A vision of statistical bioinformatics

Kanti V. Mardia

Department of Statistics, University of Leeds

1 Introduction

We open this paper with Astbury's definition of Molecular Biology - the subject which has led to the rise of Bioinformatics.

".... not so much a technique as an approach, an approach from the viewpoint of the so-called basic sciences with the leading idea of searching below the large-scale manifestations of classical biology for the corresponding molecular plan. It is concerned particularly with the forms of biological molecules and .... is predominantly three-dimensional and structural - which does not mean, however, that it is merely a refinement of morphology - it must at the same time inquire into genesis and function" W.T. Astbury (1952, 1961)

Bioinformatics is a new and rapidly developing field; its importance has been highlighted, for example, by the Human Genome Project. It includes mathematical, statistical and computing methods which aim to solve biological problems using DNA, amino acid sequences, and related information. The amount of available data, and the pool of human knowledge, is rapidly growing, and there are many non-trivial statistical problems to solve.

Broadly speaking, Bioinformatics is concerned with the organization and the function of cells as well as the underlying molecular interactions. This is a new discipline that brings together the traditional fields of biology (including molecular biology and biochemistry), biophysics, physiology, physics, molecular medicine, informatics and statistics. The work is very interdisciplinary (see also Mardia *et al.*, 2003 and Mardia, 2004).

Indeed, large amounts of numerical information arise from studies of the genome and proteins. The data come in various forms including genome sequences, protein co-ordinates and related chemical information; and microarray information on gene and protein expression.

Examples of problems include:

(i) Microarrays in which information is obtained on gene expression.

(ii) Non-coding of RNA prediction, regulatory region modelling.

(iii) Protein Structure in which the geometrical configuration allows docking of ligands, and binding.

(iv) Protein interaction in a functional sense: feedback, loops, cascade.

(v) Evolutionary trees/Phylogeny estimates.

(vi) Control mechanisms and dynamic activity of the cell (Biological System).

(vii) Protein folding.

Note that this subject is distinct from population genetics/inheritance laws and the more standard analysis of clinical trials and Health Informatics or even of Fisherian Genetics!

We first give some historical insight into Double Helix Modelling and on Protein Structure Computations (Section 2)! Special mention will be made of Professor William Thomas Astbury (1898-1961) whose contribution during his 33 years at Leeds (who was the Professor of

"Biomolecular Structure" ... the title he wanted was of "Molecular Biology" ... the name he coined!). Professor Astbury played an important role in the early understanding of DNA and proteins but the biggest celebrated discoveries leading to a Nobel Prize were left for Watson-Crick, Pauling etc. Section 3 describes the role of Statistics and Section 4 stresses the need for a paradigm shift by statisticians to make significant contributions to this new field.

2 Historical advances in bioinformatics

2.1 The double helix and DNA

Following Bookstein (2003), we examine here how James Watson and Francis Crick (1953) "proposed" the double helix model for DNA based on one *single*(!) X- ray image, (Figure 1), and the laws of structural chemistry. One of the very daring steps in this most celebrated work in life sciences. Indeed, it was all announced in virtually a single page!! This was sparked by the single image of Franklin and Gosling (1953) (which has many stories behind its availability to the other workers). This paper does quote the X-ray images of W. Astbury but were not of the same resolution (1 paper out of 6 papers in the references).

