CHAPTER 1

The early days of paleogenetics: connecting molecules to the planet

Steven A. Benner

1.1 Introduction

Anyone asked to write about the early days feels elderly. Fortunately, Emile Zuckerkandl's introduction shows that the ideas that led to this volume have been around for some time, at least in their basic form, and are rooted in ideas of many heroes of modern molecular biology, including Pauling, Anfinsen, and Zuckerkandl himself.

In 1980, my laboratory was unaware of the Pauling–Zuckerkandl paper (Pauling and Zuckerkandl, 1963; see Chapter 2 in this volume for a fuller discussion of the implications of this paper) when we set out to resurrect ancient proteins from extinct organisms. My group, then consisting of only Krishnan Nambiar and Joseph Stackhouse, was trained to describe the chemical structures and behaviors of enzymes. In those days technology was allowing molecular scientists to extend these descriptions to atomic resolution, the picosecond time scale, and the microscopic rate constant.

But what good were clever experiments to determine, for example, which of two hydrogens was removed by a dehydrogenase (Allemann *et al.*, 1988), or whether the replacement of carbon dioxide by a proton on acetoacetate proceeded with retention or inversion of stereochemical configuration (Benner *et al.*, 1981)? It occurred to us that we might be doing the biochemical equivalent of studying a Picasso with an electron microscope. Were we not describing biomolecular systems to resolutions far greater than they were designed? Biomolecules are not designed, however. They are the products of natural selection imposed upon random variation in their chemical structures. As the result of a combination of historical accident, selective pressures, and vestigiality, all constrained by physical and chemical law, different behaviors must be interesting at different levels. Biomolecular behaviors that influenced the ability of a host organism to survive, mate, and reproduce were especially interesting, as these had been fashioned by natural selection. Behaviors that did not, were not, because they had not. As a criterion for selecting interesting chemical features of a biomolecule to study in detail, an understanding of the relation between biomolecular structure and behavior and fitness was important.

It did not take long at Harvard to realize that this relation was going to be difficult to understand. There, Martin Kreitman, Robert Dorit, and others, including some very dialectical biologists (Levins and Lewontin, 1985), were struggling to make this connection starting from the side of biology (Kreitman and Akashi, 1995). Despite this interest, it was proving difficult to connect *any* biomolecular structure or behavior with the survival of an organism, at least in a way that would be compelling to those who chose to deny it (Lewontin, 1974; Clarke, 1975; Gillespie, 1984, 1991; Somero, 1995; Powers and Schulte, 1998). In fact, the discussion was central to the most hotly disputed dispute in molecular evolution, between neutralists and selectionists, where both sides of the dialectic were populated by individuals who were professionally intent on showing how any data interpretable in favor of one side could equally well support the other.

As chemists, we had no part in this fight. However, a review of the contending sides of these disputes (Benner and Ellington, 1988) reminded us of analogous disputes in organic chemistry.

These were often Seinfeld arguments about nothing. For example, chemists had for years discussed the non-classical carbocation problem (Brown, 1977). This was a disagreement about whether the structures of positively charged organic molecules, in general, were better modeled by a formula with dotted lines, or by two formulas without dotted lines. Rational observers realized that one model was undoubtedly better for some molecules, whereas the other was better for others. After all, similar issues had been addressed and resolved in many molecular systems. For example, the structures of benzene and many boron-containing compounds both contained dotted lines. Which model was best undoubtedly depended on the exact structure of the molecule being discussed. By 1980, this dispute had forced chemists to appreciate a certain truism about molecules: organic molecules are never productively discussed in terms of a general molecular structure; they must always be considered individually. This truism, of course, recognized that the discussion of models for the structure of *individual* molecules could nevertheless be interesting.

To chemists, the neutralist/selectionist dispute was directly analogous. This was essentially a disagreement about whether changes in the chemical structure of the generic protein would, in general, change its behavior enough to change its contribution to the fitness of the generic organism. Again, the rational answer was in some cases yes, and in other cases no, depending on the exact structure of the system. Proteins are, after all, organic molecules, suggesting that they must be considered individually. As expected by those who understood this truism, the neutralist/selectionist dispute, in its general form, melted away as soon as our ability to analyze the behavior of individual proteins improved (Hey, 1999).

Even in 1980, however, it was clear that connecting fitness to the behavior of *individual* biomolecules would always remain interesting, for many reasons. First, that understanding would certainly help us select behaviors of those biomolecules to study in detail. If a behavior was important to fitness, it might be highly optimized. Detailed study might therefore instruct us about the interaction between chemical structure and biomolecular behavior, instruction worthy of the growing armamentarium of biophysics and molecular biology.

1.2 History as an essential tool to understand chemistry

It was clear, however, that Structure Theory in chemistry would not support a deep understanding of biological molecules. With simple molecules, like methane, one does not ask about its purpose. This is not true about complex systems, or living systems, where it is appropriate to ask: *why* does it exist? History can be key to any answer to why? questions. Any system, natural or human-made, can be understood better if we understand *both* its structure *and* its history. We would not understand the QWERTY computer keyboard, the Microsoft Windows operating system, or the US Federal Reserve Bank (for example) if we simply deconstructed each into its parts. An understanding of the history of each is essential to an understanding of the systems themselves.

Structure Theory from chemistry had absolutely no historical component. Methane is how it is because of its structure. It always has been this way, and always will be. Where the methane came from and how it got to us was fully irrelevant to our understanding of this molecule. This raised the next in a series of questions leading to experimental paleogenetics: how were Structure Theory and natural history to be combined to better understand biomolecules? Fragments of the history of life on Earth are found in the geological strata, of course. But the fossil record is notoriously incomplete, and would not provide information about proteins even were it not. Molecular fossils (such as those found in petroleum) can be informative, but generally not about individual protein function. Further, any analysis of molecular function must recognize that the behaviors that confer fitness are determined by the system, including other organisms (ecology), the physical environment (planetary biology), and even the cosmos (astrobiology). This level of complexity defeats most theoretical contexts.

It was clear, however, that the chemical structures of proteins themselves contain historical information. The historical relationships between proteins related by common ancestry can be inferred by comparing their amino acid sequences, a theme that was already well developed by 1980 (Dayhoff *et al.*, 1978). Analysis of protein sequences could generate the basic elements of an evolutionary model: a multiple sequence alignment, a tree, and sequences of ancestral protein sequences inferred from these. From these, it might be possible to construct narratives connecting biomolecular structure to fitness.

This process was analogous to processes well known in the field of historical linguistics (Lehman, 1973), which Robert Breedlove had described to me when I was an undergraduate. This field infers the features in ancestral languages by analyzing the features of their descendent languages. For example, the Proto-Indoeuropean word for snow (**sneig^{w}h-*) can be reconstructed from the descendant words for snow in the descendant Indoeuropean languages (German *schnee*, French*neige*, Irish *sneachta*, Russian *sneg*, Sanskrit *snihyati*, and so on). Other features of the histories of these languages, such as the universal replacement of *sn-* by *n-* in the Romance languages, can also be inferred from this analysis. The analogous inferences about ancestral structures could also be done for proteins.

The reconstruction of ancestral languages provides paleoanthropological information as well. From the ancestral features of reconstructed ancestral languages, one can extract information about the people who spoke them. For example, the ease with which we reconstructed the Proto-Indoeuropean word for snow (with some concessions; the Sanskrit word cited above actually means "he gets wet") tells the story that the Proto-Indoeuropeans themselves lived in a locale where it snowed. In 1980, we hoped to tell analogous stories using proteins inferred to have been present in ancestral forms of life on Earth.

1.3 Swapping places: biologists become chemists and statisticians, just as chemists become natural historians

But would these be only just-so stories? The just-so story is one of the worst insults that a biologist can direct at another. This epithet accuses a professional adversary of building *ad hoc* explanations for specific facts (how the zebra got his stripes). The events behind a just-so story (an ancestral zebra took a nap under a ladder) cannot be independently verified, and are not mathematically modelable. Further, the story could easily be replaced by a different story, just as compelling, had the observations been the opposite. It was clear in 1980 that once the insult stuck, papers would be rejected, grant applications would be turned down, and tenure would be denied.

