Bioinformatics

Michael T. Clegg
Peter L. Morrell
University of California, Riverside, California, U.S.A.

INTRODUCTION: MINING MOLECULAR SEQUENCE DATA

DNA sequences diverge over time because of the gradual accumulation of mutational differences. Mutations erode DNA coding information, and if sufficient time has elapsed since the separation of lineages, it may be difficult to discern any historical similarity between two DNA sequences that once shared a common ancestor. The bioinformatician is presented with DNA sequences, protein sequences, or derivative data such as DNA hybridization signals associated with microarrays. A major challenge is to detect regions of similarity between different DNA sequences (or surrogate measures) to test the inferences of a common history. A common origin implies a shared function, or a related function, and this provides important clues in the assignment of a preliminary function to raw sequence data from a new source. A second major challenge is the management and curation of the vast stores of new data originating from the genomics enterprise. A variety of computational tools are available to organize and retrieve useful results from these vast databases. Genomics data are maintained in a number of publicly available databases. Perhaps the most important of these is the NCBI database, which can be accessed via the World Wide Web (http:www.ncbi.nlm.nih.gov) and provides a number of data analysis tools.

Bioinformatics combines computer science, genetics, and statistics to meet the challenge of mining useful information from the vast stores of new molecular sequence data. In this article we will touch on a few of the issues that confront bioinformations with a specific focus on DNA sequence data.

Several computer scripting or programming languages are used to handle DNA and protein sequence data. Perl and Python scripting are both commonly used; example code from Bioperl is available at http://bioperl.org/ and for Biopython at http://biopython.org/. Many tools that simulate DNA sequence evolution are written in C or C++. Dr. Richard Hudson and his colleagues have made source code to simulation-related tools available at http://home.uchicago.edu/~rhudson1/source.html and http://molpopgen.org/.

ANALYTICAL METHODS: THE CHALLENGE OF SEQUENCE ALIGNMENT

The proper alignment of molecular sequence data is fundamental to many aspects of bioinformatics. The raw data are usually strings of DNA, RNA, or protein sequences for a particular gene or protein drawn from a set of organisms. These data may also take the form of a position of a fragment on a gel (e.g., RFLPs or microsatellites) where identity in the location in the gel is assumed to imply identity in the underlying DNA or protein state. In what follows, we will restrict our discussion to DNA sequence data for simplicity. We use the term ''string of nucleotides'' to refer to the DNA sequence for a gene obtained from a particular organism (or a particular copy of a multigene family from within a genome). Thus, S_1 refers to the string from source 1, and S_n refers to the string from source n. The first analytical task is to align the strings of nucleotides (S_1, S_2, . . ., S_n) to minimize the number of nucleotide differences across the set {S}. This involves finding the minimum of a weighted function of the number of indel (insertion/deletion) events and nucleotide site differences over the set. The weights are assigned based on some prior assumptions about the likelihood of indels versus nucleotide site differences (Fig. 1). The final alignment is assumed to be the ''best'' representation of the number of mutational events that occurred over the evolutionary time spanned by the set {S}. We denote the aligned set as {A}. More-sophisticated alignment algorithms use a tree-fitting iteration to generate a best alignment by simultaneously estimating the phylogeny and alignment (e.g., Clustal W). Various alignment algorithms are readily available from a wide variety of DNA sequence analysis packages.

When the sequences in {S} are so diverged that it is difficult to obtain an alignment from DNA sequences, a translation into the derived amino acid sequence is useful. Protein change occurs at a slower rate than DNA sequence change, and alignment of a protein sequence may be relatively straightforward even when the underlying DNA sequences are highly diverged.