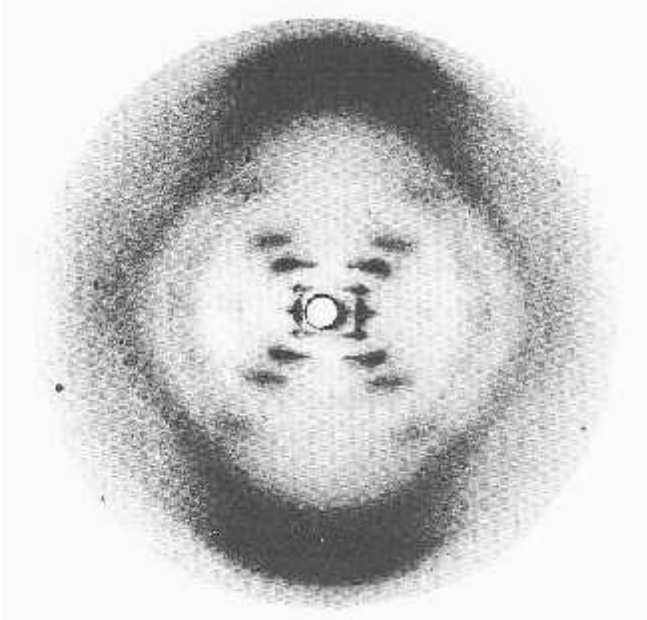

Figure 1: An X-ray photograph of DNA taken by Rosalind Franklin late in 1952.

"We wish to put forward a radically different structure." The paper goes on to introduce the double helix and then some "assumptions": "There is a residue on each chain every 3.4 A[ngstroms]. ... We have assumed an angle of 36 degrees between adjacent residues in the same chain, so that the structure repeats after ten residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fiber axis is 10 A. ... The previously published X-ray data on deoxyribose nucleic acid is insufficient for a rigorous test of our structure. ... It must be regarded as unproved until it has been checked against more exact results. Some of these are presented in the following communications [articles by Wilkins, Franklin and Gosling]. We were not aware of these results when we derived our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments."

The numbers 3.4 A., 34 A., and 10 A. of the published article thus were actually measured quantities. The article in the same issue of *Nature* by Franklin and Gosling explains these measurements explicitly:

> " The X-ray diagram of structure B shows in striking manner the features characteristic of helical structures, first worked out in this laboratory [by a team that included Francis Crick]. ... If we adopt the hypothesis of a helical structure, it is immediately possible to make certain deductions as to the nature and dimensions of the helix... In the present case the fibre-axis period is 34 A. and the very strong reflection at 3.4 A. lies on the tenth layer line. ... This suggests strongly that there are exactly 10 residues per turn of the helix.... Measurements of [the distances of those bands from the center] lead to values of the radius of about 10 A. ... We find that the phosphate groups or phosphorus atoms lie on a helix of diameter about 20 A., and the sugar and base groups must accordingly be turned inward toward the helical axis. ... The structural unit probably consists of two co-axial molecules which are not equally spaced along the fibre axis; ... this would account for the absence of the fourth layer line maxima and the weakness of the sixth. Thus our general ideas are not inconsistent with the model proposed by Watson and Crick in the preceeding communication."

The only remaining task was to figure out how the base pairs fit inside the helices. Watson noted that the like-with-like model for pairing required an impossible rotation angle between bases, and further that it offered no explanation at all for Chargaff's rules. To quote Watson's words (1968):

> "Suddenly I became aware that an adenine-thymine pair held together by two hydrogen bonds was identical in shape to a guanine-cytosine pair held together by at least two hydrogen bonds. All the hydrogen bonds seemed to form naturally; no fudging was required. ... Chargaff's rules then suddenly stood out as a consequence of a double-helical structure for DNA. In about an hour [the next day, when the metal models finally arrived] I had arranged the atoms in positions which satisfied both the X-ray data and the laws of stereochemistry."

Chargaff's rules, known since about 1949, are the experimental finding that in DNA preparations the ratios of A to T and G to C are "nearly unity." Surprisingly, neither source notes that the model supplies a testable prediction, namely, that the ratios should be not "nearly" but exactly unity.

At this point, what actually happened is what the paper says has not yet been done. From Watson's book:

> "Exact coordinates [needed to be] obtained for all the atoms. It was all too easy to fudge a successful series of atomic contacts so that, while each looked almost acceptable, the whole collection was energetically impossible. ... Thus the next several days were to be spent using a plumb line and a measuring stick to obtain the relative positions of all atoms in a single nucleotide. ... The final refinements of the coordinates were finished the following evening. Lacking the exact X-ray evidence, we were not confident that the configuration chosen was precisely correct. But this did not bother us, for we only wished to establish that at least one specific two-chain complementary helix was stereochemically possible. Until this was clear, the objective could be raised that, although our idea was aesthetically elegant, the shape

of the sugar-phosphate backbone might not permit its existence. Happily, now we knew that this was not true; ... a structure this pretty just had to exist."

