PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Structural biology

Kenneth C. Holmes and FRS

Phil. Trans. R. Soc. Lond. B 1999 **354**, 1977-1984
doi: 10.1098/rstb.1999.0537

THE ROYAL SOCIETY

Structural biology

Kenneth C. Holmes, FRS

Max-Planck-Institut für medizinische Forschung, Jahnstraße 29, 69120 Heidelberg, Germany (holmes@mpimf-heidelberg.mpg.de)

Protein crystallography has become a major technique for understanding cellular processes. This has come about through great advances in the technology of data collection and interpretation, particularly the use of synchrotron radiation. The ability to express eukaryotic genes in *Escherichia coli* is also important. Analysis of known structures shows that all proteins are built from about 1000 primeval folds. The collection of all primeval folds provides a basis for predicting structure from sequence. At present about 450 are known. Of the presently sequenced genomes only a fraction can be related to known proteins on the basis of sequence alone. Attempts are being made to determine all (or as many as possible) of the structures from some bacterial genomes in the expectation that structure will point to function more reliably than does sequence. Membrane proteins present a special problem. The next 20 years may see the experimental determination of another 40 000 protein structures. This will make considerable demands on synchrotron sources and will require many more biochemists than are currently available. The availability of massive structure databases will alter the way biochemistry is done.

Keywords: protein crystallography; synchrotron radiation; multiwave anomalous diffraction (MAD); structural genomics

1. REDUCTIONISM

The first half of the 20th century saw the unravelling of the basic metabolic processes in yeasts and animals. Disparate life forms were discovered to share common metabolic pathways mediated by similar catalytic enzymes. One began to see the underlying unity of life at the level of molecules. J. D. Bernal and Dorothy Hodgkin obtained detailed X-ray diffraction patterns from crystals of an enzyme. This was in fact an epoch-making experiment. As well as founding the discipline of X-ray protein crystallography it demonstrated that all enzyme molecules of a certain type were identical (else they would not crystallize). This idea was hotly disputed at the time—some could not countenance the counter-intuitive idea that giant molecules could be assembled with such precision. Subsequently it became clear that cells are indeed bags of precisely determined catalytic macromolecules (mostly proteins) interacting in diverse but highly orchestrated ways.

The second half of this passing century yielded an explanation for this precision: it was shown that all life shares common genetic mechanisms for determining macromolecules and, moreover, that the machinery for the synthesis of proteins is common to all forms of life. The ideas of genotype (DNA) and phenotype (proteins) took on molecular form.

These discoveries were the fruits of an analytical reductionist approach to biology. Nearly four centuries ago René Descartes ago already foresaw such possibilities:

> si on connoissoit bien quelles sont toutes les parties de la semence de quelque espèce d'animal, on pourroit déduire de cela seul, par de raisons entièrement mathématiques et certaines, toute la figure et conformation de chacun de ses membres.

Today we might be a little more humble in our aspirations. Nevertheless, the genome-sequencing projects have already provided us with the genotypic structures of all the macromolecules in a number of diverse organisms. Thus we already possess the canonical inventory of a living organism, thereby fulfilling Descartes's condition. The Cartesian programme may now be realized. It will not be fulfilled with the mathematical purity Descartes envisaged. Notwithstanding, integrating and understanding the mass of detail arising from the genome projects presents mankind with a major intellectual challenge for the new century.

Two techniques are the pillars of our burgeoning understanding of biology at a molecular level: recombinant DNA technology and X-ray crystallography. As was foreseen by Bernal, structural studies of macromolecules provide the essential molecular anatomy that is the basis for an understanding of cell physiology. X-ray crystallography enables us to 'see' the positions of thousands of atoms that form the three-dimensional (3D) structures of proteins, nucleic acids and their complexes. Since the local arrangements of atoms in these molecular systems determine their biological function and specificity, knowledge of the structures allows us to understand how these systems work in physicochemical terms. Structure therefore provides the link between biology and chemistry.

2. PROTEIN STRUCTURE

(a) *The folding problem*

The product of genetically (DNA) controlled protein synthesis is a linear polypeptide made from the 20 standard amino acids. However, the biologically active protein is a folded globule. One essential step in the Cartesian programme would therefore be to predict the

3D structure of a folded protein from its sequence. The next step would be to predict its function. We are still far from being able to carry out an *ab initio* calculation of protein folding, which demonstrates the intellectual vanity of such a pure approach. Writing a decade earlier than Descartes, Francis Bacon 'urgently advocated new ways by which men might establish a legitimate command over nature to the glory of God and the relief of man's estate'. Bacon recommended the empirical method, which in the present context may be interpreted as using X-ray crystallography or other experimental methods to determine the structure of spontaneously folded proteins. One expects that we do not need to determine experimentally the structures of all proteins, since protein structures fall into classes. Rather, by comparison of new sequences with data banks of known structures, in many cases one hopes to be able to identify the folding class. The expectation is that one will be able to identify function from the 3D structure and thereby bridge the gap from genotype and phenotype in the many situations where a simple comparison of sequences yields no information.

