BRIEFINGS IN BIOINFORMATICS. VOL 14. NO 5. 599–609
Advance Access published on 2 July 2013
doi:10.1093/bib/bbt046

Mathematics and evolutionary biology make bioinformatics education comprehensible

John R. Jungck and Anton E. Weisstein

Submitted: 11th February 2013; Received (in revised form): 13th May 2013

Abstract

The patterns of variation within a molecular sequence data set result from the interplay between population genetic, molecular evolutionary and macroevolutionary processes—the standard purview of evolutionary biologists. Elucidating these patterns, particularly for large data sets, requires an understanding of the structure, assumptions and limitations of the algorithms used by bioinformatics software—the domain of mathematicians and computer scientists. As a result, bioinformatics often suffers a 'two-culture' problem because of the lack of broad overlapping expertise between these two groups. Collaboration among specialists in different fields has greatly mitigated this problem among active bioinformaticians. However, science education researchers report that much of bioinformatics education does little to bridge the cultural divide, the curriculum too focused on solving narrow problems (e.g. interpreting pre-built phylogenetic trees) rather than on exploring broader ones (e.g. exploring alternative phylogenetic strategies for different kinds of data sets). Herein, we present an introduction to the mathematics of tree enumeration, tree construction, split decomposition and sequence alignment. We also introduce off-line downloadable software tools developed by the BioQUEST Curriculum Consortium to help students learn how to interpret and critically evaluate the results of standard bioinformatics analyses.

Keywords: *bioinformatics education; discrete mathematics; quantitative reasoning; off-line downloadable free and open-source software; evolutionary problem solving*

NARRATIVE

Mathematics and evolutionary biology have contributed enormously to the understanding of biological problems; in the past decade, a number of national and international societies have called for a renewed emphasis on mathematics within biology education [1–3]. Nonetheless, the challenge is still enormous even in interdisciplinary areas such as bioinformatics. In particular, although biology curricula often include courses in probability, statistics and calculus, few encompass topics such as discrete mathematics, geometry and graph theory—some of the concepts most useful in bioinformatics. Sandvik [4] sees an analogous problem in terms of the importance of evolutionary reasoning: 'Tree thinking is an integral part of modern evolutionary biology, and a necessary precondition for phylogenetics and comparative analyses'. He reported that not one of the undergraduate or graduate students in his class could correctly read and interpret a phylogenetic tree. Although new instruments [5–9] have helped measure and improve students' phylogenetic skills, these authors report a continued need for additional tools. Herein, we introduce such tools and accompanying teaching

Corresponding author. John R. Jungck, Departments of Biological Sciences and Mathematics, Interdisciplinary Science and Engineering Laboratory, University of Delaware, Newark, DE 19716, USA. Tel.: +1-302-831-6400; Fax: +1-302-831-6477; E-mail: jrjungck@gmail.com

John R. Jungck is Professor of Biological Sciences at the University of Delaware with joint appointments in the Department of Mathematics and the Delaware Bioinformatics Institute. He is the Director of the Interdisciplinary Sciences Learning Laboratories. He was the founder of the BioQUEST Curriculum Consortium, and is currently Editor of Biology International and the BioQUEST Library.

Anton E. Weisstein is an Associate Professor of Biology at Truman State University. His specialties are population genetics, bioinformatics and mathematical modeling. He is the primary developer of numerous software packages in the Biological ESTEEM Project. He is a mathematical biology researcher, science educator and curriculum developer.

strategies developed and piloted as part of the Biological ESTEEM Collection (*Excel* Simulations and Tools for Exploratory, Experiential Mathematics) [10], an online suite of *Excel*-based modules for open-ended investigations in introductory-level mathematical and computational biology.

The software tools and teaching strategies described in this article have grown out of faculty development workshops administered through the BEDROCK program (Bioinformatics Education Dissemination: Reaching Out, Connecting and Knitting-together). These 4–10-day workshops primarily attract faculty members who teach biology and computer science courses in the first 2 years of the undergraduate curriculum, and also involve a smaller number of high school biology teachers, graduate students interested in bioinformatics education and research scientists seeking to master bioinformatics concepts and tools to aid in their own investigations. Since 1996, we have collaborated with BioQUEST Consortium colleagues to facilitate >50 such workshops at venues across the United States and in Thailand. At these workshops, we demonstrate a pedagogical approach that combines evolutionary problem solving, collaborative inquiry learning and quantitative reasoning [11]. For example, rather than framing a discussion of phylogenetic methods as a lecture containing one or two short activities, we simply give participants a set of nucleotide or amino acid sequences and allow them to devise and test their own tree-building strategies. The 'Eureka' moment, when participants work out some key principle on their own, or even when they see how and why a particular approach fails, helps build the confidence needed to incorporate mathematical content and/or active pedagogies into their own classes.

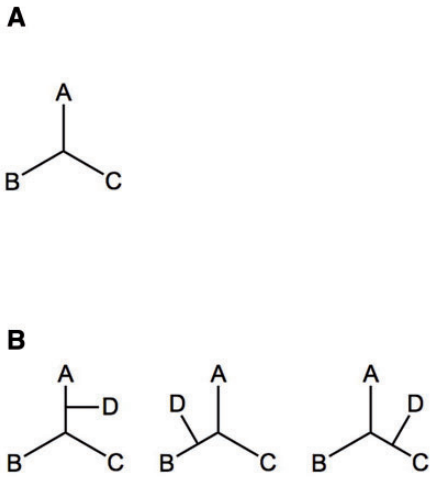
We also emphasize exploring each problem via multiple complementary approaches: for example, building a multiple sequence alignment and a corresponding phylogenetic tree, projecting the alignment onto a 3D protein structure to identify areas of conservation and mapping individual structural changes onto the tree to reconstruct a lineage's evolutionary history. This pedagogical technique mirrors the 'rule of four' from the reform calculus movement, which advocates a mixture of graphical, numerical, algebraic and verbal approaches to the same problem [12]. Similarly, we explore systems biology problems by demonstrating how graph-theoretical properties like connectedness and degree distribution can help quantify key biological principles such as co-expression, binding and spatial proximity. As Hayes [13] notes: 'When you look at a graph drawing, it's hard not to focus on the arrangement of the dots and lines, but in graph theory all that matters is the pattern of connections: the topology, not the geometry'. Thus, translating a complex biological data set into a formal mathematical abstraction can help biologists find new ways to test hypotheses [14], explore causal mechanisms [15] and infer evolutionary constraints [16].