Indeed Watson and Crick's discovery wasn't actually confirmed until the 1980s. As Crick has noted (1988, p.73) "It took over twenty-five years for our model of DNA to go from being only rather plausible, to being *very* plausible ... and from there to being virtually certainly correct."

2.2 X-ray diffraction

Wilkins *et al.* (1953) in the same issue of *Nature* also have a figure but is not as good quality as of Franklin and Gosling (1953). We quote from Wilkins *et al.* "While the biological properties of deoxypentose nucleic acid suggest a molecular structure containing great complexity, X-ray diffraction studies described here (cf. Astbury) show the basic molecular configuration has great simplicity. The purpose of this communication is to describe, in a preliminary way, some of the experimental evidence for the polynucleotide chain configuration being helical, and existing in this form when in the natural state."

The x-ray diffraction pattern is interpreted as follows in Franklin and Gosling (1953) "The innermost maxima on the first, second, third and fifth layer lines lie approximately on straight lines radiating from the origin. For a smooth single-strand helix the structure factor on the n^{th} layer line is given by:

$$F_n = J_n(2\pi r R) \exp in(\Psi + \pi/2)$$

where $J_n(u)$ is the nth-order Bessel function of u, r is the radius of the helix, and R and Ψ are the radial and azimuthal co-ordinates in reciprocal space; this expression leads to an approximately linear array of intensity maxima of the type observed, corresponding to the first maxima in the functions J_1, J_2, J_3, etc."

Similar comments are in Wilkins *et al.* (1953). A helical pattern in their Fig 2 (pattern C) from Bessel functions roughly matches x-ray pattern.

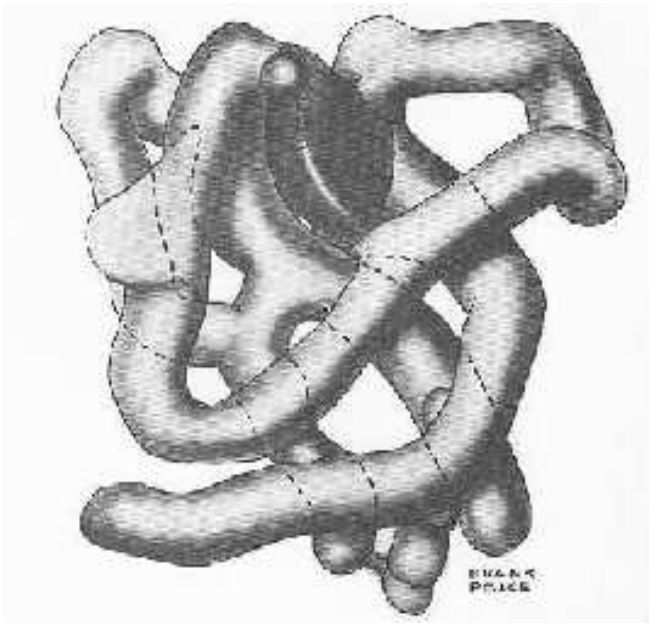

Figure 2: The first model of myoglobin at 6Å (Kendrew *et al.*, 1958)

"It may be shown that the intensity distribution in the diffraction pattern of a series of points equally spaced along a helix is given by the squares of Bessel functions. A uniform continuous helix gives a series of layer lines of spacing corresponding to the helix pitch, the intensity distribution along the n^{th} layer line being proportional to the square of J_n, the n^{th} order Bessel function. A straight line may be drawn approximately through the innermost maxima of each Bessel function and the origin. The angle this line makes with the equator is roughly equal to the angle between an element of the helix and the helix axis. If a unit repeats n times along the helix there will be a meridional reflexion (J_0^2) on the n^{th} layer line. The helical configuration produces side-bands on this fundamental frequency, the effect being to reproduce the intensity distribution about the origin around the new origin, on the n^{th} layer line, corresponding to C (in a figure)."

