OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Review

# Bringing Molecules Back into Molecular Evolution

**Claus O. Wilke***

Section of Integrative Biology, Center for Computational Biology and Bioinformatics, and Institute of Cell and Molecular Biology, The University of Texas at Austin, Austin, Texas, United States of America

**Abstract:** Much molecular-evolution research is concerned with sequence analysis. Yet these sequences represent real, three-dimensional molecules with complex structure and function. Here I highlight a growing trend in the field to incorporate molecular structure and function into computational molecular-evolution work. I consider three focus areas: reconstruction and analysis of past evolutionary events, such as phylogenetic inference or methods to infer selection pressures; development of toy models and simulations to identify fundamental principles of molecular evolution; and atom-level, highly realistic computational modeling of molecular structure and function aimed at making predictions about possible future evolutionary events.

This is an "Editors' Outlook" article for *PLoS Computational Biology*.

## Introduction

The field of molecular evolution investigates how genes and genomes evolve over time. It has its origin in the late 1960s, when the first DNA and protein sequences were becoming available. With rapid progress in sequencing technologies came ever increasing demand for computational tools to study molecular evolution. Today, molecular evolution is among the largest subfields of evolutionary biology, and arguably one of the most computationally advanced. Thousands of person years have gone into developing sophisticated alignment algorithms, phylogenetic-tree reconstruction methods, or statistical tests for positive selection.

A side effect of the strong emphasis on developing sophisticated methods for sequence analysis has been that the underlying biophysical objects represented by the sequences, DNA molecules, RNA molecules, and proteins, have taken a back-seat in much computational molecular-evolution work. The vast majority of algorithms for sequence analysis, for example, incorporate no knowledge of biology or biochemistry besides that DNA and RNA sequences use an alphabet of four letters, protein sequences use an alphabet of 20, and the genetic code converts one into the other. The choice to treat DNA, RNA, and proteins simply as strings of letters was certainly reasonable in the late 20th century. Computational power was limited and many basic aspects of sequence analysis were still relatively poorly understood. However, in 2012 we have extremely powerful computers and a large array of highly sophisticated algorithms that can analyze strings of letters. It is now time to bring the molecules back into molecular evolution. Several groups have embarked on this path, and I will highlight some of the work that has been done and speculate on future developments we may see.

In this article, I focus on the evolution of protein-coding genes, the area I am most familiar with myself. However, my overall message, that it is time to bring the molecules back into molecular evolution, similarly applies to other genetic sequences, such as intergenic regions, RNA genes, or the various forms of short RNAs. I will consider three broad areas, corresponding to three distinct research goals: (i) reconstructing and interpreting past evolutionary events; (ii) identifying fundamental principles of molecular evolution; and (iii) predicting probable evolutionary trajectories.

## Reconstructing and Interpreting Past Evolutionary Events

A major goal of comparative sequence analysis is to reconstruct and/or interpret past evolutionary events. For example, we may have sequences from multiple species and want to know how they relate to each other, which specific sequence changes caused them to diverge, and whether certain sites were under particularly strong selective pressure. The standard analysis pipeline for such questions is to align sequences, build trees, and run scans for positive or other types of selection, and/or for recombination. This analysis pipeline uses nothing but sequences as input. Only once the analysis is completed may the researchers take sites of interest they have identified, map them back onto the structure of the protein they are studying, and carry out further experimentation. (However, increasingly the initial sequence analysis is only the prerequisite for a successful study, and the value of the study is defined by the follow-up work; see e.g., [1,2].)

