Bioinformatics: from molecules to systems. A Discussion Meeting held at The Royal Society on 4 and 5 April 2005

Simon C Lovell and Balázs Papp

References	**This article cites 8 articles, 3 of which can be accessed free** http://rsif.royalsocietypublishing.org/content/2/4/397.full.html#ref-list-1
Email alerting service	Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click **here**

To subscribe to *J. R. Soc. Interface* go to: **http://rsif.royalsocietypublishing.org/subscriptions**

REPORT

Bioinformatics: from molecules to systems. A Discussion Meeting held at The Royal Society on 4 and 5 April 2005

Simon C. Lovell[†] and Balázs Papp

Faculty of Life Sciences, The University of Manchester, Michael Smith Building, Oxford Road, Manchester M13 9PT, UK

1. FROM PROTEIN SEQUENCE TO STRUCTURE

Has the protein folding problem been solved? On the first day of this Discussion Meeting, John Moult (University of Maryland, USA) pointed out that commentaries announcing the solving of the protein folding problem have a long history and, so far, these have all been the product of over-enthusiasm. However, it was clear that David Baker (University of Washington, USA) has produced the most promising advances for several years.

The protein folding problem is one of the most important and difficult facing biology. In the late 1950s, Christian Anfinsen showed that purified ribonuclease could spontaneously fold to an active, native structure (Sela *et al.* 1957; Anfinsen 1973). The purification is the key: it demonstrates that the information contained within the protein itself, specifically the amino-acid sequence, is sufficient to specify the folded three-dimensional structure. Since then there has been a large, concerted, and so far unsuccessful, effort to learn the rules by which sequence specifies structure. This is important, because protein structure in turn specifies function.

Baker's method starts with a 'low resolution' search, with proteins represented as simplified models, and then a refinement step, with proteins represented in full detail. In the refinement step, a high-quality rotamer library (Dunbrack & Karplus 1993) is used to repack side chains, and the evaluation function is dominated by van der Waals terms and hydrogen bonding. In other words, the detail of the internal packing must be correct. Because the energy well of the native state is

very narrow, it could easily be skipped over without refinement.

The excitement comes from the 'nose' of the graph on the right-hand side (figure 1*a*), indicating not only that structures close to the correct one are produced, which is relatively easy, but also that they are the ones with lowest energy, which is much more difficult. Baker gives another indication of the success of his method: the structures in figure 1*b* (prediction compared to experimentally determined structure) are superimposed. When predictions are less accurate, it is very difficult to make intelligible slides with the two structures superimposed and so they are often shown side-by-side. An informal rule of thumb is that a researcher is onto a good method if the slides look like that in figure 1*b*.

Frustratingly, there is no 'Aha!' moment. The difference between Baker's success and everyone else's failure is doing everything just a little better. The scoring and search functions are better than most, allowing the identification of the correct internal packing, but they are an evolution rather than a revolution.

Baker himself was at pains to point out that the results he presented only applied to fairly small, soluble proteins. Moreover, they were not true blind predictions. They were, in the terminology of the field, 'post-dictions'—because the structure of the protein being predicted was already known. Blind predictions are essential when developing a technique so that you can tell when progress is being made. Historically, however, post-dictions have proved to be a poor way of comparing methods. Typically, a new method would be compared with those published in the literature. If we were to produce a new method for prediction, it could seem to be more accurate than published ones. However, if our method used information in the sequence and structural databases (as the majority do) any improvement could be due to the increase in database size in the time taken for our rival to write her paper and get it through the publication process. Thus all methods were apparently improvements, but it was not clear whether progress was really being made.

This problem was largely solved in the early 1990s by John Moult and colleagues, when they proposed and ran the first CASP (critical assessment of methods in structure prediction; Protein Structure Prediction Issue 1995). The method for CASP is conceptually simple, if difficult logistically: the organizers solicit the community of experimental structural biologists for structures that are about to be solved. They send the sequences to predictors, who then have a few weeks to return their predictions (or a few days if they use fully automated methods). Predictions are evaluated by the assessors, and the results are given in a meeting (previously held in California but, for the latest round of CASP, moved to Italy because of the current restrictions on travel to America). Most predictors do not know how successful they have been before the meeting, and there is a definite air of theatre as the slide showing the degree of success is shown. For all of that, Moult insists that, contrary to common parlance, CASP is an experiment, not a competition. The stated