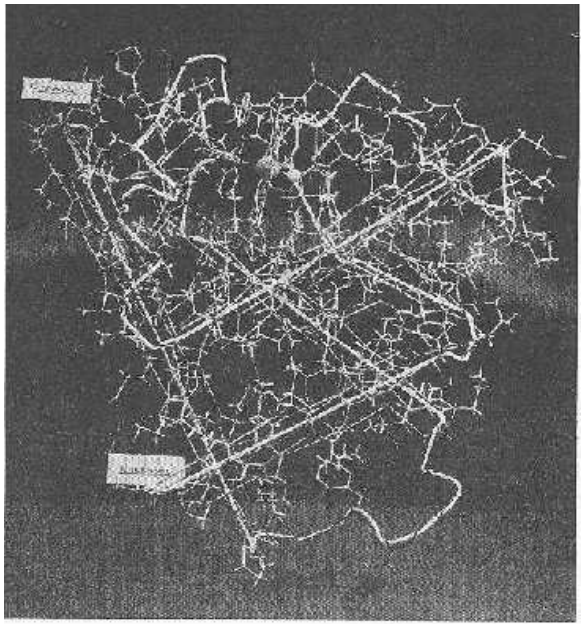

Figure 3: Model of myoglobin including side-chains (Kendrew *et al.*, 1961)

These various papers in the *Nature* of 1953, together with reviews of Watson of 1958, are well covered in Watson (1980).

2.3 Myoglobin

Bennett and Kendrew (1952) is the first paper published in a scientific journal on the application of an electronic computer to computational biology; the conference paper Bennett and Kendrew (1951) gives a briefer summary. Their conference paper can be considered as the first paper in Protein Bioinformatics. The paper starts with

"The basic task of the crystallographer is to determine the structure of molecules, organic and inorganic, from photographic patterns which result when regular arrangements of these molecules, i.e. crystals, are irradiated with X-rays. .... The geometrical form of a crystal is the consequence of the regular arrangement of the molecules of which it is built up; the regularity is that of a three-dimensional pattern which is repeated over and over again in space. It has been recognised for many years that the geometry of such space patterns, together with the observed facts of crystal form and symmetry, could give much information about the internal structure of the molecules of which the crystal is built up; but it was not until the development

of X-ray crystallography that it became possible to obtain information which could lead to the determination of the actual arrangement of the molecular units...."

The technique boils down to computing some Fourier series expansions! (see Booth, 1948)

Though the full 3-D structure image of myoglobin was published in 1958 (Kendrew, *et al.*, 1958) at 6 Angstrom resolution (1 Angstrom $=10^{-8}$ cm; see Fig 2), it does not show any atomic structure or a discernible pattern. In 1961, Kendrew *et al.* (1961) gave the first almost complete structure (see Fig 3). This structure was fully completed by our Astbury Professor, Simon Phillips in 1980! See the pdb site!! It has 153 residues and 1601 atoms so one can understand the hard task in 1951 for a computer! It has 8 alpha helices and rests are loops.

Astbury (1960), in one of his last papers, has given his own interpretation of work related to Kendrew. "And in a way, one of the strangest results of this long investigation is that about two-thirds of the myoglobin chain turns out to be in the alpha-configuration, that earliest example of protein chain-folding demonstrated over thirty years ago in X-ray studies of the structure of wool."

Myoglobin's function is to store oxygen (originally supplied by haemoglobin) in the tissues. It is particularly important for diving animals, such as whales, which need huge amounts of it to satisfy their needs during periods when atmospheric oxygen is not available to them. This explains the choice of the sperm whale as source for the protein that was used for the X-ray analysis.