Curiously, this issue also had a parallel in organic chemistry. Chemists are well known for their ability to use Structure Theory to explain a set of facts, only to be told that the facts are opposite, and then to explain the counter-facts using the same theory. Chemists are rarely defensive about this. In part, this is because Structure Theory as a heuristic has been so successful. If one can make petrochemicals and pharmaceuticals (and much in between) using a theory based on plastic tinker toy models, who can argue?

The success of non-mathematical Structure Theory from chemistry makes a larger point about human knowledge; that it is intrinsically heuristic and intuitive. This is true even for knowledge that is cast in the language of mathematics. This conclusion had been reached by the last century of epistemology as well (Suppe, 1977). Nothing is "proven" (Galison, 1987); the perception of proof is only a function of the number of logical steps that must be taken to premises that are intuitive and heuristic. Experiments end when a burden of proof is met, where that burden is defined by the culture, not by logic.

This point is not fully appreciated by many modern biologists. Many modern biologists seek to avoid the just-so story epithet, and the perception of theirs being a heuristic and/or intuitive theory, by placing a mathematical formalism on top of their models. This drives them towards statistics, which analyzes collections of things. Statistics, in turn, nearly always requires the statistician to deny the truism in chemistry that there is no such thing as general molecular behavior. This, in turn, means that statisticians, in their pursuit of general models framed in mathematical language, are not able to exploit the only research paradigm that has shown itself to be successful in understanding molecules.

In fact, the barrier between mathematics and molecular science is still higher. Statisticians are taught that a model is not scientific if it is *not* formulated in the language of mathematics. Therefore, statisticians are perplexed that a field like chemistry can be successful. And even as they acknowledge that proteins are chemicals, statisticians insist that unless protein sequences are studied as collections, the studies are "unscientific" (Robson and Garnier, 1993). Thus, statisticians actively work to deny to all other scientists the one research paradigm that has been successful to understand molecules.

This cultural phenomenology has set up a role reversal of a sort. With their training in heuristic science, chemists may have been better prepared to make the connection between chemical behavior and biological fitness than biologists. As physical scientists, chemists were trained in mathematics and statistics. Because they understood heuristic models, however, they used statistics and mathematics as tools, not as the way to respond to the complaint, "You are not doing real science".

Further, chemical theory grows by accretion, rather than revolution; it adds theories, ideas, and perspectives to its heuristic theory. This is exactly what is needed to understand the broader picture in contemporary biology. Here, by the end of the current century, we expect to see a global view of reality that combines chemical models, systems models, physiological models, paleontological models, and geological models. If the output is still dissatisfying, then the global view will add still more models. We expect (or, perhaps better, hope) that over time, an increasingly dense set of models, all interconnecting, would eventually converge upon a global picture for biology, just as it has for chemistry over the past century.

1.4 Managing heuristic science

This discussion should not be viewed as a defense of so-called soft science. Rather, it is simply an observation of how human knowledge really works. The observation need not be viewed pejoratively. Human scientists can be creative and productive *because* human understanding is intuitive and heuristic. Thus, while the scientific method taught in middle school emphasizes the importance of making unfiltered observations, analyzing data without prejudice, and doing value-neutral experiments, the productivity of scientists does not depend on the extent to which they meet this largely fictitious ideal, but rather how they manage the closedness of mind, the values, and the filters that come naturally with human cognition.

This concept of management is important. Chemistry does not ignore the natural tendency of humans to convince themselves that data contain patterns that they do not, or that patterns compel models when they need not, or that models are reality, which they are not. Rather, chemistry establishes processes that manage this tendency.

Key to this management is the use of experiment on systems that have been synthesized (Benner and Sismour, 2005). The use of synthesis to create new forms of matter, whose behavior is expected from a heuristic theory, provides an opportunity for an independent test of the heuristic. *De novo* synthesis in not available in many other disciplines. For example, planetary scientists cannot synthesize a new planet to test their theories on how planets work. If they could, the field would be dramatically transformed.

How could we use synthesis and experiment to manage the development of our historical view of biomolecules? In 1980, the answer was materializing before our eyes. Jeffrey Miller, Michael Brown, Alan Fersht, and others were developing the technology to create a protein having any sequence that might be desired. Most protein engineering was targeted to replace single amino acids in extant proteins for the purpose of understanding their role in a protein's catalytic behavior. But it was clear that protein engineering technology could also be used to synthesize ancestral proteins whose sequences had been inferred using ideas outlined by Pauling and Zuckerkandl, where the resurrected proteins could then be experimentally studied in the laboratory.

This is how the idea for experimental paleogenetics (which we originally called paleobiochemistry) began in our laboratories in 1980. We wanted to bring ancient proteins back to life to examine their behaviors. This would use synthesis to add an experimental component to our understanding of the history of biomolecules. The historical component was necessary to understand biomolecules, just as it was to understand the US Federal Reserve banking system. This experimental component would also manage the problems intrinsic to a heuristic science. Through this combination, we hoped to understand how an interaction between chance, history, vestigiality, selection, and physical law determined the structures and behaviors of individual protein families. From there, we could perhaps make inferences about how these were related to fitness and physiological function. Then, perhaps, we could select interesting biomolecular behaviors to study.

1.5 Selecting proteins to begin experimental paleoscience

But what individual protein should we look at? While the Maxam–Gilbert and Sanger papers on DNA sequencing made clear that the sequencing of the human genome was only a matter of time, databases in 1980 contained very few protein sequences. The only families of proteins that were sufficiently well represented to support experimental paleogenetics were the cytochromes, the hemoglobins, and the ribonucleases (RNases). Cytochromes were, of course, substrates for cytochrome oxidases. With no funding for this project (the National Institutes of Health routinely disapproved our proposals in this area) we could not possibly resurrect ancestral cytochromes, only to then need to resurrect their ancient oxidases. Hemoglobins were complicated to express, a problem solved only later. This left the RNases.

Fortunately, Jaap Beintema and his colleagues in Groningen had done the yeoman's job of sequencing RNases (at the level of the protein) from a wide range of ruminants and closely related nonruminant mammals (Cho *et al.*, 2005). They had, Dayhoff-style (Dayhoff *et al.*, 1978), inferred the sequences of the ancestral proteins throughout the recent history of the digestive enzymes. Barnard had raised an interesting hypothesis suggesting that digestive RNases might be unique to ruminants, and be an adaptation to their unique ruminant digestive physiology (Barnard, 1969). And so, we had a place to start.

The story of RNase resurrections is told in a separate chapter in this volume (see Chapter 18). This story illustrates well the value of paleomolecular resurrections for creating an understanding of the relation between organismic and molecular biology on one hand, and the changing ecosystems wrought by a changing planet on the other. It also, in the process, showed how we might use paleogenetics to select *in vitro* biomolecular behaviors to study in a way that considers physiological relevance (Nambiar *et al.*, 1984, 1987; McGeehan and Benner, 1989; Benner and Allemann, 1989; Stackhouse *et al.*, 1990; Allemann *et al.*, 1991; Jermann *et al.*, 1995).

But our work with RNases, and work in other laboratories in other systems, also showed that experimental paleogenetics could create contentious disputes of its own. Many of these relate to the reliability of statistically grounded tools to infer the structures of ancestral proteins, and how the outcome of paleogenetics experiments should be interpreted. These issues will be addressed in this chapter, and as they are in other chapters in this volume. I will describe the use of paleogenetic experiments to manage them in one system, the alcohol dehydrogenases.

1.6 Mathematical models are nevertheless important

Mathematical formalism is useful in the inference of ancestral sequences from the sequences of their descendants. Protein sequences lend themselves to representations as linear strings of letters. As organic molecules, such linear representations do not capture much of their organic chemical behavior, of course. Nor do such linear representations capture the behavior of protein sequences during divergent evolution. Homoplasy, correlated change, and a host of other features reveal protein sequences for what they really are: poor models of the structures of real organic molecules.

