise way in which bioinformatics is embedded into existing initiatives and the percentage of resources allocated to bioinformatics activities differs substantially between the efforts. For example, the North-East Structural Genomics (NESG) consortium relies entirely on bioinformatics to select the targets that are experimentally pursued. In contrast, the first European structural genomics project hardly explored the potentials of bioinformatics.

The most obvious initial application of bioinformatics tools to aid structural genomics has been the task of ranking the experimental targets. The task at hand is to somehow cluster all known proteins into families and to label all those families for which we do not yet have high-resolution information about the associated structures. However, computational biology is also required to optimally explore the information gained by solving a single structure. Examples are the transfer of structural information to homologues through comparative modelling and/or threading techniques. Another set of tools that aim to profit from the wealth of structures added through structural genomics aims at predicting aspects of function and at identifying functionally similar proteins. The problem of how functional specificity relates to protein structure is of very complex nature. Studying this

adaptation requires combining experimental and predictive methods. A particular example of such combinations could be the predictions of structural consequences for different sequence variants as for instance found in single nucleotide polymorphisms (SNPs). Another example could be the case of predicting protein–protein interactions (proteomics). Since the large-scale determination of large complexes of interacting proteins will not be part of the initial activities in structural genomics, we might venture to aim at obtaining a coarse-grained picture of protein–protein interactions by combining structures determined in context of structural genomics, interactions determined in context of functional genomics with tools from computational biology such as docking programs.

It a very early phase of structural genomics projects, the Juan March foundation (<http://www.march.es>) hosted a meeting on ‘Structural genomics and bioinformatics’ in Madrid (March, 2001). This meeting served as a forum to discuss some of the controversial issues at the interface of experimental structural genomics and bioinformatics; it addressed the challenges for bioinformatics resulting from structural genomics in two ways: (1) How can bioinformatics help structural genomics initiatives? (2) How can bioinformatics profit from the flood of new structures? The talks included reports from different experimental perspectives (C.M. Dobson, Oxford; J.M. Carazo, Madrid; C.D. Lima, New York; A. McDermott, New York) and a representative collection of topics on bioinformatics and computational biology, including the analysis of sequence space (M. Linial, Jerusalem; S.I. O’Donoghue, Heidelberg; B. Rost, New York; C. Sander, Boston), the distribution of protein structures and folds (L. Holm, Cambridge; A. Murzin, Cambridge; C. Orengo, London), the current status of the protein structure prediction methods in homology modelling (M. Peitsch, Basel; A. Sali, New York), threading (D. Jones, London) and protein interactions (M. Kanehisa, Kyoto; A. Valencia, Madrid). At the same time a number of presentations addressed the issues related with the prediction of protein function at various levels (T. Gaasterland, New York; F. Gago, Madrid; B. Honig, New York; M. Orozco, Barcelona; M. Sippl, Salzburg; J. Thornton, London). Almost all speakers are currently actively involved with one of the existing structural genomics initiatives.

For this issue of Bioinformatics, we selected five papers addressing the key problems in structural genomics mentioned above: (1) target selection (3 papers), (2) homology modelling and (3) structural basis of protein function. J. Liu and B. Rost present a re-estimate of the number of proteins that are targets for structural genomics on eukaryotes (‘Target space for structural genomics revisited’). E. Portugaly and M. Linial present a refined version of their original method to cluster

Table 1. Current initiatives in structural genomics

Initiative (PI, country)	URL	Focus
USA (NIH)	www.nigms.nih.gov/funding/psi/psi_research_centers.html	
BSGC (S.-H. Kim, USA)	www.strgen.org	<i>Mycoplasma pneumoniae</i>
CESG (J.L. Markley, USA)	www.uwstructuralgenomics.org	<i>Arabidopsis thaliana</i>
JCSG (I. Wilson, USA)	www.jcsg.org	<i>Caenorhabditis elegans</i>
MCSG (A. Joachimiak, USA)	www.mcsg.anl.gov	Disease related and 'easy' proteins
NESG (G. Montelione, USA)	www.nesg.org	Eukaryotes
NYSGRG (S.K. Burley, USA)	www.nysgrc.org	Enzymes
SECSG (B.-C. Wang, USA)	secsg.org	<i>Pyrococcus furiosus</i>
SGPP (W.G.J. Hol, USA)	depts.washington.edu/sgpp	Pathogenic protozoa
TBSGC (T. Terwilliger, USA)	www.doe-mpi.ucla.edu/TB/	<i>Mycobacterium tuberculosis</i>
Non-US		
PSF (U. Heinemann, Germany)	www.rzpd.de/psf	<i>Homo sapiens</i>
Spine (D. Stuart, EU)	europa.eu.int/comm/research/press/2002/pr1803en.html#ann3	500 targets of medical interest
SRG (S. Yokoyama, Japan)	www.rsgi.riken.go.jp/	<i>Thermus thermophilus</i> , <i>Mus musculus</i>
Toronto (C. Arrowsmith, Canada)	www.uhnres.utoronto.ca/proteomics/	Bacteria, Archae, Yeast
YSG (J. Janin, France)	genomics.eu.org/	<i>Saccharomyces cerevisia</i>

protein sequence space through pair-relations ('Selecting targets for structural determination by navigating in a graph of protein families'). F. Abascal and A. Valencia describe a clustering scheme applicable to fine-grained classifications of protein families ('Clustering of proximal sequence space for the identification of protein families'). E. Portugaly and M. Linial present a refined version of their original method to cluster the space of all proteins through pair-relations ('Selecting targets for structural determination by navigating in a graph of protein families'). M. Peitsch describes how comparative modelling can extend the impact for a single experimental structure ('Use of protein models'). Finally, X. Fradera, X. De La Cruz, C. H.T.P. Silva, J. L. Gelpi, F. J. Luque and M. Orozco explore the adaptation of binding sites by comparing protein bound to different substrates ('How dependent are binding sites on the bound ligand?').

Burkhard Rost^{1,2,*}, Barry Honig^{2,3} and Alfonso Valencia⁴

¹Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street, New York, NY 10032, USA

²Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 St Nicholas Avenue, New York, NY 10032, USA

³Howard Hughes Medical Institute, Department of Biochemistry and Molecular Biophysics, Columbia University, 630 West 168th Street, New York, NY 10032, USA

⁴Protein Design Group, CNB-CSIC, Cantoblanco, Madrid 28049, Spain

*To whom correspondence should be addressed.

Tel: +1 212 305 3773; Fax: +1 212 305 7932;

Email: rost@columbia.edu; <http://cubic.bioc.columbia.edu/>