2.4 Astbury

In 1934, Astbury speculates that "fibrous structures might be the basis for the crystallirity (of pepsin as discovered by Bernal and Crowfeet): globular proteins in general might be folded from elements essentially like elements of fibrous proteins!" (Tanford and Reynolds, 2001).

How right he proved to be, but that was much later. Bernal (1963) in his biography of Astbury considered this recognition that there might be no radical difference between fibrous and crystalline proteins as one of his great contributions.

Overall in Bernal's words Astbury 'influenced everybody's thinking about large biological molecules' and 'was the father of all those who since then interpreted other types of fibrous structure ... and who can recognize types of twist from the pattern of blurs on rather obscure fields.' (Bernal, 1963). The two principal discoveries missed by Astbury (Lydon, 2005) are:

THE ALPHA-HELIX FOR PROTEINS

Reasons why he missed it:

(1) Like Bragg and most others in the field, he was looking for rational (i.e. integral) helical models.

(2) He did not realise/know that the peptide bond was planar and rigid. His models were too flexible whereas those of Linus Pauling had a much more limited conformational flexibility making it easier to construct the correct model.

Why he was almost there:

(1) Astbury said repeatedly, in public, that a helix was most likely structure for linear polymers. His colleague, Ian Macarthur had obtained diffraction patterns from keratin showing the 1.5 A repeat almost impossible to explain in any other terms than a slow helix.

DNA AS THE GENETIC MATERIAL AND THE DOUBLE HELICAL STRUCTURE

Reasons why he missed it:

(1) Astbury seems to have clung to the idea that the genetic material must be protein rather than nucleic acid. He was not looking for a molecule which could carry the genetic message and replicate.

(2) He did not realise that the structure of DNA fibres is dependent on the humidity. Consequently his diffraction photographs were mixtures of the A and B forms and hence difficult to interpret.

Why he was almost there:

(1) His "pile of pennies" model for DNA was ok as far as it went. It explained the 3.4A reflection and had the bases stacked perpendicular to the axis. (It was in fact, half of the correct structure)

(2) He clearly had the concept of a linear message in his mind when he wrote of the long scroll on which is written the pattern of life. (Apparently a different concept from the earlier naive idea of genetics by spatial-fitting of some kind).

Olby (2005) comments on Astbury and his attitude on interdisciplinary research: "There were several problems. ... Could other Leeds researchers have helped him at the time? They had much respected strength in electron microscopy, and in organic chemistry, but I found no evidence of the collaborative spirit that crosses disciplines in his work.... As for mathematics — yes, if he had asked for help over Fourier theory to interpret the fibre diagrams of alpha keratin, and if he had questioned Pauling about the peptide bond." This is in quiet a contrast to Francis Crick and James Watson! For a further excellent coverage, we refer to Olby (1999).

Portugal and Cohen (1977) make an important point. "When the history of molecular biology comes to be written it will be seen that the work of Astbury from its beginnings in 1926 was, so to speak, the main line of progress of molecular biology. It started with his appreciation of the alpha fold, as he called it, *the alpha helix as we call it now*. All the way through he was guided by a profound sense of analogy, on the one hand with biological principles and on the other with textile techniques."

Astbury's first paper on DNA was in 1938 (Astbury and Bell, 1938) and comments (1963). "It is worthwhile remembering that subsequent to Astbury's pioneering, if limited, X-ray study on DNA published in 1938, there was no further work on this subject until that of Wilkins and his coworkers in the early 1950s. It should not detract from the work and originality of the Nobelists — Pauling, Perutz, Kendrew, Wilkins, Watson, and Crick — to say that they owed a great deal to their unorthodox predecessor, Astbury."

2.5 Protein bioinformatics

One cannot improve on the following quotes from Bryson (2004) in summarizing this field:

"The genome, as Eric Lander of MIT has put it, is like a parts list for the human body: it tells us what we are made of, but says nothing about how we work. What's needed now is the operating manual — instructions for how to make it go. We are not close to that point yet.

