Statistics and Molecular Structure of Biological Macromolecules

E. Demchuk[1,2], H. Singh[3], V. Hnizdo[1], K.V. Mardia[4] and D.S. Sharp[1]

[1]Heath Effects Laboratory Division, National Institute for Occupational Safety and Health,
Morgantown, WV 26505, USA;
[2]School of Pharmacy, West Virginia University, Morgantown, WV 26506, USA;
[3]Department of Statistics, West Virginia University, Morgantown, WV 26506, USA;
[4]Department of Statistics, University of Leeds, LS2 9JT, UK.

The proteins are biological macromolecules that are of primary importance to all living organisms. These macromolecules do the major work in building and controlling cells and tissues. There are more than 30,000 diferent kinds of proteins in our bodies. For example hemoglobin which carries oxygen to our tissues, insulin that regulates sugar level in the blood, the antibodies that fight infection, the actin and myosin that allows our muscles to contract, and keratin in skin, hair and finger nails are all proteins.

Proteins are polymers made up of many amino acids (often called residues) linked together to form a chain. Amino acids are molecules that consist of three functional groups, a basic amino group (NH_2), acidic (COOH) group and a side chain. There are 20 different kinds of amino acids. Two or more amino acids can merge together to form amide bonds which are also known as peptide bonds. The repeating chain of amide linkages to which the side chains are attached is called the *backbone*.

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 undelying physical-chemical principals 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. For more details about protein structures and protein misfolding the reader is referred to [11], [6].

With all the excitement generated by gene sequencing, it is easy to forget that the primary purpose of most genes is to code for proteins. Indeed, DNA is a molecule of life and 23 such very long macromolecules (namely chromosomes) code full information necessary to reproduce a human being. This information is read and interpreted depending on the conformation of a particular part of DNA molecule on which this information is recorded; i.e., depending on the conformation of DNA, genetic information may become readable or become silent. That is how different tissues (e.g. brain, liver and muscles) are developed from the identical set of information stored in the DNA molecule. Abnormal interpretation of genetic information may create tumor tissue. Thus, all biological processes are ultimately driven by the information and conformation (state) of DNA. Conformation of any polyatomic molecule including DNA and proteins is essentially determined by a set of its internal dihedral angles which result from rotations around molecular bonds. That is why it is important to study populations of these angles in the molecules and their coherent change from one conformation to another. To sum up, if gene sequencing is like the recording of music, then proteins are like the playback.

To understand factors that are involved in the stability of a given molecular conformational state (such as in a protein) and in changing from one conformation to another, it is important to determine the conformational entropy of the state. In molecular mechanics, the bond lengths, bond angles and rotational angles around the bond (torsional angles) constitute internal coordinates of the molecule. The bond lengths and bond angles are "rigid" degrees of freedom. Thus, the entropy of conformational ensemble is mainly determined by random rotational fluctuations around the bonds. Therefore, probabilistic modeling of torsional angles in molecules is a paramount goal on the way to full contol over the molecular processes.

Let $\Phi_1, \Phi_2, \ldots, \Phi_m$ be the m torsional angles of a a conformation, and $f(\phi_1, \phi_2, \ldots, \phi_m)$ be the joint probability density function of $\Phi_1, \Phi_2, \ldots, \Phi_m$. Then the internal configurational (conformational) entropy of the molecule is given by

$$S_c = -k_B E[\log(f(\Phi_1, \Phi_2, \ldots, \Phi_m))]$$

$$= -k_B \int\int \cdots \int \log(f(\phi_1, \phi_2, \ldots, \phi_m)) f(\phi_1, \phi_2, \ldots, \phi_m) d\phi_1 d\phi_2 \ldots d\phi_m$$

where E denotes expectation (or mean) and k_B is the Bottzman's constant. Note, that the configurational entropy in statistical thermodynamics is k_B times Shannon entropy defined in statistics.