Nevertheless, mathematical formalisms that treat proteins as linear strings of letters turn out to be useful (Benner *et al.*, 1997). Any model that treats protein sequences as linear strings diverging via a Markovian process provides a null hypothesis, a description of protein evolution that *would have happened* if proteins *were* formless, functionless linear strings. By observing how proteins divergently evolve, and comparing this reality to the null hypothesis, one extracts a signal about form and function (Benner *et al.*, 1997).

The null hypothesis provides a serviceable starting point for ancestral reconstruction as well. The underlying Darwinian process is, of course, semi-random. Its departures from randomness, arising from biases in the DNA-polymerization or error-repair mechanisms, nucleosome structure, or other features of the DNA molecule itself, are not likely to be strongly correlated with protein structure and behavior (again analogous to language; the conversion of *sn-* to *n-* is largely unrelated to the dictional meaning of the word snow). Hence, it is not surprising that respectable inferences of ancestral states can be made using the linear string model.

There is nevertheless an ongoing dispute over which methods are precisely best for inferring ancestral sequences (Yang *et al.*, 1995; Zhang and Nei, 1997; Pagel, 1999; Nielsen, 2002; this volume, see Chapters 4 and 8 for example). From a practical perspective, these disputes do not have a large impact on the practice of experimental paleoscience. In practice, the principal ambiguities do not generally arise from subtleties in models for inferring ancestral sequences. Rather, they arise from incomplete sequence data-sets, uncertain gap placement in multiple sequence alignments, uncertain tree topology, and too much sequence divergence relative to tree articulation. This creates uncertainties in inferred ancestral character states long before the choice of the model becomes determinative.

Thus, if an evolutionary tree is highly articulated, the branching topology is secure, and the overall extent of sequence divergence is small, then different mathematical models infer more or less the same ancestral sequences. When the tree is not highly articulated, the branching topology is not secure, and the overall extent of sequence divergence is large, even the most mathematically sophisticated analysis cannot help much.

Today, a practicing paleogeneticist is advised to apply mathematical models at all levels of sophistication to build many different candidate multiple sequence alignments, candidate evolutionary trees, and candidate ancestral sequences. Additional information, such as crystallographic and paleontological data, should be both used and not used. From this will come a view of the ambiguity in the ancestral sequences that arises from the ambiguity and bias in the input.

Four strategies can then be considered to manage this ambiguity. The first relies on improving the statistical models of sequence divergence, in the hope that an increase in the sophistication of the mathematical formalism will resolve ambiguity. The second involves collecting more sequences in the hope of eliminating the ambiguity. The third ignores the ambiguity, in the hope that the ambiguity occurs only at sites that are not critical for the biological interpretation.

The fourth involves synthesizing and studying many candidate ancestral sequences to cover all plausible alternative reconstructions, or to sample among the plausible alternative reconstructions. The experimentalist then asks whether the behavior that supports a biological interpretation (and therefore the interpretation itself) is robust with respect to the ambiguity arising from uncertainties in the models, insufficient data, poorly articulated trees, or other issues in practice. This is our preferred method. The preference reflects a belief (perhaps better described as a faith) about a feature of protein chemistry that is presently unknown, but is not unknowable in principle. If the hypersurfaces relating protein behavior to protein sequence were extremely rugged, and if every amino acid replacement caused a significant change in behavior, then ambiguity would defeat the paleogenetic research approach in all but the most ideal cases. Fortunately, biochemical reality appears to be different. For nearly all proteins, some amino acid replacements at some sites have large impacts on functional behaviors. Replacements at other sites have only a modest impact on those behaviors, and replacements at still other sites have even less impact on most behaviors.

These facts would tend to ameliorate the extent to which ambiguity compromises the interpretation of data extracted from paleoscience experiments. Ambiguities in inferred ancestral characters generally are found at sites that have suffered many amino acid replacements. Multiple replacements often (but not always) reflect the possibility of neutral drift at a site. Neutral drift implies that the choice of a residue at the site does not have a significant impact on fitness. This generally (but not always) means that replacement of an amino acid at that site does not have any impact on the behavior of a protein.

Stringing these together, we might expect that biologically interpretable behavior will not differ greatly between ancestral sequences that differ only at ambiguous sites. To the extent that the premises are true, ambiguity in general will not limit our ability to draw inferences about the behavior of ancestral proteins by experimental analysis of ancestral sequences, even if our analysis does not capture all of the ambiguity in those sequences. This, in turn, means that we will generally be able to use those behaviors to generate interesting biological interpretations. In fact, this is the case with the dozen or so examples of experimental paleogenetics where the issue has been examined over the past two decades.

1.7 Alcohol dehydrogenase

The ultimate goal of molecular paleoscience is to connect the molecular records for all proteins from all organisms in the modern biosphere with the geological, paleontological, and cosmological records to create a broadly based, coherent narrative for life on Earth (Benner *et al.*, 2002). Because much of natural selection is driven by species–species interactions, developing this narrative will require tools that broadly connect genomes from different species, as well as interconnect events within a single species. It remains an open question, of course, how much of the record has been lost through extinction, erosion, and poor fossil preservation.

The literature so far contains only the very first case studies where such a broad interconnection is conceivable. For example, modern yeast living in modern fleshy fruits rapidly convert sugars into bulk ethanol via pyruvate (Figure 1.1). Pyruvate

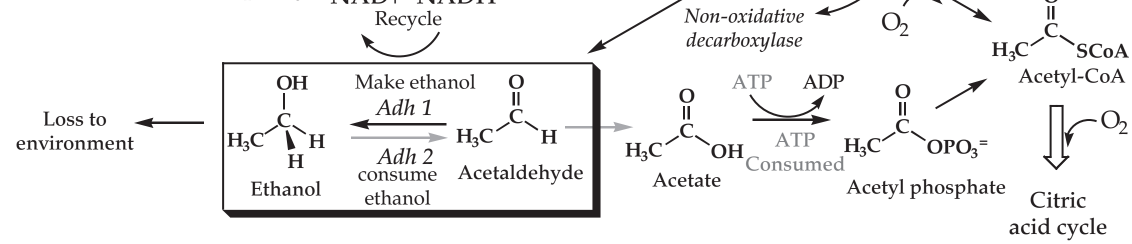

Figure 1.1 The formation of ethanol from glucose by the yeast *Saccharomyces cerevisiae* is an energetically expensive diversion of carbon in the overall degradation of glucose to give acetyl-CoA for the citric acid cycle. The yeast genome has two genes that catalyze the ethanol–acetaldehyde interconversion. One (Adh 1) is used to make ethanol; the other (Adh 2) is used to consume ethanol. Why does the yeast genome have these two in the genome, as either can catalyze this reaction in both directions? Enzymes in italics are associated with gene duplications that, according to the transition redundant exchange (TREx) clock (Benner *et al.*, 2002), arose nearly contemporaneously. The make–accumulate–consume pathway is boxed. Note that the shunting of the carbon atoms from pyruvate into (and then out of, open arrows) ethanol is energy-expensive, consuming a molecule of ATP for every molecule of ethanol generated. This ATP is not consumed if pyruvate is oxidatively decarboxylated directly to give acetyl-CoA to enter the citric acid cycle directly (open arrow to the right). If dioxygen is available, the recycling of NADH does not need the acetaldehyde-to-ethanol reduction. Reprinted from Benner, S.A. and Sismour, A.M. (2005) Synthetic biology. *Nat. Rev. Genet.* **6**: 533–543.

then loses carbon dioxide to give acetaldehyde, which is reduced by alcohol dehydrogenase 1 (Adh 1) to give ethanol, which accumulates. Yeast later consumes the accumulated ethanol, exploiting Adh 2 and Adh 1 homologs differing by 24 (of 348) amino acids.

Generating ethanol from glucose in the presence of dioxygen, only to then re-oxidize the ethanol, is energetically expensive (Figure 1.1). For each molecule of ethanol converted to acetyl-CoA, a molecule of ATP is used. This ATP would not be wasted if the pyruvate that is made initially from glucose were delivered directly to the citric acid cycle.

