PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2006) **361**, 525–527
doi:10.1098/rstb.2005.1797
Published online 3 February 2006

Bioinformatic challenges for the next decade(s)

David Eisenberg[1,*], Edward Marcotte[2], Andrew D. McLachlan[3] and Matteo Pellegrini[4]

[1]*UCLA-DOE Institute for Genomics and Proteomics, Howard Hughes Medical Institute, Box 951570 UCLA, Los Angeles, CA 90095-1570, USA*
[2]*Institute for Cellular and Molecular Biology, University of Texas, Austin, TX 78712-0159, USA*
[3]*MRC-LMB, Hills Road, Cambridge CB2 2QH, UK*
[4]*Department of Molecular, Cellular and Developmental Biology, UCLA, Los Angeles, CA 90095-1570, USA*

The science of bioinformatics has developed in the wake of methods to determine the sequences of the informational macromolecules—DNAs, RNAs and proteins. But in a wider sense, the biological world depends in its every process on the transmission of information, and hence bioinformatics is the fundamental core of biology. We here give a consideration of some of the key problems of bioinformatics in the coming decade, and perhaps longer.

Keywords: bioinformatics; genome; proteome

1. INTRODUCTION

Biology may be viewed as the study of transmission of information: from mother cell to daughter cell, from one cell or tissue type to another, from one generation to the next, and from one species to another. This informational viewpoint is termed *bioinformatics*.

The beginnings of bioinformatics may be traced to the discovery by Sanger & co-workers that the protein insulin has a definite amino acid sequence. Sanger (1952) noted that, 'It has frequently been suggested that proteins may not be pure chemical entities but may consist of mixtures of closely related substances with no absolute unique structure. The chemical results so far obtained suggest that this is not the case and that a protein is really a single chemical substance, each molecule of one protein being identical with every other molecule of the same pure protein. Thus, it was possible to assign a unique structure to the… chains of insulin. Each position in the chain was occupied by only one amino acid and there was no evidence that any of them could be occupied by a different residue…. These results would imply an absolute specificity for the mechanisms responsible for protein synthesis… (Sanger 1952)'. So, he had concluded that proteins are perfectly ordered, and that there must be a mechanism responsible for this order. It was not a huge step from this conclusion to the idea of a genetic code.

Soon after, Sanger & his co-workers (Ryle *et al.* 1955) found that the sequences of insulin from sheep and pig differ from that of beef insulin in positions 8–10 of the A-chain, although the B-chains are identical. This set of differences in sequence found among the insulin of different species was the discovery of homologues. The display of these differences was essentially the earliest alignment of a sequence family.

In any case, with the advent of whole genome sequencing, and other tools that offer genome-wide information, bioinformatics has grown into the scientific field of management and analysis of biological information.

In this paper, we make no attempt to catalogue the past achievements of bioinformatics or to classify its subfields. This is done effectively in the other papers in this issue. Rather we present several challenges for exploration in the field of bioinformatics in the future. A few of these questions are specific. Others are general. Our goal is to provoke, not to inform. The questions with a few comments follow.

2. GENOME SEQUENCES AND THEIR EXPRESSION

RNA. New forms of RNA with new structures and new functions continue to be discovered. One example of newly discovered RNAs is riboswitch RNAs, containing aptamer domains, typically 70–170 nucleotides in length. They tend to be located within the 5′-untranslated regions of the main coding region of a particular mRNA, and are capable of sensing metabolites, and then regulating gene expression (Winkler & Breaker 2003; Tucker & Breaker 2005).

(i) In each genome, what are the informational and structural RNAs?

Transcription and splicing. With the astonishing discovery that humans have fewer protein-coding genes than do many plants, scientists are speculating that complex behaviour may be encoded partly in multiple protein forms that result from alternative splicing of RNAs. This introduces a new element of difficulty in

* Author for correspondence (david@mbi.ucla.edu).

One contribution of 15 to a Discussion Meeting Issue 'Bioinformatics: from molecules to systems'.

understanding the transmission of information from the DNA code to functioning organism (Xu *et al*. 2002).