The standard analysis approach has been highly successful. Yet it ignores most of the biochemistry that ultimately determines the fitness landscape in which sequences evolve. Thus, methods that combine sequence data with additional information, such as protein structure, should yield more sensitive and more accurate estimates than methods based on sequence data alone. On the basis of this premise, a few groups have started to develop such methods. For example, some authors have developed models of coding-sequence evolution that incorporate interactions among sites mediated by protein structure [3–5]. (See also this review:

**Citation:** Wilke CO (2012) Bringing Molecules Back into Molecular Evolution. PLoS Comput Biol 8(6): e1002572. doi:10.1371/journal.pcbi.1002572

**Editor:** Sergei L. Kosakovsky Pond, University of California San Diego, United States of America

**Published** June 28, 2012



**Funding:** Funding was received from NIH grant R01 GM088344 and NSF Cooperative Agreement No. DBI-0939454. The funders had no role in the preparation of the manuscript.

**Competing Interests:** The author has declared that no competing interests exist.

* E-mail: wilke@austin.utexas.edu

[6].) Similarly, some authors have incorporated knowledge of protein structure in methods of ancestral state reconstruction [7]. Finally, in phylogenetic-tree inference, evidence is accumulating that independence of sites may not be a good assumption [8] for protein-coding and even more so for RNA-coding sequences. Thus, future methods of phylogenetic tree reconstruction may also incorporate structural information in some form. Coarse-grained models of protein-sequence evolution are being developed that may be useful for this purpose [9].

The development of methods that integrate molecular structure into sequence analysis is still in its infancy. While several groups are exploring a variety of approaches, none of these approaches is well established at this time. Comparative analyses that use nothing but sequence data remain state of the art. My expectation for the near future is that we will continue to see efforts to extend comparative analyses beyond sequence data alone. Eventually, some of these efforts will prove sufficiently useful that it will become commonplace to combine sequence data with structural, functional, or other molecular data in comparative analyses.

## Identifying Fundamental Principles of Molecular Evolution

Besides understanding and interpreting specific evolutionary events, evolutionary biologists also aim to identify fundamental principles of molecular evolution. Fundamental principles are insights that apply to many different biological systems; a classical example would be the finding that codon usage bias correlates with gene expression level [10,11].

The search for fundamental principles tends to require somewhat different computational approaches than the analysis of past evolutionary events. It often involves developing toy models (either in the form of mathematical equations or of simulations) to explore possible system dynamics under different modeling assumptions or parameter choices. The specific toy models to be explored are usually inspired by observations from past evolutionary events. To give an example from my own research, starting about 10 years ago many groups found that highly expressed proteins evolve slowly [12]. This observation prompted several authors to develop models of varying complexity that might explain the pattern [13–17].

Toy models of evolution have been studied for over a century. And much of this work has not considered the underlying biochemistry of the evolving organism. For example, the population-genetics literature contains plenty of abstract, mathematical models (such as two-locus, two-allele models) that make absolutely no assumptions about the mechanisms that connect different allelic states with different fitness values. These abstract mathematical models are valuable, of course, yet they can get us only so far. Most importantly, they cannot explain how, mechanistically, genotype maps to phenotype and fitness.

As we try to get a better understanding of the genotype-phenotype map, we have to build more realistic models. For example, virtually all the models trying to explain the relationship between evolutionary rate and expression level make concrete assumptions about mechanisms of protein folding and function [13–17]. Many implement an actual (though simplified) protein-folding model in which actual amino-acid sequences are computationally folded, using either a lattice [14,15] or an off-lattice [16] approach.

More generally, we are seeing an increased trend towards integrating some biophysical realism into toy models of evolution. Several groups are regularly working with models that incorporate some aspect of protein biochemistry, such as protein fold stability [9,14–16,18–22] or protein–protein interactions [22–24]. Models may represent individual evolving proteins [18,20,21,23] or entire cells [19,22,24]. Finally, some groups elucidate the molecular fitness landscapes that underlie adaptive events [25–27]. Works such as these aim to identify the biophysical mechanisms that drive molecular evolution.

I believe that we have only scratched the surface of what is possible with simple, biophysically inspired models of molecular evolution. I expect that we are going to see more of this modeling approach in the coming years, and that it will help us to develop a deeper understanding of fundamental principles of molecular evolution.

## Predicting Probable Evolutionary Trajectories

For many real-world applications, it would be useful to be able to predict future evolutionary events. For example, we know that H5N1 avian influenza could potentially cause a deadly pandemic if it ever evolved the ability to effectively spread between humans. What we do not know [28] is the likelihood that it will evolve this ability, nor whether it might possibly become less pathogenic as it evolves more effective human-to-human transmission capabilities. As a second example, some authors have proposed treating infectious diseases with interfering particles (e.g., [29]). Because of the potential for transmission of these particles among infected patients, the safety of such treatments stands and falls with our ability to accurately predict how such therapeutic particles might evolve once released.

Since evolution is a stochastic process, we cannot expect to ever predict which specific mutations will accumulate in a given lineage. However, at least in principle, we should be able to make probabilistic predictions of the form "Outcome A is the most likely, and has a 37% probability of occurring; outcome B is the second most likely, and has a 24% probability of occurring." It would be tremendously useful if we could make such predictions reliably, in particular for rapidly evolving pathogens. Therefore, there is growing interest among evolutionary biologists to develop predictive frameworks [30–33]. In my opinion, successful approaches in this area will most likely involve realistic, atom-level computational modeling of the system of interest.

With rapid increases in computational power over the last two decades, realistic modeling of biological systems is becoming increasingly feasible. At the molecular level, obvious applications of realistic modeling are atom-level predictions of protein structure [34] or protein-folding dynamics [35,36], and computational enzyme design [37–39]. The accuracy of these computational models, when they work, is getting quite good. For example, in computational enzyme design, where the goal is to design catalytically active enzymes de novo, crystal structures of successfully designed enzymes are often very close to the computationally predicted ones [39]. However, it is common that only a small number of the computational designs actually work as expected. In a recent study, 84 computationally designed enzymes were evaluated experimentally [39]. Of those, 50 were soluble and only two catalyzed the desired reaction.

At present, atom-level modeling of proteins is not commonly used in applications of evolutionary biology (but see [40]). However, it seems to me that as our modeling capability improves, a logical next step will be to apply these models to predicting evolution. If we can predict computationally which mutants will be able to carry out specific functions, then we should also be able to predict which mutants are likely to arise under specific, well-defined selection pressures. While I cannot imagine that we will ever be able to solve open-ended problems, such as, for example,

to predict all the sequence changes an invasive species will undergo as it is introduced into a new environment, we should have reasonable success for well-defined problems, such as to find the mutations an animal virus would require to bind to the human form of the receptor it uses for cell entry in its host species.

An alternative to atom-level modeling can be statistical inference of biophysically important sites from large sequence alignments. For example, in a recent paper Bloom and Glassman [41] proposed a method to infer the effect of point mutations on protein stability from the distribution of mutations in a dense phylogeny. This method performed better in predicting measured $\Delta\Delta G$ values than alternative methods based on protein structure and atomic force fields. Bloom and co-workers then used this method to identify mutations that were likely to be involved in the evolution of oseltamivir resistance in influenza [42].

Regardless of whether one uses atom-level modeling or statistical approaches, computational predictions are not going to be perfect. Thus, computational methods to predict evolution are most likely going to be useful in generating candidate scenarios. These candidate scenarios will include many false positives and will have to be screened experimentally to separate false from true positives.

## Author's Biography

**Claus O. Wilke** is an Associate Professor in the Section of Integrative Biology at The University of Texas at Austin. He is also a member of the Center for Computational Biology and Bioinformatics and the Institute for Cell and Molecular Biology at The University of Texas at Austin. Dr. Wilke received his PhD in theoretical physics from the Ruhr-University Bochum, Germany, in 1999. From 2000 to 2004, he was a postdoc at the California Institute of Technology, working on viral evolution and artificial life. After his postdoc, he spent a year as a Research Assistant Professor at the Keck Graduate Institute of Applied Life Sciences, Claremont, before joining The University of Texas in the fall of 2005. His current research interests are in molecular evolution, structural biology, and biostatistics. Dr. Wilke is the author of approximately 100 scientific papers. He serves as Associate Editor for *PLoS Computational Biology* and *PLoS Pathogens*, and as Section Editor for *BMC Evolutionary Biology*. In 2011, Dr. Wilke was recognized as a Leading Texas Innovator by The Academy of Medicine, Engineering, and Science of Texas.

## Summary

There is a growing trend in widely differing subfields of molecular evolution to increase biophysical realism in computational models of sequence evolution. Some subfields are further along this path than others. Among groups developing simple toy models of evolution, models incorporating some biophysical realism have been quite popular in recent years. By contrast, statistical models of sequence evolution incorporating biophysical realism are being developed by some groups but are not being routinely applied in sequence-analysis applications. A major impediment to more routine use of such models is likely the lack of widely available, easy-to-use implementations. Hopefully, we will see progress in this area soon. Methods to predict future evolutionary trajectories do not really exist at this time. However, there is a growing interest in developing them. I believe that the computational methods required for this type of prediction are falling into place in the protein-design field; we may soon see a first, small-scale demonstration that computational prediction of evolutionary trajectories is actually possible.

## References


1. Sawyer SL, Wu LI, Emerman M, Malik HS (2005) Positive selection of primate TRIM5α identifies a critical species-specific retroviral restriction domain. Proc Natl Acad Sci U S A 102: 2832–2837.
2. Balakirev ES, Anisimova M, Ayala FJ (2011) Complex interplay of evolutionary forces in the *ladybird* homeobox genes of *Drosophila melanogaster*. PLoS ONE 6: e22613. doi:10.1371/journal.pone.0022613
3. Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL (2003) Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol 20: 1692–1704.
4. Lartillot NN, Bryant D, Philippe H (2005) Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene 347: 207–217.
5. Rodrigue N, Kleinman CL, Philippe H, Lartillot N (2009) Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons. Mol Biol Evol 26: 1663–1676.
6. Rodrigue N, Philippe H (2010) Mechanistic revisions of phenomenological modeling strategies in molecular evolution. Trends Genet 26: 248–252.
7. Choi SC, Stone EA, Kishino H, Thorne JL (2009) Estimates of natural selection due to protein tertiary structure inform the ancestry of biallelic loci. Gene 441: 45–52.
8. Nasrallah CA, Mathews DH, Huelsenbeck JP (2011) Quantifying the impact of dependent evolution among sites in phylogenetic inference. Syst Biol 60: 60–73.
9. Grahnen JA, Nandakumar P, Kubelka J, Liberles DA (2011) Biophysical and structural considerations for protein sequence evolution. BMC Evol Biol 11: 361.
10. Gouy M, Gautier C (1982) Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res 10: 7055–7074.
11. Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. J Mol Evol 34: 280–291.
12. Pál C, Papp B, Hurst LD (2001) Highly expressed genes in yeast evolve slowly. Genetics 158: 927–931.
13. Wilke CO, Drummond DA (2006) Population genetics of translational robustness. Genetics 173: 473–481.
14. Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134: 341–352.
15. Yang JR, Zhuang SM, Zhang J (2010) Impact of translational error-induced and error-free misfolding on the rate of protein evolution. Mol Syst Biol 6: 421.
16. Lobkovsky AE, Wolf YI, Koonin EV (2010) Universal distribution of protein evolution rates as a consequence of protein folding physics. Proc Natl Acad Sci U S A 107: 2983–2988.
17. Cherry JL (2010) Expression level, evolutionary rate, and the cost of expression. Genome Biol Evol 2: 757–769.
18. Williams PD, Pollock DD, Blackburne BP, Goldstein RA (2006) Assessing the accuracy of ancestral protein reconstruction methods. PLoS Comput Biol 2: e69. doi:10.1371/journal.pcbi.0020069
19. Zeldovich KB, Chen PQ, Shakhnovich EI (2007) Protein stability imposes limits on organism complexity and speed of molecular evolution. Proc Natl Acad Sci U S A 104: 16152–16157.
20. Bloom JD, Raval A, Wilke CO (2007) Thermodynamics of neutral protein evolution. Genetics 175: 255–266.
21. Mendez R, Fritsche M, Porto M, Bastolla U (2010) Mutation bias favors protein folding stability in the evolution of small populations. PLoS Comput Biol 6: e1000767. doi:10.1371/journal.pcbi.1000767
22. Heo M, Shakhnovich EI (2010) Interplay between pleiotropy and secondary selection determines rise and fall of mutators in stress response. PLoS Comput Biol 6: e1000710. doi:10.1371/journal.pcbi.1000710
23. Bloom JD, Wilke CO, Arnold FH (2004) Stability and the evolvability of function in a model protein. Biophys J 86: 2758–2764.
24. Zhang J, Maslov S, Shakhnovich EI (2008) Constraints imposed by non-functional protein-protein interactions on gene expression and proteome size. Mol Syst Biol 4: 210.
25. Wagner A (2008) Neutralism and selectionism: a network-based reconciliation. Nat Rev Genet 9: 965–974.
26. Ferrada E, Wagner A (2008) Protein robustness promotes evolutionary innovations on large evolutionary time-scales. Proc R Soc B 275: 1595–1602.
27. Rajon E, Masel J (2011) Evolution of molecular error rates and the consequences for evolvability. Proc Natl Acad Sci U S A 108: 1082–1087.
28. Perdue ML, Swayne DE (2005) Public health risk from avian influenza viruses. Avian Dis 49: 317–327.
29. Weinberger LS, Schaffer DV, Arkin AP (2003) Theoretical design of a gene therapy to prevent AIDS but not human immunodeficiency virus type 1 infection. J Virol 77: 10028–10036.