This implies that yeast has a reason, transcending simple energetic efficiency, for rapidly converting available sugar in fruit to give bulk ethanol in the presence of dioxygen. One just-so story to explain this inefficiency holds that yeast, which is relatively resistant to ethanol toxicity, may accumulate ethanol to defend resources in the fruit from competing microorganisms (Boulton *et al.*, 1996). While the ecology of wine yeasts is certainly more complex than this simple hypothesis implies (Fleet and Heard, 1993), fleshy fruits do offer a large reservoir of carbohydrate. This resource must have value to competing organisms as well as to yeast. For example, humans have exploited the preservative value of ethanol since prehistory (McGovern, 2004).

The timing of Adh expression in *Saccharomyces cerevisiae* and the properties of the expressed proteins are both consistent with this story. The yeast genome encodes two major Adhs that interconvert ethanol and acetaldehyde (Figure 1.1; Wills, 1976). The first (Adh 1) is expressed at high levels constitutively. Its kinetic properties optimize it as a catalyst to make ethanol from acetaldehyde (Fersht, 1977; Ellington and Benner, 1987). In particular, the Michaelis constant (K_m) for ethanol in Adh 1 is high (17 000–20 000 μM), consistent with ethanol being a product of the reaction. After the sugar concentration drops, the second dehydrogenase (Adh 2) is derepressed. This paralog oxidizes ethanol to acetaldehyde with kinetic parameters suited for this role. The K_m for ethanol for Adh 2 is low (600–800 μM), consistent with ethanol at low concentrations becoming its substrate.

Adh 1 and Adh 2 are homologs differing by 24 of 348 amino acids. Their common ancestor, termed ADH_A, had an unknown role. If ADH_A existed in a yeast that made, but did not accumulate, ethanol, its physiological role would presumably have been the same as the role of lactate dehydrogenase in mammals during anaerobic glycolysis: to recycle NADH generated by the oxidation of glyceraldehyde 3-phosphate (Figure 1.1; Stryer, 1995). Lactate in human muscle is removed by the bloodstream; ethanol would be lost by the yeast to the environment. If so, ADH_A should have been optimized for ethanol synthesis, as is modern Adh 1. The kinetic behaviors of

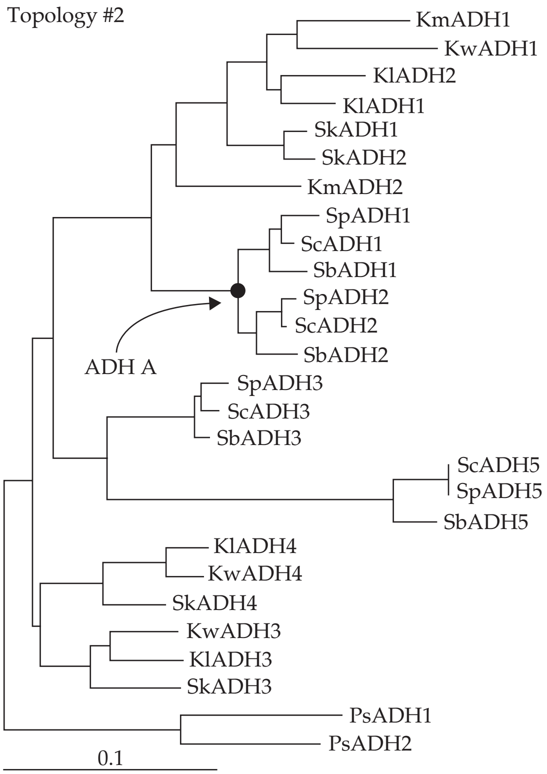

Figure 1.2 Maximum-likelihood trees interrelating sequences determined in this work with sequences in the publicly available database. Shown are the two trees with the best (and nearly equal) maximum-likelihood scores using the following parameters estimated from the data. Substitutions A–C, A–T, C–G, and G–T have a score of 1.00, A–G has a score of 2.92, and C–T has a score of 5.89; empirical base frequencies, and proportion of invariable sites and the shape parameter of the gamma distribution are set to 0.33 and 1.31, respectively. The scale bar represents the number of substitutions/codon per unit of evolutionary time. Reprinted from Benner, S.A. and Sismour, A.M. (2005) Synthetic biology. *Nat. Rev. Genet.* **6**: 533–543.

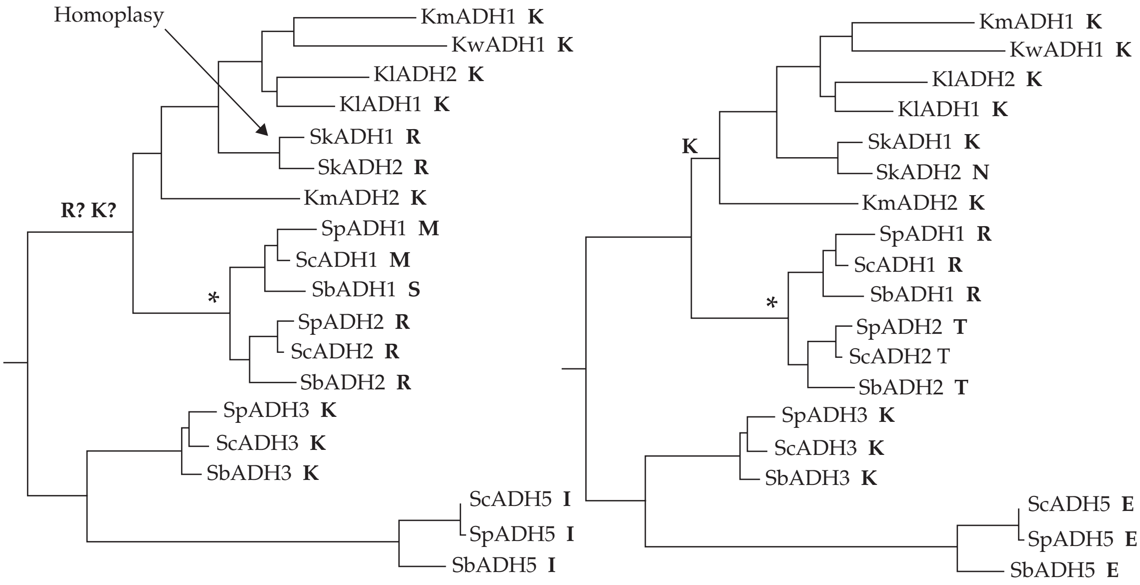

Figure 1.3 The distribution of amino acids at site 168 (left) and site 211 (right) in a set of 19 fungal alcohol dehydrogenases. The node of interest is at the right end of the branch marked by *. Note the difficulty in reconstructing the amino acids at these sites at the node at the right end of the red branch.

$\mathrm{ADH_A}$ should resemble those of modern Adh 1 more than Adh 2, with a high K_m for ethanol.

To add paleobiochemical data to convert this just-so story into a more compelling scientific narrative, a collection of Adhs from yeasts related to *S. cerevisiae* was cloned, sequenced, and added to the existing sequences in the database (Thomson *et al.*, 2005). A maximum-likelihood evolutionary tree was constructed using PAUP*4.0 (Figure 1.2; Swofford, 1998). Maximum-likelihood sequences for $\mathrm{ADH_A}$ were then reconstructed using both codon and amino acid models in PAML (Yang, 1997). When the posterior probability that a particular amino acid occupied a particular site was >80%, that amino acid was assigned at that site in $\mathrm{ADH_A}$.

When the posterior probability was <80% and/or the most probabilistic ancestral state estimated using the codon and amino acid models were not in agreement, the site was considered ambiguous, and alternative ancestral genes were considered. For example, the posterior probabilities of two amino acids (methionine and arginine) were nearly equal at site 168 in $\mathrm{ADH_A}$, three amino acids (lysine, arginine, and threonine) were plausibly present at site 211, and two (aspartic acid and asparagine) were plausible for site 236.

Figure 1.3 shows some of the reason for the ambiguities at sites 168 and 211. As can be seen by the placement of characters (here, amino acids) on the leaves of the trees, it is difficult to infer the ancestral character for the node at the right end of the branch in the tree marked by * in Figure 1.3, representing the last common ancestor of Adh 1 and Adh 2. In part, the difficulties arise because of homoplasy, a historical phenomenon where the same amino acid replacement occurred more than once at different times in the family's history. This suggested that selective constraints were influencing the selection of amino acids at those sites. This ambiguity could therefore not be ignored.