So now the quest is to crack the human proteome — a concept so novel that the term proteome didn't even exist a decade ago. The proteome is the library of information that creates proteins. 'Unfortunately', observed Scientific American in the spring of 2002, 'the proteome is much more complicated than the genome.'

That's putting it mildly. Proteins, you will remember, are the workhorses of all living systems; as many as a hundred million of them may be busy in any cell at any moment. That's a lot of activity to try to figure out. Worse, protein behavior and functions are based not simply on their chemistry, as with genes, but also on their shapes. To function, a protein must not only have the necessary chemical components, properly assembled, but then must also be folded into an extremely specific shape. "Folding" is the term that's used, but it's a misleading one as it suggests a geometrical tidiness that doesn't in fact apply. Proteins loop and coil and crinkle into shapes that are at once extravagant and complex. They are more like furiously mangled coat hangers than folded towels."

"... As the late French geneticist Jacque Monod put it, only half in jest: "Anything that is true of E. coli must be true of elephants, except more so." ... It cannot be said too often: all life is one. That is, and I suspect will forever prove to be, the most profound true statement there is."

3 Role of statistics: Past, present and future

The approaches used in this discipline range from the statistical analysis and organisation of the abundant experimental data to independent mathematical models for cellular and molecular processes, which are not directly accessible by experimental techniques. Of key interest is the quantitative analysis of cell function and the underlying molecular interactions.

Fortunately, there are many groups in the University of Leeds working on problems related to Bioinformatics.

In many areas of Bioinformatics, Statistics plays a key role because the data are incomplete, occasionally erroneous and very voluminous. Overall there is the need to extract a signal from a very noisy set of observations. In Statistics, there have been significant studies in Gene-expression microarray data, but very little work in Protein Biology. Shape Analysis, Directional Statistics and False Discovery Rates have very high potential to make significant breakthroughs in the understanding of Protein Structure and Prediction. Data Mining is another challenging area of statistical research in view of the rapidly growing biological databases. Thus, we foresee significant advances in Protein Biology, especially statistical breakthroughs in the area of Protein Structure and Prediction.

Some of the key problems today, in which Statistics is vital, include the following.

A Matching different sequences in the genome. The purpose here is to recognise the similarities in genes across different species and through evolution.

B In Microarrays the data contain information on gene expression and a key statistical issue is the huge number of multiple comparisons which could so easily lead to false discoveries. Typically there are 20-50 observations with 2,000-4,000 variables; a classic example of "fat data".

C Matching protein co-ordinates, "the alignment problem". Again the goal is to match proteins across different species and through evolution, and more generally to recognise parts of proteins that may have evolved separately but perform similar functions.

D One of the greatest challenges in protein analysis is the prediction of 3D protein structure from the sequence information of the amino acids. Initial methods of prediction are based on databases and concepts of similarity, but this approach does not provide direct understanding on how proteins fold.

E There is a pressing need for statistical understanding of complex pathways through protein expressions for critical diseases. In addition, gene expression data in microarrays present many challenges in further understanding of origins of diseases and the effectiveness of their treatments.

The challenges in System Biology are definitely many. We quote Costas (2002)

"One of the greatest related challenges will be combining all genetic, molecular, geometrical and environmental effects into a model with a coherent set of spatiotemporal dynamics, and how to estimate experimentally the involved parameters ... . While it is important to bear in mind that evolution will very likely not be explained as the development of optimal shapes for performing specific tasks implied by the environment, the shape-function paradigm will still essentially be involved with adaptation and fitness. In this sense, powerful approaches are required not only to characterise biological morphology, but also to estimate the degree of fitness of specific morphologies with respect to specific tasks. This important possibility involves the whole range of spatial and temporal scales, from proteins to environment, passing through cells, organs, members and individuals. Such endeavours will imply the application of concepts and tools from physics, image and shape analysis and dynamical systems."

