BIOINFORMATICS: PAST, PRESENT AND FUTURE

Susan R. Wilson
Mathematical Sciences Institute, Australian National University, Australia

Keywords: Bioinformatics, biological sequence analysis, sequence alignment, hidden Markov models, evolutionary models, phylogenetic reconstruction, gene expression analysis, microarray data, design and analysis of microarray experiments, systems biology, federated data integration, bio-grids, proteomics, functional genomics, comparative genomics

Contents

UNESCO – EOLSS SAMPLE CHAPTERS

Summary

Bioinformatics is a newly emerging field that is increasingly being widely viewed as a more fundamental discipline at the intersection of the biological sciences with the mathematical, statistical, and physical sciences and chemistry and information technology. This review chapter touches briefly on those aspects of bioinformatics that will be of interest to biometricians, including biological sequence analysis and alignment (Section 2), applications of hidden Markov models in bioinformatics (Section 3), evolutionary models and phylogenetic reconstruction (Section 4), gene expression analysis and microarrays (Section 5), proteomics (Section 6) systems biology (Section 7) and federated data integration and bio-grids (Section 8), finishing with a brief discussion highlighting the exciting future of the field during the coming decades.

1. Introduction

Bioinformatics is an emerging field that was once considered to be the part of computational biology that explicitly dealt with the management of the increasing number of large databases, including methods for data retrieval and analyses, and algorithms for sequence similarity searches, structural predictions, functional predictions and comparisons and so forth. There has been phenomenal growth of life science databases. For example, the most widely used nucleotide sequence database is Genbank that is maintained by the National Center for Biotechnology Information (NCBI) of the US National Library of Medicine; as of February 2003 it contained 28.5 billion nucleotides from 22.3 million sequences. Its size continues to grow exponentially as more genomes are being sequenced. However, there is a very large gap (that will take a long time to fill) between our knowledge of the functioning of the genome and the generation (and storing) of raw genomic data.

Bioinformatics involves the analysis of biological data. So very recently, the field of bioinformatics has been rapidly evolving, not only due to the impact of the various genome projects, but also with the development of experimental technologies, such as microarrays for gene expression analyses and mass spectrometry for detection of protein-protein interactions. Currently bioinformatics is being increasingly widely viewed as a more fundamental discipline that also encompasses mathematics, statistics, physics and chemistry. Further, the field is already looking forward to a 'systems biology' approach and to simulations of whole cells with incorporation of more levels of complexity. A recent editorial in the journal *Bioinformatics* noted that 'a major upcoming challenge for the bioinformatics community [is] to adopt a more statistical way of thinking and to interact more closely with statisticians'.

The stated goal for many researchers is for developments in Bioinformatics to be focused at finding the fundamental laws that govern biological systems, as in physics. However, if such laws exist, they are a long way from being determined for biological systems. Instead the current aim is to find insightful ways to model limited components of biological systems and to create tools which biologists can use to analyze data. Examples include tools for statistical assessment of the similarity between two or more DNA sequences or protein sequences, for finding genes in genomic DNA, for quantitative analysis of functional genomics data, and for estimating differences in how genes are expressed in say different tissues, for analysis and comparison of genomes from different species, for phylogenetic analysis, and for DNA sequence analysis and assembly. Tools such as these involve statistical modeling of biological systems. Although the most reliable way to determine a biological molecule's structure or function is by direct experimentation, there is much that can be achieved *in vitro*, i.e. by obtaining the DNA sequence of the gene corresponding to an RNA or protein and analyzing it, rather than the more laborious finding of its structure or function by direct experimentation.

Much biological data arise from mechanisms that have a substantial probabilistic component, the most significant being the many random processes inherent in biological evolution, and also from randomness in the sampling process used to collect the data. Another source of variability or randomness is introduced by the

UNESCO – EOLSS SAMPLE CHAPTERS

biotechnological procedures and experiments used to generate the data. So the basic goal is to distinguish the biological 'signal' from the 'noise'. Today, as experimental techniques are being developed for studying genome wide patterns, such as expression arrays, the need to appropriately deal with the inherent variability has been multiplied astronomically. For example, we have progressed from studying one or a few genes in comparative isolation to being able to evaluate simultaneously thousands of genes (or expressed sequence tags) Not only must methodologies be developed which scale up to handle the enormous data sets generated in the post-genomic era, they need to become more sensitive to the underlying biological knowledge and understanding of the mechanisms that generate the data. For biometricians, research has reached an exciting and challenging stage at the interface of computational statistics and biology. The need for novel approaches to handle the new genome-wide data (including that generated by microarrays) has coincided with a period of dramatic change in approaches to statistical methods and thinking. This 'quantum' change has been brought about, or even has been driven by, the potential of ever more increasing computing power. What was thought to be intractable in the past is now feasible, and so new methodologies need to be developed and applied.