Finally, the diversity of venues in which we have held workshops has led us to appreciate the dependability and accessibility of downloadable off-line tools. Whether due to lack of funds, unreliable infrastructure or cumbersome security protocols, individual instructors cannot always guarantee that their students will have access to well-maintained computers and reliable internet connections. These classrooms can nonetheless use the software tools we describe by simply downloading them in advance (perhaps to a personal computer) and using them to generate data sets for pencil-and-paper analysis. In classrooms with dependable internet access and a sufficient number of computers, the instructor can combine such activities with direct student use of the software tools themselves.

COMPUTATIONAL COMPLEXITY

Why should a biologist develop some intuitive sense of computational complexity? Often in a workshop with college and university faculty as well as with students, we will hear that: 'the supercomputer is down because we haven't gotten a result in several minutes'. When you ask them how many sequences they submitted in their batch job to build trees, a not atypical response is: 'Oh, about 300'. At this point, we often go through an inductive proof of the combinatorial complexity of the number of trees for a particular number of sequences. When we compute that there are more than Avogadro's number of distinct tree topologies (considering only strictly bifurcating, unrooted trees) for 23 sequences, they usually do not request us to push the point home any further. We also note that trees can in fact include hundreds of different sequences, but that this relies on methods that can efficiently exclude enormous numbers of possible trees rather than exhaustively computing each.

In workshops, we begin by noting that an unrooted tree containing $n = 3$ taxa has a unique topological solution (Figure 1A). We then ask where a



C

EvolSeq 1.3.xls

	A	B	C	D	E	F
1			Unrooted Trees		Rooted Trees	
2	# of sequences	# Pairwise distances	# trees	# branches / tree	# trees	# branches / tree
3	3	3	1	3	3	4
4	4	6	3	5	15	6
5	5	10	15	7	105	8
6	6	15	105	9	945	10
7	7	21	945	11	10,395	12
14	14	91	3.16E+11	25	7.91E+12	26
15	15	105	7.91E+12	27	2.13E+14	28
16	20	190	2.22E+20	37	8.20E+21	38
17	25	300	2.54E+28	47	1.19E+30	48
18	30	435	8.69E+36	57	4.95E+38	58
19	35	595	7.30E+45	67	4.89E+47	68
20	40	780	1.31E+55	77	1.01E+57	78
21	45	990	4.56E+64	87	3.97E+66	88
22	50	1225	2.84E+74	97	2.75E+76	98
23	N	$\frac{N(N-1)}{2}$	$\frac{(2N-5)!}{2^{N-3}(N-3)!}$	$2N-3$	$\frac{(2N-3)!}{2^{N-2}(N-2)!}$	$2N-2$

Figure 1: (**A**) The only unrooted tree topology for three sequences. (**B**) Adding a fourth sequence (equivalent to adding a vertex and edge to any of the three edges in the previous tree) increases the number of distinct unrooted topologies to three. (**C**) Extending this process to N taxa yields inductive formulas for the number of rooted and unrooted tree topologies. For a data set of just 25 sequences, the number of possible topologies greatly exceeds Avogadro's number (6.02×10^{23}). Table reproduced from EvolSeq, an interactive spreadsheet in the Biological ESTEEM Project's online collection.

fourth taxon could be added to this tree. If trees are strictly bifurcating, participants easily see that the new taxon can be added to any one of three existing branches. Doing so adds a new branch to the tree while also splitting an existing branch into two segments, thereby increasing the total number of branches by two. Thus, for $n = 4$ taxa, each of the three possible unrooted tree topologies contains five branches (Figure 1B). Next, we ask participants to work in groups to count the possible topologies for $n = 5$ taxa. Through a combination of brute-force enumeration and induction from the previous case, they soon arrive at the general conclusion that $3 \times 5 \times 7 \times 9 \times \ldots \times (2n - 5)$ topologies are possible for an unrooted tree with n taxa. This formula can be rewritten as $(2n - 5)! / [2^{n-3}(n - 3)!]$, or as $(2n - 5)!!$, where the double factorial notation indicates multiplication of every other term.

We next turn to rooted trees, asking participants to count the number of rooted topologies for $n = 3$ and for $n = 4$ taxa, and to explain their results. With minimal prompting, they realize that rooting a tree at a specific point is topologically equivalent to adding a taxon at that same point. Therefore, the number of rooted tree topologies for n taxa is equal to the number of unrooted topologies for $n + 1$ taxa and can be written as $(2n - 3)! / [2^{n-2}(n - 2)!]$ or $(2n - 3)!!$. We use these formulas to show that, for both tree types, the number of possible topologies increases rapidly with the number of taxa (Figure 1C); determining all possible tree topologies can thus be an onerous process even on a supercomputer. For users of bioinformatics software, we also show that similar considerations apply for run times versus the number of genomes [17].