To handle these ambiguities, all 12 (all $2 \times 2 \times 3$ combinations) candidate $\mathrm{ADH_A}$s were resurrected by constructing genes that encoded them, transforming these genes into a strain of *S. cerevisiae* from which both Adh 1 and Adh 2 had been deleted, and expressing them from the Adh1 promoter. All of the ancestral sequences could rescue the double-deletion phenotype in the expression yeast.

Table 1.1 Kinetic properties of Adh 1, Adh 2, and candidate ancestral ADH$_A$s. Reprinted from Benner, S.A. and Sismour, A.M. (2005) Synthetic biology. *Nat. Rev. Genet.* **6**: 533–543.

Sample[a]	K_m (μM) Ethanol	NAD$^+$	Acetaldehyde	NADH
Adh1	20 060	218	1492	164
MKD	17 280	511	1019	144
MKN	13 750	814	1067	1106
MRD	11 590	734	1265	287
MRN	10 960	554	1163	894
MTD	10 740	467	959	190
MTN	N/A	N/A	N/A	N/A
RKD	8497	449	1066	142
RKN	7238	407	1085	735
RRD	7784	400	1074	203
RRN	8403	172	1156	1142
RTD	6639	254	1083	316
RTN	7757	564	1158	477
Adh1[b]	24 000	240	3400	140
Adh1[c]	17 000	170	1100	110
Adh2[b]	2700	140	45	28
Adh2[c]	810	110	90	50
Adh3[c]	12 000	240	440	70
Adh1[c] (*S. pombe*)	14 000	160	1600	100
Adh1(M270L)[c]	19 000	630	1000	80
KlP20369[d]	27 000	2800	1200	110
KlX64397[d]	23 000	2200	1700	180
KlX62766[d]	2570	310	100	20
KlX62767[d]	1560	200	3100	30

[a] The three letters in the sample names designate the amino acids at positions 168, 211, and 236; thus MKD is Met-168, Lys-211, Asp-236. The remaining residues were the same as in Adh 1, except for the following changes (using sequence numbering of Adh1 from *S. cerevisiae*): Asn-15, Pro-30, Thr-58, Ala-74, Glu-147, Leu-213, Ile-232, Cys-259, Val-265, Leu-270, Ser-277, and Asn-324. Kl, *Kluyveromyces lactis*; N/A, not applicable; *S. pombe*, *Schizosaccharomyces pombe*.

[b] From Thomson (2002).

[c] From Ganzhorn *et al.* (1987).

[d] From Bozzi *et al.* (1997).

Table 1.1 lists kinetic data from the candidate ancestral ADH$_A$s (Thomson *et al.*, 2005). Simple kinetic metrics were then used to assess the quality of the data. In particular, the Haldane equation relates the equilibrium constant for the Adh reaction with various of the measured kinetic parameters according to the equation (Segal, 1975).

$$K_{eq} = V_f K_{iq} K_p / V_r K_{ia} K_b$$

where V_f and V_r are forward and reverse maximal velocities, K_{ia} and K_{iq} are disassociation constants for NAD$^+$ and NADH, and K_b and K_p are Michaelis constants for ethanol and acetaldehyde, respectively. These parameters were calculated from the experimental data. The Haldane equation reproduced the literature equilibrium constant for the reaction to within a factor of two. One variant, termed MTN (for the amino acids at sites 168, 211, and 236), had very low catalytic activity in both directions. This suggested that this particular candidate ancestor was not the true ancestor.

Significant to the hypothesis, the kinetic properties of the candidate ancestral ADH$_A$s resembled those of Adh 1 more than Adh 2 (Table 1.1). This included the high K_m for ethanol, a sign of an ancestor that did not have ethanol at low concentrations as its physiological substrate. From this observation, Thomson *et al.* (2005) inferred that the ancestral yeast did not have an Adh specialized for the consumption of ethanol, like modern Adh 2, but rather had an Adh specialized for making ethanol, like modern Adh 1. This, in turn, suggested that the ancestral yeast prior to the time of the duplication did not consume ethanol. This implied that the ancestral yeast also did not make and accumulate ethanol under aerobic conditions for future consumption, and that the make–accumulate–consume strategy emerged after Adh 1 and Adh 2 diverged. These interpretations were robust with respect to the ambiguities in the reconstructions.

Several details are worthy of further comment. For modern Adh 1, the ranges of literature K_m values were 17 000–24 000 μM for ethanol, 170–240 μM for NAD$^+$, 1100–3400 μM for acetaldehyde, and 110–140 μM for NADH (Ganzhorn *et al.*, 1987). These comparisons, together with the Haldane analysis, provide a view of the experimental error in the kinetic parameters reported in the paleoreconstruction. The interpretations about the kinetic behavior of the ancient ADH$_A$ are based on differences well outside of experimental error.

Further, when paralogs are generated by duplication, many believe that the duplicate that then evolves more rapidly is the one that acquires the new functional role (Kellis *et al.*, 2004). If this were generally true, one might identify the functionally innovative duplicate by a simple bioinformatics analysis. Whereas this may be true for many genes, chemical principles do not obligate this outcome, and it is not true with these Adh paralogs. Here, the rate of evolution is not markedly faster in the lineage leading to Adh 2 (having the derived behavior) than in the lineage leading to Adh 1 (having the primitive behavior). The paleobiochemistry experiment was necessary to assign the primitive behavior.

Further, the Haldane ratio relates various kinetic parameters (K_{cat}, K_{m}, K_{diss}) that can change via a changing amino acid sequence to the overall equilibrium constant, which the enzyme (being a catalyst) cannot change. Thus, if a lower K_{m} for ethanol is selected, other terms in the Haldane must change to keep the ratio the same. This is observed in data for the ancestral proteins prepared here and the natural enzymes.

1.8 Interconnecting models

By the end of the current century, we can expect that the divisions between branches of biology (molecular, cell, systems, organismic, environmental, geo-, and astro-biology) will be subsumed within a broad model of the phenomenon we know as life. This will include, of course, the products of the reductionist paradigm that has placed chemical structures upon many of the phenomena unique to biology, including genetics, emergent behavior, Darwinian evolution, and functional complexity. It will also incorporate the products of the reductionist paradigm that uses mathematical models to describe the interaction between individuals in a population and different organisms within an ecosystem.

But it will also include a history of the biosphere based on the geological, paleontological, and genomic records. This history will be needed to address the questions of why and how in biology. Here, the answers will come in the form of narratives that describe historical events that fashioned the molecules, cells, systems, organisms, and environments for individual biomolecules. There will be little room for statistics in this model. Rather, the individual traits of individual systems will be understood as the products of chance, necessity, and vestigiality interacting under constraints from physical law.

Further, this model will be heuristic. It will avoid the epithet of being a just-so story by having multiple lines about many systems on Earth that interconnect and intercorrelate in a comprehensive model for the history of the planet, the life that it holds, and the chemistry behind that life. Further, it will depend on the synthesis of ancestral forms to test its heuristics, where experimental paleoscience will repeatedly require the revision of the heuristics.

It is possible to combine the data that were available before the experimental paleoscience done with the Adh system, the data on ADH_{A} from the experiments described here, and subsequently emerging information, to set us on this path to this complex, intercorrelated, and interconnected future for this individual system. We might begin by asking whether the Adh 1/Adh 2 duplication and the accumulate–consume strategy that it presumably enabled became fixed in the yeast population in response to a particular selective pressure?

Hypothetically, the emergence of a make–accumulate–consume strategy may have been driven by the domestication of yeast by humans selecting for yeast that accumulates ethanol. Alternatively, the strategy might have been driven by the emergence of fleshy fruits that offered a resource worth defending using ethanol accumulation. We might distinguish between the two by estimating the date when the Adh 1/2 duplication occurred. Even with large errors in the estimate, a distinction should be possible, as human domestication occurred in the past million years, while fleshy fruits arose in the Cretaceous, after the first angiosperms appeared in the fossil record 125 million years ago (Sun, 2002), but before the extinction of the dinosaurs 65 million years ago (Collinson and Hooker, 1991; Fernandez-Espinar *et al.*, 2003).