Large data banks of structure also provide the basis for a classification of the primitive protein folds, thereby allowing us a glimpse into early life—a molecular palaeontology. Moreover, the application of protein structure determination has already yielded abundant reward in the Baconian manner. The development of effective drugs against AIDS and the design of immunization strategies against influenza are but two from many examples. Moreover, genome projects are already producing a large number of DNA sequences of individual genes, many of interest per se, since they allow us to locate mutations leading to genetic diseases. The 3D structures of the encoded proteins provide the basis for rational drug design.

(b) *Protein crystallography*

X-ray crystallography is the major technique for establishing the 3D structures of macromolecules. X-ray diffraction from crystals yields a pattern of spots (Bragg reflexions), each of which is a section through the 3D Fourier transform of the crystal. Each orientation of the crystal gives a different pattern. If a full 3D set of Bragg reflexions is collected by rotating the crystal and collecting the pattern for each orientation, then by means of a Fourier synthesis the 3D structure (electron density) of the molecule may be calculated. This elegantly simple recipe is marred by the fact that, experimentally, one registers the intensities of the Bragg reflexions but not their phases; without the phases one cannot carry out a meaningful Fourier summation. Bernal foresaw that the structure of protein molecules could in principle be obtained by crystallography but was rather vague about how one might actually do this. When Max Perutz and Lawrence Bragg resumed the study of crystalline haemoglobin after the Second World War, initially they did not know how to solve the phase problem. An elegant theoretical study produced some low-resolution phases which yielded the overall shape of the molecule (Bragg & Perutz 1952)—a method based on the properties of the Fourier transform of the time of arrival of trains at Cambridge treated as a set of random numbers! Enduring success came when Perutz showed that the method of isomorphous replacement, until then applied but rarely in the field of crystallography and never to macromolecules, was in fact ideally suited to the protein problem (Green *et al*. 1954). His first successful application of the method to haemoglobin provided the basis for all subsequent work in the field. The method consists of binding one or two heavy atoms (e.g. mercury) to the protein and observing the changes in intensity of the Bragg reflexions. From the differences one can calculate the phases. The differences are quite small and there are few atoms heavier than mercury. Therefore, the limiting molecular weight that can be phased successfully depends on the accuracy of the data. This is one reason why protein crystallography has driven data collection technology.

In the seven years following Perutz's wonderful discovery, John Kendrew and his associates 'beavered away' at collecting 25 000 reflexions from myoglobin (a small relative of haemoglobin) five times (for five isomorphous heavy atom derivatives). Data were recorded on photographic film and intensities measured by densitometry. Kendrew had been a leader in operational research during the Second World War (where he met Bernal and became infected with the protein crystallography bug), therefore the task of organizing this massive undertaking fell within the natural orbit of his abilities and experience. Everything was being done for the first time. For example, the first computer programs for summing Fourier series (Bennett & Kendrew 1952) (in the early days protein crystallography was very computer limited). At the end of this Herculean task the atomic structure of myoglobin, with its numerous α-helices (a regular secondary structure element of polypeptides that had been predicted by Linus Pauling), was revealed (Kendrew *et al*. 1960). Afterwards Kendrew wrote that he would never do it again, nor did he.

At the same time Perutz published the structure of haemoglobin at 5.5 Å resolution (Perutz *et al*. 1960). Atomic resolution took longer because the molecule is much bigger (it is essentially four myoglobins put together with tetrahedral symmetry). The job of haemoglobin (and myoglobin) is to bind oxygen reversibly to its haem groups. Whereas myoglobin simply binds molecular oxygen for storage in muscle, haemoglobin transports oxygen from the lungs and the affinity for oxygen is regulated. This regulation is achieved by a cooperative interaction between the four haem groups in which each haem group senses whether the other haem groups have an oxygen bound or not, although they are not close to each other. Through this interaction the tetrameric haemoglobin can switch cooperatively between high-affinity and low-affinity forms. On the basis of atomic resolution of the two forms, Perutz was able to explain this interaction in terms of mechanical linkages provided by elements of the protein linking the four haems (Perutz 1972; Perutz & TenEyck 1972). This explanation is very important for demonstrating how atomic structure can reveal the workings of molecular machines. Although originally disputed as being too simplistic, this model has stood the test of time (Perutz *et al*. 1998) and is an object lesson for the success which can pertain to mechanistic interpretations. Now there are numerous examples of protein crystallography providing the basis for understanding molecular

machines, in particular molecular motors, notably the F1-ATPase (a circular motor) (Abrahams *et al*. 1994) and muscle (a linear motor; for a review, see Geeves & Holmes 1999). For their work, Perutz and Kendrew shared the Nobel Prize in Chemistry in 1962.

