

# What is statistical bioinformatics?

Kanti V. Mardia

Department of Statistics, University of Leeds

## 1 Definitions

We start with defining what is statistics followed by what is Bioinformatics, then yet this “undefined” subject of Statistical Bioinformatics.

Statistics is of course well defined now!

“Statistics is the detective work of extracting meaning from data! Though it is underpinned by the science of probability!!”

The detective work by Sherlock Holmes in “The Dancing Men” is a very good example! But the recent popular work on solving the puzzles of “The Da Vinci Code” is not a good example.

Coming to Bioinformatics, perhaps the original definition is of Astbury (1952):

“... 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”.

In the same vein, the physicist approach in Schrödinger (1944) has influenced many pioneers such as James Watson and Francis Watson (Gould, 1995, p25). However, the subject is very much now dominated by databases related to the genome and the proteome and other “eomes”! A whole new zoo!

Maybe we can formulate our definition as:

“Bioinformatics comprises the study of the DNA, proteins, RNA, etc., of organisms and their interaction, evolution and function”.

Microarrays, gene expression, protein expression, protein folding problems are all part of it. The subject focuses more on algorithms than statistical models. Thus biologists and computer scientists have made considerable headway in the field. The large size of the data sets have been the driving force. Indeed, in LASR 2002 (Mardia and Westhead, 2002), we suggested the definition:

“Bioinformatics is the science of managing and analysing genomic (molecular) data”.

Most of the work so far has not been model based but the exceptions where statistics has come into play are the p-values and e-values, multiple comparisons, false discovery rates, hidden Markov models for sequence analysis. Also, the work on microarrays has dominated the statistical scene! Thus at present there are very few statisticians at the heart of the subject!!

Perhaps the first text with statistical methods in bioinformatics is by Durbin et al (1998), followed by Ewens and Grant (2001). However, we found in 2000 that the statistics of protein

structure (3-D) had not received much attention. Then followed a series of articles in our LASR issues starting from LASR 2001 (eg. Demchuk et al, 2001). To quote, from the preface of LASR 2001:

“We all realise that further developments are taking place in bioinformatics related to DNA sequencing and proteins analysis, and have some review as well as forward looking papers in the Workshop. However, with all the excitement generated by genetics and genome sequencing, it is easy to forget that the primary purpose of most genes is to code proteins. Indeed, Fred Sanger, for instance, won his first Nobel Prize for sequencing a protein (insulin) in 1958 and his second for DNA sequencing techniques much subsequently in 1980. The proteins do the major work in building and controlling cells and tissues. Indeed, if gene sequencing is like the recording of music, the proteins are like the playback”.

Since then, at least in the UK, there has been a move to define the subject of “statistical bioinformatics”. From the web, one statistics department says:

“Statistical bioinformatics is the study of large biological data sets obtained by new micro-technologies by means of proper statistical methods”

Another describes the challenges:

“Modern biological assay techniques, derived from a combination of advances in biological science itself, the development of high-throughput measurement equipment, and the power and storage capacity of modern computers, are revolutionising our ability to examine the genetic make-up of humans, animals and plants from tissue samples. Huge quantities of data are becoming available, potentially of great value in aiding scientific understanding and promoting prevention and cure of disease. However, data of these kinds show very complex patterns of variation, from a variety of sources, both biological and technical, and there are therefore fascinating challenges for statisticians wishing to contribute to this area.”

On the other hand, we could say what it is not! (Mardia, 2005)

“This subject is distinct from population genetics/inheritance laws and the more standard analysis of clinical trials and health informatics, or even Fisherian Genetics!”

Although in 2006 the picture has already become less clear. There is growing interaction of genetics and genomics, eg. Hap Map project (2005), and there is increasing involvement of statisticians in that project.

The subject sits on the interface of statistics and biochemistry. A “true” definition will no doubt evolve but for the time being our working definition is:

“Statistical Bioinformatics is the art and the science of statistically modelling and analysing genomic and proteomic data, while keeping the biochemistry context prominent”.

