Dimensionality and dependence problems in statistical genomics

Pietro Liò
EMBL-EBI European Bioinformatics Institute,
Wellcome Trust Genome Campus,
Hinxton, UK
E-mail: pietrol@ebi.ac.uk

INTRODUCTION

There is no unique criterion for describing even the genome of the simplest organism. Genome sequences have been investigated on the basis of several characteristics; for example linguistics, percentage of shared homologous genes and patterns associated to chemical properties of sequences. The situation can be well represented by the analogy from engineering drafting where blueprints of some complicated machinery are usually replaced with an exploded view of the same, with each component shown from many different angles or in holographic view. Similarly, statistical biases related to coding properties, gene duplications and repeats are usually represented in the form of several colour-coded wheels or bars. Very dark lines may be used to highlight all values that are more than \( \pm 3 \) standard deviations from the average (see for instance Pedersen et al.

Although genome studies have become central to bioinformatics, in most research there is a growing need to use statistics to complement bioinformatics. Two problems are of particular concern in statistical genomics. The first is dimensionality, i.e. the number of variables can exceed the number of responses by factor of tens. This is often the case for gene expression data. The analysis of gene expression and biochemical data, such as concentrations and kinetic parameters of DNA, RNA, protein and other metabolites, requires the development of statistical methods to reduce the dimensionality.

Connected with the problem of dimensionality is the problem of dependence. Although functional analysis techniques, such as principal and independent component analysis, usually represent valid approaches, phylogenetic methods provide optimality in solving the dependence problem in sequence analysis. The assessment of the phylogenetic relationships comes from solving the statistical dependencies among sequence data that are due to the common ancestry.

The present paper is not intended to serve as a complete literature survey of methods in statistical genomics, rather to introduce the reader to some of the most recent aspects of this field, and to warn of major pitfalls associated with genome analysis. The discussion covers few broad areas: using phylogenetic methods, using methods for variable selection and for

Abstract

Genome studies have become central to a wide range of biological areas. This paper is intended to discuss the state of art and potentiality of using statistics in genome comparison. The problem of dependence and dimensionality in genome data is addressed first. The focus is on using phylogenetic methods in genome comparison, on combining sequence and gene expression data to find transcription sites, and on multiple hypothesis testing in gene expression data. The paper concludes with speculations on the future course of statistical and bioinformatics analyses.
combining sequence and gene expression data to find transcription sites as examples of reducing dimensionality. The paper briefly describes multiple hypothesis testing procedure in high dimensional data and concludes by discussing current perspectives and future directions in genome comparative analysis.

**USING EVOLUTIONARY INFORMATION IN GENOME STUDIES**

DNA and amino acid sequences contain information regarding both the phylogenetic relationships among species and the evolutionary processes that have caused the sequences to diverge. Phylogenetics at genome level is the study of phylogenies and processes of evolution by the analysis of genome sequences. The analysis of the statistical dependency among sequences allows each species to contribute its own quantity to all estimates. Ignoring or simplifying the phylogenetic relations among sequences confounds the information available from sequence analysis with the patterns induced by the evolutionary process. In addition, the evolutionary distance is an important consideration when distinguishing between sequence similarities resulting from strong selective pressures and those resulting from inadequate divergence time. This is especially critical as more closely related DNA regions are more likely to share functional and structural characteristics. As a result, it is difficult to unravel which patterns of conservation come from the shared characteristics and which arise from the limited number of mutations that have occurred. Maybe most importantly, it is difficult or impossible to evaluate the statistical significance of the observed patterns of conservation and variation without an underlying model that explicitly includes the evolutionary relationships between the sequences and the manner in which they change.

Among several statistical approaches for phylogenetic inference, that of maximum likelihood (ML) is the most mathematically robust and provides a framework to incorporate simple and complex models of evolution. A great attraction of the likelihood approach in phylogenetics is the existence of several robust statistical hypothesis tests. Therefore, when more than two genome sequences are available, it is better to use phylogenetic methods than simple sequence pairwise comparison and it is better using maximum likelihood than other methods. Phylogenetic methods also make appropriate corrections for multiple hits. In pairwise comparisons of sequences, the multiple hit corrections differ for each pair because the evolutionary separation of two human sequences is not the same as the evolutionary separation of human and mouse or mouse and rat. Moreover contiguous regions might have different evolutionary separations.