(ii) What are the signals recognized by the splicing machinery for controlling alternate splicing of mRNAs? Can we computationally establish: the set of mature mRNAs; the conditions for their expression; and the effect on the cell of their differential expression?

3. THE HUMAN GENOME

(iii) By comparison of primate and other genomes, which genes define humanness? Are these in just a few loci, such as those encoding for proteins dealing with speech and brain development, or are they encoded in many loci? (see Enard & Paabo 2004).

(iv) What is the evolutionary history of Homo sapiens? How and at what stages of pre-history have the major variations in human genes arisen?

4. EXTENSIONS OF THE GENETIC CODE: COVALENT MODIFICATIONS, THE ENVIRONMENT AND EPIGENETICS

The genetic code specifies not only the sequences of proteins, but also a vast range of variation of these proteins through covalent modifications. These modifications include both cutting by proteolysis and splicing through several mechanisms. They also include dozens of sidechain reactions catalyzed by enzymes themselves specified by the genetic code. It has been estimated that at least 10% of the proteins encoded by the human genome code for enzymes that modify proteins (Marcotte 2001). Especially interesting are the many enzymes that catalyze the condensation of sugars with proteins, modifying the surface displayed to the environment, essentially offering proteins a coat of camouflage.

(v) To what extent is there genetic determination of phenotypes of organisms, and to what extent do environmental factors operate? Can the environmental influences be described quantitatively by a 'metabolomics approach' to bioinformatics, in which the molecular composition of the cell or organism is specified?

(vi) More specifically, can we infer which kinases phosphorylate which substrate proteins, which ubiquitin ligases are responsible for ubiquitinating which targets, and so forth, and in this way define the post-translational regulatory relationships in a cell?

(vii) Can we infer the histone code, the set of cellular signals that loosen histones about chromatin to permit its transcription? (See Jenuwein & Allis 2001.)

(viii) As a specific example of interaction of environment with genetics in humans, what accounts for the occurrence of schizophrenia in monozygotic twins being only about 60% (Klaning 1999)?

5. HOW IS THE ONE-DIMENSIONAL GENETIC CODE REALIZED IN THE THREE-DIMENSIONAL WORLD?

Upon transcription into RNA and translation into protein, the linear sequence of DNA is transformed into three-dimensional structures and machines. Can we compute the structures from the sequence? For the past 40 years, many biophysicists have believed this is possible. Their position is based on the work of Anfinsen & co-workers who demonstrated that the enzyme RNase A folds spontaneously in aqueous solution and acquires enzymatic activity. That is, sequence determines structure. Anfinsen (1973) stated his *thermodynamic hypothesis* as: 'the three-dimensional structure of a native protein in its normal physiological milieu… is the one in which the Gibbs free energy of the whole system is the lowest.'

(ix) What are the significant non-covalent forces to be considered in computing the lowest free energy structure and how are they best described? Why is accurate design of a protein sequence to form a given structure more feasible than accurate prediction of the structure that is acquired by a given protein sequence? (See Kuhlman *et al*. 2003.)

(x) How can genomic information be combined with energetic considerations to compute structures ('Darwin to the rescue of Schroedinger').

(xi) Does the thermodynamic hypothesis apply to amyloid-like fibrils and other misfolded proteins? (See Sambashivan *et al*. 2005). What is the role of kinetic barriers in protein folding?

6. CELLULAR FUNCTION AND SIMULATION

Proceeding to the next step in the transmission of genetic information into the three-dimensional world, we must consider the interactions of proteins with themselves and with other proteins and with other molecules.

(xii) To what extent can we infer complexes of proteins and protein networks from sequences? What other types of genomic and proteomic information can contribute to understanding of protein networks? How do these networks change with cell cycle, and how do they differ in different cells and organelles?

(xiii) Is it possible to simulate the metabolism of a cell in terms of its molecular constituents, assuming that all are known? Do new insights emerge from the simulation about the nature of life? How does the behaviour of a single cell differ from the average behaviour of a cell population?