The stress is in “analysing” using some plausible “models” holistically (Mardia and Gilks, 2005)!

## **2 How can this subject evolve?**

We started setting up dialogues through LASR Workshops for the last six years. This has led partly to our new Centre for Statistical Bioinformatics. In recent years, the LASR mission has been to act as a focus for statistical-bioinformatic interdisciplinary research bearing in mind

the international scene. Further, training of Ph.D. and M.Sc. students is essential. We already have the first Ph.D student (Vysaul Nyirongo) completing his thesis exclusively in Statistical Bioinformatics, and many are researching away! Our M.Sc. in Statistical Bioinformatics is also to start in 2007; some biologically motivated topics for this course are:

- sequence analysis
- phylogeny
- structure analysis
- function analysis.

Modern statistical tools have a lot to offer in this area: hidden Markov model, false discovery rates, directional statistics, shape analysis, Markov chain Monte Carlo methods, pattern recognition, fat data analysis, and so on. Computationally efficient methods, genomic data base understanding, visualisation tools etc are also inbuilt into the syllabus.

There are few statisticians deeply committed to the area except for gene expression in microarrays. Why? Because the biological jargon is so very vast. Abstraction and transfer of ideas from biological frame of reference to statistical framework is a painful exercise. Indeed, interaction between statisticians and scientists can at least take the following three forms:

- Consultancy type problem needing only standard statistical tools
- Incremental research problems
- Research problems which can lead to a “quantum leap” in biology.

### **3 Specific research problems**

Here is a description of some experiences where the Department here has made the shift. Perhaps a chronological development of some statistical bioinformatics problems might be the right order.

#### **3.1 Conformational entropy**

From the Department, the first paper (Demchuk et al, 2001) joint with the late Harsinder Singh in LASR 2001 came after my visit to National Institute for Occupational Safety and Health, Morgan Town, West Virginia. Looking back, the title of the paper

*“Statistics and Molecular Structure of Biological Macromolecules”*

was a bit too broad! But it takes on a problem of molecular modelling of torsional angles through directional statistics (of methanol molecule as an example). The idea was to understand factors that are involved in stability of a given conformational state of a protein (say), through conformational entropy of the state. First exposure to some biological jargon! To quote from Demchuk et al (2001):

“Proteins are compact polymers. Like shoelaces (usually schematically represented by ribbons), their polypeptide chains loop about each other in a variety of ways (i.e. they fold). Only one of these many ways allows the protein to function properly. Protein misfolding can lead to

insoluble lumps and can either cause or promote many deadly diseases such as the Alzheimer's disease, mad cow disease, cystic fibrosis and some types of cancer. In order to cure protein misfolding diseases, it is important to understand protein stability and the underlying physical-chemical principles of the process. One day we may witness a development of small molecules (drugs) that can correct or prevent misfolding problems, or new genetic therapies that substitute for them."

We are writing a follow up paper started with the late Dr Singh - leading to a deeper study of conformational entropy through multivariate von Mises distributions - these distributions are in the area of directional statistics, to which we have contributed a great deal at Leeds.

### 3.2 Matching and Assignment problems (Homology)

With the help of Nicola Gold and Dave Westhead, we were exposed in the same period to active sites of proteins and the problem of homology. If two active sites have "many" atoms in common then we gain the knowledge that the two may have the same function. Potentially, the knowledge may lead to developing a new drug. Also it could give some idea of the evolution tree. It occurred to us that the Department had already developed methods to electrophoretic gels (data from Chris Glasbey, Leeds University Ph.D. thesis of Gary Walker) published in Dryden and Mardia (1998). Statistically, these problems involve matching unlabelled configurations under unknown transformations. We started looking at a few simple examples of active sites (17-beta hydroxysteroid dehydrogenase and carbonyl reductase). We learnt of a computationally powerful method of Nicola Gold and Dave Westhead, it was though not statistical. The joint work first started with Charles Taylor (which has a new concept of so called "Coffin bin"! ) using regression type model which was implemented by an EM algorithm. My first talk was given on the topic at the local RSS Meeting in Nottingham, 16th May 2002. The framework was fully defined in Section 3.1 of LASR 2002 (Mardia and Westhead, 2002). The first preliminary approach is in LASR 2003 (Taylor, Mardia and Kent, 2003).