For some years after the publication of the atomic structure of myoglobin no new protein structures were solved and one began to wonder if the method was really general or limited to the highly α-helical globins. Then in 1965 came David Phillips's determination of the structure of lysozyme (Blake *et al*. 1965). This work in fact first demonstrated the power of protein crystallography to explain biological function in terms of physics and chemistry. The structure showed the complete path of the polypeptide chain folded into both α-helices, as seen in myoglobin, and β-sheet, (a second regular secondary structure element that had been predicted by Pauling but not hitherto observed in three dimensions). Lysozyme has antibacterial activity due to its ability to hydrolyse the polysaccharide component of the bacterial cell wall. On the basis of the structure, Phillips was able to produce an explanation for lysozyme's catalytic mechanism (Johnson *et al*. 1968). Here was the first structural explanation of how an enzyme speeds up a chemical reaction. It changed forever the way that enzymes are studied and marked the start of the integration of protein crystallography into biochemistry.

In the ensuing 35 years, over 8000 protein structures have been solved by X-ray crystallography, first slowly, now in a flood. A host of enzyme mechanisms and cellular control functions have been elucidated. Structural biology has taken its place as an important part of cell biology, much as was foreseen in Bernal's early vision of the role of structure in biology.

(c) *Synchrotron radiation*

What Bernal could not foresee was that structural biology would become symbiotic with high-energy physics. When Perutz and Kendrew took up the study of proteins in Cambridge, commercial X-ray sources were so weak that one X-ray diffraction photo of a myoglobin crystal could require a month of exposure. In Bernal's department at Birkbeck College London, W. Ehrenberg had developed an elegant microfocus X-ray tube of high brilliance which was very successfully used for photographing DNA fibres and muscle fibres, but which was not of much service for protein crystals. The MRC Unit for Molecular Biology at the Cavendish Laboratory in Cambridge invested considerable effort in the development of rotation anode X-ray tubes. This made it possible to collect diffraction data in a reasonable time and spurred the application of X-ray diffraction to amenable living systems such as muscle. Early success on muscle was tantalizing. With another factor of 1000 or so one would be able to see the muscle molecules move during a contraction. But rotating anode tubes were at the end of the line.

Initially driven by the muscle experiments, attention turned to synchrotron radiation. Synchrotron radiation is an expensive by-product of circular electron accelerators. Since the electrons are continuously being accelerated towards the middle of the ring such machines liberate a broad spectrum of radiation, including X-rays. Because the electrons are highly relativistic the beam of radiation comes out in a very narrow cone around the direction of flight of the electron. Thus synchrotrons throw off radiation tangentially. The higher the energy the more you get. In the 1960s, a number of 6 GeV synchrotrons were being commissioned. Such machines dump their beam into a target 50 times a second. A little later the interest of high-energy physicists shifted to storage rings, where the electrons circulate continuously for some hours and are used for positron–electron collision experiments. To be useful for a diffraction experiment an X-ray monochromator is needed to pick out a single wavelength. This was successfully done with a standard quartz crystal monochromator and in 1970 the first synchrotron X-ray diffraction pattern was obtained from a muscle fibre using the DESY synchrotron in Hamburg as a source (Rosenbaum *et al*. 1971). Thus the feasibility of this technology was established. A decade later synchrotron radiation actually did make it possible to see the muscle molecules moving (Huxley *et al*. 1981).

(d) *Synchrotron radiation becomes routine for structure determination*

The laser-like optical properties and high intensity of synchrotron radiation gradually made it the source of choice for all kinds of protein crystal data collection. Initially it was reserved for exotic problems but the widespread adoption of cryotechniques (Garman & Schneider 1997) (freezing the crystals at liquid-nitrogen temperatures) with the resultant ease of handling and the dramatic reduction of radiation damage made an important contribution to the application of synchrotron radiation as a routine method. Now one has custom-built storage rings (the so-called third-generation sources) equipped with special devices in which linear arrays of magnets are inserted into the electron beam to make it undulate and 'shake out' more or less monochromatic synchrotron radiation. Undulators yield copious quantities of X-rays if the energy of the storage ring is high enough. For protein crystallography the high brilliance and small size of the X-ray beam from an undulator permits accurate data collection even from tiny crystals (20–50 μ). This development is very important for structural biology since tiny crystals are often the only ones to be had.

For many years detectors were a serious problem. Originally, photographic film was introduced into X-ray crystallography by Bernal because he did not have the patience to measure intensities with a gold-leaf electroscope. Film continued to be used almost exclusively for the ensuing 70 years. In the last decade it was replaced by imaging plates, which are much more accurate. However, reading the intensities off imaging plates takes time and this now provided the bottleneck in the rate of data collection from synchtrotron sources. In the last couple of years, charge-coupled detectors (CCDs) of sufficient size and resolution have become available (Westbrook 1997). Such detectors allow fast read-out of the diffraction data and come close to removing the data read-out bottleneck. Custom-built solid-state detectors should completely remove this restriction.