The topology of the evolutionary tree in Figure 1.2 suggests that the Adh 1/Adh 2 duplication occurred before the divergence of the *sensu strictu* species of *Saccharomyces* (Fernandez-Espinar *et al.*, 2003), but after the divergence of *Saccharomyces* and *Kluyveromyces*. The date of divergence of *Saccharomyces* and *Kluyveromyces* is unknown, but might be estimated to have occurred

80±15 million years ago (Berbee and Taylor, 1993). This date is consistent with a transition-redundant exchange (TREx) clock (Benner, 2003), which exploits the fractional identity (f_2) of silent sites in conserved 2-fold-redundant codon systems to estimate the time since the divergence of two genes. Between pairs of presumed orthologs from *Saccharomyces* and *Kluyveromyces*, f_2 is typically 0.82, not much lower than the f_2 value (0.85) separating Adh 1 and Adh 2 (Benner *et al.*, 2002), but much lower than paralog pairs within the *Saccharomyces* genome that appear to have arisen by more recent duplication (approx. 0.98; Lynch and Conery, 2000).

Interestingly, Adh 1 and Adh 2 are not the only pair of paralogs where $0.80 < f_2 < 0.86$ (Benner *et al.*, 2002). Analysis of approximately 350 pairs of paralogs contained in the yeast genome (considering pairs that shared at least 100 silent sites, and diverged by less than 120 point-accepted replacements per 100 sites) identified 15 pairs having $0.80 < f_2 < 0.86$ (Figure 1.4). These represent eight duplications that occurred near the time of the Adh 1 and Adh 2 duplication, if f_2 values are assumed to support a clock.

These duplications are not randomly distributed within the yeast genome. Rather, six of the eight duplications involve proteins that participate in the conversion of glucose to ethanol (Table 1.2). Further, the enzymes arising from the duplicates are those that appear, from expression analysis, to control flux from hexose to ethanol (Schaaff *et al.*, 1989; Pretorius, 2000). These include proteins that import glucose, pyruvate decarboxylases that generate the acetaldehyde from pyruvate, the transporter that imports thiamine for these decarboxylases, and the Adhs (the italicized proteins in Figure 1.1). If the f_2 clock (within its expected variance) is assumed to date paralogs in yeast, this cluster suggests that several genes other than Adh duplicated as part of the emergence of the new make–accumulate–consume strategy, near the time when fleshy fruit arose.

The six gene duplications proposed to be part of the emergence of the make–accumulate–consume strategy (in the $0.80 < f_2 < 0.86$ window) are *not* associated with one of the documented blocks of genes were duplicated in ancient fungi, possibly as part of a whole-genome duplication (Wolfe and Shields, 2001). Two duplications in genes that are

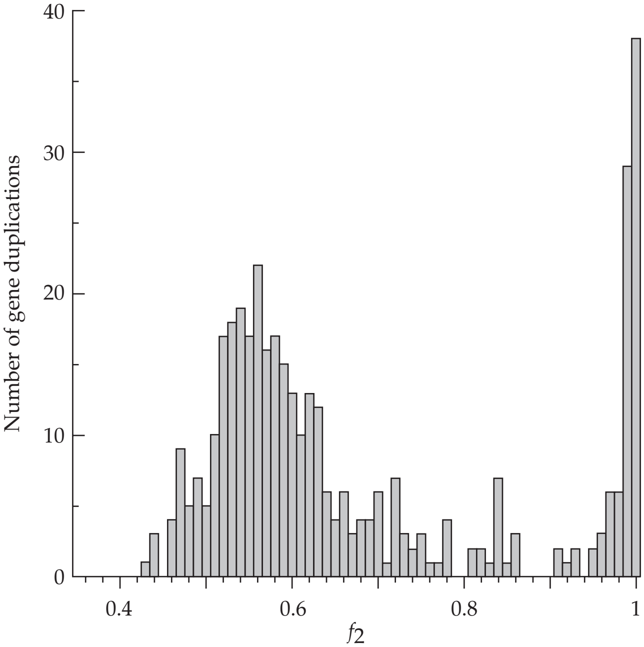

Figure 1.4 A histogram showing all of the pairs of paralogs in the *S. cerevisiae* genome, dated using the transition redundant exchange (TREx) tool (Benner, 2003). The episode of gene duplication where $0.80 < f_2 < 0.86$ is isolated from more ancient duplications (the mode of the distribution at the left) and more recent duplications (represented by the bars at the very right of the plot). Paralog pairs are considered only if they have with at least 100 aligned silent sites, and are not separated by more than 120 point-accepted mutations per 100 aligned amino acid sites (PAM units). Reprinted with permission from Benner *et al.* (2002) Planetary biology: paleontological, geological, and molecular histories of life. *Science* **296**: 864–868, © 2002 AAAS.

Table 1.2 Duplication in the *S. cerevisiae* genome (SGD), where $0.80 < f_2 < 0.86$. Reprinted from Benner, S.A. and Sismour, A.M. (2005) Synthetic biology. *Nat. Rev. Genet.* **6**: 533–543.

SGD name	gi number	Trivial name	Annotation and comments
Inosine-5′-monophosphate dehydrogenase family (3 paralogs, 3 pairs, 2 duplications)[a]			
$f_2 = 0.803$[c]; pair associated with Wolfe duplication blocks 1 and 44			
YAR073W	gi|456156	IMD1	Nonfunctional homolog, near telomere, not expressed
YLR432W	gi|665971	IMD3	Inosine-5′-monophosphate dehydrogenase
$f_2 = 0.825$[c]			
YLR432W	gi|665971	IMD3	Inosine-5′-monophosphate dehydrogenase
YHR216W	gi|458916	IMD2	Inosine-5′-monophosphate dehydrogenase
Subfamily pair: YHR216W/YAR073W, $f_2 = 0.93$ (proposed recent duplication creating a pseudogene)			
Sugar transporter family A (4 paralogs, 4 pairs, 3 duplications)[b]			
$f_2 = 0.805$[c]; pair not associated with any duplication block			
YJR158W	gi|1015917	HXT16	Sugar transporter repressed by high glucose levels
YNR072W	gi|1302608	HXT17	Sugar transporter repressed by high glucose levels
$f_2 = 0.806$[c]; pair not associated with any duplication block			
YDL245C	gi|1431418	HXT15	Sugar transporter induced by low glucose, repressed by high glucose
YNR072W	gi|1302608	HXT17	Sugar transporter repressed by high glucose levels
$f_2 = 0.809$[c]; pair not associated with any duplication block			
YJR158W	gi|1015917	HXT16	Sugar transporter repressed by high glucose levels
YEL069C	gi|603249	HXT13	Sugar transporter induced by low glucose, repressed by high glucose
$f_2 = 0.810$[c]; pair not associated with any duplication block			
YEL069C	gi|603249	HXT13	Sugar transporter induced by low glucose, repressed by high glucose
YDL245C	gi|1431418	HXT15	Sugar transporter
Subfamily pair: YEL069C/YNR072W, $f_2 = 0.932$ (proposed recent duplication)			
Subfamily pair: YJR158W/YDL245C, $f_2 = 1.000$ (proposed very recent duplication)			
Chaperone family A (2 paralogs, 1 pair, 1 duplication)[a]			
$f_2 = 0.81$; pair associated with Wolfe duplication block 48			
YMR186W	gi|854456	HSC82	Cytoplasmic chaperone, induced 2–3–fold by heat shock
YPL240C	gi|1370495	HSP82	Cytoplasmic chaperone, pheromone signaling, Hsf1p regulation
Phosphatase/thiamine transport family A (2 paralogs, 1 pair, 1 duplication)[b]			
$f_2 = 0.818$; pair not associated with any duplication block			
YBR092C	gi|536363	PHO3	Acid phosphatase implicated in thiamine transport
YBR093C	gi|536365	PHO5	Acid phosphatase, one of three repressible phosphatases
Pyruvate decarboxylase family A (2 paralogs, 1 pair, 1 duplication)[b]			
$f_2 = 0.835$; pair not associated with any duplication block			
YLR044C	gi|1360375		PDC1 Pyruvate decarboxylase, major isoform
YLR134W	gi|1360549		PDC5 Pyruvate decarboxylase, minor isoform