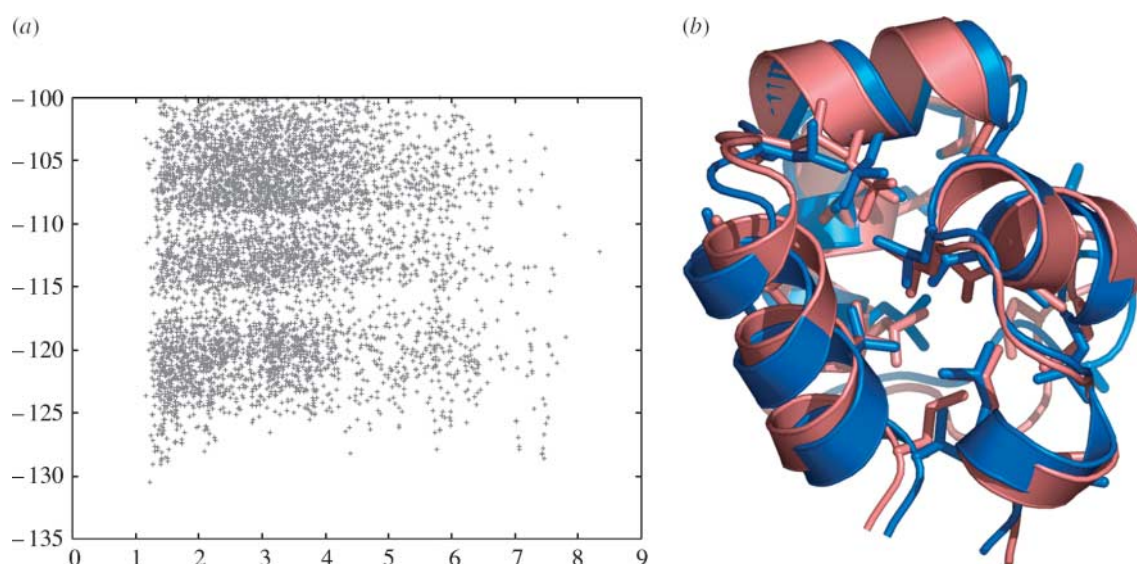[†]Author for correspondence (simon.lovell@manchester.ac.uk).

Figure 1. (*a*) The relationship between root mean square deviation between models and the correct structure (*x*-axis) and the scoring function (*y*-axis). There are some models (lower left-most points) that have both the lowest score and are the most accurate. Panel (*b*) shows the lowest scoring model (blue) superimposed onto the real structure (gold). In addition to the correct secondary structure being in the correct position, the side chains also substantially superimpose. (Figure reproduced courtesy of D. Baker, University of Washington, Seattle, USA.)

aims of CASP are to determine the state of the art, to identify progress and bottlenecks, and therefore to show where effort may be best focussed.

The direction of the prediction field has moved in the direction of using evolutionary relationships and fragments of known proteins. Moult described this as the triumph of knowledge over physics. Some of the effects of CASP can be a little perverse. As pointed out by Tom Blundell (University of Cambridge, UK), developers of new methods tend to perform less well than those who apply the methods of others.

There are many problems outstanding. Moult divided the predictions into different 'zones' of difficulty. Zone 1 is the easiest in which to produce an accurate model because closely related homologues of known structure are available. Nevertheless, there are still problems because it is much easier to predict how two proteins will be similar than to predict how they will differ; some sort of refinement of the structure is required, and existing techniques like molecular dynamics do not do the job accurately. Zone 2 is more difficult, because only distantly related homologues of known structure are available. Refinement is still required but, additionally, getting accurate sequence alignments is problematic. In zone 3, there are no known homologues of known structure. The most commonly used approach to modelling is to produce many thousands of models, and then to try and select the most accurate ones. It is this selection that is the most difficult problem. Zone 3 is where most progress has been made in recent years, and these are the types of predictions with which David Baker has had most success.

Zone 4 is altogether different. Zone 4 is where membrane proteins live, and this was the subject of David Jones' (University College London, UK) talk. Approximately 35% of human proteins have one or more trans-membrane domains, but only approximately 0.5% have known structure. Moreover, these proteins are extremely important: around 25% of drugs target G-protein-coupled receptors, which are a class of α-helical membrane-spanning proteins. The disparity arises because there are many technical difficulties with solving the structures of membrane proteins experimentally.

Membrane protein structures are also harder to predict than those of globular proteins. In part, this is due to the very heterogeneous environment in which they are found. A trans-membrane protein must interact not only with the hydrophobic centre of the lipid bilayer, but also the polar head groups of the phospholipids and the surrounding water. The difficulty also arises from the problems with experimental methods. There are so few membrane proteins in the structural database, that it is extremely difficult to develop rules from which to produce models. The rules that Jones uses in some cases are similar to those developed for globular proteins (e.g. amino-acid propensities to be in different types of structures), in some cases they are specific to membrane-spanning helices (e.g. the longer a helix is, the more tilted it is in the membrane), whereas in other cases rules used for globular proteins, such as solvation potentials, are omitted.

