Journal of Molecular Evolution (2021) 89:157–164
https://doi.org/10.1007/s00239-021-09993-1

COMMENTARY

Ancestral Sequence Reconstruction: From Chemical Paleogenetics to Maximum Likelihood Algorithms and Beyond

Avery G. A. Selberg[1] · **Eric A. Gaucher**[2] · **David A. Liberles**[1]

Received: 4 October 2020 / Accepted: 4 January 2021 / Published online: 24 January 2021

Abstract
As both a computational and an experimental endeavor, ancestral sequence reconstruction remains a timely and important technique. Modern approaches to conduct ancestral sequence reconstruction for proteins are built upon a conceptual framework from journal founder Emile Zuckerkandl. On top of this, work on maximum likelihood phylogenetics published in Journal of Molecular Evolution in 1996 was one of the first approaches for generating maximum likelihood ancestral sequences of proteins. From its computational history, future model development needs as well as potential applications in areas as diverse as computational systems biology, molecular community ecology, infectious disease therapeutics and other biomedical applications, and biotechnology are discussed. From its past in this journal, there is a bright future for ancestral sequence reconstruction in the field of evolutionary biology.

Introduction

Modern sequencing technologies enable us to know the sequence of any protein-encoding gene in any extant species. Ancestral sequence reconstruction offers the opportunity to infer what that protein sequence was at various ancestral points in a phylogenetic tree, providing a window into molecular functions encoded in ancient genomes and how they might differ from those observed in present day genomes. A recent overview of such insights has been provided by Gumulya and Gillam (2017). The history of ancestral protein sequence reconstruction begins with a seminal paper in 1963 from Journal of Molecular Evolution founding editor Emile Zuckerkandl (Pauling et al. 1963) and continues with co-publication of the first maximum likelihood algorithm for ancestral protein sequence reconstruction (Yang et al. 1995; Koshi and Goldstein 1996). The notion that extant sequences and phylogenies could be used to not only infer the topological history of evolution, but also to make inference about the functional history of proteins continues to be an important concept.

Zuckerkandl and Pauling were one of the first to build on the idea that recent genes evolved from previous (homologous) ancestral genes. They noted that using aligned sequences at the tips of a phylogenetic tree, it is possible to determine the amino acid sequence of the ancestral gene and determine where on the tree specific mutations occurred. Not only did their work provide evidence that homologous genes derive from a common gene ancestor, but they also conceptualized a framework that led to the first methods for ancestral sequence reconstruction. Although Zuckerkandl and Pauling noted that the number of mutations between an ancestral gene and a daughter gene is correlated with time, the first widely used method of ancestral sequence reconstruction, parsimony, was notably time-independent (Fitch 1971). Maximum likelihood was introduced as an algorithm a decade later in Journal of Molecular Evolution (Felsenstein 1981) but it would take another 15 years before widely used models of protein evolution in a maximum likelihood framework were developed for ancestral protein sequence reconstruction. This was innovative work published in Journal of Molecular Evolution (Koshi and Goldstein 1996) and contemporaneously by others in Genetics (Yang et al. 1995).

Handling editor: **Aaron David Goldman**.

✉ David A. Liberles
daliberles@temple.edu

[1] Department of Biology and Center for Computational Genetics and Genomics, Temple University, Philadelphia, PA 19122, USA

[2] Department of Biology, Georgia State University, Atlanta, GA 30303, USA

Early Methodological Development

Following prior methods that used the principle of parsimony (Fitch 1971), maximum likelihood was used to infer protein ancestral states in a phylogeny. Maximum likelihood allowed for a more accurate characterization of ancient sequences with an appropriate model of sequence evolution. Just like parsimony, maximum likelihood requires a phylogenetic tree and the sequences at the tips of the tree. Unlike parsimony methods, maximum likelihood now requires a substitution matrix (which can be a mixture of models) and other evolutionary model components for proteins as well as branch lengths under the model to find the most likely ancestral sequence for nodes throughout the tree.

Using marginal likelihoods that integrated over probabilities of specific amino acids in other nodes in a tree, the probability vector could be generated for all aligned positions at each node in a tree. This is calculated from the equation below, that uses three knowns: a given mutation matrix M, a given (unrooted) evolutionary tree T, and given amino acids at the tips of the tree $\{A_r\}$. Using these three known values, we can use the equation below to find A_r, the ancestral sequences at ancestral nodes (Koshi and Goldstein 1996).

$$P\left(A_r \middle| \{A_i\}', M, T\right) = \frac{P\left(\{A_i\}' \middle| A_r, M, T\right)\left(P(A_r)\right)}{P\left(\{A_i\}' \middle| M, T\right)}$$

Using this equation, we can calculate the maximum likelihood at one particular node. To find the maximum likelihood ancestor for an arbitrary node in the tree, we select that node as the root, with application of the pulley principle in likelihood-based phylogenetics for time reversible (equilibrium) models. We then sum over all the possibilities (substitutions) from that particular ancestral node and its descendents and can do so for any internal node (Yang et al. 1995) (Fig. 1).