Table 1.2 (*Continued*)

SGD name	gi number	Trivial name	Annotation and comments
By ortholog analysis, *Saccharomyces bayanus* (gi|515236) diverged from *S. cerevisiae* after the $f_2 = 0.835$ duplication, and *Kluyveromyces* diverged before			
Glyceraldehyde-3-phosphate dehydrogenase family (3 paralogs, 3 pairs, 2 duplications)[b]			
$f_2 = 0.845$[c]; pair not associated with any duplication block			
YJL052W	gi|1008189	TDH1	Glyceraldehyde-3-phosphate dehydrogenase
YGR192C	gi|1323341	TDH3	Glyceraldehyde-3-phosphate dehydrogenase
$f_2 = 0.845$[c]; pair not associated with any duplication block			
YJL052W	gi|1008189	TDH1	Glyceraldehyde-3-phosphate dehydrogenase
YJR009C	gi|1015636	TDH2	Glyceraldehyde-3-phosphate dehydrogenase
Subfamily pair: YJR009C/YGR192C, $f_2 = 0.991$, proposed very recent duplication			
Alcohol dehydrogenase family (2 paralogs, 1 pair, 1 duplication)[b]			
$f_2 = 0.848$; pair not associated with any duplication block			
YMR303C	gi|798945	ADH2	Alcohol dehydrogenase, glucose-repressible
YOL086C	gi|1419926	ADH1	Alcohol dehydrogenase, constitutive
Spermine transporter family (2 paralogs, 1 pair, 1 duplication)[a]			
$f_2 = 0.86$; pair associated with Wolfe duplication block 34			
YGR138C	gi|1323230	TPO2	Spermine-transporter activity
YPR156C	gi|849164	TPO3	Spermine-transporter activity
Sugar transporter family B (3 paralogs, 3 pairs, 2 duplications)[b]			
$f_2 = 0.847$[c]; pair not associated with any duplication block			
YDR343C	gi|1230670	HXT6	Sugar transporter, high affinity, high basal levels
YDR345C	gi|1230672	HXT3	Sugar transporter, low-affinity glucose transporter
$f_2 = 0.854$[c]; pair not associated with any duplication block			
YDR342C	gi|1230669	HXT7	Sugar transporter, high affinity, high basal levels
YDR345C	gi|1230672	HXT3	Sugar transporter, low affinity
Subfamily pair: YDR342C/YDR343C, $f_2 = 0.994$, proposed very recent duplication			

[a] Not associated with fermentation. These are associated with duplication blocks within the yeast genome (Kellis *et al.*, 2004), where the high value of f_2 (typically equilibrated in block paralog pairs) may reflect either variance, or selective pressure to conserve silent sites in individual codons.

[b] Associated with the pathway to make, accumulate and consume ethanol. Genes involved in the fermentation pathway that are not rate-limiting (Schaaff *et al.*, 1989; Pretorius, 2000), generally do not have duplicates in the yeast genome by (e.g. hexokinase, glucose-6-phosphate isomerase, phosphofructokinase, aldolase, triose phosphate isomerase, and phosphoglycerate kinase are all present in one isoform). Enolase has two paralogs (ENO1 and ENO2), where $f_2 = 0.946$. These are distantly related to a homolog known as ERR1, with the silent sites equilibrated. Phosphoglycerate mutase has three paralogs, GM1, GM2 and GM3, with silent sites that are essentially equilibrated.

[c] These pairs represent a family generated with a single duplication with $0.80 < f_2 < 0.86$, and subsequent duplication(s) in the derived lineages. Paralog pairs are considered only if they have with at least 100 aligned silent sites, and are not separated by more than 120 point-accepted mutations per 100 aligned amino acid sites (PAM units). $f_2 =$ fraction of nucleotides conserved at 2-fold-redundant codon sites only, and only at sites where the amino acid is identical.

not associated with fermentation that fall in the $0.80 < f_2 < 0.86$ window *are* part of a duplication block (see Table 1.2). The silent sites for most gene pairs associated with blocks are nearly equilibrated (with the prominent exception of ribosomal proteins), and therefore suggest that most blocks arose by duplications more ancient than duplications in the $0.80 < f_2 < 0.86$ window. Therefore, the hypothesis that a set of six time-correlated duplications (Table 1.2) generated the make–accumulate–consume strategy in yeast near the time when fermentable fruit emerged is not inconsistent with the whole-genome-duplication hypothesis.

This bioinformatics extends the paleoscience experiments across the yeast genome. As of today, this may not convert the story that the paleoscience experiments told into an acceptable narrative. But having a dozen gene duplications correlating with the result from the paleoexperiment helps. The next step is to extend this narrative. If the yeast genome shows evidence of this ecosystem innovation, then so should the genomes of the plants making the fruit, the fruit flies laying eggs in the fermentable fruit, moving down and up in the ecosystem. The end of the Cretaceous saw, in addition to the emergence of fruits, the extinction of the dinosaurs and the emergence of mammals and fruit flies (Baudin *et al.*, 1993; Ashburner, 1998; Barrett and Willis, 2001). For example, females of different fruit fly species position their eggs in fruits with different levels of fermentation (Hougouto *et al.*, 1982). Further, the impact of the introduction of alcohol into the ecosystem should have had impact on microorganisms other than *S. cerevisiae* that had contact with alcohol-rich media.

Likewise, many organisms other than *S. cerevisiae* participate in alcoholic fermentation before yeast takes over. In rotting fruits, *S. cerevisiae* becomes dominant after fermentation begins, while osmotic stress and pH, as well as ethanol, appear to inhibit the growth of competing organisms (Pretorius, 2000). Whereas the genomes of organisms that participate in the initiation of fermentation are not yet available, should they become so, they too can be examined for evidence of adaptive change.

Adding more information will not provide proof. Again, proof is not accessible in the real world. This means that at no point will the narrative evolve to the point where someone who is committed to disagreeing with the narrative not have the option to find a reason not to believe it. This was shown by the experiences with the non-classical carbocation and neutralist/selectionist dispute. But there is only one history of life on Earth. As enough lines of evidence converge on the model, interconnecting enough threads from chemistry, systems biology, ecology, and planetary science, the model will converge. And eventually the model building will end (Galison, 1987).

References

Allemann, R.K., Hung, R., and Benner, S.A. (1988) A stereochemical profile of the dehydrogenases of *Drosophila melanogaster*. *J. Am. Chem. Soc.* **110**: 5555–5560.

Allemann, R.K., Presnell, S.R., and Benner, S.A. (1991) A hybrid of bovine pancreatic ribonuclease and angiogenin. An external loop as a module controlling substrate specificity? *Prot. Eng.* **4**: 831–835.

Ashburner, M. (1998) Speculations on the subject of alcohol dehydrogenase and its properties in *Drosophila* and other flies. *Bioessays* **20**: 949–954.

Barnard, E.A. (1969) Biological function of pancreatic ribonuclease. Nature **221**: 340–344.

Barrett, P.M. and Willis, K.J. (2001) Did dinosaurs invent flowers? Dinosaur angiosperm coevolution revisited. *Biol. Rev.* **76**: 411–447.

Baudin, A., Ozier-Kalogeropoulos, O., Denouel, A., Lacroute, F., and Cullin, C. (1993) A simple and efficient method for direct gene deletion in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **21**: 3329–3330.

Benner, S.A. (2003) Interpretive proteomics: finding biological meaning in genome and proteome databases. *Adv. Enzyme Regul.* **43**: 271–359.

Benner, S.A. and Ellington, A.D. (1988) Interpreting the behavior of enzymes. Purpose or pedigree? *CRC Crit. Rev. Biochem.* **23**: 369–426.

Benner, S.A. and Allemann, R.K. (1989) The return of pancreatic ribonucleases. *Trends Biochem. Sci.* **14**: 396–397.

Benner, S.A. and Sismour, A.M. (2005) Synthetic biology. *Nat. Rev. Genet.* **6**: 533–543.