Karplus and Kushik [5], and Levy et al. [7] have modeled internal torsional coordinates of macromolecules using multivariate normal distribution. Assuming that the Φ_i's are distributed according to a multivariate normal distribution with variance-covariance matrix Σ, the configurational torsional entropy is given by

$$S_c = \frac{mk_B}{2} + \frac{k_B}{2} \log[(2\pi)^m \mid \Sigma \mid]. \tag{1}$$

The authors have used this approach for entropy calculations on butane and decaglycine.

Reliance on a multivariate normal distribution that does not take in to consideration the circular nature of the torsional angles is a major drawback of this approach. It does not provide a good fit when either the fluctuations around rotatable bonds are large or if anglular distributions are multipeaked. Refer to [9] and [1] for a comprehensive review in circular statistics.

In order to solve this problem, Demchuk and Singh [2] introduced a probabilistic approach to molecular modelling that is based on circular probability distributions rather than a linear Gaussian approach. Assuming that Φ_i's are independent and Φ_i follow a von Mises distribution with the concentration parameter κ_i, they derived the following expression for configurational entropy of the molecule

$$S_c = k_B \left[m \log(2\pi) + \sum_{i=1}^{m} \log(I_0(\kappa_i)) - \sum_{i=1}^{m} \kappa_i \frac{I_1(\kappa_i)}{I_0(\kappa_i)} \right]. \tag{2}$$

where I_0 and I_1 are the modified Bessel functions of order 0 and 1 respectively.

As a case study, they modelled the torsional angle of methanol molecule (Figure 1) by the 3-mode von Mises distribution with a probability density function given by

$$f(\phi) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos[3(\phi - \phi_0)]}, \quad -\pi \leq \phi \leq \pi,$$

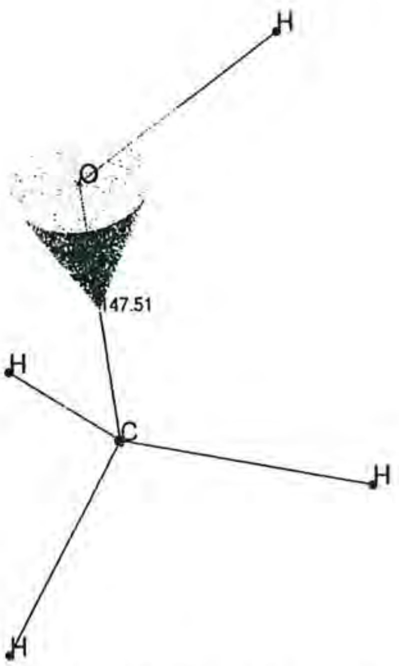

Figure 1: Methanol molecule. The cone denotes a rotatable bond.

where $\kappa > 0$ and $\phi_0 = -2\pi/3$. These three modes have a physical explanation. Let us fix in Figure 1 one of the three hydrogen atoms at CH_3 group. Then a hydrogen atom at the OH bond is free to rotate, whereas the other two hydrogen atoms at the CH_3 group are not. However, there exist three points of minimum along the path of this 'free' hydrogen atom because of repulsion by the three CH bonds. Further, θ is a dihedral angle between the planes HOC and OCH1, where H1 denotes the selected hydrogen atom at the CH_3 group.

With the assumption of 3-mode von Mises distribution, the torsional potential energy of the molecule is given by

$$V = \frac{V_0}{2} \{1 - \cos[3(\Phi - \phi_0)]\} \tag{3}$$

where V_0 is the maximum torsional potential energy of the molecule.

Demchuk and Singh [2] also derived the following bathtub shaped probability density function of the torsional potential energy V of the molecule

$$g(v) = \frac{1}{\pi I_0(\kappa)} e^{\kappa\left(1 - \frac{2v}{V_0}\right)} v^{-1/2} (V_0 - v)^{-1/2}, \; 0 \leq v \leq V_0. \tag{4}$$

The three mode von Mises distribution provided an excellent fit to the data of torsional angles in methanol obtained by molecular dynamics simulations. The bathtub shaped probability distribution provided an excellent fit to the torsional energy data (Figure 2).