This exploration also allows us to point out an interesting paradox within evolutionary bioinformatics: the pairwise distances simultaneously underdetermine the tree's topology and overdetermine its branchlengths. This is a simple consequence of the fact that the number of branches increases linearly with n, whereas the number of pairwise distances (here defined as the number of nucleotide or amino acid differences between sequences) increases quadratically and the number of topologies increases factorially. Therefore, we introduce best-fit methods for estimating branchlengths in situations where simple arithmetic approaches no longer hold.

The seemingly simple and abstract task of enumerating phylogenetic trees thus motivates active exploration of key topics such as combinatorial explosions, unrooted versus. rooted trees, recursion formulas and approximation methods. These concepts, along with a clearer appreciation of the overwhelming size of tree and sequence space [18], form the foundation for subsequent exploration of more concrete problems, such as the construction of phylogenetic trees.

TREE CONSTRUCTION

Many modern software packages for constructing phylogenetic trees include a bewildering array of different procedures: neighbor joining, maximum parsimony, maximum likelihood and Bayesian algorithms are just some of the more widely used techniques. However, we have found that two older methods—UPGMA and Fitch–Margoliash—introduce core phylogenetic principles in a way that learners can extend to more complex procedures. During faculty workshops, we begin with a simple matrix of pairwise distances between nucleotide sequences (Figure 2A) and ask participant groups to use these data to infer a phylogenetic tree (Figure 2B). Individual groups typically experiment with a variety of intuitive approaches: joining sequences to the tree in order of increasing or decreasing similarity; representing branches as straight, slanted or curved lines; and solving the tree's topology and branchlengths simultaneously or sequentially. As participants discuss their trees, they unpack and collectively examine their evolutionary assumptions and mathematical methods.

At this point, we elicit and formalize participants' intuitive notions regarding genetic distance. In particular, we note that any measurement of distance must be non-negative and symmetric (i.e. the distance from A to B is equal to the distance from B to A), and must satisfy the triangle inequality $d_{AC} \leq d_{AB} + d_{BC}$ for any set of three sequences. Moreover, if the sequences are evolving via a strict molecular clock, the distance matrix will also satisfy the three-point condition [19]: for any three sequences, the two largest pairwise distances will be equal. Trees satisfying this condition are described as ultrametric. In such trees, the distance between any two sister taxa can be divided equally between their two branches, making tree construction much simpler.

In practice, sequence evolution often deviates from a strict molecular clock: substitution rates may vary over time or among sequences (perhaps because of changes in selective pressure or effective population size). Phylogenetic trees of such sequences will no longer be ultrametric, but may instead satisfy the weaker condition of additivity. In an additive tree, the genetic distance between any two points can be found by adding together the lengths of all branches since those sequences' most recent common ancestor [Figure 3]. Reconstruction of additive trees is complicated by the fact that, unlike in an ultrametric tree, sister taxa need not be equidistant from their most recent common ancestor. Instead, branchlengths must be inferred by solving a system of simultaneous linear equations. For a three-taxa tree, these equations have a single exact solution [20]: if d_{AB}, d_{AC} and d_{BC} are the three pairwise distances between sequences, the distances from their central vertex V are given by $d_{AV} = \frac{1}{2}(d_{AB} + d_{AC} - d_{BC})$, $d_{BV} = \frac{1}{2}(d_{AB} + d_{BC} - d_{AC})$ and $d_{CV} = \frac{1}{2}(d_{BC} + d_{AC} - d_{AB})$. Larger trees can be handled by solving for three sequences at a time, collapsing the two closest sequences into their common vertex, calculating the average distance from that vertex to each remaining sequence in the distance matrix, adding the next sequence to the remaining two and solving again.

Thus far, we have measured genetic distance as the number of nucleotide or amino acid differences between sequences. However, this simple metric does not consider that multiple substitutions may have occurred at the same position, and thus tends to underestimate actual amount of divergence. Therefore, the next logical refinement is the

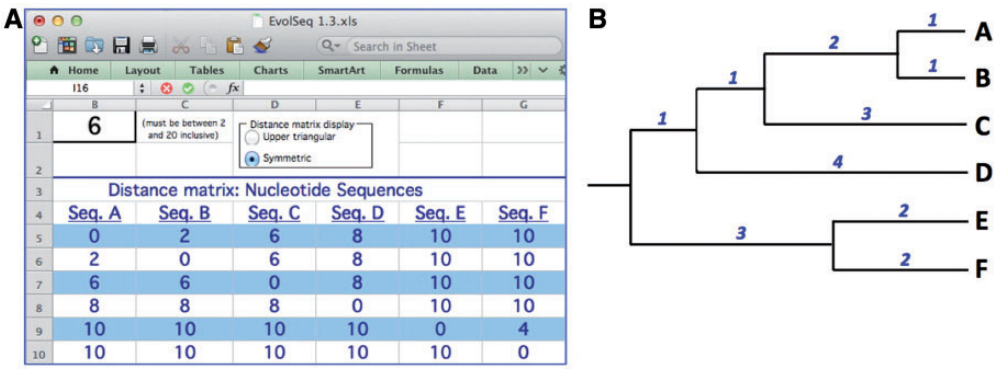


Figure 2: (**A**) Genetic distance matrix for nucleic acid sequences generated by EvolSeq; (**B**) Corresponding phylogenetic tree illustrating the ultrametric condition.