Using storage rings, undulator beam lines and CCD detectors one can now collect a full high-resolution data set from a medium-sized protein in about 10 min. Using

multiwavelength anomalous diffraction (MAD) phase determination (see §2(f)) and automatic model-building programs, an atomic model of a protein has been obtained from scratch in a few hours. Kendrew needed six years.

(e) *Expression of proteins in bacteria*

The first reports of the successful translation and expression of eukaryotic proteins by insertion of specially constructed plasmids containing the appropriate eukaryotic DNA sequences into *Escherichia coli* were published in 1978 (Chang *et al*. 1978). Besides leading to the establishment of a number of biotech companies, this new technology was of great importance for structural biology. To grow crystals one requires at least a few milligrams of protein. Nuclear magnetic resonance (NMR) spectroscopy requires a few hundred milligrams. Many important proteins exist only in tiny amounts in the cell and a conventional purification from an organ or tissue source would often be impossible. Moreover, the natural product will not contain the appropriate ^{13}C and ^{15}N isotopes necessary for NMR spectroscopy nor the selenomethionine substitutions so important for crystallography (see §2(f)). Therefore the ability to produce any protein in quantity from its DNA in bacteria by a process akin to fermentation is of great importance in structural biology. Furthermore, this methodology is central to the strategies being developed in structural genomics.

Some eukaryotic proteins will not express as soluble proteins in bacteria, rather they precipitate. For their folding, these proteins require the presence in the cell of specific eukaryotic helper proteins (chaperons) which are missing in bacteria. In this case the expression has to be done in eukaryotic cells. Yeast, cellular slime mould and insect cells have been found suitable.

(f) *Routine phase determination by MAD*

Synchrotron radiation, which is inherently polychromatic and therefore tunable through the use of monochromators, has enabled the development of MAD for phase determination in protein crystallography (Hendrickson 1991). Here data is collected with X-radiation of various wavelengths around the wavelength of a suitable absorption edge of a heavy metal attached to the protein. The contribution of the out-of-phase scattering of the heavy atom (the 'anomalous' part) changes sign as the absorption edge is crossed. This alteration can be used for phase determination. Since all measurements are made on the same crystal under more or less identical conditions, this method removes a number of serious sources of systematic error. The method is fairly accurate and lends itself to automation. Any of the conventional heavy atoms used in isomorphous replacement, e.g. mercury, uranium or platinum, can also be used for MAD but so can a number of lighter elements. The greatest impact has come from selenium, which can systematically be incorporated into proteins as selenomethionine to replace methionine in bacterial protein expression. The use of seleno-methionine with MAD provides a more or less automatic way of obtaining accurate phases and generally yields better electron density maps. It will probably become the method of choice for the next decade and will enable structure determination to become much more an assembly line activity. One notable limitation is the necessity for bacterial expression.

(g) *The limits of the method*

In terms of molecular size, larger complexes are more difficult to crystallize, but the limits of the method are set by the phasing power of heavy atoms. In practice, it looks as if molecules up to about 10^6 Da will be solvable by the methods in use. The use of resonance nuclear scattering (Mössbauer scattering) for phasing might improve this situation dramatically (the resonance nuclear scattering from one iron nucleus is equivalent to about 500 electrons—a super heavy atom). Beams of Mössbauer radiation of adequate intensity may be available by exciting an iron-containing crystal with the intense X-ray beams from a free electron laser.

However, crystals must be of a certain minimum size for X-ray diffraction to work. The diffraction pattern arises from the coherent scattering of X-rays by electrons (Thompson scattering), which is a tiny effect. It is much weaker than the absorbing of photons by the specimen—the photoelectric effect. Therefore specimens suffer considerable radiation damage. The method only works because of the enormous number of constructively scattering molecules contained in even small crystals. Crystals have been likened to scattering amplifiers. In spite of this amplification the diffraction pattern in fact fades away on irradiation with X-rays. Therefore, X-ray crystallography of proteins is not a good method if judged from the ratio of signal-to-radiation damage (dose). The absorbed photons give rise to a cloud of free radicals. Since most of a protein crystal is water, most of the free radicals are produced outside the protein and have to diffuse to cause damage. Hence the great success of cryocrystallography. If the crystals are frozen to liquid-nitrogen temperature the free radicals cannot diffuse and the protein remains intact. However, even this wonder has its limits. In practice, one finds that the minimum size of a crystal from which data can be collected, even with the best undulator beam lines, is 20–30 μ. It is possible that at very high rates of data collection (10–20 ns), which can be obtained from the pulsed structure of synchrotron radiation, this situation might improve somewhat. However, the crystal would have to be replaced for every new shot, which would complicate data collection considerably. Realistically, we have probably already reached the limits of the method. Nor would more intensity help. The available beam lines often have to be attenuated for small crystals to avoid destroying the crystals. It appears as if the onset of radiation damage may not be a linear function of dose.