Meanwhile in my seminar at Bristol on 28th February 2003, Peter Green mentioned that our formulation could be improved.

Old formulation: Regression type

New formulation: Bayesian hierarchical model

The MCMC implementation of our model was presented at LASR 2004 by Charles in his talk, but was left out in Kent, Mardia and Taylor (2004). Meanwhile the approach with Peter Green has a lot to offer in new methodology and in decision making. The following new ideas have emerged.

- Probabilistic model generating configurations
- Matching parameter
- Noise component
- Rigid transformation and plausible distribution of orthogonal matrices
- Distribution of matching matrix  $M$
- Posterior distribution of matches.

The problem is high dimensional but a slick MCMC implementation has been introduced.

The next question was how to summarize the result?

- Appropriate loss function
- Marginal posterior probability of matches through linear programming
- New visualization tools (comb presentation)

Surprises: Sequence order not used in matching but that order is preserved in the matched output!

This work was presented by Peter Green last year (Green and Mardia, 2005) and published in Green and Mardia (2006). Since then the goal has been to determine what new insight it provides from a biological view point! The work has led to a computational method leading to better pattern of matches, i.e. glycine rich motif extended in glyceraldehyde-3 phosphate dehydrogenase structure (3dbv-3) as an example. This leads neatly to the idea of Bayesian refinement as a new principle (Mardia et al, 2006). The other papers in the LASR Proceedings on this matching problem include Mardia et al (2005), Mardia and Nyirongo (2004), Nyirongo et al (2005), and Schmidler (2003, 2004). The topic is also recently treated by Dryden et al (2006).

### 3.3 Ramachandran plots

These are basic scatter plots of the torsional angle ( $\phi$ ,  $\psi$ ) giving the idea of the modes of the distribution corresponding to the different secondary structures, eg. alpha-helix, beta sheets and loops. The original work was done in 1960s by Ramachandran and colleagues. Their importance is summarized clearly by the following quote from Rose (2001):

“No biochemistry textbook is complete without a  $\phi$ ,  $\psi$ -plot... This plot ranks alongside the double helix and the alpha-helix among fundamentals of structural biochemistry”.

But a statistical model has never been proposed. Subramaniam in his thesis here, made the first attempt to this problem and presented in LASR 2003 (Mardia, Taylor and Subramaniam, 2003). This has led to a deeper study of distributions on a torus. Indeed, it turned out that a mixture of such distributions provides a deep insight into the variability (Mardia et al, 2006). Such work has a great future since the angles play a key role in Bioinformatics. See for example, the simulation of secondary structure in Kent and Hamelryck (2005), and in Boomsma, Kent, Mardia, Taylor and Hamelryck (p. 91, this volume).

### 3.4 Future

We will take one sub-field of matching problems in Protein Bioinformatics. New effective tools are required not only in modelling but how to incorporate a variety of extra information such as:

- Colour information: eg. amino acid types
- Directions:  $C_\alpha$  and  $C_\beta$
- Lock and key: Docking problem
- Surface interactions.

Also the matching is to be incorporated under varying constraints:

- Chemical bonds (Van der Waals constraints)
- Order information: sequence order of amino acids
- Geometrical structures:  $\alpha$ -Helices,  $\beta$ -sheets, loops, ...

This is all a cry for Unlabelled Statistical Shape Analysis!

## 4 Discussion

There is definitely need for wake up call for statisticians (See also, Gilks, 2005; Mardia and Gilks, 2006). As John Tukey (also the inventor of the terms like software, hardware, bit) said (remark attributed to him in Rao and Szekely, 2000):

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

We also have obsessions with models but keeping an open mind is what is required. Also be aware of computing power - though Efron (2002) posed the following interesting scenario reminding that computer power is not enough!

