on ‘regime shift’ or step change in assemblage structure and function, but is it the case that gradual reorganisation of systems is a much more widespread phenomenon than abrupt change? There are still many unresolved issues in the quantification and interpretation of biodiversity change.

It is not just temporal data that are patchy. A further difficulty is that assessments of biodiversity change are thwarted by data gaps, with rich tropical areas and invertebrate assemblages being particularly underrepresented in biodiversity databases. New technological innovations, including sampling of environmental DNA (eDNA), remote sensing, networks of camera traps and sensors, and acoustic surveys have the potential to fill these gaps and automate sampling and analyses. A non-trivial challenge will be to provide continuity with existing biodiversity time series collected using traditional methods and to interpret change in the alpha/beta/gamma framework that underpins all biodiversity measurement. It is also important to report uncertainty in assessments of biodiversity change. The IPBES (Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services, www.ipbes.net) does this by stating whether trends are ‘well-established’, ‘established but incomplete’, or ‘inconclusive’.

Looking ahead

Protecting the world’s ecosystems over decades to come is a formidable challenge, but from the perspective of the measurement of biodiversity there are some encouraging developments. Data gaps will increasingly be resolved; the IPBES 2030 Work Programme has as one of its objectives the strengthening of the knowledge base on which its assessments are made, and it will be supported in this venture by emerging technologies. Citizen science will also play a growing role in this endeavour. Ecological theory can still only incompletely predict the consequences for ecosystem structure and function of the current rapid reconfiguration of natural systems, but this is an important research focus with exciting developments on the horizon. New additions to the biodiversity measurement toolkit will increase the precision and information content of assessments of biodiversity change. Better reporting of these biodiversity metrics to reflect the different facets and scaling properties of biodiversity trends will support conservation efforts. A more nuanced view of how this biodiversity change plays out will also reinforce the message that, while it is right to be gravely concerned about the fate of our planet’s ecosystems, we can still act to safeguard Darwin’s ‘endless forms most beautiful and most wonderful’ for generations to come.

FURTHER READING

BioTIME database of assemblage time series, https://biotime.st-andrews.ac.uk.

Chao, A., Gotelli, N.J., Hsieh, T.C., Sander, E.L., Ma, K.H., Colwell, R.K., and Ellison, A.M. (2014). Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. Ecol. Monogr. *84*, 45–67.

Das Gupta Review (Economics of Biodiversity), www.gov.uk/government/publications/final-report-the-economics-of-biodiversity-the-dasgupta-review.

Díaz, S., Kattge, J., Cornelissen, J.H.C., Wright, I.J., Lavorel, S., Dray, S., Reu, B., Kleyer, M., Wirth, C., Prentice, I.C., *et al*. (2016). The global spectrum of plant form and function. Nature *529*, 167–171.

Dornelas, M., Gotelli, N.J., McGill, B.J., Shimadzu, H., Moyes, F., Sievers, C., and Magurran A.E. (2014). Assemblage time series reveal biodiversity change but not systematic loss. Science *344*, 296–299.

GBIF—the Global Biodiversity Information Facility, www.gbif.org.

Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services, https://www.ipbes.net/.

Living Planet Index, www.livingplanetindex.org.

Magurran, A.E., Dornelas, M., Moyes, F., Gotelli, N.J., and McGill, B. (2015). Rapid biotic homogenization of marine fish communities. Nat. Commun. *6*, 8405.

PREDICTS — Projecting Responses of Ecological Diversity In Changing Terrestrial Systems, www.predicts.org.uk.

Sala, E., Mayorga, J., Bradley, D., Cabral, R.B., Atwood, T.B., Auber, A., Cheung, W., Costello, C., Ferretti, F., Friedlander, A.M., *et al*. (2021). Protecting the global ocean for biodiversity, food and climate. Nature *592*, 397–402.

The Convention of Biological Diversity, www.cbd.int/convention/.

Tucker, C.M., Cadotte, M.W., Carvalho, S.B., Davies, T.J., Ferrier, S., Fritz, S.A., Grenyer, R., Helmus, M.R., Jin, L.S., Mooers, A.O., *et al*. (2017). A guide to phylogenetic metrics for conservation, community ecology and macroecology. Biol. Rev. *92*, 698–715.