A subsequent methodological development involved joint reconstruction across nodes instead of the original marginal reconstruction algorithm (Pupko et al. 2000). Here, instead of maximizing the likelihood of states at an individual node while integrating over all others, all nodes are considered together in maximizing the likelihood across the tree. While joint reconstruction has not been widely used, conceptually it provides a maximum likelihood method for providing a complete evolutionary history of each site. In practice, marginal and joint reconstructions at any site give very similar sequences, although differences do occur (Pupko et al. 2007).

One early important computational application of ancestral sequence reconstruction was in finding specific episodes of positive or negative selection on various ancestral branches of a given tree. After ancestral sequences at internal nodes were generated, nonsynonymous (dN) and synonymous (dS) nucleotide differences were calculated from the inferred substitution between ancestral nodes to obtain the ratio of dN/dS (then known as K_A/K_S). Nodes with dN/dS values less than 1 show evidence for negative selection, while dN/dS values greater than one show evidence for positive selection. This application was first performed on lysosomes in primates (Messier and Stewart 1997) and in leptin and the leptin receptor's extracellular domain across mammals (Benner et al. 1998), and these analyses were able to find specific adaptive and purifying episodes localized to

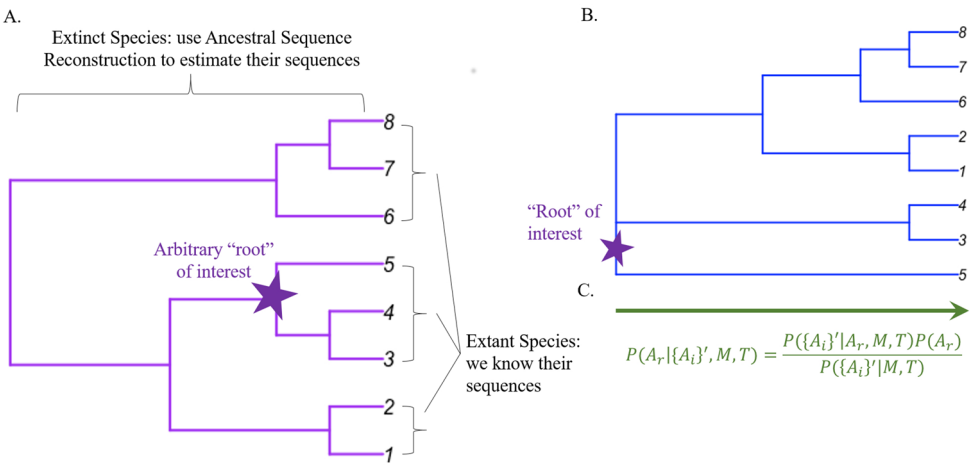

Fig. 1 The procedure for conducting model-based Ancestral Sequence Reconstruction is depicted. **a** An ancestral node to declare as the "root" of interest is selected. **b** The tree is re-rooted with this node. Based on the pulley principle for a time reversible model, any node can arbitrarily be declared the "root" without changing the likelihood. **c** Inference of the maximum likelihood ancestor is made by summing over all possible amino acid substitutions. This is done by comparing evolutionary trajectories from the ancestral sequence with extant sequences at the tips

specific nodes on the phylogenetic tree. One such episode demonstrated that leptin had evolved significantly during early primates, following the most recent common ancestor of rodents and primates– and suggested that leptin may not be functionally associated with obesity in humans as it is in mice (Benner et al. 1998). It was also demonstrated that short adaptive episodes can be masked by long-term negative selection, like in lysozyme evolution studied in primates (Messier and Stewart 1997). These methods spurred the subsequent development of maximum likelihood methods that integrated across ancestral sequence probabilities in estimating branch-specific dN/dS values (Yang 1998), methods that are still widely used today based upon the Goldman-Yang model (Goldman and Yang 1994).

Early Experimental Applications of Computational Ancestral Sequence Reconstruction

In finding bouts of positive selection, ancestral sequence reconstruction generated experimentally testable hypotheses for studying molecular evolutionary history with potential protein functional change. While other recent reviews have examined this direction more systematically (Gumulya and Gillam 2017; Liberles et al. 2020), an overview of key developments from a historical perspective is presented. The first experimental study using ASR involved the replacement of three amino acid positions in a modern lysozyme protein with inferred ancestral residues at these positions (Malcolm et al. 1990), and proceeded to dissect possible intermediate pathways for how these amino acid positions evolved under selective constraints during an episode of functional divergence. It took an additional 5 years before the first full-length ancestral sequence was inferred and generated in the laboratory (DNA synthesis technology had improved enough to allow synthesis of full-length genes). An ASR study generating 13 resurrected ribonucleases revealed episodes of functional divergence during artiodactyl evolution (Jermann et al. 1995).