One of the most useful membrane protein specific metrics is 'variphobicity' which combines hydrophobicity and sequence variability. For soluble globular proteins, the proteins' exterior is both accepting of sequence substitutions (variable) and hydrophilic. The interior, conversely, is hydrophobic and conserved. Membrane proteins, by contrast, have regions that interact with solvent, and so are variable and hydrophilic, regions that interact with the lipid bilayer and so are variable and hydrophobic, and regions that pack in the proteins' core, and so are conserved and hydrophobic. Fold recognition software can be written that

gives high scores when conserved hydrophobic residues are packed together and variable hydrophobic residues are in the membrane-spanning region where they would interact with the lipid.

The paucity of data leads to more problems, though. There are very few structures that it is difficult to test the methods on a wide range of membrane proteins to determine the accuracy of the prediction methods. This is made even more problematic by the necessity of keeping the test set of proteins separate from the set of proteins used to develop the method. Still, the rate at which membrane protein structures are being solved is increasing and, with computational methods developing in parallel, each structure has a dual utility.

2. FROM PROTEIN STRUCTURE TO FUNCTION AND DRUG

Applications of both computationally and experimentally derived knowledge of protein structures were discussed by Michael Sternberg (Imperial College, UK) and Tom Blundell. Sternberg wants to take protein sequences, predict their structures, and use this information to identify function. In contrast, Blundell wants to predict structures from sequences and use these structures to produce drugs. Both use evolutionary relationships for the first step, and use knowledge of protein structure and the constraints it places on evolution to identify distant evolutionary relationships and so identify likely structures. Blundell then identifies functional sites from evolutionary constraints, and docks small molecules into these sites. From here, experimental methods take over: the results of the docking predictions of which molecules are likely to bind are tested by soaking these likely molecules into protein crystals. The structures are solved and the process iterated with larger, and so more tightly binding, molecules. In this way inhibitors are produced, which, in the majority of cases, have biological activity as drugs.

Sternberg uses the information of evolutionarily conserved residues to assign function by comparing conserved residues in the protein of interest to a database of conserved residues. The database is based around the Gene Ontology (GO) database, which gives a hierarchical description of function. The use of GO is an attempt to get away from traditional methods that assign function on the basis of overall sequence similarity and instead to use specific residues that produce function as an indicator of the correct assignment.

3. EVOLUTION OF GENE REPERTOIRE AND GENOMIC ANATOMY

It is generally agreed that bacteria can adapt to new environments by expanding their protein repertoire via gene duplication or horizontal gene transfer, but what are the factors opposing the process of genome expansion? Selection for fast reproduction rate could have an important impact on genome size, but it might not be the sole factor: it has been suggested that there could be an increasing regulatory cost associated with increasing the number of genes (Bird 1995). Christine Orengo (University College London, UK) argued that this latter explanation should be invoked to understand the evolution of prokaryotic genome size. To investigate this issue, she calculated protein domain family occurrences for 100 bacterial genomes by exploiting structural data and identified a set of universal families present in most of the analysed species for further analysis. Importantly, a domain family can be represented by several relatives in a genome had the family undergone extensive functional diversification. Thus it is not surprising that domain families involved in the most conserved cellular processes (e.g. translation) are represented by similar numbers of relatives in all bacterial genomes. In contrast, the sizes of domain families with metabolic and regulatory functions show strong correlations with genome size, corroborating the idea that increasing metabolic and regulatory complexity leads to larger genomes in prokaryotes. Most importantly, while metabolic families expand linearly with genome size, the sizes of regulatory families increase more rapidly (as a power law), suggesting that increasing complexity could be limited by the inflated relative cost of gene regulation in larger genomes. Combinatorial increases in the number of potential gene interactions that should be regulated in a larger genome might be responsible for the intensified regulatory burden, but further studies are needed to dissect the exact nature and magnitude of this hypothesized logistic cost.

Although it is well established that the gene repertoire of an organism can expand by gene duplications, it is less obvious how new interactions between different genes arise during evolution. For example, how often are the births of novel interactions linked to gene duplication events? Sarah Teichmann (Medical Research Council, Cambridge, UK) addressed this question by studying protein–protein interactions and paralogs in yeast. She presented evidence that there are many more interactions between paralogous pairs than expected by chance and over one third of the protein complexes contain homologous subunits, suggesting that evolution from homo- to hetero-oligomers by means of gene duplication might be a common process. Novel interactions, however, will not exclusively occur between duplicate copies if one or both of two non-homologous interacting partners, say A and B, undergo duplication. Indeed, up to half of the protein–protein interactions can be traced back to these sorts of gene duplication events. Interestingly, most of the duplications affect only one of the interacting partners, thus giving rise to A–B and A′–B complexes (only A is duplicated), but only rarely to A–B and A′–B′ (both A and B are duplicated) despite the potential gene dosage imbalance it involves in the short term (Veitia 2004).