It is often the case that several distinct genomic segments are available for all or some of the species. Cao and colleagues showed that phylogeny based on different vertebrate genes could suggest wrong trees. This happens for both short and long-range phylogenetic distances and under a variety of models. It is usually better to concatenate the sequences and analyse as a single data set. Currently available programs such PAUP and PAML make it possible to analyse the concatenated sequences allowing for heterogeneity of evolutionary patterns. This heterogeneity may be as simple as modelling all genome regions and genes as having the same patterns of replacement but different rates throughout a common (inferred) tree. Otherwise, it allows each genomic region to evolve with different replacement patterns, and with different rates of replacement in all lineages of the inferred tree, which now has a common topology for each genome segment/gene, but quite unrelated branch lengths from one region to the next. For example, it is possible to estimate the transition/transversion ratio of each lineage of each segment of the phylogenetic tree.
PHYLOGENETIC ANALYSIS USING DUPLICATED GENES

Eukaryotic genomes contain a large number of duplicated fragments. There are different extents of duplication: a small part of a gene, a module, an exon, a coding region, a full gene, a cluster, part of a chromosome, a full chromosome. Whole genome duplications are widely believed to have played an important role in the evolution of the yeast, plant and vertebrate genomes. Large-scale gene duplication is a major force driving the evolution of genetic functional innovation. Genes exist predominantly as families with related structures and functions. A gene family is subject to concerted evolution, which homogenises the member genes of the family. However, protein homogeneity can also be attained by strong purifying selection. Although gene duplications are very frequent, the rate-limiting step is getting a new function. The use of evolutionary trees to analyse the history of gene duplication and estimate duplication times provides a powerful tool for studying genome phylogeny. Repeats can also be used. There is a good chance that gene conversion has taken place between the duplicates if they are near each other. Gene conversion will make both copies look more like each other than like the ancestral sequence. Insertions may break up the gene conversion event, but in a long unit, short repeats may not form an inhibitive block. There are several methods for testing gene conversion among duplicated genes and repeats. Most of these methods derived from the test proposed by Sawyer. This test allows one to detect identical segments of DNA between clustered genes that may indicate gene conversion events. Sawyer’s gene conversion test, when applied to pairs of sequences, measures the sum of the squared total lengths of these fragments (SSL). If gene conversion has been operating, fragment lengths should be more varied and SSL would be larger than if variant sites were distributed at random along the sequence. Therefore the same number of variant sites is then randomly permuted along the sequence many times and the SSL for each replication calculated, and compared with the observed value.

REPEATS AND PHYLOGENETIC ANALYSIS

Repeats account for a substantial percentage of the vertebrate sequence but tend to have GC content and dinucleotide composition somewhat different from that of the DNA into which they are inserted. Including them in the phylogenetic analysis makes it difficult to interpret the results; see for instance Webster et al. Interspersed repeats are usually masked before doing genome alignment, or at least analysed separately. It is known that a number of repeats, mainly SINES, do have CpG island-like sequences which tend to be found by the standard algorithms, but they are probably methylated after insertion into the genome and so the CpGs tend to disappear in time; ie they are not true unmethylated islands. Use of interspersed repeats can provide upper and lower boundaries for the age of individual duplications. When an interspersed repeat is present in one but absent in the other element, the duplication must be older than the repeat. If both units share the element or if one unit interrupts an interspersed repeat, it has to be younger than that interspersed repeat. Sometimes there are some risks in relying exclusively on repeats to assess the dates of gene duplication. A gene conversion event between a repeat in the duplicated segment and one outside the segment could make the repeat age estimate incorrect and the duplication may be far older than it appears from the divergence level. In particular, tandem repeated units cannot be reliably dated in such a way.

For example, similarities would often drop dramatically after insertion of transposable elements, the insert forming the new boundary of gene conversion.
process. A gene conversion event could bring a repeat in one of the duplicate copies over to the other one, so the presence of the repeat in both copies does not prove it is older than the original duplication event. It is possible to recognise the event has occurred because the two repeats should then be more similar to each other than either is to the ancestral sequence. A deletion event could take out a repeat from just one of the two duplicates, so its presence in just one copy does not prove it is younger than the duplication event; such a deletion event is not likely to be exact so it is possible to recognise it has occurred. A gene conversion event between a repeat in the duplicated segment and one outside the segment could make your repeat age estimate incorrect.

FUNCTION AND SELECTION USING PHYLOGENETIC ANALYSIS

An important reasoning for using phylogenetic methods instead of pairwise sequence analysis is the investigation of function conservation in families of paralogous genes or in conserved non-coding regions. Note that there are an increasing number of gene functions inferred from pairwise sequence similarities to genes whose functions have been inferred from pairwise sequence similarities. Paralogous genes need not necessarily have the same function and the assignment of function simply by pairwise sequence similarity is misleading. Recent work takes into account the phylogeny and the differing substitution rates between amino acids to detect functional divergence between paralogous genes.18 Gu has developed a quantitative measure for testing the function divergence within a family of duplicated genes. The method is based on measuring the decrease in mutation rate correlation between gene clusters of a gene family; the hidden Markov model (HMM) procedures allow the amino acid residues responsible for the functional divergence to be identified.18 Phylogenies can be used to infer the effects of natural selection on molecular functions. Directional selection is known to increase the ratio of non-synonymous ($n$) to synonymous ($s$) nucleotide substitutions; this ratio ($\omega = dn/ds$) can be used to detect lineages in a molecular phylogeny along which selection has occurred.19,20 Positively selected amino acid sites are usually identified by an elevated ratio of non-synonymous (amino acid changing)–to–synonymous substitution; a ratio much greater than 1 implies strong evidence of selection at that site. This is in contrast to sites that are highly conserved owing to functional constraints and thus have few non-synonymous substitutions, and to sites evolving neutrally, which would tend towards more equal ratios of non-synonymous to synonymous substitutions. If the estimates of $\omega$ for the branches of interest are not greater than 1, one might not be able to conclude positive selection based on the likelihood ratio test alone. Relaxed selective constraint along the lineages of interest is an alternative compatible explanation. Furthermore, purifying selection is more effective in large populations than in small populations, and so differences in population sizes along lineages provide another compatible hypothesis. If amino acid changes are slightly deleterious, we expect them to be removed from the population at a higher rate in a large population than in a small one. As a result, we expect to see a smaller $\omega$ in a large population than in a small one, even if there is no difference between the two lineages in selective pressure or gene function.

WHAT ELSE IN THE GENOME SHOULD BE CONSIDERED IN PHYLOGENOMICS?

Current phylogenetic approaches to genome comparison still regard genomes as bags of genes and use simple models of evolution. The ultimate goal of phylogenomics is to consider different models of evolution for genome regions.
with different base composition and evolution dynamics. In particular the GC content is an important characteristic of DNA sequences. The different parts of the genome (protein genes, RNA genes and spacers, promoter and regulatory regions, micro- and minisatellites) reveal different positive linear correlations between GC content and the G + C percentage of their genomic DNA. Vertebrate genomes are made up of long DNA segments, the isochores, which are homogeneous in GC content and differ in gene content. For example, the human genome can be subdivided into five families, L1, L2, H1, H2 and H3, which are characterised by increasing GC levels and gene concentrations; the two GC-poor families (L1 and L2) represent 30 per cent and 33 per cent of the genome, respectively, and three GC-rich families (H1, H2 and H3) represent 24, 7.5 and 4–5 per cent of the genome, respectively. Gene concentration reaches an up to 20-fold higher level in H3 than in L1 isochores. The isochrome organisation of the avian genome contains an additional, GC-richest, H4 isochrome family. It is known that isochores are mainly in genomes of warm-blooded vertebrates while only very few cold-blooded vertebrates have isochrome-like genomes. Since a 10 °C increase in temperature in vitro corresponds to an increase the rate of cytosine deamination 5.7-fold, Fryxell and Zuckerkandl proposed that the deamination of methyl-cytosine and cytosine might be the cause of GC isochores. Other hypotheses for the arising of isochores are reviewed in Eyre-Walker and Hurst.

Base composition and evolution dynamics of isochores can be estimated by segmenting the genome sequences (coded as C, G = 1; A, T = −1) in different sequence windows and computing the $\chi^2$ statistics of the GC-content of each window with respect to the average value of the whole genome. Nekrutenko and Li have developed a compositional heterogeneity index, $H_{GC}$, to quantify and compare the compositional differences within and between genome sequences. These authors considered a sequence that can be divided into $n$ windows of length $L$ and calculated the genomic average GC percentage, $GC_{av}$, and the GC content of each window (eg $GC_1$, $GC_2$, ..., $GC_n$). Then they computed the average GC content difference between two adjacent windows:

$$H_{GC} = \frac{1}{n} \sum_{i=2}^{n} |GC_i - GC_{i-1}|$$

$$\sqrt{\frac{GC_{av}(1 - GC_{av})}{L}}$$

where the term in the denominator represents the standard error (SE); the SE attains the maximum value at $GC_{av} = 0.5$. Note that, if the window is small, its GC content is subject to strong fluctuations, whereas if the window is large, it may conceal heterogeneity. Usual choices of windows are in the range 10–50 kb. Liò and Vannucci used wavelets to smooth G + C profiles and locate specialised islands in genome sequences. See Liò for an introduction to wavelets for biologists. First they performed a $\chi^2$ statistic on the wavelet coefficients of a G + C profile. Secondly, they generated a wavelet scalogram: the presence of several peaks suggested the presence of specialised islands. An inverse wavelet transform using the scales corresponding to the peaks allows the location of the genomic island to be found. Figure 1 shows the wavelet-denoised G + C plot of the MHC region in human chromosome 6, which contains the boundary of an isochrome. Recently, Nicolas and collaborators proposed a new statistical genome segmentation method based on an HMM and estimating parameters using maximum likelihood. Their model has several hidden states and it enables different compositional classes to be separated, for example non-coding RNAs, coding regions for each DNA strand (+, −), highly expressed coding regions, genes coding for hydrophobic proteins, and intergenic regions. Their approach is not based on sliding windows.

---

**G + C content**

Vertebrate genomes are made up of long DNA segments, the isochores, which are homogeneous in GC content and differ in gene content. For example, the human genome can be subdivided into five families, L1, L2, H1, H2 and H3, which are characterised by increasing GC levels and gene concentrations; the two GC-poor families (L1 and L2) represent 30 per cent and 33 per cent of the genome, respectively, and three GC-rich families (H1, H2 and H3) represent 24, 7.5 and 4–5 per cent of the genome, respectively. Gene concentration reaches an up to 20-fold higher level in H3 than in L1 isochores. The isochrome organisation of the avian genome contains an additional, GC-richest, H4 isochrome family. It is known that isochores are mainly in genomes of warm-blooded vertebrates while only very few cold-blooded vertebrates have isochrome-like genomes. Since a 10 °C increase in temperature in vitro corresponds to an increase the rate of cytosine deamination 5.7-fold, Fryxell and Zuckerkandl proposed that the deamination of methyl-cytosine and cytosine might be the cause of GC isochores. Other hypotheses for the arising of isochores are reviewed in Eyre-Walker and Hurst.

**Isochore**

**Wavelet analysis**

Base composition and evolution dynamics of isochores can be estimated by segmenting the genome sequences (coded as C, G = 1; A, T = −1) in different sequence windows and computing the $\chi^2$ statistics of the GC-content of each window with respect to the average value of the whole genome. Nekrutenko and Li have developed a compositional heterogeneity index, $H_{GC}$, to quantify and compare the compositional differences within and between genome sequences. These authors considered a sequence that can be divided into $n$ windows of length $L$ and calculated the genomic average GC percentage, $GC_{av}$, and the GC content of each window (eg $GC_1$, $GC_2$, ..., $GC_n$). Then they computed the average GC content difference between two adjacent windows:

$$H_{GC} = \frac{1}{n} \sum_{i=2}^{n} |GC_i - GC_{i-1}|$$

$$\sqrt{\frac{GC_{av}(1 - GC_{av})}{L}}$$

where the term in the denominator represents the standard error (SE); the SE attains the maximum value at $GC_{av} = 0.5$. Note that, if the window is small, its GC content is subject to strong fluctuations, whereas if the window is large, it may conceal heterogeneity. Usual choices of windows are in the range 10–50 kb. Liò and Vannucci used wavelets to smooth G + C profiles and locate specialised islands in genome sequences. See Liò for an introduction to wavelets for biologists. First they performed a $\chi^2$ statistic on the wavelet coefficients of a G + C profile. Secondly, they generated a wavelet scalogram: the presence of several peaks suggested the presence of specialised islands. An inverse wavelet transform using the scales corresponding to the peaks allows the location of the genomic island to be found. Figure 1 shows the wavelet-denoised G + C plot of the MHC region in human chromosome 6, which contains the boundary of an isochrome. Recently, Nicolas and collaborators proposed a new statistical genome segmentation method based on an HMM and estimating parameters using maximum likelihood. Their model has several hidden states and it enables different compositional classes to be separated, for example non-coding RNAs, coding regions for each DNA strand (+, −), highly expressed coding regions, genes coding for hydrophobic proteins, and intergenic regions. Their approach is not based on sliding windows.
and it enables different genome compositional classes to be separated without prior knowledge of their content, size and localisation. The results from segmentation algorithms can then be used in a phylogenetic analysis and appropriate models of evolution can be used for different genome regions.

**Microarray analysis**

**DIMENSIONALITY AND GENE EXPRESSION DATA**

Data from gene expression experiments represent challenging problems for statisticians because they usually consist of many correlated variables (gene expressions) and relatively few samples. Microarray statistics can be supervised or unsupervised. In the unsupervised case, only the expression data are available and the goal is mainly to identify distinct sets of genes with similar expressions, suggesting that they may be biologically related. In supervised problems a response measurement is also available for each sample and the goal of the experiment is to find sets of genes that, for example, relate to different kind of diseases, so that future tissue samples can be correctly classified. Statistical traditional methods for both clustering and classification have been extensively applied to gene expression data. See for instance Eisen et al.\(^{28}\) and Pan.\(^{29}\)

Recent work reports on a method based on using a probit model for classification. This method uses latent variables to specialise the model to a regression setting and a Bayesian mixture priors to perform a variable selection and identify the important genes.\(^{30}\) see Shoemaker et al.\(^{31}\) for a biologist-oriented introduction to Bayesian statistics. The model assigns a posterior probability to all possible subsets of genes and then proceeds by searching for those subsets that have high posterior probability. This method is able to locate small sets of genes that lead to good classification results.

Linked to both dimensionality and statistical dependence is the problem of multiple hypotheses testing. The procedures for multiple comparison aim at controlling the probability of committing even a single type-I error within the tested family of hypotheses. The main problem with classical procedures, such the Bonferroni correction, which hinder many applications in biology, is that they tend to have substantially less power than uncorrected procedures. The false discovery rate (FDR) is a new approach to multiple hypothesis testing.\(^{32}\) The FDR is the expected proportion of true null hypotheses rejected out of the total number of null hypotheses rejected. Multiple comparison procedures controlling the FDR are very powerful and are especially suited to large multiple comparison problems in which existing procedures lack power. Recently Efron, Tibshirani and collaborators\(^{33}\) have implemented FDR in microarray analysis and devised software, SAM.\(^{34}\)

**COMBINING SEQUENCE AND GENE EXPRESSION DATA TO FIND TRANSCRIPTION SITES**

The search for transcription sites often starts from the assessment of the statistical
significance of over- or under-representation of oligomers in complete genomes may indicate phenomena of positive or negative selection. The analysis of the bias of oligomers requires a model of how sequences are generated and the statistical distribution of the oligomer.

An oligomer appears with an unexpected frequency in a sequence if the number of its occurrences is significantly different from an estimator of the expected count under the considered model. The standard methodology uses Markov chains to obtain a theoretical expectation for the count of an oligomer. The order \( k \) of the Markov chain model means that the probability that a base occurs at a given position in the sequence depends on the \( k \) previous bases. The choice of the order of the Markov model depends on sequence length, because of the data requirements in estimation.

There are two kinds of statistics to compare the theoretical expectation with the real observed count. For non-rare oligomers, a Gaussian approximation is appropriate and the statistic is simply the \( z \)-score. For rare oligomers a compound Poisson approximation is generally used. More specialised methods, which are used in gene prediction, implement HMMs. Although this methodology seems to give interesting results with bacterial and low-eukaryotic genomes, it does not help so much with the great complexity of the transcription mechanisms in high-eukaryote genomes. Gene expression data, combined with sequence analysis, seem to represent a promising alternative method to identify transcription sites. Bussemaker et al. consider both yeast sequences and expression data and outputs statistically significant motifs: the regions upstream genes contribute additively to the log-expression level of a gene.

Recently Conlon et al. have proposed a similar method. They have first identified a number of sequence patterns upstream the yeast genes using MDscan and then tested for candidate transcription sites. Their algorithm is based on fitting a linear regression with gene expression as response and a score based on transcription site candidates as covariates: 

\[
Y_g = \alpha + \beta_m S_{mg} + \varepsilon_g,
\]

where \( Y_g \) is the log2 expression value of the gene \( g \), \( S_{mg} \) measures how well the upstream sequence of a gene \( g \) matches a motif \( m \), in terms of both degree matching and number of sites, and \( \varepsilon_g \) is the gene-specific error term. The baseline expression \( \alpha \) and the regression coefficient \( \beta_m \) are estimated from the data. Then they performed stepwise regression starting with a model with intercept-only and adding/removing motifs at each step. The candidate motif with a significant \( p \)-value (say for example 0.01) for the simple linear regression coefficient \( \beta_m \) are retained and used by the stepwise regression procedure to fit a multiple regression model:

\[
Y_g = \alpha + \sum_{m=1}^{M} \beta_m S_{mg} + \varepsilon_g.
\]

Again, this stepwise regression begins with only the intercept term and adds at each step the motif that gives the largest reduction in residual error. After adding each new transcription sites, the model is checked to remove the ones whose effects have been sufficiently explained by the last added.

PERSPECTIVES: STATISTICAL GENOMICS TO MEET STATISTICAL MODELLING OF CHROMATIN

Once we have algorithms to analyse combined data of sequences, gene expression data and proteomic data, what comes next? Perhaps all the information on sequence, gene expression and protein expression will need to be combined with the growing amount of data on chromatin organisation. An important attempt to link sequence and chromatin organisation was made by Audit et al. using the wavelet transform to investigate the fractal scaling
properties of coding and non-coding human DNA sequences. They provided evidence that long-range correlations in the small-scale regime (10–200 base pairs), observed in eukaryotic sequences, depend on nucleosome patterns.

Experiments have shown that gene density influences chromosomes compartmentalisation in the nucleus and gene-rich regions appear looping outside of chromosome territories. For example, the territories of the human chromosome 19, which contain mainly gene-dense, GC-rich isochore and early replicating chromatin, are located toward the nuclear centre, whereas the territories of chromosome 18, which consist mainly of gene-poor, low GC content, and later replicating chromatin, is located close to the nuclear border. Both chromosomes have a similar DNA content, 85 and 67 Mb, respectively. All currently available chromatin models do not include sequence patterns. Basically, two different models of higher-order chromatin structure up to the structure of the entire chromosome territories have been proposed. Yokota et al. and Sachs et al. constructed a random-walk/giant-loop quantitative model for the overall geometric structure of a human chromosome in interphase nuclei by measuring the distance between specific genomic sequences labelled by fluorescence in situ hybridisation (FISH). They found that the relationship between the mean-square interphase distance and genomic separation has two linear phases. On one level, chromatin appears to be arranged in large flexible loops averaging approximately 3 Mb in size. Within each loop, chromatin is randomly folded. On the second level, specific loop-attachment sites are arranged to form a supple, backbone-like structure, which also shows random walk behaviour. Instead the multi-loop subcompartment model assumes that 1 Mb chromatin domains are built up like a rosette from a series of chromatin-loop domains with sizes of 100 kb, again assuming a random organisation for each loop. These models make different predictions about the interphase distances between genes and other DNA segments that are located along a given chromosome. Since DNA persistence length and geometry depend on sequence, it is very likely that bioinformatics will be soon challenged in combining data sequence with chromatin location and functioning.

Acknowledgments
I thank N. Goldman and M. Vannucci for critical reading of the manuscript. The author is supported by a BBSRC grant.

References


34. URL: www-stat.stanford.edu/~tibs/SAM/


38. URL: http://bioprospector.stanford.edu/MDscan/