The above two inaugural experimental ASR studies used parsimony to infer ancestral character states. The advancement of robust statistical approaches during the 1990s (i.e., maximum likelihood) paved the way for more sophisticated experimental studies capable of probing deeper (in time) and more divergent evolutionary questions. The first study to accomplish such a feat involved the resurrection of ancestral rhodopsin proteins from a group of archosaurs that included birds and dinosaurs, and suggested these ancestors were able to best see in dim lighting (Chang et al. 2002). The next study to achieve a similar goal involved the resurrection of proteins used to infer the environmental temperature of the last common ancestor of bacteria, inferred to have lived billions of years ago on early Earth (Gaucher et al. 2003). The third study to achieve this goal involved the resurrection of steroid receptor proteins and demonstrated that the earliest steroid receptors likely bound estrogen (Thornton et al. 2003).

This trifecta of studies opened the door for a diversity of experimental ASR studies that have spanned numerous periods of evolutionary history and have probed a plethora of biomolecular functionalities. This has all been achieved from a seed planted by Pauling and Zuckerhandl in 1963, and we anticipate that a similar level of growth will occur for experimental ASR over the next ~ 60 years.

Methodological Improvements

One of the criticisms of ancestral sequence reconstruction approaches is concerned with potential bias of the likelihood statistical framework. Maximum parsimony and maximum likelihood approaches pick the most parsimonious/likely ancestor for experimental reconstruction and this has the potential to show functional bias by excluding rare variants that are likely to be present in any individual and/or likely to be slightly deleterious. One way of adding expected rare variants to the ancestors is through Bayesian Sampling, sampling of multiple sequences from the posterior distribution (Williams et al. 2006). This framework was shown computationally to have functional effects on traits like thermostability, although such biases have been experimentally shown to have minimal effect on actual protein thermostability (Gaucher et al. 2008) or actual protein fluorescence (Alieva et al. 2008). Another experimental extension of Bayesian Sampling involved serum paraoxonases (PONs) using a library of alternative PONs. This was created to consider alternative ambiguously predicted ancestors, evaluating the effects of inclusion of this uncertainty. Through the library approach, these authors were able to find certain predictions made by maximum likelihood were very unlikely to reflect the actual ancestral sequences (Bar-Rogovsky et al. 2015).

A different approach was developed to consider the inconsistency of using different models for alignment and phylogenetics. BaliPhy (Suchard and Redelings 2006) uses the same model to simultaneously generate the tree and perform the alignments through a Markov Chain Monte Carlo method. This method is dependent upon the underlying substitution model and includes a model for insertion and deletion events. This consistency is meant to solve any problems that are caused by the ad hoc nature of models and parameters for multiple sequence alignment (Anisimova et al. 2010).

Another of the limitations of ancestral sequence reconstruction approaches is the simplicity of models used for protein evolution. Developments in the protein model also

began stepping away from the 20 × 20 amino acid matrix. This was initially done by using mixtures of substitution models (Koshi and Goldstein 1995) and models that did not assume the same mutational process for all sites in a mutation-selection framework (Halpern and Bruno 1998; Lartillot and Philippe 2004). The CAT models have been extended to include temporal shifts in amino acid fitnesses (CAT-BP) and have spawned work on the related mutation-selection models, including towards relaxing assumptions of an equilibrium process (Blanquart and Lartillot 2008; Teufel et al. 2018). Variants of the mutation-selection framework remain at the cutting edge of amino acid substitution models, but have not been widely used for ancestral sequence reconstruction yet. The strategy utilizing mixtures of substitution matrices, including while explicitly considering an attribute of protein structure (position solvent accessibility), for ancestral sequence reconstruction has recently been revisited with promising results (Moshe and Pupko 2019). Specifically, an improvement in the log-likelihood describing fit to empirical datasets was found together with the observation that the mixture of matrices resulted in major differences in inferred ancestral sequences in those datasets.

Another class of models explicitly considered the protein's tertiary structure (Robinson et al. 2003; Rodrigue et al. 2006; Kleinman et al. 2010; Grahnen et al. 2011; Arenas et al. 2017; Chi et al. 2018). Such models do not yet describe structural constraints on protein sequences well. Many biological processes affecting protein biochemistry and evolution affect selective constraints that dictate which amino acids are substituted and which are not, but these are ignored in current ancestral sequence reconstruction methods. There is ripe future ground for further model development in this direction, where has recently been reviewed (Chi and Liberles 2016). This represents a different new direction in modeling.

Overall, the state of the art of protein models has progressed from PAM-style models of increasing sophistication (Dayhoff et al. 1978; Jones et al. 1992; Whelan and Goldman 2001; Le and Gascuel 2008) to CAT models (Lartillot and Philippe 2004) to CAT models with breakpoints (Zhou et al. 2010) to mutation-selection models (Teufel et al. 2018). Breakpoints and covarion-type models enable rates to shift at a site over a tree (Wang et al. 2007). Another direction where important developments are improving our ability to model sequences is with models that combine inter-specific with intra-specific processes (Wilson et al. 2011; Hey et al. 2018). This can provide a formal mechanism to characterize segregating sequence variants that more informally were modeled with Bayesian sampling from an inter-specific model.

This discussion of modeling has mostly focused on sequence substitution. The standard likelihood-based methods didn't account for indel positions and ancestral sequences grew in length as one progressed back the tree (Pupko et al 2007). GASP (Edwards and Shields 2004) coupled model-based sequence reconstruction with parsimonious reconstruction of indel positions, treating each position independently. POY is a parsimony-based simultaneous alignment and tree method that treats both substitutions and indels using parsimony (Wheeler et al. 2015). As previously mentioned, Baliphy has a simple indel model in a likelihood framework to extend these types of models (Suchard and Redelings 2006). Two additional statistical treatments of gaps include a sequence and length-based model that generated a zipfian distribution and a fuller set of propensities of indel occurrence (Chang and Benner 2004) and an evolutionary HMM for alignment that can be employed for ancestral sequence reconstruction (Rivas and Eddy 2015). As with sequence substitution models, the future is ripe for development of integrated models for insertion and deletion, coupled to substitution that will improve our ability to reconstruct ancestral sequences. Without integrated models for insertions and deletions together with substitutions, there exists bias in current methods that has been shown to lead to too long ancestral proteins (Vialle et al. 2018). To take a different step towards reducing this bias, one method for dealing with alignment error integrates over alignments (Aadland and Kolaczkowski (2020) and this reduces the number of gapped positions. Towards the future, it is also the case that many of the most sophisticated models do not have software implementations and filling this gap will also be important in the future.

Extending Ancestral Sequence Reconstruction into Systems Biology

Most ASR studies examine individual proteins. However, differentiating between inter-molecular compensatory processes and directional selection acting on multiple proteins in a pathway requires an extension of these techniques to multiple members of a pathway (Orlenko et al. 2016b; Olson-Manning 2020). To reconstruct a pathway, one could use ancestral sequence reconstruction for every protein in an entire pathway, insert these into an organism, and measure flux. This however would be time consuming, even with newer models and methods. Another approach would be to reconstruct molecular phenotypes in different ways. The simplest phenotypic approach would be to reconstruct pathway flux as a continuous single value. However, assuming the pathway structure is conserved over the tree, individual parameters can be reconstructed independently subject to thermodynamic constraints and fit into differential equations to model pathway fluxes at particular ancestral nodes. While the ancestral reconstruction approaches are futuristic, glycolysis is one pathway where comparative analysis of kinetic

models across species has been performed (Orlenko et al. 2016a). This work established trends consistent with mutation-selection-drift balance where mutation occurs at the gene/protein level and selection occurs at the pathway level (Orlenko et al. 2016b) giving rise to gene-specific context-dependent evolution, which is increasingly well understood as a general observation in comparative data (Eguchi et al. 2019).

Extending Ancestral Sequence Reconstruction Further into Molecular Ecology

To the extent that ASR is linked to uncovering environmental adaptation associated with positive directional selection for changes in protein function, a key feature of this is the relationship of ancient organisms to their environment and ecosystem. Including archosaur vision (Chang et al. 2002, discussed above) and bacterial protein thermostability (Risso et al. 2013), insights into ancient environments and how organisms interacted with them have been gained. In the latter case, direct effects from individual proteins (Risso et al. 2013) interplay with proteome-wide effects driven by temperature on substitution (Goldstein, 2007; Zeldovich et al. 2007; Gromiha et al. 2013). The next level beyond, extending protein interaction analysis to those involving inter-specific interactions, whether host–pathogen interactions, with Dobzhansky–Muller incompatibilities during speciation, or across a food web. In these cases, ancestral sequence reconstruction can be used not only to make protein functional inference, but also to make ecological inference about what species are interacting with what other species in a community. Reconstructing parts of the proteome widely across the tree of life has the potential to identify changes in community structure and species interactions over time. This is a direction that we have only seen the tip of the iceberg.

Ancestral Sequence Reconstruction and Infectious Disease

Ancestral sequence reconstruction can be used to understand viral evolution and towards therapeutic applications (Arenas 2020). An understanding of the evolutionary histories of these viruses can lead to applications in detecting targeted regions for future therapeutics, and to assist in predicting new viral resistance against current drugs.

Ancestral sequence reconstruction is also of emerging interest for vaccine technologies, especially for the development of vaccines to combat rapidly evolving viruses such as HIV and influenza strains (Gaschen et al. 2002; Ducatez et al. 2011). Using ancestrally derived sequences to create vaccine reagents takes advantage of the evolutionary history of the virus. This strategy contrasts with other methods which construct a consensus sequence from different viral strains, ignoring phylogenetic structure. A vaccine reagent can be based on the last common ancestral sequence of all the strains that are circulating, or from other points in the tree. For example, when the phylogenetic topology is skewed, the "center of tree" method may be implemented. The center of tree method considers the ancestral sequence that minimizes the evolutionary distance between different viral strains of interest (Nickle et al. 2003).

In the age of the SARS-CoV-2, ancestral sequence reconstruction has become of immediate interest to assist in vaccine development (Zhou et al. 2020). Like the rapidly evolving RNA virus influenza and retrovirus HIV, SARS-CoV-2 is also an RNA virus. However, a recent study used ancestral sequence reconstruction to demonstrate that unlike other RNA viruses, mutations in SARS-CoV-2 are rare, as the evolution rate is slower than the transmission rate. Because of the slow evolution of SARS-CoV-2, only one vaccine candidate may be necessary to match all currently circulating SARS-CoV-2 variants (Dearlove et al. 2020).

Aside from disease causing viruses, viruses are also developed to serve as a vehicle for gene therapy (Ivics et al. 1997). The Adeno-associated Virus (AAV) has been considered an efficient gene therapy for both inherited and infectious diseases. However, the complex structure and diversity associated with different target receptor binding for AAV make the virus difficult to properly structurally assemble when designed. Using ancestral sequence reconstruction, Zinn et al. (2015) were able to provide a virus with a structure that would remain evolutionarily resilient to future mutations and maintain broad clinical applicability.

Biomedical and Biotechnological Directions for Ancient Proteins

In addition to all the insights ASR reveals about natural evolutionary processes, it turns out that ancient proteins also have applied functions in biotechnology and biomedicine (Randall et al. 2016). Ancestral variants have been used to develop clinical treatments for type 2 diabetes (Skovgaard et al. 2006), gout (Kratzer et al. 2014), hemophilia (Zakas et al. 2017), tyrosinemia (Hendrikse et al. 2020) and others. It is anticipated that this trend in biomedicine will continue as ASR generates proteins having expanded biomolecular functionalities with lower immunogenic responses in human patients compared to their modern protein counterparts. Further, ancestral variants are being used in the biotechnology sector due to their unique and desirable properties. Companies such as nanoGUNE (Manteca et al. 2017), Syngenta, New England Biolabs (Zhou et al. 2012), DuPont (Ladics

et al. 2020) and others have developed or integrated ancient proteins into their biotechnology product development pipelines, while some ancient proteins have even been tested for their value in the cosmetic industry (Perez-Jimenez et al. 2011).

The irony of ancient proteins having an applied utility to the development of therapeutics and industrial enzymes is clear. It is reasonable to expect that this utility will expand within the public and private sectors as more examples are discovered in the coming years. Sometimes one must explore the past in order to navigate the future.

Concluding Thoughts

Starting with the vision of Journal of Molecular Evolution founding editor Emile Zuckerkandl together with Linus Pauling, through the maximum likelihood method of Felsenstein to the application of this method to protein ancestral sequences by Koshi and Goldstein, Journal of Molecular Evolution has been an important home for the development of the field. As models and statistical frameworks for characterizing protein evolution over a phylogenetic tree continue to improve, these developments will continue to impact the field of ancestral sequence reconstruction, with downstream applications in fields as disparate as biomedicine and community ecology.

Acknowledgements Avery Selberg receives partial support from the National Science Foundation under Grant Number 2037635 awarded to David Liberles.

References

Aadland K, Kolaczkowski B (2020) Alignment-integrated reconstruction of ancestral sequences improves accuracy. Genome Biol Evol 12:1549–1565. https://doi.org/10.1093/gbe/evaa164

Alieva NO, Konzen KA, Field SF et al (2008) Diversity and evolution of coral fluorescent proteins. PLoS ONE 3:e2680. https://doi.org/10.1371/journal.pone.0002680

Anisimova M, Cannarozzi G, Liberles DA (2010) Finding the balance between the mathematical and biological optima in multiple sequence alignment. Trends Evol Biol 2:e7. https://doi.org/10.4081/eb.2010.e7

Arenas M (2020) Protein evolution in the flaviviruses. J Mol Evol 88:473–476. https://doi.org/10.1007/s00239-020-09953-1

Arenas M, Weber CC, Liberles DA, Bastolla U (2017) ProtASR: an evolutionary framework for ancestral protein reconstruction with selection on folding stability. Syst Biol 66:1054–1064. https://doi.org/10.1093/sysbio/syw121

Bar-Rogovsky H, Stern A, Penn O et al (2015) Assessing the prediction fidelity of ancestral reconstruction by a library approach. Protein Eng Des Sel 28:507–518. https://doi.org/10.1093/protein/gzv038

Benner SA, Trabesinger N, Schreiber D (1998) Post-genomic science: converting primary structure into physiological function. Adv Enzym Regul 38:155–180. https://doi.org/10.1016/s0065-2571(97)00019-8

Blanquart S, Lartillot N (2008) A site- and time-heterogeneous model of amino acid replacement. Mol Biol Evol 25:842–858. https://doi.org/10.1093/molbev/msn018

Chang MSS, Benner SA (2004) Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments. J Mol Biol 341:617–631. https://doi.org/10.1016/j.jmb.2004.05.045

Chang BSW, Jönsson K, Kazmi MA et al (2002) Recreating a functional ancestral archosaur visual pigment. Mol Biol Evol 19:1483–1489. https://doi.org/10.1093/oxfordjournals.molbev.a004211

Chi PB, Liberles DA (2016) Selection on protein structure, interaction, and sequence. Protein Sci 25:1168–1178. https://doi.org/10.1002/pro.2886

Chi PB, Kim D, Lai JK, Bykova N, Weber CC, Kubelka J, Liberles DA (2018) A new parameter-rich structure-aware mechanistic model for amino acid substitution during evolution. Proteins: Struct Funct Bioinform 86:218–228. https://doi.org/10.1002/prot.25429

Dayhoff M, Schwartz R, Orcutt B (1978) 22 a model of evolutionary change in proteins. Atlas Protein Seq Struct 5:345–352

Dearlove B, Lewitus E, Bai H et al (2020) A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. Proc Natl Acad Sci USA 117:23652–23662. https://doi.org/10.1073/pnas.2008281117

Ducatez MF, Bahl J, Griffin Y et al (2011) Feasibility of reconstructed ancestral H5N1 influenza viruses for cross-clade protective vaccine development. Proc Natl Acad Sci USA 108:349–354. https://doi.org/10.1073/pnas.1012457108

Edwards RJ, Shields DC (2004) GASP: gapped ancestral sequence prediction for proteins. BMC Bioinform 5:123. https://doi.org/10.1186/1471-2105-5-123

Eguchi Y, Bilolikar G, Geiler-Samerotte K (2019) Why and how to study genetic changes with context-dependent effects. Curr Opin Genet Dev 58–59:95–102. https://doi.org/10.1016/j.gde.2019.08.003

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376. https://doi.org/10.1007/BF01734359

Fitch WM (1971) Toward defining the course of evolution: minimum change for a specific tree topology. Syst Biol 20:406–416. https://doi.org/10.1093/sysbio/20.4.406

Gaschen B, Taylor J, Yusim K et al (2002) Diversity considerations in HIV-1 vaccine selection. Science 296:2354–2360. https://doi.org/10.1126/science.1070441

Gaucher EA, Govindarajan S, Ganesh OK (2008) Palaeotemperature trend for Precambrian life inferred from resurrected proteins. Nature 451:704–707. https://doi.org/10.1038/nature06510

Gaucher EA, Thomson JM, Burgan MF, Benner SA (2003) Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. Nature 425:285–288. https://doi.org/10.1038/nature01977

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725–736. https://doi.org/10.1093/oxfordjournals.molbev.a040153

Goldstein RA (2007) Amino-acid interactions in psychrophiles, mesophiles, thermophiles, and hyperthermophiles: insights from the quasi-chemical approximation. Protein Sci 16(9):1887–1895

Grahnen JA, Nandakumar P, Kubelka J, Liberles DA (2011) Biophysical and structural considerations for protein sequence evolution. BMC Evol Biol 11:361. https://doi.org/10.1186/1471-2148-11-361

Gromiha MM, Pathak MC, Saraboji K, Ortlund EA, Gaucher EA (2013) Hydrophobic environment is a key factor for the stability of thermophilic proteins. Proteins 81:715–721. https://doi.org/10.1002/prot.24232

Gumulya Y, Gillam EMJ (2017) Exploring the past and the future of protein evolution with ancestral sequence reconstruction: the "retro"approach to protein engineering. Biochem J 474:1–19

Halpern AL, Bruno WJ (1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol Biol Evol 15:910–917. https://doi.org/10.1093/oxfordjournals.molbev.a025995

Hendrikse NM, Holmberg Larsson A, Svensson Gelius S et al (2020) Exploring the therapeutic potential of modern and ancestral phenylalanine/tyrosine ammonia-lyases as supplementary treatment of hereditary tyrosinemia. Sci Rep 10:1315. https://doi.org/10.1038/s41598-020-57913-y

Hey J, Chung Y, Sethuraman A et al (2018) Phylogeny estimation by integration over isolation with migration models. Mol Biol Evol 35:2805–2818. https://doi.org/10.1093/molbev/msy162

Ivics Z, Hackett PB, Plasterk RH, Izsvák Z (1997) Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. Cell 91:501–510. https://doi.org/10.1016/s0092-8674(00)80436-5

Jermann TM, Opitz JG, Stackhouse J, Benner SA (1995) Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. Nature 374:57–59. https://doi.org/10.1038/374057a0

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275–282. https://doi.org/10.1093/bioinformatics/8.3.275

Kleinman CL, Rodrigue N, Lartillot N, Philippe H (2010) Statistical potentials for improved structurally constrained evolutionary models. Mol Biol Evol 27:1546–1560. https://doi.org/10.1093/molbev/msq047

Koshi JM, Goldstein RA (1995) Context-dependent optimal substitution matrices. Protein Eng 8:641–645. https://doi.org/10.1093/protein/8.7.641

Koshi JM, Goldstein RA (1996) Probabilistic reconstruction of ancestral protein sequences. J Mol Evol 42:313–320. https://doi.org/10.1007/BF02198858

Kratzer JT, Lanaspa MA, Murphy MN et al (2014) Evolutionary history and metabolic insights of ancient mammalian uricases. Proc Natl Acad Sci USA 111:3763–3768. https://doi.org/10.1073/pnas.1320393111

Ladics GS, Han K-H, Jang MS et al (2020) Safety evaluation of a novel variant of consensus bacterial phytase. Toxicol Rep 7:844–851. https://doi.org/10.1016/j.toxrep.2020.07.004

Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095–1109. https://doi.org/10.1093/molbev/msh112

Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. Mol Biol Evol 25:1307–1320. https://doi.org/10.1093/molbev/msn067

Liberles DA, Chang B, Geiler-Samerotte K et al (2020) Emerging frontiers in the study of molecular evolution. J Mol Evol 88:211–226. https://doi.org/10.1007/s00239-020-09932-6

Malcolm BA, Wilson KP, Matthews BW et al (1990) Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. Nature 345:86–89. https://doi.org/10.1038/345086a0

Manteca A, Schönfelder J, Alonso-Caballero A et al (2017) Mechanochemical evolution of the giant muscle protein titin as inferred from resurrected proteins. Nat Struct Mol Biol 24:652–657. https://doi.org/10.1038/nsmb.3426

Messier W, Stewart CB (1997) Episodic adaptive evolution of primate lysozymes. Nature 385:151–154. https://doi.org/10.1038/385151a0

Moshe A, Pupko T (2019) Ancestral sequence reconstruction: accounting for structural information by averaging over replacement matrices. Bioinformatics 35:2562–2568. https://doi.org/10.1093/bioinformatics/bty1031

Nickle DC, Jensen MA, Gottlieb GS et al (2003) Consensus and ancestral state HIV vaccines. Science 299:1515–1518. https://doi.org/10.1126/science.299.5612.1515c

Olson-Manning CF (2020) Elaboration of the corticosteroid synthesis pathway in primates through a multistep enzyme. Mol Biol Evol 37:2257–2267. https://doi.org/10.1093/molbev/msaa080

Orlenko A, Hermansen RA, Liberles DA (2016a) Flux control in glycolysis varies across the tree of life. J Mol Evol 82:146–161. https://doi.org/10.1007/s00239-016-9731-2

Orlenko A, Teufel AI, Chi PB, Liberles DA (2016b) Selection on metabolic pathway function in the presence of mutation-selection-drift balance leads to rate-limiting steps that are not evolutionarily stable. Biol Direct 11:31. https://doi.org/10.1186/s13062-016-0133-6

Pauling L, Zuckerkandl E, Henriksen T, Lövstad R (1963) Chemical paleogenetics. Acta Chem Scand 17:S9–S16

Perez-Jimenez R, Inglés-Prieto A, Zhao Z-M et al (2011) Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. Nat Struct Mol Biol 18:592–596. https://doi.org/10.1038/nsmb.2020

Pupko T, Doron-Faigenboim A, Liberles DA, Cannarozzi G (2007) Probabilistic models and their impact on the accuracy of reconstructed ancestral protein sequences. In: Liberles DA (ed) Ancestral sequence reconstruction. Oxford University Press, Oxford, pp 47–51. https://doi.org/10.1093/acprof:oso/9780199299188.003.0004

Pupko T, Pe'er I, Shamir R, Graur D (2000) A fast algorithm for joint reconstruction of ancestral amino acid sequences. Mol Biol Evol 17:890–896. https://doi.org/10.1093/oxfordjournals.molbev.a026369

Randall R, Radford C, Roof K et al (2016) An experimental phylogeny to benchmark ancestral sequence reconstruction. Nat Commun 7:12847. https://doi.org/10.1038/ncomms12847

Risso VA, Gavira JA, Mejia-Carmona DF, Gaucher EA, Sanchez-Ruiz JM (2013) Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian β-lactamases. J Am Chem Soc 135:2899–2902. https://doi.org/10.1021/ja311630a

Rivas E, Eddy SR (2015) Parameterizing sequence alignment with an explicit evolutionary model. BMC Bioinform 16:406. https://doi.org/10.1186/s12859-015-0832-5

Robinson DM, Jones DT, Kishino H et al (2003) Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol 20:1692–1704. https://doi.org/10.1093/molbev/msg184

Rodrigue N, Philippe H, Lartillot N (2006) Assessing site-interdependent phylogenetic models of sequence evolution. Mol Biol Evol 23:1762–1775. https://doi.org/10.1093/molbev/msl041

Skovgaard M, Kodra JT, Gram DX et al (2006) Using evolutionary information and ancestral sequences to understand the sequence-function relationship in GLP-1 agonists. J Mol Biol 363:977–988. https://doi.org/10.1016/j.jmb.2006.08.066

Suchard MA, Redelings BD (2006) BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. Bioinformatics 22:2047–2048. https://doi.org/10.1093/bioinformatics/btl175

Teufel AI, Ritchie AM, Wilke CO, Liberles DA (2018) Using the mutation-selection framework to characterize selection on protein sequences. Genes 9:409. https://doi.org/10.3390/genes9080409

Thornton JW, Need E, Crews D (2003) Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. Science 301:1714–1717. https://doi.org/10.1126/science.1086185

Vialle RA, Tamuri AU, Goldman N (2018) Alignment modulates ancestral sequence reconstruction accuracy. Mol Biol Evol 35:1783–1797. https://doi.org/10.1093/molbev/msy055

Wang H-C, Spencer M, Susko E, Roger AJ (2007) Testing for covarion-like evolution in protein sequences. Mol Biol Evol 24:294–305. https://doi.org/10.1093/molbev/msl155

Wheeler WC, Lucaroni N, Hong L et al (2015) POY version 5: phylogenetic analysis using dynamic homologies under multiple optimality criteria. Cladistics 31:189–196. https://doi.org/10.1111/cla.12083

Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691–699. https://doi.org/10.1093/oxfordjournals.molbev.a003851

Williams PD, Pollock DD, Blackburne BP, Goldstein RA (2006) Assessing the accuracy of ancestral protein reconstruction methods. PLoS Comput Biol 2:e69. https://doi.org/10.1371/journal.pcbi.0020069

Wilson DJ, Hernandez RD, Andolfatto P, Przeworski M (2011) A population genetics-phylogenetics approach to inferring natural selection in coding sequences. PLoS Genet 7:e1002395. https://doi.org/10.1371/journal.pgen.1002395

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568–573. https://doi.org/10.1093/oxfordjournals.molbev.a025957

Yang Z, Kumar S, Nei M (1995) A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141:1641–1650

Zakas PM, Brown HC, Knight K et al (2017) Enhancing the pharmaceutical properties of protein drugs by ancestral sequence reconstruction. Nat Biotechnol 35:35–37. https://doi.org/10.1038/nbt.3677

Zeldovich KB, Berezovsky IN, Shakhnovich EI (2007) Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput Biol 3:e5

Zhou Y, Brinkmann H, Rodrigue N et al (2010) A Dirichlet process covarion mixture model and its assessments using posterior predictive discrepancy tests. Mol Biol Evol 27:371–384

Zhou Y, Asahara H, Gaucher EA, Chong S (2012) Reconstitution of translation from *Thermus thermophilus* reveals a minimal set of components sufficient for protein synthesis at high temperatures and functional conservation of modern and ancient translation components. Nucleic Acids Res 40:7932–7945

Zhou P, Yang X-L, Wang X-G et al (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579:270–273. https://doi.org/10.1038/s41586-020-2012-7

Zinn E, Pacouret S, Khaychuk V et al (2015) In silico reconstruction of the viral evolutionary lineage yields a potent gene therapy vector. Cell Rep 12:1056–1068. https://doi.org/10.1016/j.celrep.2015.07.019