Unfortunately too many of the current practices in the biological sciences rely on methods developed when computational resources were very limiting and are often either (a) simple extensions of methods for working with one or a few outcome measures, and do not work well when there are thousands of outcome measures, or (b) ad-hoc methods (that are commonly referred to as 'statistical' or 'computational', or more recently 'data mining'! methods) that make many assumptions for which there is often no (biological) justification. The challenge now is to creatively combine the power of the computer with relevant biological and stochastic process knowledge to derive novel approaches and models, using minimal assumptions, and which can be applied at genomic wide scales. Such techniques comprise the foundation of bioinformatic methods in the future.

2. Biological Sequence Analysis

2.1. Background

With the advent of whole genomes becoming available for many species, as well as many other (usually extremely large) databases, such as protein sequence databases, increasingly biologists are asking questions about whether some *query* sequence of interest to her or him is significantly similar to some other sequence/s in one (or more) of these databases. If some of these similar sequences are likely to correspond to genes or proteins with known functions, then by association it is inferred that the function of the query sequence is related.

Essentially an evolutionary model is assumed where the two sequences have diverged from some common ancestor by the process of mutation and selection. Mutations can be such as to change one nucleotide to another in a DNA sequence or change one amino acid to another in a protein. Natural selection is a type of screening process that favors neutral or advantageous mutations, insertions and deletions compared with more deleterious mutations, insertions and deletions.

UNESCO – EOLSS SAMPLE CHAPTERS

Consider the following two sequences

$u_1, u_2, \ldots\ldots\ldots\ldots, u_{n_1}$

$v_1, v_2, \ldots\ldots\ldots\ldots, v_{n_2}$

where u_i and v_j represent the elements of the set {A,G,C,T} in the case of DNA sequences, and have the letters from the set of 20 amino acids in the case of protein sequences. Comparisons of two sequences usually cannot distinguish between whether a deletion has occurred in one sequence or an insertion in the other. Insertions and deletions are referred to as gaps, and in scoring an alignment these are penalized by assigning them a 'cost'. The most widely used program is BLAST (Basic Local Alignment Search Tool) and its variants, and the more heuristic algorithm FASTA is also widely used. This method starts with an initial word search - finding the best and longest continuous set of ungapped words in a database - as a starting point. The length of this - the initial *k*-tuple or word size - is the primary determinant of the speed and sensitivity of the search. In the remainder of this section we overview sequence analysis from the viewpoint of basic BLAST search, separately discussing the three steps, namely (i) finding a scoring system for comparing elements of two sequences (ii) finding optimal alignments, (iii) assessing significance.

UNESCO – EOLSS SAMPLE CHAPTERS

2.2. Scoring Systems

For comparing DNA sequences, simple nucleotide scores can be used, such as $+m$ for a match, $-m'$ otherwise (that may be equal to m), corresponding to a simple evolutionary model in which all nucleotides are equally common and all substitutions are equally likely. However, if the sequences of interest code for protein, it is usually better to compare the protein translation to amino acids, since, after only a small amount of evolutionary change, if simple nucleotide substitutions are used there is less information with which to deduce homology. Substitution matrices are used for amino acid alignments. These are matrices in which each possible residue substitution is given a score reflecting the probability that it is related to the corresponding residue in the query. The alignment score will be the sum of the scores for each position in the aligned sequence.

These scores are obtained as follows. The null hypothesis is that u_i and v_j occur independently, and hence the probability of the two sequences is $\prod p_{u_i} \prod p_{v_j}$. The alternative hypothesis is that the two sequences have diverged from the same ancestor. Denoting the probability that u_i and v_j have evolved independently from w_k as $p_{u_i v_j}$ then the probability for the whole alignment is $\prod p_{u_i v_j}$. The ratio of these two probabilities gives the likelihood ratio comparing the two hypotheses, and to obtain an additive 'scoring' system, logarithms are taken, so

$$s(u_i, v_j) = \log\left(p_{u_i v_j} / p_{u_i} p_{v_j} \right) \qquad (1)$$

For proteins we have a 20×20 matrix, and the commonly used scoring matrices are PAM and BLOSUM matrices.