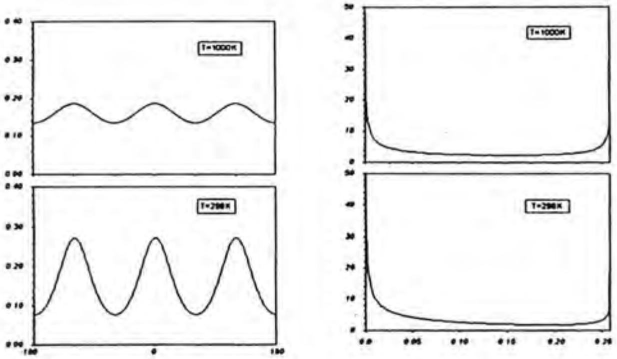

Figure 2: Histograms of torsion angles and energy of methanol. Angles at (a) 298 K and (b) 1000 K, energy at (c) 298 K and (d) at 1000 K. Fitted lines are solid.

In general, molecules have more than one torsional angle and these are often dependent. Macromolecules such as proteins have a very large number of torsional angles. See Figure 3 for picture of a small protein, human TNF-β factor, which is one of the most important human cytokines regulating the life and death of human cells, and it is one of the key mediators of AIDS pathogenesis. It consists of 144 amino acids or 2200 atoms and contains 707 torsional angles.

Singh, Hnizdo and Demchuk [12] introduced a new probability distribution on the torus for modeling the distribution of two dependent torsional angles of a molecule. The proposed distribution belongs to a general class of distributions introduced earlier by Mardia [8], [9]. Let Θ_1 and Θ_2 be two circular random variables in the range of $[-\pi, \pi]$. Singh et al. introduce a joint probability distribution for Θ_1 and Θ_2 with probability density function given by

$$f(\theta_1, \theta_2) = Ce^{\kappa_1 \cos(\theta_1-\mu_1)+\kappa_2 \cos(\theta_2-\mu_2)+\lambda \sin(\theta_1-\mu_1)\sin(\theta_2-\mu_2)}, \tag{5}$$

$-\pi \leq \theta_1, \theta_2 \leq \pi$, where $\kappa_1, \kappa_2 \geq 0, -\infty < \lambda < \infty, -\pi \leq \mu_1, \mu_2 \leq \pi$ and C is a normalizing constant, so that $f(\theta_1, \theta_2)$ is a probability density function.

They showed that for small fluctuations, the above distribution can be approximated by a bivariate normal distribution, and they obtained expressions for the normalizing constant and marginal circular variances. They also showed that conditional distributions are von Mises. The marginal distributions are either unimodal von Mises like or bimodal symmetric. They defined a generalization of the above model which allows an arbitrary number of peaks in the marginal distributions. They illustrated the utility of this distribution by modeling two angular variables in methanol and in a linear peptide described by Demchuk et al [3].

Mardia, Singh, Hnizdo and Demchuk [10] introduced a multivariate version of the above distribution and discussed moment and pseudo likelihood methods for estimation of the parameters of the distribution.

The above models assume marginal distributions to be symmetric. In general, at least some of the torsional degrees of freedom of macromolecules can have skewed marginal distributions. Hnizdo, Singh and Demchuk [4] proposed a fourier series expansion of the potential function

Figure 3: Human TNF-β factor.

approach for modeling such a distribution. For a large number of dependent torsional angles, this approach poses a serious computational challenge in the parameter estimation procedure.

Acknowledgements

Kanti Mardia would like to express his thanks to Kevin de Souza for his help.
K.V.M. (visit to NIOSH) and H.S. were partially supported by the Biostatistics Branch of NIOSH Health Effects Laboratory Division, #HELD99B0030.

References

[1] Batschelet, E. B. (1981), *Circular Statistics in Biology* (New York: Academic Press).

[2] Demchuk, E. and Singh, H. (2001), "Statistical thermodynamics of hindered rotation from computer simulations," *Molecular Physics*, Vol. 99, pp. 627-636.