3. STRUCTURAL GENOMICS

(a) *Assigning structure to sequences*

A number of genomes have been sequenced. Many more, including the human genome, are nearing completion. The human genome will contain coding sequences for about 100 000 proteins. Scanning the sequence of (say) the already completed yeast genome shows that the majority of all the predicted proteins are without recognizable function. Thus the genome projects are generating large amounts of ignorance. A systematic attempt to

come to terms with this problem via structural biology has become known as structural genomics (Shapiro & Lima 1998). The purpose of structural genomics is to assign 3D structures to the protein products (known as proteomes) and to investigate the biological implications of these assignments. The assignment of structures to proteomes can be carried out on two levels—experimental and computational. The experimental level involves the directed, large-scale determination of the protein structures using X-ray crystallography or, for smaller proteins, NMR spectroscopy. One foresees solving the structures of 20 000 proteins in the next ten years. The computational level involves the assignment of structures to proteins using calculations that mostly involve demonstrating homology to proteins of known structure.

Three classes of computational methods are presently used to assign structures to genome sequences: the detection of distant sequence homologies (this usually involves pairwise or multiple-sequence comparisons); fold recognition (which tries to determine whether the sequence of a new protein fits a fold from a known structure); and predictions of secondary structure (α-helix, β-sheet or loops) based on statistical rules derived from known structures. Pairwise sequence comparisons can presently match only between 11 and 20% of the proteins from sequenced genomes to known structures. However, much of the low scoring can be attributed to the limitations of the method. Pairwise sequence comparisons detect only about half of the relationships between proteins with 20–30% sequence identity, although experience shows that two proteins with 20–30% sequence homology will have nearly identical structures. Improved computational techniques have allowed protein folds to be assigned to between a quarter (*Caenorhabditis elegans*) and a half (*Mycoplasma genitalium*) of the protein sequences in these two genomes (Hubbard *et al*. 1998). The results of such analyses begin to give insight into questions as to whether structure–function relationships are conserved across species, e.g. do like metabolic pathways use the same protein folds in disparate organisms (see e.g. Gerstein & Levitt 1997; Teichmann *et al*. 1999)?

(b) *How many protein folds are there?*

Proteins are made of a few hundred amino acids strung together. There are 20 amino acids. One can quickly ascertain that the total number of possible protein sequences is greater than the number of particles in the universe. However, to form a protein the polypeptide chains must fold and pack its side chains to form a dense, well-defined structure. This is a diabolical 3D jigsaw puzzle (Levitt *et al*. 1997). In the last 3500 million years only a limited number of solutions, probably about 1000, have been discovered. Moreover, once Darwinian selection started working there was probably not much opportunity for organisms to develop new folds. Old folds mutated to new closely related folds or fused with other folds to produce new multi-domain proteins with new functions. The performance of the new enzyme was refined by point mutation. A genetic modification that altered the fold would lead to an inactive protein. Therefore the primeval folds are maintained. If one can catalogue these folds one has a base set for the prediction of structures of unknown proteins.

How far along are we with identifying primeval folds? One example of a catalogue of primeval folds is represented by the structural classification of proteins (SCOP) database (Hubbard *et al*. 1997). The SCOP database aims to provide insight into the structural and evolutionary relationships between all proteins whose structure is known. The SCOP classification has been constructed manually by visual inspection and comparison of structures, but with the assistance of computational methods. Proteins are classified in a hierarchy that reflects both structural and evolutionary relatedness. The principal levels are family, superfamily and fold. Proteins clustered together into families are clearly evolutionarily related. Proteins in a superfamily probably share a common evolutionary origin. Proteins have a common fold if they have the same major secondary structures in the same arrangement and with the same topographical connections. To date we know the structures of about 8000 proteins. This is clearly still a small sample of all proteins. Based on this sample the SCOP database now contains about 450 folds.

4. PLANNING FOR THE NEXT TEN YEARS

(a) *More structures*

Clearly to get further we need many more experimentally determined structures. Without a much more substantial experimental database we will be able neither to catalogue all the primeval folds nor estimate the reliability of a structure arrived at on the basis of prediction. However, it will be impossible for protein crystallography to keep pace with the sequencing speed of the genome projects. Structures will lag behind, which is particularly unfortunate since structure is probably the best way of getting an estimate of function without doing a great deal of biochemistry. Therefore relational structural genomic databases (such as PRESAGE; Brenner *et al*. 1999) will fill up with functional tags rather slowly. In terms of understanding the genome, structural genomics will be a bottleneck.