Elements in the accepted point mutation PAM*n* matrices are computed from a Markov chain model of the replacement of one amino acid by another following its initial occurrence by mutation, taking into account the frequencies of the amino acids, their respective mutabilities and the probabilities for replacement of amino acids. A PAM1 substitution matrix has the property that it is derived from a Markov chain for which the 'average probability' of a change from one amino acid to another in the chain is 0.01. A PAM*n* substitution matrix is found from the n^{th} power of the Markov chain transition matrix that gave the PAM1 substitution matrix.

Now the elements in the (symmetric) BLOSUM*x* matrix are found by clustering protein sequences into 'blocks' of aligned sequences such that they have *x*% identity, and then an estimated and rounded log likelihood ratio is determined. The BLAST web page at NCBI gives the matrices for x = 45, 62 & 80. The 20 diagonal elements are all positive, while the 190 distinct off-diagonal elements (i.e. pairwise comparisons) are generally negative. Note, the larger *n* is for a PAM matrix the longer is the evolutionary distance, whereas for BLOSUM matrices, smaller values of *x* correspond to longer evolutionary distance.

In choosing to use, say a PAM*n* (or BLOSUM*x*) substitution matrix, an assumption is being made about the value of *n* (or *x*) and hence an implicit assumption about how long in the past the most recent common ancestor existed. Such an assumption is most unlikely to be correct, especially since databases contain information about many species, and the most recent ancestor of these various species with the species of the query sequence might vary markedly. Commonly chosen values are $x = 62$, $n = 120$ or 250. Problems can arise if too small a value of *n* is chosen. For example, suppose that with the correct choice n' the probability that u_i and v_j have evolved independently from w_k is $p'_{u_i v_j}$, then the mean score is proportional to

$$\sum_{i,j} p'_{u_i v_j} \log\left(p'_{u_i v_j} / p_{u_i} p_{v_j} \right) \qquad (2)$$

As $n' \to \infty$, the mean score becomes negative, and the more negative this value, the more likely it is that the null hypothesis will be accepted. For example, in the simple symmetric model (described later), if the value $n = 100$ is chosen, the mean score is negative for $n' \geq 193$. Alternatively if too small a value of *n* is chosen the test is less powerful. Note that trying to overcome this problem by choosing a variety of substitution matrices leads to one of the many multiple testing problems inherent in the use of BLAST (and discussed further below).

When comparing non-coding DNA sequences, a more complex model than the above simple evolutionary model, in which transitions are more likely than transversions, yields different 'mismatch' scores for transitions and transversions. The best scores to use will depend on whether one is comparing relatively diverged or closely related sequences.

UNESCO – EOLSS SAMPLE CHAPTERS

2.3. Sequence Alignment

Given a scoring system, next we need an algorithm to find an optimal alignment for a pair of sequences. Alignments can be global or local. To find the global alignment of protein sequences, Needleman & Wunch developed a dynamic programming algorithm, and there have been many extensions. The most widely used local alignment method is that given by Smith & Waterman (S-W) by first constructing a matrix *M* indexed by *i* and *j* corresponding to u_i and v_j in our sequences. Let S_{ij} be the score of the optimal alignment between the subsequence up to u_i, and the subsequence up to v_j. Then the S-W algorithm is given by

$$\begin{aligned} &S_{ij} = \\ &\max(0,\ S_{i-1,j-1} + s(u_i, v_j), \\ &\quad S_{i-1,j} + \text{gap penalty},\ S_{i,j-1} \\ &+\text{gap penalty}) \end{aligned} \tag{3}$$

The score *S* can never become negative, and hence always there will be areas of similarities even if there are long mismatches or gaps in between.

2.4. Assessment of Significance

2.4.1. Overview of Basic Theory

Having obtained an optimal alignment, the next step is to assess its significance, namely is this a significantly good alignment, i.e. match? In other words how does one test the null hypothesis that there is no significant homology between the two sequences against the alternative hypothesis that there is significant homology? The basic theory underlying the answer to this question was due to Sam Karlin, and is based on random walk theory, on renewal theory and on asymptotic distribution theory.