It should be clear from this limited discussion that the alignment step is a crucial operation, and all subsequent

DOI: 10.1081/E-EPCS 120006085

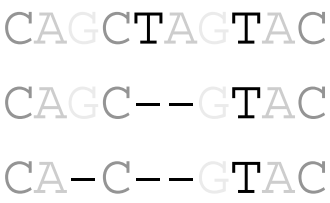

Fig. 1 A nucleotide sequence alignment with two indels. Alternative alignments could include a single indel but would require a nucleotide substitution.

calculations (e.g., distance measures, see below) depend on the accuracy of this operation. An operation that assumes a very high penalty (weight) for indels will force a larger number of nucleotide site differences on the alignment; conversely, assuming a very low penalty for indels will force a large number of small indels at the expense of nucleotide site differences (Fig. 1). These distortions will affect all subsequent inferences based on the sequence data. Indels are much more common in noncoding regions. When they occur in coding sequence, they often involve nucleotide triplets and preserve the gene's reading frame (Fig. 2). As indels become superimposed on one another during the course of evolution (this is especially common in noncoding sequence), it is increasingly difficult to determine the boundaries of individual indel events. As a consequence, it is usually not possible to count the number of indels that separate two sequences, and because the process of insertion/deletion is not stochastically regular, it is also not feasible to construct mathematical models of the process to provide a basis for estimation of the number of events. One practical way out of this dilemma is to select a set {S} where the time of separation between each S_i (i = 1, 2,. . ., n) is sufficiently small so that at most one indel event will have occurred in a region. It is then possible to count the number of events across the set {S}.

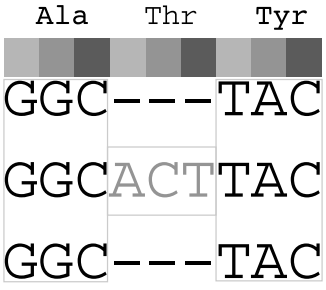

Fig. 2 In this alignment, first-, second-, and third-position nucleotides are indicated above each base. A three-base-pair indel results in an amino acid sequence with threonine inserted between the alanine and tyrosine in the first and third sequence.

ESTIMATION OF DISTANCE METRICS

Once an alignment is obtained, various calculations can be made on the aligned set {A}. One of the most common and useful calculations is to estimate the evolutionary distance between a pair of sequences (see Ref. 1 for a detailed discussion of distance metrics). The simplest distance measure is the percent divergence between two sequences, where divergence is measured as nucleotide site differences and indels are omitted from the calculation. This is a satisfactory measure for sequences that have been separated for a ''sufficiently brief'' period of evolutionary time. In operational terms, ''sufficiently brief'' means that the likelihood of two mutations hitting the same site is small enough to be neglected. Over longer periods of time, the likelihood of two or more hits at a site cannot be neglected, so a model of the substitution process must be introduced. Mathematical models permit estimation of the total number of events both observed and unobserved.

A number of mathematical models of nucleotide substitution have been introduced over the years. They all have the following fundamental assumptions in common: 1) The probability of a substitution event per site is assumed to be small so that multiple events per site have close to a zero probability over small intervals of time; 2) statistical independence of mutational events over time and over sites is assumed; and 3) an assumption must be made about the equilibrium frequency of nucleotides in the sequence (usually a uniform frequency of 25% per

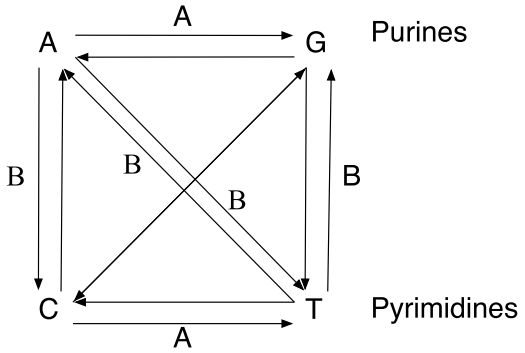

Fig. 3 Under the simplest (one-parameter) model of nucleotide substitution, the rate of change between any pair of nucleotides is equivalent. Under the two-parameter model (pictured here), the rate of transitions (A) may differ from the rate of transversions (B).

nucleotide type is assumed [A = 0.25, T = 0.25, G = 0.25, C = 0.25]). The simplest model of nucleotide change assumes that all possible nucleotide interchanges occur with an identical probability,[2] which requires that only a single mutation parameter be estimated.