(b) *Structural genomics*

The smallest bacterial genome contains only 450 coding sequences. It is quite feasible to determine the structure of each gene product. A number of initiatives in this direction are already underway (Montelione & Anderson 1999). Automatic methods for expressing, purifying and even for crystallizing proteins are being developed. In general, the aim would be to identify the coding sequences in a genome and build each into an expression system for *E. coli*, express each soluble gene product and then crystallize it. If the protein should prove recalcitrant, gentle trypsin treatment or the expression of subdomains would be used to break the problem down to amenable subdomains. Structure determination would follow in the hope that function will be predictable from the structure. Thermophilic bacteria are a favoured species for getting started because of the ease of purification of the expressed gene product (boiling).

If such projects get going they will generate a large requirement for protein biochemistry. Standardized chemical treatments that are adequate for handling DNA in the sequencing work are not appropriate for folded and

structured proteins. Traditionally, a new protein is treated with individual care and often becomes a lifetime friend. As Sidney Brenner once pointed out, 100 000 gene products mean 100 000 chairs of biochemistry. The pressure of numbers is about to sweep away this traditional approach. Unfortunately, this will not happen without loss. The assembly-line approach may not turn out to be suitable for the majority of proteins.

(c) *How many protein structures?*

Such strategies select soluble proteins. Moreover, since there will be pressure to produce results quickly, difficult proteins (which denature or get degraded easily) will get passed over. Therefore there could be a bias towards looking at the proteins we already know, since these have already been operationally selected for their solubility, stability and ease of handling. Therefore many protein structures will be solved many times (as is presently the case) leading to useful redundancy but without contributing anything new to the canonical list of structures.

Such automatic strategies also leave out all the membrane proteins since these are generally insoluble if expressed in *E. coli*. Membrane proteins make up about one-quarter of a genome. Special methods that take account of the membrane-targeting signals present in such proteins will have to be used. Moreover, obtaining 3D crystals from membrane proteins is still far from routine.

Another class of problem not addressed by the structural genomics initiatives is protein–protein interaction. Clearly not all proteins interact with each other, but quite a lot do. X-ray crystallography of protein complexes has provided remarkable insight into, for example, G-protein signalling pathways. There is considerable interest in cell biology in the specifics of protein–protein interactions.

Another use of structure is mapping the effects of point mutations that have been introduced into a protein in order to modify and elucidate function. Each point mutation that has a significant functional correlate is worth another structure determination. The number of protein structures that need to be determined in order to understand cell biology is, therefore, essentially unlimited. At present 'structure saturation' is not a factor we need to consider in allocating resources.

The question of how many structures can actually be determined depends on a number of factors, including funding policy. Numbers such as 20 000 in the next ten years have been discussed. The implications of such targets in terms of human resources and training are only just beginning to be considered. From the point of view of the technique of protein crystallography, 20 000 in the next ten years seems to be a feasible target. It is, however, by no means certain that the biochemical–crystallization part of this programme can be realized.

What should we do with all these structures? Structural biochemists who have the experience and insight to look at a structure with, say, 8000 atoms and get some sense out of it in terms of function are at present limited to a few hundred in the world. Moreover, such insight always leans heavily on a background of biochemical and functional analysis. Semi-automatic structure determination as is foreseen in structural genomics is therefore liable initially to expand our ignorance rather than diminish it. However, in the long term it should prove invaluable. It will lead to a very different emphasis on the way biochemistry is done. When choosing a new project one will go to the database and choose a protein with a known structure. The database will also contain lots of useful information about preparation, stability and guessed substrate specificity. Knowledge of the structure should then allow focused experiments to proceed quickly. Probably within a decade or two it will be difficult for a young biochemist to appreciate how biochemistry used to be done blind without a 3D model around to guide one's experiments.

(d) *Planning undulator beam lines*

The dependence of structural biology on undulator beam lines has forced a certain degree of thought about the planning of electron storage rings to act as X-ray sources for the coming deluge. High-energy electron storage rings suitable for undulator insertion devices costs around US$100 000 000 (a rather global estimate). Not only are such machines expensive but also they cannot be purchased from a catalogue. Planning times of ten years are normal. Therefore one has to estimate now how many protein structures are going to be solved in the next 10–20 years. If an undulator beam line can collect data at the rate of one data set every 30 min and if the beam line is operational round the clock for 200 days a year, then each beam line could collect about 10 000 sets of data. Data sets have to be collected a number of times to solve the phase problem. Moreover, not all data sets will be good. However, it looks as if it might be possible to get a throughput of about 500 structures a year if other factors such as data retrieval and crystal handling are carefully optimized. Taking the goal of 20 000 structures in ten years (or 2000 in one year), it would seem that five to ten beam lines in the world are about right. The combined resources of Europe, the USA and Japan are slightly above this level. Therefore, for the next decade we will probably get by if the beam line infrastructure is strengthened.