Consider Table 1 that gives a simple case of two aligned DNA sequences (of equal length). The bold position numbers are those in which the same nucleotide occurs in both sequences (i.e. 'matches' occur). Suppose we give a score +*m* for a match and –*m* where the nucleotides are different. Comparing the two sequences (starting from the left, with value zero) we can find the accumulated value of the scores, namely *m*, 2*m*, *m*, 0, –*m*, ... (given in bottom row of Table 1; alternatively a graph representation could be used). This is a simple random walk with steps $\pm m$. Ladder points are those that are lower than any previously reached point, and occur when the values 0, –*m*, –2*m*, … are *first* reached – in this example at positions $L_1 = 4$, $L_2 = 5$, $L_3 = 8$, ... (all shown in bold in the bottom row).

The term 'upwards excursion' describes that part of the walk between two consecutive ladder points, L_i and L_{i+1}, and interest is in the maximum height starting from the first ladder point L_i. (Where ladder points fall at consecutive positions, such as 8 & 9 above, then this value is zero.) The maximum height achieved by the various excursions is the

UNESCO – EOLSS SAMPLE CHAPTERS

basis for the test statistic to evaluate the homology of the two sequences. In the above example the heights of the various upward excursions are, respectively, 0, 1, 0, 0, 3, 4, with a maximum value of 4.

1	**2**	3	4	5	**6**	7	8	9	10
G	G	T	A	C	T	G	G	G	G
G	G	G	G	C	C	T	T	C	C
m	2*m*	*m*	**0**	***-m***	0	*-m*	***-2m***	***-3m***	***-4m***

11	**12**	**13**	14	15	16	**17**	**18**	19	20
A	A	C	T	T	T	T	T	C	C
A	A	C	C	G	G	T	T	A	A
-3*m*	-2*m*	*-m*	-2*m*	-3*m*	-4*m*	-3*m*	-2*m*	-3*m*	-4*m*

21	**22**	**23**	**24**	25	**26**	**27**	28	29	30
C	G	G	G	T	A	A	A	A	T
G	G	G	G	T	A	T	C	C	C
-5m	-4*m*	-3*m*	-2*m*	*-m*	-2*m*	-3*m*	-4*m*	-5*m*	***-6m***

Table 1: Two aligned DNA sequences

For comparison of two protein sequences, the scores are obtained from the appropriate (or selected) 20×20 substitution matrix, and the random walk is less 'regular' than in the above example. Suppose the substitution matrix used allocates a score $S(i, j)$ to a match of amino acids *i* and *j*. The null hypothesis is that the two sequences are random with respect to one another and under this hypothesis the mean score is $\sum_{i,j} S(i, j) p_i p_j$ where p_i is the frequency of amino acid *i*. For the BLAST procedure it is necessary that this mean score be negative, and we assume it is, so that the general trend of the random walk (starting from the left) is downward, passing through a sequence of increasingly negative ladder points. The test statistic is the height $Y_{\max}$ of the largest upwards excursion following a ladder point relative the position of that ladder point, before the walk reaches the next ladder point. It can be shown, using standard random walk theory, that if *Y* is the maximum height achieved by the walk after reaching any ladder point and before reaching the next, then

$$\Pr(Y \geq y) \sim Ce^{-\lambda y} \tag{4}$$

where λ is the unique positive solution of the moment generating equation

$$\sum_{i,j} p_i p_j e^{\lambda S(i,j)} = 1 \tag{5}$$

This is referred to as a 'geometric-like' distribution. For the simple random walk above, letting *p* (<1/2) be the probability of a positive step, we obtain $\lambda = \log\left[(1-p)/p\right]$ and $C = 1 - e^{-\lambda}$. The formula for C for a given substitution matrix is more complicated (and

UNESCO – EOLSS SAMPLE CHAPTERS

not given here). We are interested in $Y_{\max}$, the largest of these maximum heights. Unfortunately there is no limiting probability distribution for $Y_{\max}$, although bounds can be found. Next consider the number of ladder points reached by the walk before it finishes (after n steps have been taken where n is the length of the sequences being compared). The theory is complex, but for our purposes the number of ladder points can be taken as n/A where A is the mean number of steps taken from one ladder point to the next (and can be found by random walk theory or by application of Wald's identity), and also is taken here as given. Then it can be shown that the *P*-value associated with an observed value $y_{\max}$ of $Y_{\max}$ is approximately

$$1-\left(1-Ce^{-\lambda y_{\max}}\right)^{n/A} \qquad (6)$$

and in BLAST this is approximately $1-\exp(-e^{-s})$, where $s=\lambda y_{\max}-\log nK$ and $K=Ce^{-\lambda}/A$. When s is large, $P\text{-value}\sim e^{-s}$. Note that if the elements in the chosen substitution matrix are all multiplied by some constant k, then it can be shown that the *P*-value is not altered. Hence the quantity s is called a 'normalized score'. The BLAST software also gives an 'Expect' value. This is the mean number of walks to reach a height equal to or higher than the maximum height observed when the null hypothesis is true. When the *P*-value is small, this is close to the *P*-value, otherwise it is close to $-\log(1-P\text{-value})$. Note that the *P*-values are quite sensitive to the somewhat arbitrary numerical values of K and λ.

-
-
-

TO ACCESS ALL THE **24 PAGES** OF THIS CHAPTER,
Visit: http://www.eolss.net/Eolss-sampleAllChapter.aspx

Bibliography

Baldi, P. and Brunak, S. (2001). *Bioinformatics: The Machine Learning Approach*, 2nd Ed., 452 pp, USA: The MIT Press. [This gives more background to the material of Sections 3 and 4, and covers areas of bioinformatics that have applied the machine learning approach.]

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (2000). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, 356 pp, UK: Cambridge University Press. [This concentrates on applications of hidden Markov models in bioinformatics.]

Elston, R., Olson, J., and Palmer, L. (2002). (eds) *Biostatistical Genetics and Genetic Epidemiology*, 831 pp, UK: John Wiley. [This includes articles on bioinformatics, on DNA sequences, and on sequence analysis; updates will appear in the *Encyclopedia of Biostatistics* 2nd Edition, 2004.]

Ewens W.J. and Grant, G.R. (2001). *Statistical Methods in Bioinformatics*. 476 pp, USA: Springer-Verlag. [This includes more background on the material in Sections 2, 3 and 4, especially the theory underpinning BLAST.]

Maindonald, J.H., Pittelkow, Y.E. and Wilson, S.R. (2003). Some considerations for the design of microarray experiments. *Science and Statistics*, Vol. 40 (ed. D.R. Goldstein), 367-390. USA: Institute of Mathematical Statistics. [This discusses issues relevant for the design of microarray experiments with emphasis on uses of replication and the importance of identifying major sources of variation.]

Speed, T. (2003). (ed.) *Statistical Analysis of Gene Expression Microarray Data*, 222 pp, UK: Chapman & Hall. [This is a small collection of papers mainly concerned with analysis of gene expression data.]

Biographical Sketch

Sue Wilson was born is Sydney, Australia; she obtained her B.Sc. from the University of Sydney (1969), followed by her Ph.D. from the ANU (awarded 1973). Sue then spent two years as a Lecturer in the Department of Probability and Statistics at Sheffield University. She returned to ANU towards the end of 1974 and has since held a variety of positions there, both in some of the Statistical groupings, as well as at the National Centre for Epidemiology and Population Health. She is currently Professor, Statistical Science Program, Centre for Mathematics and its Applications, Mathematical Sciences, Institute and Co-Director, Centre for Bioinformation Science (joint with John Curtin School of Medical Research), at the Australian National University (ANU). Sue has over 150 publications in biometry and applied statistics, with a particular emphasis recently on statistical genetics/genomics and bioinformatics. These papers have arisen from her extensive consulting experience in the biological, social, and medical sciences, leading to statistical modeling developments to answer substantive research questions in these disciplines. Sue is an elected member of the International Statistical Institute, a Fellow of the American Statistical Association and a Fellow of the Institute of Mathematical Statistics (IMS). She was President, International Biometric Society, 1999 & 2000 (Vice President, IBS, 1998, 2001). Currently she is Associate Editor, *Annals of Human Genetics*; Associate Editor, *Computational Statistics and Data Analysis*; Member, Editorial Board, *Statistical Methods in Medical Research*; Member, Editorial Board, *Behavior Genetics*.

UNESCO – EOLSS SAMPLE CHAPTERS