“Suppose that you could buy a really fast computer, one that could do not a billion calculations per second, not a trillion, but an infinite number. So after you unpacked it at home, you could numerically settle the Riemann hypothesis, the Goldbach conjecture, and Fermat’s last theorem (this was a while ago), and still have time for breakfast. Would this be the end of mathematics?”

or even end of statistics? Perhaps not!

We again end as last year.

“Statisticians need to be more open, more ready to learn “molecular biology”, more computationally aware, more ready to understand databanks, ...”

But above all, we always need great scientist friends!! This all is a part of solving great questions in life sciences of taming the nature and immortality, etc!

## Acknowledgements

I wish to thank Stuart Barber, Wally Gilks, John Kent and Charles Taylor for their helpful comments.

## References

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

Demchuk, E., Hnizo, V., Mardia, K.V., Sharp, D.S. and Singh, H. (2001) Statistics and Molecular Structure of Biological Macromolecules. *Functional and Spatial Data Analysis*, 9-14. Edited by Mardia, K.V., Kent, J.T. and Aykroyd, R.G. Leeds University Press.

- Dryden, I.L., Hirst, J.D. and Melville, J.L. (2006). Statistical analysis of unlabelled point sets: comparing molecules in chemoinformatics. *Biometrics*, to appear.
- Dryden, I.L. and Mardia, K.V. (1998). *Statistical Shape Analysis*. Chichester, Wiley.
- Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998). *Biological sequence analysis. Probabilistic models of proteins and nucleic acids*. Cambridge, Cambridge University Press.
- Efron, B. (2002). The bootstrap and modern statistics. *Statistics in the 21st Century*, 325-332. Edited by Raftery, A.E., Tanner, M.A. and Wells, M.T. London, Chapman and Hall.
- Ewens, W.J. and Grant, G.R. (2001). *Statistical Methods in Bioinformatics - an Introduction*. New York, Springer-Verlag. (Second edition, 2005.)
- Gilks, W. (2004). Bioinformatics: new science- new statistics. *Significance*, **1**, 7-9.
- Gould, S.J. (1995). What is Life? as a problem in history. *What is Life? The Next Fifty Years*, 25-39. Edited by Murphy and O'Neill.
- Green, P.J. and Mardia, K.V. (2005). Matching and alignment in protein bioinformatics, using Bayesian hierarchical models. *Quantitative Biology, Shape Analysis, and Wavelets*, 29. Edited by Barber, S., Baxter, P.D., Mardia, K.V. and Walls, R.E. Leeds, Leeds University Press.
- Green, P.J. and Mardia, K.V. (2006). Bayesian Alignment Using Hierarchical Models, with Applications in Protein Bioinformatics. *Biometrika*, **93**, 235-254.
- The International Hap Map Consortium (2005). A haplotype map of the human genome. *Nature*, **437**, 1299-1320.
- Kent, J.T. and Hamelryck, T. (2005). Using the Fisher-Bingham distribution in stochastic models for protein structure. *Quantitative Biology, Shape Analysis, and Wavelets*, 57-60. Edited by Barber, S., Baxter, P.D., Mardia, K.V. and Walls, R.E. Leeds, Leeds University Press.
- Kent, J.T., Mardia, K.V. and Taylor, C.C. (2004). Matching problems for unlabelled configurations. *Bioinformatics, Images, and Wavelets*, 33-36. Edited by Aykroyd, R.G., Barber, S. and Mardia, K.V. Leeds, Leeds University Press.
- Mardia, K.V. (2004). LASR Workshops and emerging methodologies. *Bioinformatics, Images, and Wavelets*, 7-15. Edited by Aykroyd, R.G., Barber, S. and Mardia, K.V., Leeds University Press, 7-15.
- Mardia, K.V. (2005). A vision of statistical bioinformatics. *Quantitative Biology, Shape Analysis, and Wavelets*, 9-20. Edited by Barber, S., Baxter, P.D., Mardia, K.V. and Walls, R.E. Leeds, Leeds University Press.
- Mardia, K.V. and Gilks, W. (2005). Meeting the statistical needs of 21st-century science. *Significance*, **2**, 162-165.
- Mardia, K.V., Green, P.J., Nyirongo, V.B., Gold, D.N. and Westhead, D.R. (2006). Bayesian refinement of protein functional site matching, submitted.
- Mardia, K.V. and Nyirongo, V. (2004). Procrustes statistics for unlabelled points and applications. *Quantitative Biology, Shape Analysis, and Wavelets*, 137. Edited by Aykroyd, R.G., Barber, S. and Mardia, K.V. Leeds, Leeds University Press.