What about the second decade of the new millennium? With the long time-lag for planning and commissioning machines we need to estimate now what the needs might be then. This problem cannot be solved unambiguously and someone may need to set priorities. However, it is difficult to set realistic priorities for ten years hence. We can extrapolate from the time taken for doubling the number of structures solved, which is at present about four years. Hence it might be appropriate to increase present capacity by a factor of four or five in the next decade. This can be achieved either by building more machines or reallocating resources on existing machines. The seriousness with which the biological community takes this problem is demonstrated by the fact that a large fraction of the costs of the planned British 3 GeV storage ring DIAMOND will be borne by the Wellcome Foundation.

(e) *Impact of NMR*

The major alternative method to X-ray diffraction for structure determination is NMR spectroscopy. NMR can measure distances between protons that are close to each other. If the sample is enriched in the isotopes ^{13}C and

^{15}N then distances between these nuclei and protons can also be measured. If enough pairwise distances can be measured one has the basis of a structure determination. Over 1500 protein structures have now been determined by this method (see Wuthrich (1998) for review). The advantage of NMR over X-ray crystallography is that it measures proteins in solution. Its main drawback is that it only works with proteins of low molecular weight. Recently, however, ways of extending the methods to higher molecular weights have been developed (Salzmann *et al*. 1998). Furthermore, the use of magnets with ever-higher fields increases sensitivity. NMR will make an increasingly important contribution to biological structure, especially to structural genomics, since it avoids the need for crystallizing the protein. However, it is unlikely to supplant X-ray crystallography as the primary means of getting the 3D structure of biological macromolecules.

(f) *Electron microscopy*

Electrons in an electron microscope typically have energy of 100–300 keV. The interaction with the specimen is much stronger than for X-rays but the radiation damage is not so severe, which would appear to make electron microscopy the method of choice for structure determination (Henderson 1995). Since X-ray crystallography appears to give up at about 20 μ crystal size, perhaps electron crystallography could take over for really small crystals. One major drawback is that the specimen has to be viewed in high vacuum, although the use of cryotechniques has considerably mitigated this problem. The more serious drawback is that (because of the strong interaction with the specimen) the specimen has to be thin (100–200 nm), which severely limits the applicability to crystalline proteins. However, this makes it the ideal method for dealing with membrane proteins in their native state (embedded in a membrane). Such specimens are often very reluctant to form 3D crystals suitable for X-ray crystallography but can be viewed as single sheets of membrane (two-dimensional crystals) in the electron microscope. Electron microscopy can determine the structure of such two-dimensional sheets of membrane proteins at atomic resolution (Henderson *et al*. 1990). The method finds more and more application, often as a method of assembling protein subunits into complexes (Stowell *et al*. 1998). However, it is still far from being routine. Some of the problems are instrumental. A number of features can be added to electron microscopes to improve their resolution and sensitivity. In fact, one could probably make the method available for a large class of membrane proteins with a fraction of the investment that has gone into NMR. For membrane proteins which cannot be persuaded to form 3D crystals (almost certainly most of them) there is at present no better method available.

5. CONCLUSION

Structural biology is now an integral part of cell biology. The rate and ease of solving structures has reached a level where cell biologists rather than crystallographers are ready to solve structures that are relevant and central to their interests. In the next 20 years one hopes to arrive at a situation where the cell biologist/biochemist will be able to look up the 3D structure of a protein much as one nowadays looks up the sequence. Relational databases will correlate structure and function. This should lead to protein function routinely being understood in chemical and physical terms and thereby bring us closer to a realization of Descartes's vision.

REFERENCES

Abrahams, J. P., Leslie, A. G., Lutter, R. & Walker, J. E. 1994 The structure of F1-ATPase from bovine heart mitochondria determined at 2.8 Å resolution. *Nature* **370**, 621–628.

Bennett, J. M. & Kendrew, J. C. 1952 Computation of Fourier syntheses with a digital electronic calculating machine. *Acta Crystallogr.* **5**, 109–113.

Blake, C. C., Koenig, D. F., Mair, G. A., North, A. C., Phillips, D. C. & Sarma, V. R. 1965 Structure of hen egg-white lysozyme. A three-dimensional Fourier synthesis at 2 Angstrom resolution. *Nature* **206**, 757–761.

Bragg, W. L. & Perutz, M. F. 1952 The structure of haemoglobin. *Proc. R. Soc. Lond.* A **213**, 425–435.

Brenner, S. E., Barken, D. & Levitt, M. 1999 The PRESAGE database for structural genomics. *Nucl. Acids Res.* **27**, 251–253.

Chang, A. C. Y., Nunberg, J. H., Kaufman, R. J., H.A., E., Schimke, R. T. & Cohen, S. N. 1978 Phenotypic expression in *E. coli* of a DNA sequence coding for mouse dihydrofolate reductase. *Nature* **275**, 617–624.