Recent years have witnessed not only an increasing understanding of gene content evolution, but also new discoveries on the evolution of genomic anatomy (Hurst *et al.* 2004). Although bacterial operons are well-established examples of non-random gene organization, it is less well understood to what extent eukaryotic gene order deviates from random and how gene clusters appear during evolution. Ken Wolfe (Trinity College

Dublin, Ireland) added a new piece to the puzzle of eukaryotic gene order evolution by uncovering a large metabolic gene cluster in baker's yeast. Six of the eight genes of the allantoin degradation pathway are located next to one another on the same chromosome, which would be highly unlikely if gene order was completely random. Furthermore, this cluster must have been assembled recently, after yeast acquired the ability to grow anaerobically, because orthologs of the participating genes are scattered around the genome in obligately aerobic yeast species. Selection pressure for the ability to live in anaerobic environments can explain why baker's yeast has adapted to use allantoin instead of urate as a nitrogen source (degradation of urate requires oxygen), but why should genes involved in the same metabolic pathway sit close together on the chromosome? One possibility is the need for tight transcriptional co-regulation, which might be partly mediated by chromatin modification (therefore the physical proximity of genes facilitates their coordinated expression). A less frequently considered alternative possibility is direct selection for increased genetic linkage between the clustered genes, independent of their co-regulation. The observations of Wolfe are consistent with both scenarios: genes of the allantoin degradation pathway are located in a sub-telomeric chromosomal region where chromatin modification is known to operate and the local recombination rate is also low (high genetic linkage), suggesting that both selective forces were responsible for the emergence of the cluster.

4. A GROWING CHALLENGE FOR SYSTEMS BIOLOGY

The ambitious aim of systems biology is to describe cellular behaviour in a quantitative manner. One of the most fundamental cellular behaviours is growth and to build an *in silico* representation of a growing cell, we will need to understand what contribution individual genes make to the control of growth and how the coordinated behaviour of thousands of genes regulates cellular growth. Steve Oliver (University of Manchester, UK) presented genome-scale experimental approaches to generate the data required to elucidate these issues in yeast. First, employing the conceptual framework of metabolic control analysis (Kacser & Burns 1973), one can estimate the growth control coefficient of each gene by systematically altering the level of individual gene products (by manipulating gene dosage) and measuring the impact on the growth rate.

In principle, major growth rate controllers can be discovered in this way. Second, growth rate is altered and changes in gene product concentrations (i.e. mRNA levels) are measured, and thereby genes that are co-ordinately up- or down-regulated with increasing growth rates can be identified. Both categories of experimental studies were applied to yeast, and a surprising result emerged. Almost none of the genes with high growth rate control are significantly up-regulated with increasing growth rates. What is the explanation for this finding? One possibility is that the activity of these genes is primarily regulated at the post-transcriptional level (e.g. translational, post-translational or allosteric regulation), thus mRNA profiling cannot reveal their role in growth control. However, if molecular activity of a growth controller gene does not depend on its mRNA level, it remains unclear how growth rate could depend on its gene dosage in the first place. Alternatively, instead of regulating only genes that are most limiting for growth, the cell is more likely to modulate the activity of pathways at multiple sites to achieve the desired growth rate (Thomas & Fell 1998). Integration of different 'omics' approaches will hopefully resolve this issue in the near future.

REFERENCES

Anfinsen, C. B. 1973 Principles that govern the folding of protein chains. *Science* **181**, 223–230.

Bird, A. P. 1995 Gene number, noise reduction and biological complexity. *Trends Genet.* **11**, 94–99.

Dunbrack Jr, R. L. & Karplus, M. 1993 Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J. Mol. Biol.* **230**, 543–574.

Hurst, L. D., Pal, C. & Lercher, M. J. 2004 The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* **5**, 299–310.

Kacser, H. & Burns, J. A. 1973 The control of flux. *Symp. Soc. Exp. Biol.* **27**, 65–104.

Protein Structure Prediction Issue 1995. *Proteins Struct. Funct. Genet.* **23**(Suppl. 1), 295–462.

Sela, M., White, F. H. & Anfinsen, C. B. 1957 Reductive cleavage of disulfide bridges in ribonuclease. *Science* **125**, 691–692.

Thomas, S. & Fell, D. A. 1998 The role of multiple enzyme activation in metabolic flux control. *Adv. Enzyme Regul.* **38**.

Veitia, R. A. 2004 Gene dosage balance in cellular pathways: implications for dominance and gene duplicability. *Genetics* **168**, 569–574.