30. Martnez JL, Baquero F, Andersson DI (2007). Predicting antibiotic resistance. Nature Rev Microbiol 5: 958–965.
31. Deforche K, Camacho R, Van Laethem K, Lemey P, Rambaut A, et al. (2007) Estimation of an in vivo fitness landscape experienced by HIV-1 under drug selective pressure useful for prediction of drug resistance evolution during treatment. Bioinformatics 24: 34–41.
32. Bull JJ, Molineux IJ (2008) Predicting evolution from genomics: experimental evolution of bacteriophage T7. Heredity 100: 453–463.
33. Papp B, Notebaart RA, Pál C (2011) Systems-biology approaches for predicting genomic evolution. Nat Rev Genet 12: 591–602.
34. Qian B, Raman S, Das R, Bradley P, McCoy AJ, et al. (2007) High-resolution structure prediction and the crystallographic phase problem. Nature 450: 259–264.
35. Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, et al. (2010) Atomic-level characterization of the structural dynamics of proteins. Science 330: 341–346.
36. Voelz VA, Bowman GR, Beauchamp K, Pande VS (2010) Molecular simulation of *ab initio* protein folding for a millisecond folder NTL9(139). J Am Chem Soc 132: 1526–1528.
37. Röthlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, et al. (2008) Kemp elimination catalysts by computational enzyme design. Nature 453: 190–195.
38. Jiang L, Althoff EA, Clemente FR, Doyle L, Röthlisberger D, et al. (2008) De novo computational design of retro-aldol enzymes. Science 319: 1387–1391.
39. Siegel JB, Zanghellini A, Lovick HM, Kiss G, Lambert AR, et al. (2010) Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. Science 329: 309–313.
40. Kasson PM, Ensign DL, Pande VS (2009) Combining molecular dynamics with Bayesian analysis to predict and evaluate ligand-binding mutations in influenza hemagglutinin. J Am Chem Soc 31: 11338–11340.
41. Bloom JD, Glassman MJ (2009) Inferring stabilizing mutations from protein phylogenies: application to influenza hemagglutinin. PLoS Comput Biol 5: e1000349. doi:10.1371/journal.pcbi.1000349
42. Bloom JD, Gong LI, Baltimore D (2010) Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. Science 328: 1272–1275.