Other progressively more complex models are also available. For example, a slightly more complex model posits two mutation parameters—one for the probability of transition mutations and one for the probability of transversion mutations (a transition is an interchange of a purine for another purine or a prymidine for another prymidine and a transversion is a purine to prymidine interchange or the reverse) (Fig. 3). There are now computer programs available that allow researchers to choose the model of DNA substitution that best fits their data.[3,4] A number of programs implement various distance calculations from aligned sequence sets; among the best known is the PHYLIP package,[5] at http://evolution.genetics.washington.edu/phylip.html. The PHYLIP Web page provides links to many other computer programs useful for the study of sequence evolution.

At its simplest, the distance between two sequences (1 and 2), denoted $D_{1,2}$, is a function of the mutation rate μ and the time ($2t$) since the two sequences separated from a common ancestor (the factor 2 occurs because both sequences trace back t time units to their common ancestor). It follows that $D_{1,2}$ is approximately equal to $2\mu t$. This simple relation is very useful in many rough calculations such as obtaining time estimates based on a molecular clock assumption.

Molecular Clock Analyses: As noted previously, the time between separation of lineages can be estimated if mutation rates are constant per site per unit time and if no other forces intervene to alter the rate of accumulation of mutational change between molecules that trace back to a single common ancestor. The most likely force to affect the accumulation of nucleotide change is natural selection. Natural selection acts as a kind of editor of the sequence messages transmitted through time (see article on population genetics). Mutations that arise and are deleterious to function are edited out by selection because their carriers are less likely to survive and reproduce. Hence, it is important to attempt to use sequence regions that are less likely to be perceived by selection for clock calculations (see forthcoming section on partitions of DNA sequence data).

Another likely cause of rate variation is variation in generation time. Recall that we assumed that μ is constant per unit time. Our conventional measurement of time is in years (celestial time), but the more natural measure of time among organisms is in terms of generations. It seems likely that μ is closely correlated with replication cycles that in turn depend on generation time. We must therefore regard molecular clock calculations as rough approximations.

INFERRING HISTORY: THE PROBLEM OF ESTIMATING PHYLOGENETIC TREES

The sequence alignment also provides the starting point for estimating a phylogeny. We will not delve into the technical details of phylogeny estimation, but rather give a brief outline of the major methods in common use (the interested reader is referred to Ref. 1 or Ref. 6 and the PHYLIP documentation (cited previously) for a detailed discussion of this topic). All methods assume that the elements of {A} are connected through a dichotomous branching tree. The problem is to deduce the branching network from {A} subject to some constraint or optimality criterion. There are three different approaches to tree estimation in common use: 1) parsimony; 2) maximum likelihood; and 3) distance-based methods.

Parsimony attempts to minimize the number of site changes over the tree (''character state changes'' in the jargon of systematics). Thus, the parsimony optimality criterion is minimum evolution or minimum number of character state changes. Well-developed algorithms exist for parsimony calculations, and tree estimation is relatively fast for large data sets. This method tends to be the choice of workers in systematics because it conforms to a philosophy that emphasizes the importance of calculating phylogenetic trees from ancestrally derived characters. Parsimony does have at least two drawbacks. First, it does not provide a unique tree. Frequently a number of trees will give the same minimum number of character state changes (i.e., a number of trees may satisfy the optimality criterion). This is not a particularly important drawback as it is a simple consequence of the fact that the parsimony optimality criterion is based on a discrete measure (number of site or character state changes) rather than a continuous measure (as is the case with maximum likelihood and distance-based approaches discussed below). Second, parsimony can be biased when evolutionary rates differ substantially over the tree. This bias is manifested by a tendency for long edges (rapidly evolving branches) to be joined together when in fact they should join other branches on the tree. This is a more significant concern in view of the fact that rate variation appears to be common over long evolutionary distances.

Maximum likelihood is a statistically derived method that depends on a mathematical model of the nucleotide substitution process (analogous to the distance estimation models introduced above). This method has the virtue that the machinery of statistical hypothesis testing can be

implemented to discriminate among candidate trees. It also provides a direct statement about the uncertainty associated with any tree in terms of the variances of branch length estimates. Finally, assuming that the model is appropriate, maximum likelihood takes proper account of rate variation. It is possible to test the fit of various models to data (as discussed above); so, model validation is a feature of the method. Maximum likelihood has the serious drawback that it requires large amounts of computer time for moderate data sets (20–30 sequences), and larger data sets cannot be analyzed in reasonable lengths of time. Several maximum likelihood programs are available over the World Wide Web for tree estimation, and faster versions are under development (see the PHYLIP Web site).

Distance-based methods, as the name implies, begin by estimating a matrix of distances from {A}(denoted $D_{i,j}$) using the estimation methods introduced in the section on evolutionary distances. The optimality criterion seeks to minimize the total distance over the tree. A number of different algorithms have been introduced to estimate a tree from the $[D_{i,j}]$ or from some transformation of the $[D_{i,j}]$ that attempts to adjust for variation in evolutionary rates. Distance-based methods are computationally fast and can be very useful for data exploration. In the case of both parsimony and distance methods, it is easy to use resampling methods (such as the bootstrap) to evaluate the statistical support for a particular tree. Because the various estimation methods differ in assumptions and methodologies, many workers view consistency among methods as an indication that a particular phylogeny is robust to these variations.

A phylogenetic tree is the fundamental starting point for much analytical work in genomics because it provides our best estimate of the pattern of historical relationships among a set of sequences and hence among the set of organisms that donated those sequences. A wide variety of questions are addressed by using the phylogeny as an organizing framework. To illustrate this point, we mention a few important problems that rest on a phylogenetic analysis.

1. For many purposes, it is useful to deduce the pattern of duplication among members of a multigene family. This is accomplished by estimating a gene family phylogeny.
2. An important question is whether the pattern of transposon evolution is consistent with vertical transmission (and hence consistent with a phylogenetic tree).
3. Phylogenetic analysis allows us to ask which partitions of sequence along a chromosome are allelic and share a common history of transmission (inconsistent combinations of mutations that distinguish alleles might indicate intralocus recombination in the gene phylogeny).
4. It is often desirable to map other molecular or phenotypic changes on a tree to determine their time of origin and the temporal order in which major evolutionary events occurred.

ANALYSIS OF SEQUENCE DATA: PARTITIONS OF DATA SETS

It is often informative to divide {S} into various subsets that reflect natural partitions of the data. A natural partition might be untranslated regions and exons; a further partition could be 5′ untranslated regions, introns, exons, and 3′ untranslated regions. In addition, exon regions can be further partitioned into codons, and codons can be partitioned into synonymous versus replacement sites. Finally, synonymous sites can be partitioned into twofold, fourfold, and sixfold degenerate sites based on codon degeneracy patterns. All the calculations introduced above can be performed on various partitions of the data.

One partition that is especially informative is the partition of exon regions into replacement sites (sites where a substitution induces an amino acid change) and synonymous sites (sites where a substitution does not induce an amino acid change). Because synonymous changes do not change the protein, they are thought to be approximately neutral to natural selection. Let us denote D_s as the distance estimated from the synonymous partition of {A} and D_n as the distance estimated from the replacement partition of {A}. Then, the ratio D_s / D_n measures the average strength of selective constraint on amino acid change as compared to synonymous change. The ratio is one when replacement sites change with the same frequency as synonymous sites. Typically this ratio is observed to be approximately 10, indicating a tenfold selective retardation in protein change relative to the assumed mutational input. (This follows because if synonymous changes are neutral, then D_s approximates $2t\mu$ and $2t$ cancels in the ratio.) It is possible to plot this ratio on branches of a phylogenetic tree to search for indications of acceleration or retardation of protein evolution on particular branches of the tree. A plot of this kind allows the detection of regions of the tree where evolutionary patterns have shifted. A ratio of one or greater indicates an acceleration of protein evolution, which may be an indicator of an adaptive shift. For example, Jia et al.[7] have used this approach to identify regions of accelerated protein evolution for the Myb family of plant transcription factors. A D_s / D_n ratio of one or greater in population data can also indicate a balanced polymorphism as is evidently the case with the major histocompatibility polymorphism of humans.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

A different partition is a moving average of the D_s / D_n ratio across the sequence graphed as a function of a sliding window. For example, a window of 100 nucleotides may be chosen, and the averages of D_n and D_s are calculated for this window beginning with nucleotide 1 to 100 in the sequence. The window is then moved one nucleotide to the right (2 to 101), and the averages are recalculated and so forth until the end of the sequence is reached. Large changes in the plots may indicate regions of the sequence that are subject to differing evolutionary forces.

Bioinformatics is a relatively new and rapidly evolving area of research, but there are several excellent books that provide either a detailed introduction to the field[8–11] or a more in-depth exploration of particular problems.[12]

CONCLUSIONS

The mining and analysis of genomics data is still in an early phase. Only a few of the large number of potential applications have been touched on in this article. During the 1990s the technology for producing molecular sequence data grew faster than our ability to extract all useful knowledge from these data sets. Despite this lag, there has been a fortuitous correspondence between the growth of computational power and the growth of genomics databases. There is every reason to expect that the combination of increasing computational power, new analytical techniques, and the expansion of databases will lead to a continuing stream of new discoveries that will have a revolutionary impact on biology, medicine, agriculture, and environmental management during the next quarter century.

ARTICLES OF FURTHER INTEREST

Agriculture and Biodiversity, p. 1
Arabidopsis thaliana: *Characteristics and Annotation of a Model Genome*, p. 47
Arabidopsis Transcription Factors: Genome-Wide Comparative Analysis Among Eukaryotes, p. 51
Chromosome Structure and Evolution, p. 273
Crop Domestication: Fate of Genetic Diversity, p. 333
Crop Improvement: Broadening the Genetic Base for, p. 343
Genetic Diversity Among Weeds, p. 496
Genetic Resource Conservation of Seeds, p. 499
Genome Rearrangements and Survival of Plant Populations to Changes in Environmental Conditions, p. 513
Genome Size, p. 516
Molecular Analysis of Chromosome Landmarks, p. 740
Molecular Evolution, p. 748
Mutational Processes, p. 760
Polyploidy, p. 1038
Population Genetics, p. 1042

REFERENCES

1. Nei, M.; Kumar, S. *Molecular Evolution and Phylogenetics*; Oxford University Press: New York, 2000.
2. Jukes, T.H.; Cantor, C.R. Evolution of Protein Molecules. In *Mammalian Protein Metabolism*; Munro, H.N., Ed.; Academic Press: New York, 1969; 21–132.
3. Posada, D.; Crandall, K.A. MODELTEST: Testing the model of DNA substitution. Bioinformatics **1998**, *14* (9), 817–818.
4. Muse, S.; Kosakovsky Pond, S. *HyPhy. Hypothesis Testing Using Phylogenies*; 2001.
5. Felsenstein, J. *Inferring Phylogenies*; Sirauer: Massachusetts, 2003.
6. Nei, M. *Molecular Population Genetics and Evolution*; Neuberger, A., Tatum, E.I., Eds.; North-Holland: Amsterdam, 1975; Vol. 40. http://www.bio.psu.edu/People/Faculty/Nei/Lab/publications.htm.
7. Jia, L.; Clegg, M.T.; Jiang, T. Excess nonsynonymous substitutions suggest that positive selection episodes operated in the DNA-binding domain evolution of *Arabidopsis* R2R3-MYB genes. Plant Mol. Biol. **2003**, *52* (3), 627–642.
8. Baxevanis, A.D.; Ouellette Francis, B.F. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, 2nd Ed.; Wiley-Interscience: New York, 2001.
9. Krane, D.E.; Raymer, M.L. *Fundamental Concepts of Bioinformatics*; Benjamin Cummings: San Francisco, 2002.
10. Meidanis, J.; Setubal, J.C. *Introduction to Computational Molecular Biology*; PWS Publishing Company: Boston, MA, 1997.
11. Mount, D.W. *Bioinformatics: Sequence and Genome Analysis*; Cold Springs Harbor Press: Cold Springs Harbor, 2001.
12. Gusfield, D. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*; Cambridge University Press: Cambridge, 1997.