- Mardia, K.V., Patrangenaru, V. and Sugathadasa, S. (2005). Protein gels matching. *Quantitative Biology, Shape Analysis, and Wavelets*, 163-165. Edited by Barber, S., Baxter, P.D., Mardia, K.V. and Walls, R.E. Leeds, Leeds University Press.
- Mardia, K.V., Taylor, C.C. and Subramaniam, M. (2003). Applications of Circular Distributions to Conformational Angles in Proteins. *Quantitative Biology, Shape Analysis, and Wavelets*, 149-152. Edited by Aykroyd, R.G., Mardia, K.V. and Langdon, M.J. Leeds, Leeds University Press.
- Mardia, K.V., Taylor, C.C. and Westhead, D.R. (2003). Structural Bioinformatics Revisited. *Stochastic Geometry, Biological Structure and Images*, 11-18. Edited by Aykroyd, R.G., Mardia, K.V. and Langdon, M.J. Leeds, Leeds University Press.
- Mardia, K.V. and Westhead, D.R. (2002). New major challenges in bioinformatics. *Statistics of Large Data Sets: Functional and Image Data, Bioinformatics and Data Mining*, 9-15. Edited by Aykroyd, R.G., Mardia, K.V. and McDonnell P. Leeds, Leeds University Press.
- Murphy, M.P. and O'Neill, L.A.J. (1995). *What is Life? The Next Fifty Years*. Cambridge, Cambridge University Press.
- Nyirongo, V., Mardia, K.V. and Westhead, D.R. (2005). EM algorithm, Bayesian and distance approaches to matching functional sites. *Quantitative Biology, Shape Analysis, and Wavelets*, 155-156. Edited by Barber, S., Baxter, P.D., Mardia, K.V. and Walls, R.E. Leeds, Leeds University Press.
- Rao, C.R. and Szekely, G.J. (eds) (2000). *Statistics for the 21st Century*. New York, Marcel Delcker.
- Rose, G.D. (2001). In Memoriam: Professor G.N. Ramachandran (1922-2001). *Protein Science*, **10**, 1691-1692.
- Schmidler, S.C. (2003). Statistical shape analysis of protein structures families. *Stochastic Geometry, Biological Structure and Images*, 13. Edited by Aykroyd, R.G., Mardia, K.V. and Langdon, M.J. Leeds, Leeds University Press.
- Schmidler, S.C. (2004). Bayesian shape matching and protein structure alignment. *Quantitative Biology, Shape Analysis, and Wavelets*, 25. Edited by Aykroyd, R.G., Barber, S. and Mardia, K.V. Leeds, Leeds University Press.
- Schrödinger, E. (1944). *What is Life? The Physical Aspect of the Living Cell*. Cambridge, Cambridge University Press.
- Taylor, C.C., Mardia, K.V. and Kent, J.T. (2003). Matching Unlabelled Configurations Using the EM Algorithm. *Stochastic Geometry, Biological Structure and Images*, 19-21. Edited by Aykroyd, R.G., Mardia, K.V. and Langdon, M.J. Leeds, Leeds University Press.