Benner, S.A., Rozzell, J.D., and Morton, T.H. (1981) Stereospecificity and stereochemical infidelity of acetoacetate decarboxylase (AAD). *J. Am. Chem. Soc.* **103**: 993–994.

Benner, S.A., Cannarozzi, G., Chelvanayagam, G., and Turcotte, M. (1997) *Bona fide* predictions of protein secondary structure using transparent analyses of multiple sequence alignments. *Chem. Rev.* **97**: 2725–2843.

Benner, S.A., Caraco, M.D., Thomson, J.M., and Gaucher, E.A. (2002) Planetary biology: paleontological, geological, and molecular histories of life. *Science* **296**: 864–868.

Berbee, M.L. and Taylor, J.W. (1993) Dating the evolutionary radiations of the true fungi. *Can. J. Bot.* **71**: 1114–1127.

Boulton, B., Singleton, V.L., Bisson, L.F., and Kunkee, R. E. (1996) Yeast and biochemistry of ethanol fermentation. In *Principles and Practices of Winemaking*, pp 139–172. Chapman and Hall, New York.

Bozzi, A., Saliola, M., Falcone, C., Bossa, F., and Martini, F. (1997) Structural and biochemical studies of alcohol dehydrogenase isozymes from *Kluyveromyces lactis*. *Biochim. Biophys. Acta* **1339**: 133–142.

Brown, H.C. (1977) *The Nonclassical Ion Problem*. Plenum Press, New York.

Cho, S., Beintema, J.J., and Zhang, J.Z. (2005) The ribonuclease A superfamily of mammals and birds: identifying new members and tracing evolutionary histories. *Genomics* **85**: 208–220.

Clarke, B. (1975) The contribution of ecological genetics to evolutionary theory: detecting the direct effects of natural selection on particular polymorphic loci. Genetics **79**: 101–108.

Collinson, M.E. and Hooker, J.J. (1991) Fossil evidence of interactions between plants and plant-eating mammals. *Philos. Trans. R. Soc. Lond. Ser. B Biological Sciences* **333**: 197–208.

Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C. (1978) A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* (Dayhoff, M.O., ed.), pp. 345–352. National Biomedical Research Foundation, Washington DC.

Ellington, A.D. and Benner, S.A. (1987) Free energy differences between enzyme bound states. *J. Theor. Biol.* **127**: 491–506.

Fernandez-Espinar, M.T., Barrio, E., and Querol, A. (2003) Analysis of the genetic variability in the species of the *Saccharomyces* sensu stricto complex. *Yeast* **20**: 1213–1226.

Fersht, A.R. (1977) *Enzyme Structure and Mechanism*. W.H. Freeman, New York.

Fleet, G.H. and Heard, G.M. (1993) Yeast growth during fermentation. In *Wine Microbiology and Biotechnology* (Heard, G.M., ed.), pp 27–54, Harwood Academic Publishers, Chur, Switzerland.

*Galison, P.L. (1987) *How Experiments End*. University of Chicago Press: Chicago.

Ganzhorn, A.J., Green, D.W., Hershey, A.D., Gould, R.M., and Plapp, B.V. (1987) Kinetic characterization of yeast alcohol dehydrogenases. Amino acid residue 294 and substrate specificity. *J. Biol. Chem.* **262**: 3754–3761.

*Gillespie, J.H. (1984) Molecular evolution over the mutational landscape. *Evolution* **38**: 1116–1129.

Gillespie, J.H. (1991) *The Causes of Molecular Evolution*, p. 336. Oxford University Press, Oxford.

Hey, J. (1999) The neutralist, the fly and the selectionist. *Trends Ecol. Evol.* **14**: 35–38.

Hougouto, N., Lietaert, M.C., Libion-Mannaert, M., Feytmans, E., and Elens, A. (1982) Oviposition-site preference and ADH activity in *Drosophila melanogaster*. *Genetica* **58**: 121–128.

Jermann, T.M., Opitz, J.G., Stackhouse, J., and Benner, S. A. (1995) Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature* **374**: 57–59.

Kellis, M., Birren, B.W., and Lander, E.S. (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* **428**: 617–624.

Kreitman, M. and Akashi, H. (1995) Molecular evidence for natural selection. *Ann. Rev. Ecol. Syst.* **26**: 403–422.

Lehman, W.P. (1973) *Historical Linguistics*. Holt, Reinhard, Winston, New York.

Levins, R. and Lewontin, R. (1985) *The Dialectical Biologist*. Harvard University Press, Cambridge, MA.

Lewontin, R.C. (1974) *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.

Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151–1155.

McGeehan, G.M. and Benner, S.A. (1989) An improved system for expressing pancreatic ribonuclease in *Escherichia coli*. *FEBS Lett.* **247**: 55–56.

McGovern, P.E. (2004) Fermented beverages of pre- and proto-historic China. *Proc. Natl. Acad. Sci. USA* **101**: 17593–17598.

Nambiar, K.P., Stackhouse, J., Stauffer, D.M., Kennedy, W.G., Eldredge, J.K., and Benner, S.A. (1984) Total synthesis and cloning of a gene coding for the ribonuclease S protein. *Science* **223**: 1299–1301.

Nambiar, K.P., Stackhouse, J., Presnell, S.R., and Benner, S.A. (1987) Expression of bovine pancreatic ribonuclease A in *E. coli*. *Eur. J. Biochem.* **163**: 67–71.

Nielsen, R. (2002) Mapping mutations on phylogenies. *Syst. Biol.* **51**: 729–739.

Pagel, M. (1999) Inferring the historical patterns of biological evolution. *Nature* **401**: 877–884.

Pauling, L. and Zuckerkandl, E. (1963) Chemical paleogenetics molecular restoration studies of extinct forms of life. *Acta. Chem. Scand.* **17**: S9–S16.

Powers, D.A. and Schulte, P.M. (1998) Evolutionary adaptations of gene structure and expression in natural populations in relation to a changing environment: a multidisciplinary approach to address the million-year saga of a small fish. *J. Exp. Zool.* **282**: 71–94.

Pretorius, I.S. (2000) Tailoring wine yeasts for the new millennium: Novel approaches to the ancient art of winemaking. *Yeast* **16**: 675–729.

Robson, B. and Garnier, J. (1993) Protein structure prediction. *Nature* **361**: 506.

Schaaff, I., Heinisch, J., and Zimmerman, F.K. (1989) Overproduction of glycolytic enzymes in yeast. *Yeast* **5**: 285–290.

Segel, I.H. (1975) *Enzyme Kinetics*. John Wiley and Sons, New York.

Somero, G.N. (1995) Proteins and temperature. *Annu. Rev. Physiol.* **57**: 43–68.

Stackhouse, J., Presnell, S.R., McGeehan, G.M., Nambiar, K.P., and Benner, S.A. (1990) The ribonuclease from an extinct bovid. *FEBS Lett.* **262**: 104–106.

Stryer, L. (1995) *Biochemistry*, 4th edn. W.H. Freeman and Company, New York.

Sun, G. (2002) Archaefructaceae, a new basal angiosperm family. *Science* **296**: 899–904.

Suppe, F. (1977) *The Structure of Scientific Theories*, 2nd ed. University of Illinois Press, Urbana, IL.

Swofford, D.L. (1998) *PAUP* Phylogenetic Analysis Using Parsimony Version 4*. Sinauer Associates, Sunderland, MA.

Thomson, J.M. (2002) *Interpretive Proteomics: Experimental Paleogenetics as a Tool to Analyze Function and Discover Pathways in Yeast*. Dissertation, University of Florida.

Thomson, J.M., Gaucher, E.A., Burgan, M.F., De Kee, D. W., Li, T., Aris, J.P., and Benner, S.A. (2005) Resurrecting ancestral alcohol dehydrogenases from yeast. *Nat. Genet.* **37**: 630–635.

Wills, C. (1976) Production of yeast alcohol dehydrogenase isoenzymes by selection. *Nature* **261**: 26–29.

Wolfe, K.H. and Shields, D.C. (2001) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**: 708–713.

Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**: 555–556.

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

Zhang, J.Z. and Nei, M. (1997) Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J. Mol. Evol.* **44**: S139–S146.