In all these problems statistical concerns are paramount yet the discipline of statistics has not yet made as fundamental an impact in these areas as it should, except for sequence analysis and for microarrays. Thus there is a tremendous opportunity for statistics to make a stronger impact, and a fundamental need within the subject of Bioinformatics for this sort of input.

4 A need for a paradigm shift

There is definitely a need for a wake up call. In some sense as a profession, we always seem to suffer from delusion. A remark attributed to John Tukey (in Rao & Szekely, 2000) summarizes this well:

"The bulk of current statistical research appears to be finding exact solutions to wrong problems instead of approximate solutions to right problems."

It has been commented that we could be obsessed by the desire to create statistical models (see Breiman, 2001; Gilks, 2004). To the extent that Breiman in his stimulating paper claims that there are two cultures in the statistical modelling world (?) to reach conclusions from data; one of data modelling (practiced by 98% of all statisticians) and the other of algorithmic modelling (practiced by the remaining 2%). He makes a strong case for the second culture (the algorithmic modelling) which includes "model free" analysis using techniques such as CART, machine learning, data mining. The paper contains stimulating discussions by Cox, Efron and others. Efron equates this algorithmic culture to that of black boxes with lots of knobs to twiddle! Here, knobs to twiddle stands for adjusting a number of parameter values.

Indeed, Gilks (2004) emphasises that research in Bioinformatics would be helped from a shift of paradigm. There are many examples in genomics/bioinformatics where models are really only a means to an end, the end being to make predictions. For example, when comparing two proteins sequences that have diverged from a common ancestor (orthologs). We can write down probabilistic models, based on substitution rates, gap insertions etc., and statistical model based on hidden Markov models or stochastic context free grammars, but these do not really represent nature's mechanisms. The reality is much more complex, and poorly understood, and should involve the structure of the protein, which we may not even know. Also, there is the whole of natural selection acting on the sequence, which is very hard to model realistically. But all we really want from our model is some kind of scoring which we can use to predict whether two sequences really are orthologs. that is, we need some form of algorithmic modelling.

In Bioinformatics, algorithms are generally regarded as more important than statistical models or coherence (Gilks, 2004). Thus their subject in contrast has the following flavours

(i) New methods are mostly in a web-based tool (or software is freely available)

(ii) The model is only treated as a convenient part of an algorithm, eg. in a hidden Markov model of gene structure where conditional assumptions do not reflect a causative mechanism,

(iii) Coherance between submodels of a system or maximisation of proper likelihood has much lower priority than the computational efficiency or development time, bearing in mind that large data sets are involved.

That is (Gilks, 2004): "an efficient algorithm based on a heuristically justified objective function, delivered in reasonable time, is usually preferable to a principled statistical approach that takes years to develop or ages to run. Having said this, the case for a more principled approach can be made more effectively once cruder approaches have exhausted their harvest of low-hanging fruit."

To take it further with this modern age, it may be that the authors also provide their ups and downs on web related to their publication, i.e. a type of diary! As Medawar (1963) has argued that the scientific paper is a kind of fraud.

> "The neat format with its 'Introduction' followed by 'Methods', then 'Results' and finally 'Discussion' bears no relation to the way scientists actually work. While the final results must stand up to cold and objective scrutiny, the process of achieving them rarely takes the form of the calm and logical progression suggested by the telling. Purging events of all human emotion, the formal impersonal style totally fails to indicate who actually did what and why".

James Watson (1968) and Crick (1988) are definitely very good examples!

Thus statisticians need to be more open, more ready to learn "molecular biology" , more computationally aware, more ready to understand data banks, ...! But of course these require adequate resources!!

Acknowledgments

I wish to thank Wally Gilks, John Kent, John Lydon, Jeremy Norman, Vysaul Nyirongo, Robert Olby, Charles Taylor, Dave Westhead and Alistair Walder for their helpful comments.

References

Astbury, W.T. (1947). X-Ray Studies of Nucleic Acids *Symp. Soc. Exp. Biol.*, **1**, 66, Cambridge University Press.

Astbury, W.T. (1952) *The Harvey Lecture*, 1950-51. Thomas.