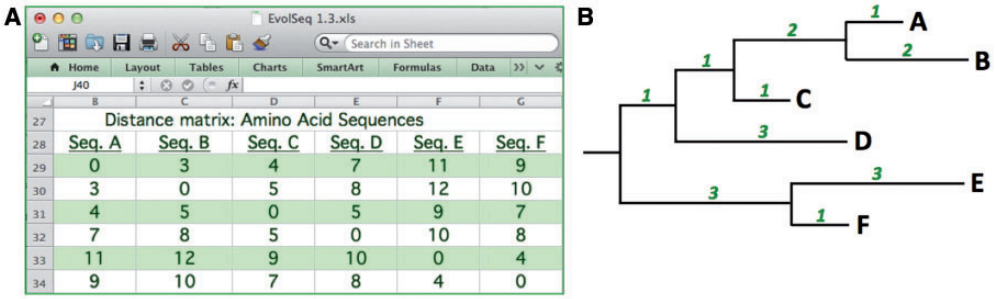


Figure 3: (**A**) Genetic distance matrix for amino acid sequences generated by EvolSeq; (**B**) Corresponding phylogenetic tree illustrating the additive criterion.

Jukes–Cantor distance metric, which includes a correction for this phenomenon. Subsequently, we introduce more sophisticated distance metrics that incorporate transition–transversion bias (Kimura 2-parameter model), unequal base frequency (HKY85) and variation in substitution rate across positions (HKY + Γ + I). We also discuss distance metrics used to analyze amino acid sequences, such as the BLOSUM80 matrix for closely related sequences and the PAM250 matrix for more distant relationships.

To produce distance matrices for ultrametric and additive trees, we use the ESTEEM package EvolSeq, an Excel workbook that simulates the molecular evolution of DNA sequences. The workbook begins with a single random sequence, and then follows that sequence through time as it reproduces and mutates. Eventually, up to 20 related sequences are generated. EvolSeq then calculates the (ultrametric) genetic distances between each pair of DNA sequences and also the (additive) distances between the associated amino acid sequences. Although EvolSeq's evolutionary model is extremely simplistic, it is useful as a tool for rapidly generating trees for classroom pencil-and-paper exercises. The central lesson for students is that simple algebraic methods can yield substantial insight into phylogenetic methods, and that additional refinements based on the evolutionary models appropriate to a particular biological system can improve those methods still further.

TREE HYPOTHESIS TESTING

As described earlier in the text, genetic distance matrices for groups of four or more taxa can seldom be fitted exactly to a phylogenetic tree. In cases where this mismatch is not the focus of study, best-fit approaches may be used to infer an approximate tree. However, substantial deviation from purely tree-like structure may itself contain valuable biological information, such as episodes of convergent evolution or violation of the evolutionary model used to infer the tree. Therefore, split decomposition may be used to resolve and display the full set of taxonomic groupings supported by the data, rather than only those groupings inferred to result from shared ancestry [21].

Split decomposition begins by dividing the set of taxa into two partitions J and K such that each partition contains at least two taxa. We can then sample taxa i and j from J, and taxa k and l from K, to calculate $\omega_{(ij)(kl)} = \max(d_{ik} + d_{jl},\ d_{il} + d_{jk}) - (d_{ij} + d_{kl})$. (This statistic simply measures the strength of the inequality described by the four-point condition.) If $\omega > 0$ for every possible quartet of taxa sampled in this way, then the dichotomy (J, K) is defined as a split with isolation index $\alpha = \frac{1}{2}\min(\omega_{(ij)(kl)})$. For example, in the distance matrix in Figure 4A, consider the partition (AB)(CD). For this partition, $\omega_{(AB)(CD)} = \max(80,\ 83) - 58 = 25$ is the only compatible quartet; therefore, it is a split with isolation index $\alpha = 12.5$. (See Figure 4B for a graphical depiction of this formula.) Similar calculations reveal that the partition (AC)(BD) is a split with index $\alpha = 1.5$, whereas the partition (AD)(BC) yields $\omega < 0$, and therefore not a split. Split decomposition eliminates the system's overdeterminacy: the six branchlengths represent the exact solution to the system of equations that describe the six pairwise distances.

SplitDecomp is an Excel workbook from the ESTEEM Collection that performs split decomposition on a set of four DNA sequences and their associated amino acid sequences (Figure 5). The user can type in the sequences or paste them in from a text file. The program then translates the DNA sequences

A

SplitDecomp – Dentist.xls

Search in Sheet

Home | Layout | Tables | Charts | SmartArt | Formulas | Data | Review | Developer

N13

	A	B	C	D	E	F	G	H	I	J	K	L
1	Nucleotide sequences											
2	Distance matrix							Split decomposition				
3			A	B	C	D						
4			Dentist	Patient E	Local control 3	Local control 22		Putative tree		Four-point condition		Isolation index
5	A)	Dentist		14	43	40		((A,B), (C,D))		58	$=d_{AB} + d_{CD}$	12.5
6	B)	Patient E			43	37		((A,C), (B,D))		80	$=d_{AC} + d_{BD}$	1.5
7	C)	Local control 3				44		((A,D), (B,C))		83	$=d_{AD} + d_{BC}$	-1.5
8	D)	Local control 22										

B

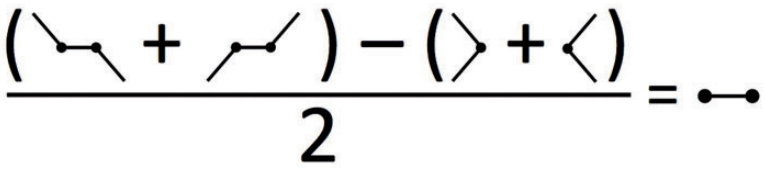


Figure 4: (**A**) Split decomposition results as produced by SplitDecomp. The isolation index values suggest strong support for the partition (AB)(CD), much weaker support for (AC)(BD) and no significant support for (AD)(BC). (**B**) A graphical interpretation of an isolation index. Addition and subtracting appropriate pairwise distances allows estimation of the internal branchlength, which is equivalent to determining which tree topology best separates one pair of sequences from the other pair.

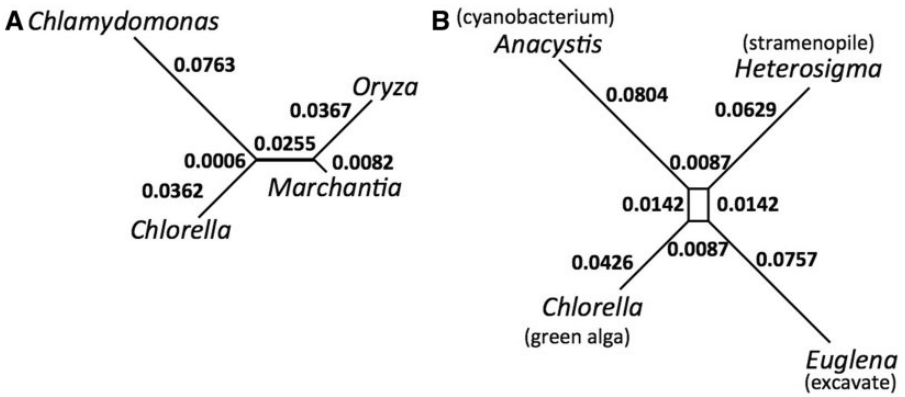


Figure 5: Split decomposition results as produced by SplitsTree [22]. Data set modified from [23], with taxon names updated.

into amino acid sequences and calculates the isolation index α for each of the three possible unrooted phylogenies (Figure 6). The user can use these indices to sassess the data's support for each of these topologies.

For greater numbers of taxa, a split decomposition network [22] can be used to simultaneously depict multiple phylogenetic hypotheses, with each split's length corresponding approximately to its support from the data. For example, the split separating tobacco and rice from the remaining taxa (weight = 0.0104) is 13 times stronger than that separating tobacco and liverworts from the others (0.0008), consistent with the hypothesis that the vascular plants represent a monophyletic group (Figure 5A). By contrast, there is nearly as much evidence for a *Euglena*-heterokont clade (0.0087) as for a *Euglena*-green algae grouping (0.0142), which leads us to infer the presence of conflicting signals within this data set (Figure 5B).

SplitDecomp can be used to address a number of overarching issues in phylogenetics. For example, students can explore the relationship between branchlength and phylogenetic confidence by generating random DNA sequences of given length and using

split indices to determine the length of the inferred central branch. Because the sequences are random, there is no central branch, but phylogenetic methods will still infer one based on chance resemblances among sequences. (This can be shown using Biology Workbench [25], EvolSeq, or any of the numerous other phylogenetic tools available online.) This exercise can reinforce the concept of phylogenetic trees as hypotheses to be tested. A similar approach can illuminate other important topics such as long-branch attraction (see Figure 6) and recombination.

SplitDecomp can also be used to explore and analyze a specific phylogenetic question. For example, a 1992 study [26] examined whether an HIV-infected dentist had inadvertently transmitted the virus to several of his patients during invasive dental surgery. This case had enormous implications for health care practice and received international attention. The study focused on sequence data from the viruses infecting the dentist, the HIV-positive patients and a number of HIV-positive local controls with no known epidemiological link to the dentist. Curricular materials already exist [27] for introducing students to phylogenetic concepts and methods using the data from this study. Students can then further analyze their findings using SplitDecomp to estimate the lengths of individual branches and determine whether they are well supported by the data (Figure 7).

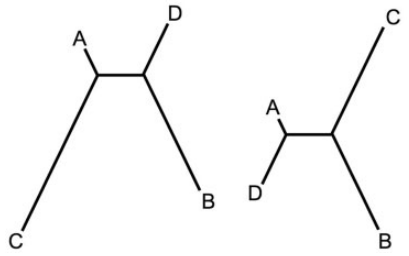


Figure 6: Long-branch attraction. The left-hand figure shows the actual phylogeny for four sequences; the right-hand figure shows the phylogeny reconstructed using ClustalW in Biology Workbench. Sequences B and C, at the end of the two longest branches, have been mistakenly clustered together. Long-branch attraction is a well-known problem for several phylogenetic methods, notably parsimony [24].

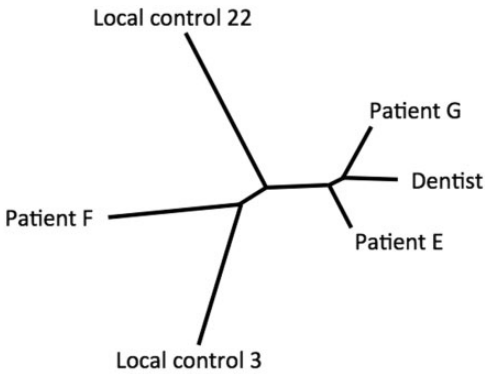


Figure 7: Phylogeny of HIV sequences from the dentist, three patients and two local controls. How would you interpret this phylogeny? Can split decomposition be used to quantify the statistical support for your conclusions? Modified from [27].

SEQUENCE ALIGNMENT: SCORING, SUBSTITUTIONS, GAP PENALTIES

Phylogenetic inferences rely on the assumption that the characters under study are homologous, representing descent from a shared ancestor rather than convergent evolution from different starting points. For example, when analyzing morphological data, the presence of eyespots on the wings of two different butterflies (of the same or different species) may result from those butterflies both having inherited the relevant genetic signals from the same ancestor [28]. Similarly, if two individuals have a thymine residue at the same position in the same gene, this may also reflect common ancestry. However, an indel mutation can shift a residue's position in one sequence relative to another, whereas the small number of possible states (4 for nucleic acid sequences, 20 for amino acids) can produce spurious matches between residues derived from different positions (Figure 8). The process of posing hypotheses about which residues in one sequence are homologous to specific residues in another is called sequence alignment.

Most alignment procedures use dynamic programming, a computational approach that recursively breaks a large problem into smaller sub-problems, the solutions of which are then plugged back into the overall problem. When aligning a pair of sequences, the problem can be represented by an alignment matrix (Figure 9A). Beginning at the upper left corner, we may take any one of three possible steps: (i) move right one cell, thereby adding a gap in sequence #1, (ii) move down one cell, thereby adding a gap in sequence #2 or (iii) move diagonally down to the right one cell, thereby

adding no gap to either sequence. We then continue taking steps in any of these directions until the lower right corner is reached. Each possible path through the alignment matrix represents a potential alignment for the two sequences (Figure 9B).

Next, each cell is assigned a score that reflects the evolutionary likelihood of the best partial alignment leading up to that cell. For example, a step that matches a nucleotide with an identical nucleotide represents evolutionary conservation and thus receives a positive match bonus. By contrast, steps that represent evolutionary change receive either a mismatch penalty (for nucleotide changes) or a gap penalty (for insertions or deletions). Gap penalties are usually assumed to be larger than mismatch penalties, reflecting the relative likelihood of these two types of mutation.

The cell in the upper left corner receives a score of zero; other cell scores are calculated along the different paths leading to each cell. Cells along the topmost row, which can be reached only by repeatedly adding gaps in sequence #1, thus receive scores that are simple multiples of the gap penalty. The same also holds for cells in the left column. By contrast, all other cells have three possible scores representing the three different directions from which they can be reached (steps i, ii and iii, earlier in text). Each cell receives the largest of those three scores, representing the best partial alignment up to that point. The final score (in the bottom right cell) is then the score of the best possible alignment between the two sequences; that alignment itself is given by the path or paths that yield that score (Figure 9C).

Sequence 1:	AAGTACCTG	AAGTACCTG
Sequence 2:	AAGTATCTA	AAGTATCTA
Sequence 3:	ATATCTG	A--TATCTG

Figure 8: An example of sequence alignment. In the alignment at left, the boldfaced Ts in the first two sequences are hypothesized to be homologous to the italicized T in sequence 3. However, if the same sequences are aligned as shown at right, the underlined T is the homologous one. This alignment maximizes sequence conservation and would, therefore, be preferred by most alignment procedures.

The aforementioned procedure is known as the Needleman–Wunsch algorithm [29] and is appropriate for global alignments (i.e. when both sequences are to be aligned in their entirety, such as when comparing two complete genes: see Figure 10A). Other applications of sequence alignment require different algorithms. In particular, semi-global alignments attempt to infer homology between an entire short sequence (the query sequence) and some portion of a much longer sequence (the subject sequence). This question may arise when aligning a single gene within an entire genome, or a single sequencing fragment to a larger assembly. In such cases, gaps before and after the query sequence simply represent regions where sequence data were not collected rather than true evolutionary events, so these gaps are not penalized within the scoring matrix (Figure 10B). Similarly, local alignments are appropriate when searching for local regions of homology between two larger sequences (e.g. to find a shared motif). This situation requires the Smith–Waterman algorithm [30], in which cell scores are restricted to non-negative values and the final alignment terminates at the maximum value occurring anywhere within the alignment matrix (Figure 10C).

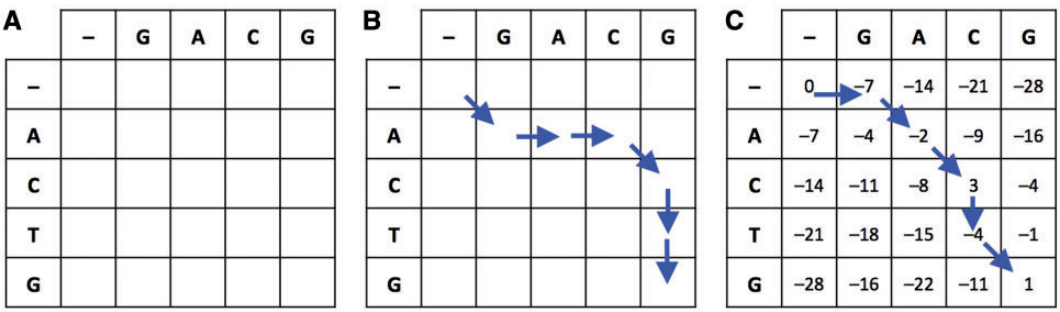


Figure 9: (**A**) Initial alignment matrix for two nucleic acid sequences. (**B**) One possible path through this matrix, showing locations of putative gaps (horizontal and vertical arrows) and matches or mismatches (diagonal arrows). (**C**) Scoring matrix produced by dynamic programming, with arrows indicating optimal alignment. Each cell's score represents the degree of evolutionary conservation in the best partial alignment up to that point in the overall matrix.

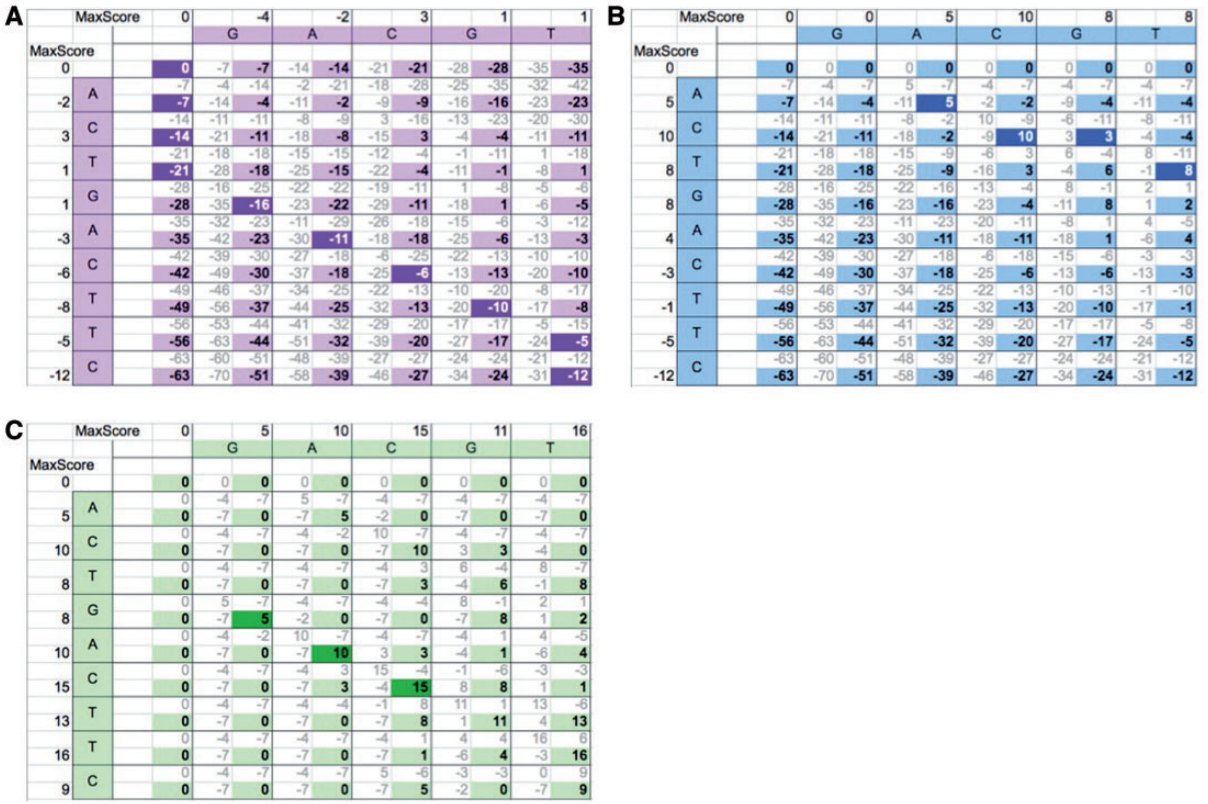


Figure 10: A comparison of (**A**) global, (**B**) semiglobal and (**C**) local alignments, as produced by PairwiseAlignment.

The ESTEEM tool, PairwiseAlignment, presents these calculations in the familiar setting of an Excel workbook. The user enters a pair of nucleic acid sequences and chooses an alignment type; the spreadsheet then computes cell scores and displays a trackback of the optimal alignment. This tool can be used to help demonstrate dynamic programming by having students work one or two simple examples by hand. Once students understand this process, PairwiseAlignment can be used to explore more realistic scenarios that illustrate deeper conceptual questions in sequence alignment. For example, next-generation DNA sequencing methods rely on assembling enormous numbers of short (35–500 bp) reads into a single contig [31]. Pairs of such reads may be locally aligned, and then the highest-scoring pair combined and the process repeated. Alternatively, individual reads may be mapped onto a reference genome via semi-global alignment. However, this alignment procedure may be influenced by the relative frequency of different types of sequencing error (e.g. substitutions versus indels) and is further complicated by the presence of repetitive elements within the region of interest. These issues can be introduced, and potential solutions assessed, through an investigation in which students assemble a contig from a set of 8–15 reads. As students find problems within the data set, such as frequent A→T miscalls, they can be asked to appropriately modify the PairwiseAlignment program (e.g. by reducing the mismatch penalty for such mutations). Thus, understanding the mathematics of sequence alignment provides a conceptual framework for diagnosing and solving many of the challenges that arise when analyzing real data.

OTHER BASIC MATHEMATICS IN BIOINFORMATICS EDUCATION

We have previously published [32] on introducing mathematical modeling to students via engaging them in (i) depicting a system by using box and arrow diagrams; (ii) writing 'word' equations that qualitatively describe the system's behavior; (iii) translating 'word' equations into mathematical symbols; (iv) implementing equations in modeling software (often a spreadsheet like Excel by entering formulas and drawing graphs); and (v) applying

their model to specific cases where actual data are available.

Through the NUMB3R5 COUNT workshop series [33], Neuhauser has introduced students into exploring microarray gene expression data with principal component analysis, single nucleotide polymorphism data with exploratory data analysis and re-sampling statistics, sub-sampling larger populations with special reference to drug testing and cancer genetics with dynamic programming and backtracking. These represent just a few of the opportunities to use simple off-line tools to interactively engage students in developing comprehension of the power of mathematics in making sense of complex data while making only a few simple assumptions.

CONCLUSION

The Mathematics Association of America held a number of conferences to identify appropriate mathematics for the first 2 years [34] for different areas of science and social science. In the volume on 'Mathematics for the Emerging Technologies', the authors emphasized that to prepare students for twenty-first century challenges, student not only 'need to be able to understand and use mathematics with formulas, graphs and tables . . . they need to communicate this understanding'. Bioinformatics students incapable of meeting these criteria might have click interactivity with current software packages, but they will lack the conceptual depth to adjust to new tools without a better mathematical foundation [35–37]. We believe that by introducing the mathematics visually, with simple spreadsheet calculations where they can follow the calculations by checking a formula bar associated with a cell, and using multiple actual data sets associated with contemporary interesting problems, most students/faculty will move beyond either mystification or estrangement to some conceptual appreciation. We have found that classroom use in regions of the world without good access to the internet or without powerful computer hardware has appreciated workshops that use tools that run on their machines and that the mathematics is explained and accessible. Achieving true quantitative literacy—the ability to translate among different representations of the same system, or to adapt strategies from one biological system to another—requires multiple approaches [12]. The ESTEEM tools we have described provide one viable set of tools to support this learning outcome.

Key points

- Understanding core mathematical and computational concepts enhances biology students' and biologists' ability to interpret and critically evaluate the results of bioinformatics analyses.
- Open-ended explorations can help biology students master and apply bioinformatics principles to research-like questions and data sets.
- Simple, but flexible, software tools are available to support these teaching strategies and learning goals.

References

1. Cohen JE. Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. *PLoS Biol* 2004;**2**:2017–23 e439.
2. Jungck JR. Mathematical biology education: modeling makes meaning. *Math Model Nat Phenom* 2011;**6**:1–21.
3. Jungck JR. Ten equations that changed biology: mathematics in problem-solving biology curricula. *Bioscene* 1997;**23**: 11–36.
4. Sandvik H. Tree thinking cannot be taken for granted: challenges for teaching phylogenetics. *Theory Biosci* 2008; **127**:45–51.
5. Baum DA, Offner S. Phylogenies and tree-thinking. *Am Biol Teach* 2008;**70**:222–9.
6. Baum DA, Smith SD, Donovan SS. The tree-thinking challenge. *Science* 2005;**310**:979–80.
7. Omland KE, Cook LG, Cris MD. Tree thinking for all biology: the problem with reading phylogenies as ladders of progress. *Bioessays* 2008;**30**:854–67.
8. Young AK, White BT, Skurtu T. Teaching undergraduate students to draw phylogenetic trees: performance measures and partial successes. *Evol Educ Outreach* 2013;**6**: 16–30.
9. Perry JE, Meir E, Herron JC, *et al*. Evaluating two approaches to helping college students understand evolutionary trees through diagramming tasks. *CBE Life Sci Educ* 2008;**7**:193–201.
10. BioQUEST Curriculum Consortium *Biological ESTEEM: Excel Simulations and Tools for Exploratory, Experiential Mathematics*. http://bioquest.org/esteem/ (10 May 2013, date last accessed).
11. Jungck JR, Donovan SS, Weisstein AE, *et al*. Bioinformatics education dissemination with an evolutionary problem solving perspective. *Brief Bioinform* 2010;**11**:570–81.
12. Knill O. *On the Harvard Consortium Calculus*. http://www.math.harvard.edu/~knill/pedagogy/harvardcalculus/ (10 May 2013, date last accessed).
13. Hayes B. Graph theory in practice: part 1. *Am Sci* 2000;**88**: 9–13.
14. Aittokallio T, Schwikowski B. Graph-based methods for analysing networks in cell biology. *Brief Bioinform* 2006;**7**: 243–55.

15. Greenland S, Brumback B. An overview of relations among causal modelling methods. *Int J Epidemiol* 2002;**31**: 1030–7.
16. Wagner A. Evolutionary constraints permeate large metabolic networks. *BMC Evol Biol* 2009;**9**:231–47.
17. Lemmon AR, Milinkovitch MC. The metapopulation genetic algorithm: an efficient solution for the problem of large phylogeny estimation. *Proc Natl Acad Sci USA* 2002;**99**: 10516–21.
18. Hillis DM, Heath TA, St John K. Analysis and visualization of tree space. *Syst Biol* 2005;**54**:471–82.
19. Van de Peer Y. Phylogenetic inference based on distance methods. In: Lemey P, Salemi M, Vandamme AM (eds). *The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. Cambridge, UK: Cambridge University Press, 2009.
20. Fitch WM, Margoliash E. Construction of phylogenetic trees. *Science* 1967;**155**:279–84.
21. Bandelt HJ, Dress AW. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol Phylogenet Evol* 1992;**1**:242–52.
22. Huson DH. *SplitsTree*. http://www.splitstree.org/ (10 May 2013, date last accessed).
23. Huson DH. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* 1998;**14**:68–73.
24. Felsenstein J. Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool* 1978;**27**: 401–10.
25. Subramaniam S. The biology workbench: a seamless database and analysis environment for the biologist. *Proteins* 1998;**32**:1–2.
26. Ou CY, Ciesielski CA, Myers G, *et al*. Molecular epidemiology of HIV transmission in a dental practice. *Science* 1992; **256**:1165–71.
27. Donovan S. Molecular forensics. In: Jungck JR, Fass MF, Stanley ED (eds). *Microbes Count!* New York: Canterbury Press, 2003;129–36.
28. Breakfield MP, Gates J, Keys D, *et al*. Development, plasticity and evolution of butterfly eyespot patterns. *Nature* 1996;**384**:236–42.
29. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 1970;**48**:443–53.
30. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981;**147**:195–97.
31. Zhang J, Chiodini R, Badr A, *et al*. The impact of next-generation sequencing on genomics. *J Genet Genomics* 2011; **38**:95–109.
32. Weisstein AE. Building mathematical models and biological insight in an introductory biology course. *Math Model Nat Phenom* 2011;**6**:198–214.
33. Neuhauser C. NUMB3R5 COUNT! *Biological Problem Solving Using Data*. http://bioquest.org/numberscount/ (10 May 2013, date last accessed).
34. Hovis MA, Kimball RL, Peterson JC. *A Vision: Mathematics for the Emerging Technologies*. Memphis: American Mathematical Association of Two-Year Colleges, 2002.
35. He M, Petoukhov S. *Mathematics of Bioinformatics: Theory, Practice, and Applications*. New Jersey: John Wiley and Sons, 2011.
36. Waterman M (ed). *Mathematical Methods for DNA Sequences*. Boca Raton: CRC Press, 1999.
37. Dress A. The mathematical basis of molecular phylogenetics. *The BioComputing Hypertext Coursebook*. 1995. http://www.dur.ac.uk/stat.web/Bioinformatics/welcomepage.html ½(10 May 2013, date last accessed).