Centre for Biological Diversity, School of Biology, University of St Andrews, St Andrews, Scotland KY16 9TH, UK.
E-mail: aem1@st-andrews.ac.uk

Primer

Phylogenomics

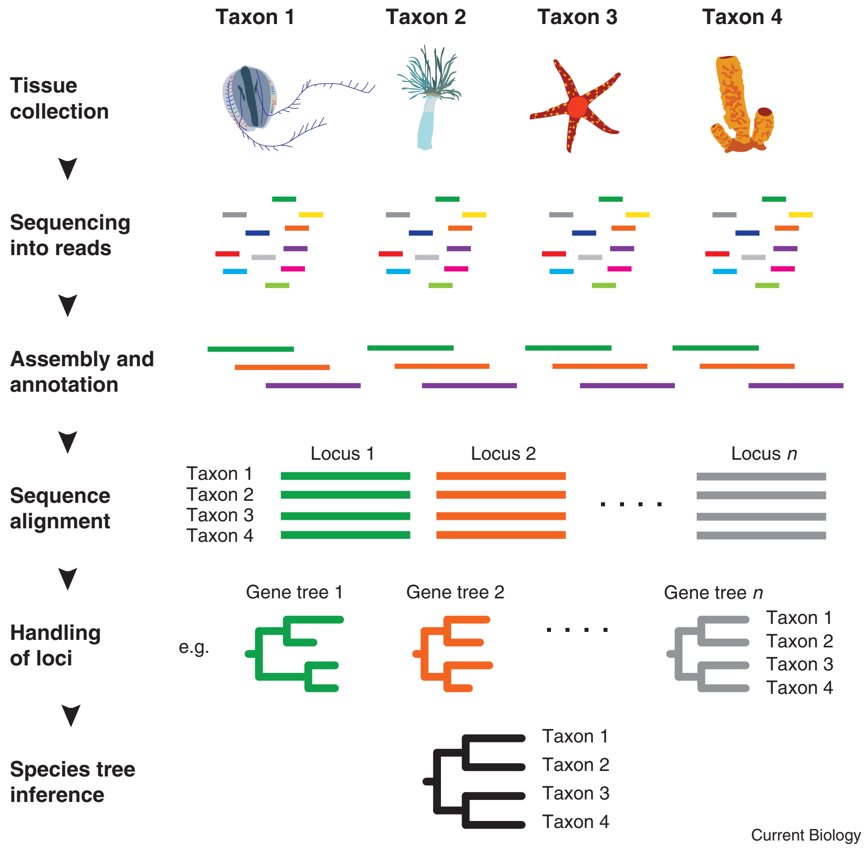
David A. Duchêne

The reconstruction of evolutionary relationships among species is fundamental for our understanding of biodiversity. Today, evolutionary relationships are closely related with the depiction of the tree of life, and research on the topic is underpinned by methods in molecular phylogenetics that have grown in popularity since the 1960s. These methods depend on our understanding of how nucleotide or amino acid sequences evolve through time and in different lineages. Armed with this knowledge, researchers can make inferences about the relationships and amount of genomic divergence among species.

The term ‘phylogenomics’ is primarily used to refer to an extension to phylogenetics that considers not only evolution of nucleotides or amino acids, but also broader processes acting on whole genomes. A dominant simplifying assumption in the field is that genomes are made up of segments that are to some degree independent, including in their evolutionary history. Examining several hundreds or thousands of genomic loci is becoming routine in the biological sciences. However, this has only been possible in the past two decades, with the increasing availability of genome-scale sequencing techniques.

An early insight from phylogenomics that has dominated debate in the field is that genomic regions are very different in terms of the information they contain about evolutionary history. It is common for phylogenomc studies to find that, among thousands of loci, every one has its own individual historical signal. Due to such striking differences in signals across genomic regions, it has become standard to distinguish the evolutionary trees showing the history of individual regions, often called ‘gene trees’, versus the tree representing the history of all of the genomic regions combined, or ‘species tree’.

The data sets used in phylogenomics allow biologists to address questions in a wide variety of fields, such as taxonomy, population genetics, comparative biology and molecular



Current Biology

Figure 1. Basic phylogenomic protocol.
Each step requires several considerations and usually contains multiple checks of the data. Examples of common decisions made by the researcher at each step include: choice of samples to best answer the phylogenetic question; choice of tissue type for data extraction; method of DNA extraction; choice of sequencing technology; removal of reads with insufficient representation in the sequencing products (or insufficient coverage); tests of regions as being truly genomic rather than sequencing artifacts; identification of homology among sequences; identification of the reading frame; extent of substitution saturation in the data; choice of analyses that assume individual signals of phylogeny across loci versus a concatenated alignment; choice of branch support metrics; choice of a statistical framework for species tree inference (e.g. maximum likelihood or Bayesian inference).

evolution. The term ‘phylogenomics’ can be used as the name of other fields of research, and was in fact originally used in reference to the prediction of gene function from related gene sequences. This primer focuses on phylogenomics as an extension to phylogenetics, which aims to infer evolutionary relationships among species.

Phylogenomic data

A phylogenomic data set includes multiple alignments of nucleotide or amino acid sequences. The data generally span a representative sample of a single or several genomic regions, rather than including the complete end-to-end chromosome-level sequence. Commonly used types of genomic regions include genes (regions that can directly code for proteins or functional RNA), individual exons, introns (non-coding regions between exons), transposable elements (non-coding ‘jumping genes’), RNA-transcribed regions and so-called ultra-conserved elements (UCEs). With continued advances in sequencing technology, studies are increasingly using multiple types of genomic regions simultaneously.

The exact genomic sequences in a phylogenomic data set are generally determined in what is known as a library, comprising a set of sequences that are likely to occur in all of the species studied. The regions in the library determine the types of data that will be sequenced. Whole-genome sequencing requires cutting fragments across the whole genome to construct a library. Many studies develop libraries by using probes that target a particular type of genomic region in the genome, excluding any remaining genomic material. Sequencing a representative portion of the genome with such methods means that the sequencing products are homogeneous to some degree, for example in their evolutionary rate and even function.

The products of sequencing are known as reads and can be assembled into contiguous genomic regions or ‘contigs’ (Figure 1). After assembly, a step known as gene annotation gives contigs an identity that can take several forms, such as molecule type, gene function, location in a genome, or homology to the genomes of other species. An efficient and common method of gene annotation in phylogenomics is to align contigs to an existing reference genome of another species. A phylogenomic data set is finalized with the careful alignment of annotated sequences to ensure the homology of each corresponding nucleotide or amino acid across species. At each step of data preparation, there are steps of data checking and cleaning. For example, contigs might be discarded if they are not found to resemble biological data, and are therefore possible artifacts of sequencing known as ‘chimaeras’.

Different data types have various qualities. Exons are often considered to be subject to complex dynamics of selection that can complicate phylogenetic analyses. Introns and transposable elements are fast evolving and are subject to weaker selective constraints. UCEs are usually slowly evolving and can contain a mix of coding and non-coding sequences, so that modelling their evolution accurately can be difficult.

Regardless of the data type used, the data will ideally have evolved at a speed that is appropriate for the depth of the phylogenetic questions being addressed. Regions that evolve too fast will have an historical signal that is eroded by too many superimposed changes, also known as ‘substitution saturation’. Conversely, regions that evolve too slowly lack historical signal. It is, therefore, appropriate to ask whether a phylogenetic method can adequately extract the historical signal in a data set. Tests of sequence alignments, such as those of substitution saturation, information content and model performance can be instrumental for assessing phylogenomic data quality.

Phylogenomic data handling

Data handling in phylogenomics involves the choice both of biological models of evolution and of the method of maximising the efficiency of computational tasks. Effective data handling also aims to minimize the uncertainty and bias that might arise in species-tree inference.

One method of phylogenomic data handling is to join the available loci into a single alignment, known as a ‘concatenated data set’ or ‘supermatrix’. Under this concatenation approach, the loci are assumed to follow a single set of evolutionary relationships among species. Concatenation analyses allow some other aspects of phylogenetic models to vary among individual loci or sets of loci. For instance, loci might vary in their gene-tree branch lengths or ratio of transitions to transversions. The species tree is then estimated as the one with the best fit to the data, and is often estimated using a maximum-likelihood statistical framework.

There are some instances where concatenating all genes leads to high confidence in an incorrect tree, so that it is preferable to handle the data differently. Concatenation assumes that the main source of error in the results is the lack of sufficient data. However, in some circumstances, such as when divergence has been relatively recent, the main source of error in the data is the difference in the historical signal among genes that arises from incomplete lineage sorting (see below). These scenarios can be addressed using a model called the ‘multi-species coalescent’, which considers each gene tree to be an independent outcome of evolution along the lineages in the species tree. A gene tree, in this case, can be taken to arise from any kind of non-recombining region, such as an exon, intron or even a single nucleotide.

Some methods provide an intermediate approach between full concatenation or coalescence analyses. For example, loci with similar evolutionary signals can be collected to create groupings or ‘bins’ of loci that are assumed to follow a single tree topology. The bins can be biologically inspired, such as by well-known non-recombining regions (e.g. mitochondrial or chloroplast genomes), or the core genomes of bacteria. Alternatively, the bins can be chosen automatically using statistical criteria.

Perhaps the most desirable, yet most computationally expensive approach to handling phylogenomic data is within a full maximum likelihood or Bayesian hierarchical framework. Bayesian analyses allow the simultaneous estimation of large numbers of parameters, often forming a hierarchy of models inside larger models, (e.g. a gene-tree model inside a multi-species coalescent species-tree model). The parameters are jointly estimated and can include those of the substitution models, gene trees, the species tree, evolutionary timescales, population-size dynamics and even biogeography. Full maximum likelihood approaches also exist but generally allow the joint estimation of far fewer parameters. While a Bayesian framework may seem ideal, it can be slow when analysing very large data sets. Methods of effectively handling genomic data in full maximum-likelihood and Bayesian frameworks are active areas of research.

One highly efficient method of data handling in phylogenomics is to perform analyses stepwise. For example, loci can be analysed individually as a first step, followed by the preferred method of species tree inference (Figure 1). Stepwise approaches are tractable because they break down computer time into smaller tasks, but the researcher must recognize the assumptions at each step and the possible bias and uncertainty that can become compounded in downstream analyses.

A method of minimizing bias in analyses, while also reducing the computational demands, is to focus on a subset of loci rather than all of the data available. This approach is inspired by the fact that increasing the amounts of data has a diminishing marginal benefit in terms of the inferences. A subset of loci can be a random selection of genomic windows, or chosen according to some criterion that minimizes a particular source of bias. Some criteria of locus choice are based on the phylogenetic information present in each locus, aiming to minimize stochastic error or maximize model performance. However, a preference should generally be placed on using the best models available to avoid the introduction of bias arising from the method of data selection.

Stochastic and systematic error

A critical ingredient for estimating a species tree reliably is a sufficient amount of data. A lack of data, for example arising from very short or very slow-evolving genomic regions, introduces stochasticity and can lead to an incorrect result. Stochastic error can also be problematic when there is an excess of evolutionary change and the signal is dominated by substitutional saturation. Phylogenomics aims to minimize this form of error by maximizing the amount of data available. Nonetheless, even some complete genomes can have very short loci, each with a very limited signal, and even whole genomes can be very short, such as the core genomes of some viruses and bacteria.

Metrics of uncertainty in inferences known as branch supports provide a window into the stochastic error affecting the results. Branch supports are measures of the statistical support for a given phylogenetic grouping, and include the non-parametric bootstrap, Bayesian posterior probabilities and measures of the conflict across the data (e.g. entropy-based metrics and concordance factors). Different measures of branch support can have very different interpretations and most are considered only indirect measures of stochastic error.

Another critical ingredient in phylogenomics is an evolutionary model that reasonably describes the process of molecular evolution that generated the data. Using a highly incorrect model can cause overconfidence in an incorrect result, also known as systematic error. Some processes are well-known to be important in phylogenetic models, such as heterogeneity in the evolutionary process across

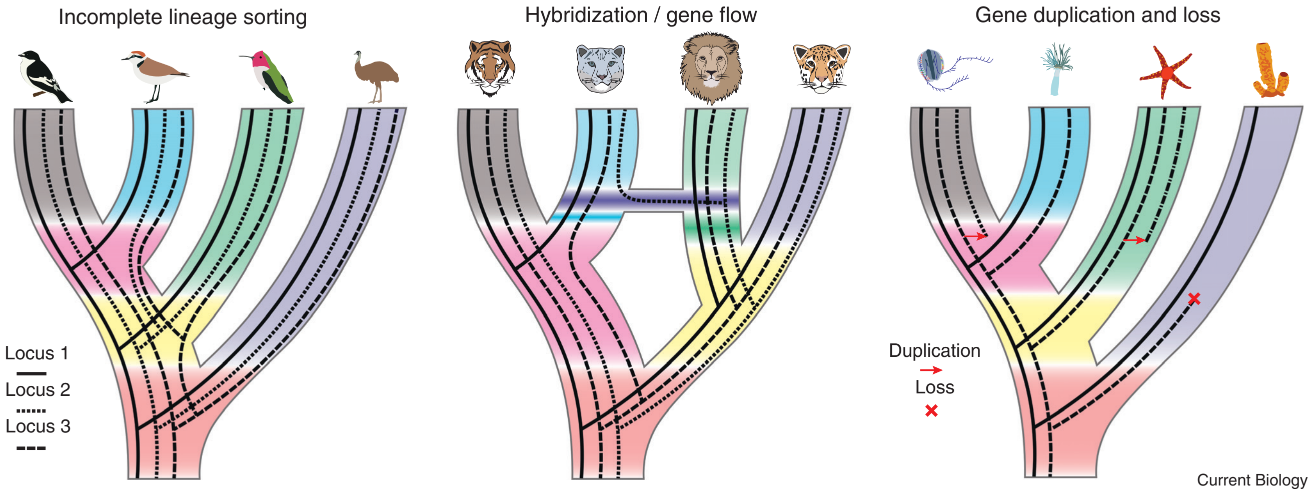

Incomplete lineage sorting
Hybridization / gene flow
Gene duplication and loss
Locus 1
Locus 2
Locus 3
Duplication
Loss
Current Biology

Figure 2. Biological sources of discordance among gene trees.
(Left) Incomplete lineage sorting can drive incongruence among gene trees and is modelled by the multispecies coalescent. The example shows basal splits among modern birds. (Middle) Hybridization or gene flow can result in gene-tree incongruence due to genomic exchange between species after speciation. The example shows a group of modern felines that have experienced such an exchange. (Right) Gene duplication leads to paralogous loci and gene loss leads to missing data. The example shows some of the earliest divergences among extant metazoans. These processes are rarely problematic when gene duplicates are effectively identified before phylogenomic analyses are performed. Broad tree branches show species lineages, with colours indicating mixing populations with possibly different sizes and dynamics. Black lines represent gene trees within the species tree.

nucleotides in a sequence alignment. However, some processes that can be problematic are rarely addressed in phylogenomic analyses, such as heterogeneity in base composition across sequences. It is always advisable to choose the best model available and assess whether the model is realistic enough. There are several ways to do this: for example, it is common to evaluate phylogenetic models of nucleotide substitution for their ability to predict several of the attributes of the empirical data.

Explicit models of gene-tree discordance

Gene trees can vary across the genome due to several historical processes. Recombination allows each locus to take a distinct evolutionary path, allowing loci to diverge in a population before a speciation event occurs (Figure 2). The result can be a gene tree that differs from the species tree. This phenomenon, known as ‘incomplete lineage sorting’, is more likely to occur when ancestral population size has been relatively large or when time between speciation events has been relatively short.

If incomplete lineage sorting has occurred very often across a set of species, gene trees will seldom resemble the species tree. In this case, it can be misleading to assume that only a single tree describes the evolutionary history of a whole genome. One way to address incomplete lineage sorting is by allowing loci to have independent histories, all of which are, however, embedded in a single underlying species tree. This process is explicitly modelled by the multi-species coalescent. Under such a model, it is important to obtain estimates of gene trees that minimize stochastic error.

Horizontal gene transfer is another potential source of discordance among gene trees. It involves the movement of segments of the genome from one species into another and is comparable to species hybridization and introgression. Horizontal gene transfer can lead to divergence events in some gene trees occurring later than the divergences among species (Figure 2). It is ubiquitous in unicellular microbes and in the early evolution of life on earth, and similar processes are increasingly being recognized in phylogenetic studies of eukaryotes (Figure 2).

As lineages split via speciation as well as join via hybridization, it is important to recognize that a species tree is not necessarily fully bifurcating, but can instead look like a network of relationships (Figure 2). A network model accommodates events in which species arise via the merging of populations that were ancestrally distinct. Models that describe this process are often extensions of the multi-species coalescent, such as the multispecies network coalescent.

Gene duplication and loss are other sources of difficulty when estimating species-trees (Figure 2). Even when loci are homologous across species, they might be either orthologous, such that their divergence represents a speciation event, or they might be paralogous, such that their divergence represents an independent gene duplication event. In addition to identifying homology, several methods exist for identifying whether genes might be orthologous or paralogous. It is also possible to implement models of gene diversification and loss for inferring the history of these gene-specific events and their association with species divergences.

Prospects in phylogenomics

Phylogenomics can be summed up as a number of exciting theoretical and methodological advances that greatly expand traditional phylogenetics in two directions: first, by exploiting the increasing availability of genomic sequencing products and computational power; second, in modelling the multiple sources of heterogeneity in signals across genomic loci. The primary bottleneck in the field is still computation time and efficiency, rather than data availability. With its advances, phylogenomics is revolutionising inferences of taxonomy, epidemiology, demographic history, biogeography, divergence time estimation, comparative analysis, genome and trait evolution, among other fields.

Novel methods of estimating the values of a large number of parameters are now allowing the use of highly complex models in analyses of genome-scale data sets. This is also leading to increasingly effective model selection and model averaging, maximising our power to learn from genomic data. Regardless of the directions taken by the upcoming advances in the field, phylogenomics will for the foreseeable future bring dramatic improvements to our understanding of the evolutionary history of life on Earth.

FURTHER READING

Bleidorn, C. (2017). Phylogenomics (Cham, Switzerland: Springer International Publishing).
Bravo, G.A., Antonelli, A., Bacon, C.D., Bartoszek, K., Blom, M.P.K., Huynh, S., Jones, G., Knowles, L.L., Lamichhaney, S., Marcussen, T., *et al*. (2019). Embracing heterogeneity: Building the Tree of Life and the future of phylogenomics. PeerJ *7*, e6399.
Bromham, L. (2016). An Introduction to Molecular Evolution and Phylogenetics (Oxford: Oxford University Press).
Degnan, J.H., and Rosenberg, N.A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. *24*, 332–340.
Edwards, S.V. (2016). Phylogenomic subsampling: a brief review. Zool. Scr. *45*, 63–74.
Liu, L., Xi, Z., Wu, S., Davis, C.C., and Edwards, S.V. (2015). Estimating phylogenetic trees from genome-scale data. Ann. N. Y. Acad. Sci. *1360*, 36–53.
Mirarab, S., Nakhleh, L., and Warnow, T. (2021). Multispecies coalescent: theory and applications in phylogenetics. Annu. Rev. Ecol. Evol. Syst. *52*, https://doi.org/10.1146/annurev-ecolsys-012121-095340.
Reddy, S., Kimball, R.T., Pandey, A., Hosner, P.A., Braun, M.J., Hackett, S.J., Han, K.-L., Harshman, J., Huddleston, C.J., Kingston, S., *et al*. (2017). Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. Syst. Biol. *66*, 857–879.
Scornavacca, C., Delsuc, F., and Galtier, N. (2020).Phylogenetics in the Genomic Era (No commercial publisher), https://hal.inria.fr/PGE/.
Yang, Z., and Rannala, B. (2012). Molecular phylogenetics: principles and practice. Nat. Rev. Genet. *13*, 303–314.

Centre for Evolutionary Hologenomics, University of Copenhagen, Øster Farimagsgade 5A, 1352 Copenhagen, Denmark.
E-mail: david.duchene@sund.ku.dk

Primer

Morphospace

Graham E. Budd

In his famous (if uncharacteristic) burst of lyricism at the end of the *Origin* Darwin described biodiversity as “endless forms most beautiful and wonderful”. It is easy to agree with him when one considers red-lipped batfish or pelagic holothurians. But *are* they endless, or are there limitations to the variety of forms — and if there are, where do they come from? Can morphological evolution be described by Brownian motion of a gas, slowly diffusing to fill up all the space of possible forms, or does it operate within a certain set of constraints? And if there are constraints, where do they come from? The concept of morphospace is an attempt to map out the products of evolution within a quantitative framework to try to shed light on these questions.

Historically, the first important attempt to map the outcomes of evolution was Sewall Wright’s ‘adaptive landscape’, in which the fitness of certain genes was conceptually mapped. In the 1940s, this concept seems to have been adapted by G.G. Simpson by considering not the map of the selective value of genes, but of morphology. Morphospace occupancy takes the process of abstraction of the pattern from the underlying causes one pace further by considering not the fitness of particular forms, but their abundance, or indeed their presence at all. Although there are thus clear historical precedents in the work of Sewall Wright and Simpson, and indeed in the even earlier work of D’Arcy Thompson, the concept of morphospace itself seems to have developed in the 1960s and 1970s, above all from the groundbreaking work of David Raup, who mapped the morphospace occupancy of shelled invertebrates (incidentally pioneering computer graphics in the process via an ingenious use of an oscilloscope). Raup himself stressed the importance of the practical application of his methods, but inevitably, these early beginnings have led to considerable exploration of the theoretical underpinnings of morphospace.

Defining morphospace

A morphospace can be considered to be a type of configuration space, in which objects (in this case, organisms) are placed at points within the space according to a particular set of their properties. An often desirable property of this space is the concept of distance, i.e. the transformation required to map one point onto another. In order for distance to be meaningful in a quantifiable sense, it is necessary for the space to be metric, which means that such transformations conform to a set of axioms (Figure 1). Formally defined configuration spaces do not need to be metric, and in such cases, features such as distance have no meaning, although looser but still useful concepts such as proximity may.

Many familiar spaces are, in addition to being metric, also Euclidean, so that distances can be calculated from extensions of the Pythagorean properties of right-angled triangles (considered in a vector

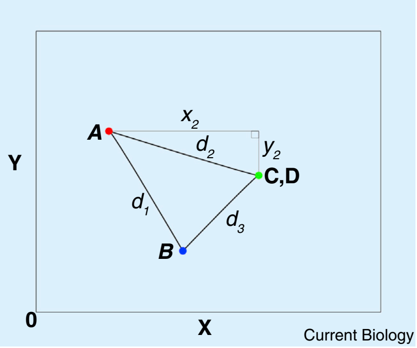


Figure 1. Metric morphospace.
The foundational concept of a metric (morpho-)space, illustrated by four points (A,B,C,D) in a two-dimensional space with distances (d) between them. For the space to be metric, several axioms must be fulfilled: i) if the distance between two points is 0, they occupy the same place (C,D); ii) symmetry: the distances d are identical no matter which direction they are measured in; iii) the triangle inequality: d_1 is always less than $d_2 + d_3$. In a Euclidean space, a distance d is given by the Pythagorean measure (here, $d_2 = \sqrt{x_2^2 + y_2^2}$). In the more generalised affine space, the Pythagorean distance measure may not apply (i.e. angles and distances are not defined, but collinearity is). Other types of morphospace may lack even more components of metricity but still possess the more general feature of proximity.