[3] Demchuk, E., Bashford, D. and Case, D. A. (1997), "Dynamics of a type VI reverse turn in a linear peptide in aqueous solution," *Folding and Design*, 2, pp. 35-46.

[4] Hnizdo, V., Singh, H. and Demchuk, E. (2001), "A Fourier series expansion of the potential function approach to probabilistic modeling of torsional angles," preprint.

[5] Karplus, M. and Kushick, J. N. (1981), "Method for estimating the configurational entropy of macromolecules," *Macromolecules*, **14**, 325-332.

[6] Leach, A. R. (1997), *Molecular Modeling*, Longman.

[7] Levy, R. M., Karplus, M., Kushick, J. and Perahia, D. (1984), "Evaluation of the configurational entropy for proteins: Application to molecular dynamics simulations of an α-helix," *Macromolecules*, **17**, 1370-1374.

[8] Mardia, K. V. (1975), "Statistics of directional data (with discussion)," *Journal of Royal Statistical Society, Series B*, Vol. 37, pp. 349-393.

[9] Mardia, K. V. and Jupp, P. E. (1999), *Directional Statistics*, Wiley.

[10] Mardia, K. V., Singh, H., Hnizdo, V. and Demchuk E. (2001), "Multivariate von Mises distributions", Preprint.

[11] McMurry J. and Fay, R. C. (2001), *Chemistry*, Prentice Hall.

[12] Singh, H., Hnizdo, V. and Demchuk, E. (2001), "Probabilistic model for two dependent circular variables," communicated for publication.

Identification of differential gene expression from DNA microarrays

C.A. Glasbey and C.D. Mayer
Biomathematics and Statistics Scotland
JCMB, King's Buildings, Edinburgh, EH9 3JZ, Scotland

1 Introduction

Microarray (or DNA chip)-based hybridisation analyses using high density DNA probe arrays is a powerful tool for a broad and diverse set of genetic applications, including gene expression monitoring, sequence information, and genotype analysis (see, for example, Chipping Forecast, 1999). It allows new biomolecular approaches that promise to revolutionise our understanding of physiology and disease. A good starting point for finding out more about this rapidly developing field is the web site: `www.gene-chips.com`

To study the difference in gene expression between two samples, each gene is placed as a spot on a glass slide, and the slide is then hybridized with the two samples, each labelled with a fluorescent marker. Finally, the microarray is scanned at high spatial resolution at the two wavelengths. For example, Fig 1 shows an array comparing two strains of Human Cytomegalo Virus (HCMV). Red spots reveal genes only expressed by virus strain one, green spots show genes only expressed by strain two, yellow spots show genes expressed by both strains and dark spots by neither. The first stages in the analysis of such data are estimation of expression of each gene, and identification of differential gene expression. We consider them in the following two sections.

2 Image analysis

Image analysis is the first step in analysing microarray data. Methods of noise reduction, background correction and segmentation are needed before the integrated intensity of individual spots can be obtained. Work is reported at the US National Human Genome Research Institute web page `www.nhgri.nih.gov/DIR/LCG/15K/HTML/img_analysis.html` and in Yang et al. (2000). However, there remain needs and opportunities for exploration of alternative, improved methods and for extensive, empirical and theoretical comparisons between methods.

In the talk we illustrate the use of median filters to reduce the effects of speckle noise. We also use morphological operators such as the top-hat filter to correct the images for background trend, as proposed by Yang et al. (2000). Then we consider alternative segmentation methods to isolate the spots in the images.

To estimate the difference in spot intensity between the two samples, we model pixel values by bivariate log-normal distributions. The maximum likelihood estimator for this model, and

FUNCTIONAL AND SPATIAL DATA ANALYSIS

International Conference, held in Leeds, UK, 9-11 July 2001,

incorporating the 20th Leeds Annual Statistical Research Workshop

Sponsored and organized by the Department of Statistics, University of Leeds, UK.

Edited by

K.V. Mardia and R.G. Aykroyd
Department of Statistics,
University of Leeds, Leeds LS2 9JT, UK.