7. TRANSMISSION OF INFORMATION BETWEEN DIFFERENT CELLS

Still another step in the transmission of information is between different cell types, either in the same organism or between organisms. Many types of questions could be formulated; the following is one example:

(xiv) Is the susceptibility to infection of one cell (say a human macrophage cell) by another cell type (say a *Mycobacterium tuberculosis* cell)

determined by a single genetic locus of the host cell, or by many loci?

8. GENOMIC MEDICINE

A bioinformatics frontier of vast importance is genomic medicine: the application of bioinformatics to diagnosis, prognosis and therapy. A very general question is:

(xv) What types of genomic and proteomic molecular profiles from a patient are necessary to extract useful classifiers? The classifiers sought are those that distinguish patients who will respond well to a drug to those who will not, and which drugs are likely to have adverse side effects in given patients.

9. SPECIES AND EVOLUTION

Among the most intensely studied questions to which bioinformatics can contribute are those that relate to the evolution of life on earth. Much work is presently directed towards constructing an accurate tree of life, and towards understanding the molecular events that underlie evolution. Some specific questions in this area are:

(xvi) Can a species be defined at the genomic (molecular) level? Is the barrier to interbreeding of species defined by a small number of genetic loci, or by many distributed changes throughout the genome? Does speciation come about through changes at a small number of loci, or many loci? How do species sub-types coevolve under differing selection pressures?

(xvii) What was the origin of the great profusion of protein sequences? Why are there so many apparent singlet ('ORFan') sequence families?

10. THE CHALLENGE TO BIOINFORMATICS IN THE TWENTY-FIRST CENTURY

The questions above are a small sample of those that face our nascent field as we enter a new century. We expect that genome-wide information will continue to expand in types, and to grow in abundance. As DNA sequences for individuals become available, along with RNA and protein profiles, medical informatics will blossom as a field, and undoubtedly divide into as many subfields as there are medical specialties. Along with medicine, we expect that other sciences will become increasingly involved, such as ecology and public health, among others.

REFERENCES

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

Enard, W. & Paabo, S. 2004 Comparative primate genomics. *Annu. Rev. Genomics Hum. Genet.* **5**, 351–378.

Jenuwein, T. & Allis, C. D. 2001 Translating the histone code. *Science* **293**, 1074–1080. (doi:10.1126/science.1063127)

Klaning, U. 1999 Greater occurrence of schizophrenia in dizygotic but not monozygotic twins. Register-based study. *Br. J. Psychiatry* **175**, 407–409.

Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. 2003 Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368. (doi:10.1126/science.1089427)

Marcotte, E. M. 2001 Measuring the dynamics of the proteome. *Genome Res.* **11**, 191–193. (doi:10.1101/gr.178301)

Ryle, A. P., Sanger, F., Smith, L. F. & Kitai, R. 1955 The disulphide bonds of insulin. *Biochem. J.* **60**, 541–556.

Sambashivan, S., Liu, Y., Sawaya, M. R., Gingery, M. & Eisenberg, D. 2005 Amyloid-like fibrils of ribonuclease A with three-dimensional domain-swapped and native-like structure. *Nature* **437**, 266–269. (doi:10.1038/nature03916)

Sanger, F. 1952 The arrangement of amino acids in proteins. *Adv. Protein Chem.* **VII**, 1–67.

Tucker, B. J. & Breaker, R. R. 2005 Riboswitches as versatile gene control elements. *Curr. Opin. Struct. Biol.* **15**, 342–348. (doi:10.1016/j.sbi.2005.05.003)

Winkler, W. C. & Breaker, R. R. 2003 Genetic control by metabolite-binding riboswitches. *ChemBioChem* **4**, 1024–1032. (doi:10.1002/cbic.200300685)

Xu, Q., Modrek, B. & Lee, C. 2002 Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res.* **30**, 3754–3766. (doi:10.1093/nar/gkf492)