Astbury, W.T. (1960). X-ray diffraction studies of protein structure. *The New Scientist*, **8** , pp 93-95.

Astbury, W.T. (1961) Molecular biology or Ultrastructural biology, *Nature*, **190**, p 1124.

Astbury, W.T. and Bell, F.O. (1938). Some recent developments in the X-ray study of proteins and related structures. *Cold Spring Harbor Symp.*, **6**, 109-122.

Crick, F. (1988). *What Mad Pursuit*. Basic Books, New York.

Bernal, J.D. (1963). William Thomas Astbury. in : *Biographical Memoirs of Fellows of the Royal Society*, **9**, pp1-36.

Bennett, J.M. and Kendrew, J.C. (1951). The computation of Fourier syntheses with a digital electronic calculating machine. *Manchester University Computer Conference*, pp35-37. Tilloston, Bolton.

Bennett, J.M. and Kendrew, J.C. (1952). The computation of Fourier syntheses with a digital electronic calculating machine. *Acta Crystallographica*, **5**, 109-116.

Bookstein, F.L. (2003). *Quantitative reasoning and the double helix*. Personal communication.

Booth, A.D. (1948). *Fourier Technique in X-ray Organic Structure Analysis*. Cambridge University Press.

Bryson,B. (2004). *A Short History of Nearly Everything*. Black Swan, London.

Breiman, L. (2001). Statistical modelling: Two cultures *Statistical Science*, **16**, 199-231.

Costa, L. da F. (2002). Bioinformatics Today and Tomorrow. *Bioinformatics: Pharmatel*, 102-104.

Franklin, R.E. and Gosling, R.G. (1953). Molecular Configuration in Sodium Thymonucleate. *Nature*, **171**, 740-741.

Gilks, W. (2004). Bioinformatics: new science- new statistics. *Significance*, **1**, 7-9.

Green, P.J. (2003). Diversities of gifts, but the same spirit. *The Statistician*, **52**, 423-438.

Gribbin,J (1985). *In Search of the Double Helix*. McGraw-Hill, New York.

Lydon, T. (2005). Personal communication.

Mardia, K.V. (2003) Structural bioinformatics revisited. *Proceedings in Stochastic Geometry, Biological Structure and Images*, 11-18. Edited by Aykroyd, R.G., Mardia, K.V. and Langdon, M.J. Leeds University Press.

Mardia, K.V. (2004) LASR Workshops and emerging methodologies. *Proceedings in Bioinformatics, Images, and Wavelets* , 7-15. Edited by Aykroyd, R.G., Barber, S. and Mardia, K.V. Leeds University Press.

Medawar, P.B. (1963). Is the scientific paper a fraud? *The Listener*, 12th September issue.

Olby, R. (1999). *The Path to the Double Helix*. 2nd ed. Power publications, New York.

Olby, R. (2005). personal communication.

Phillips S.E.V. (1980). Structure and Refinement of Oxymyoglobin at 1.6A Resolution. *Journal of Molecular Biology*, **142**, 531-554.

Portugal, F.H. and Cohen, J.S. (1977). *A Century of DNA*. The MIT press. Cambridge, Mass.

Rao, C.R. and Szekely, G.J. (2000). (eds) *Statistics for the 21st Century* . Marcel Dekker. New York.

Tanford, C. and Reynolds, J. (2001). *Nature's Robots: A History of Proteins*. Oxford University Press.

Watson, J.D. (1968). *The Double Helix*. Atheneum, New York.

Watson, J.D. (1980). *The Double Helix*. Edited by G.S. Stent. W.W. Norton & Co. New York.

Watson, J.D. and Crick, F.H.C. (1953). A Structure for Deoxyribose Nucleic Acid. *Nature*, **171**, 737-738.

Wilkins, M.H.F., Stokes, A.R. and Wilson, H.R. (1953). Molecular Structure of Deoxypentose Nucleic Acids, *Nature*, **171**, 738-740.