Garman, E. F. & Schneider, T. R. 1997 Macromolecular crystallography. *J. Appl. Crystallogr.* **30**, 211–237.

Geeves, M. A. & Holmes, K. C. 1999 Structural mechanism of muscle contraction. *A. Rev. Biochem.* **68**, 687–727.

Gerstein, M. & Levitt, M. 1997 A structural census of the current population of protein sequences. *Proc. Natl Acad. Sci. USA* **94**, 11 911–11 916.

Green, D. W., Ingram, V. M. & Perutz, M. F. 1954 The structure of haemoglobin IV. Sign determination by the isomorphous replacement method. *Proc. R. Soc. Lond.* A **225**, 287–307.

Henderson, R. 1995 The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Q. Rev. Biophys.* **28**, 171–193.

Henderson, R., Baldwin, J. M., Ceska, T. A., Zemlin, F., Beckmann, E. & Downing, K. H. 1990 Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *J. Mol. Biol.* **213**, 899–929.

Hendrickson, W. A. 1991 Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science* **254**, 51–58.

Hubbard, T. J. P., Murzin, A. G., Brenner, S. E. & Chothia, C. 1997 SCOP: a structural classification of proteins database. *Nucl. Acids Res.* **25**, 236–239.

Hubbard, T. J., Ailey, B., Brenner, S. E., Murzin, A. G. & Chothia, C. 1998 SCOP, Structural Classification of Proteins database: applications to evaluation of the effectiveness of sequence alignment methods and statistics of protein structural data. *Acta Crystallogr.* D **54**, 1147–1154.

Huxley, H. E., Simmons, R. M., Faruqi, A. R., Kress, M., Bordas, J. & Koch, M. H. J. 1981 Millisecond time-resolved changes in X-ray reflections from contracting muscle during rapid mechanical transients, recorded using synchrotron radiation. *Proc. Natl Acad. Sci. USA* **78**, 2297–2301.

Johnson, L. N., Phillips, D. C. & Rupley, J. A. 1968 The activity of lysozyme: an interim review of crystallographic and chemical evidence. *Brookhaven Symp. Biol.* **21**, 120–138.

Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R., Davies, D. R., Phillips, D. C. & Shore, V. C. 1960 Structure of

myoglobin: a three dimensional Fourier synthesis at 2 Å resolution. *Nature* **185**, 422–427.

Levitt, M., Gerstein, M., Huang, E., Subbiah, S. & Tsai, J. 1997 Protein folding: the endgame. *A. Rev. Biochem.* **66**, 549–579.

Montelione, G. T. & Anderson, S. 1999 Structural genomics: keystone for a human proteome project. *Nature Struct. Biol.* **6**, 11–12.

Perutz, M. F. 1972 Nature of heme–heme interaction. *Nature* **237**, 495–499.

Perutz, M. F. & TenEyck, L. F. 1972 Stereochemistry of cooperative effects in hemoglobin. *Cold Spring Harb. Symp. Quant. Biol.* **36**, 295–310.

Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G. & North, A. C. T. 1960 Structure of haemoglobin: a three dimensional fourier synthesis at 5.5 Å resolution obtained by x-ray analysis. *Nature* **185**, 416–422.

Perutz, M. F., Wilkinson, A. J., Paoli, M. & Dodson, G. G. 1998 The stereochemical mechanism of the cooperative effects in hemoglobin revisited. *A. Rev. Biophys. Biomol. Struct.* **27**, 1–34.

Rosenbaum, G., Holmes, K. C. & Witz, J. 1971 Synchrotron radiation as a source for x-ray diffraction. *Nature* **230**, 434–437.

Salzmann, M., Pervushin, K., Wider, G., Senn, H. & Wuthrich, K. 1998 TROSY in triple-resonance experiments: new perspectives for sequential NMR assignment of large proteins. *Proc. Natl Acad. Sci. USA* **95**, 13 585–13 590.

Shapiro, L. & Lima, C. D. 1998 The Argonne Structural Genomics Workshop: Lamaze class for the birth of a new science. *Structure* **6**, 265–267.

Stowell, M. H., Miyazawa, A. & Unwin, N. 1998 Macromolecular structure determination by electron microscopy: new advances and recent results. *Curr. Opin. Struct. Biol.* **8**, 595–600.

Teichmann, S., Chothia, C. & Gerstein, M. 1999 Advances in structural genomics. *Curr. Opin. Struct. Biol.* **9**, 390–399.

Westbrook, E. 1997 CCD-based area detectors. *Meth. Enzymol.* **276**, 244–268.

Wuthrich, K. 1998 The second decade—into the third millenium. *Nature Struct. Biol.* **5**(Suppl), 492–495.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES