PERGAMON

Advan. Enzyme Regul. 43 (2003) 271–359

ADVANCES IN
ENZYME
REGULATION

www.elsevier.com/locate/advenzreg

Interpretive proteomics—finding biological meaning in genome and proteome databases

Steven A. Benner[a,b,c]

[a] Department of Chemistry, University of Florida, Gainesville FL 32611, USA
[b] Department of Anatomy and Cell Biology, University of Florida, Gainesville FL 32611, USA
[c] Foundation for Applied Molecular Evolution, P. O. Box 13174, Gainesville FL 32604, USA

Introduction

The 1990s was the decade of the genome. Today, complete draft sequences are available for the genomes of many eubacteria, several archaebacteria, several unicellular eukaryotes, several plants, and a growing collection of animals, including *C. elegans* (a worm), the fruit fly and mosquito, the mouse, and human.

As these sequences have accumulated, it has become increasingly apparent that new methods are needed to exploit the information that they (must) contain. Organic chemistry has always been driven by the discovery of new natural products, elucidation of their structures, and exploration of their behaviors. Gene sequences are no more (and no less) than the structures of natural products responsible for inheritance. The genome database, and the corresponding protein sequence database, therefore provide a new collection of natural product structures to study. These display every behavior of interest to chemists: conformation, supramolecular organization, combinatorial assembly, and catalysis are just a few.

At the same time, biomedical scientists are hoping that new insights into biology, disease, and treatment will be extracted from this collection of sequence data. They are adding to the data, comparing the expressed genetic inventory of diseased and normal tissues, and attempting to correlate genomic data with physiological function. Biomedical science *should* be revolutionized by genomic data. But how?

Genomic projects also present opportunities for the emerging fields of Geobiology, Planetary Biology and Astrobiology. Geobiology and Planetary Biology seeks to understand the relation between living organisms and their global environment. The history of life on Earth cannot be separated from the history of the planet. Each has defined the structure of the other. A major, almost visionary (at this

E-mail address: benner@chem.ufl.edu (S.A. Benner).

0065-2571/03/$ - see front matter © 2003 Elsevier Science Ltd. All rights reserved.
doi:10.1016/S0065-2571(02)00024-9

time) avenue for research seeks to combine the disparate traditions in molecular and physical sciences with a wealth of data from Natural History.

Astrobiology is defined as the study of the origin, evolution and distribution of life (including life on Earth) within the context of cosmic evolution. With the recent advances in planetary science, including landing on Mars and close inspection of the moons of Jupiter, specific features of terrean biochemistry have become important to astrobiology. In particular, it is important now to distinguish features of terrean life that reflect unique solutions to problems presented by life (in general) from those that do not. The first are likely to be mirrored in life that originated independently on other planets; the second are not. Unique solutions are likely to arise from constraints imposed by fundamental chemical reactivity (assumed to be universal) and Darwinian processes that drive organisms to optimize chemical behavior, also assumed to be universal. Astrobiological research with terrean genomes is needed to identify and distinguish chemical features of terrean life that reflect selection, neutral drift, and origins (Hey, 1999).

One consequence of large genomic databases is the ready availability of the evolutionary histories of families of proteins represented within the global proteome. After all of the genomes of all of the organisms on Earth have been sequenced, all of the encoded proteins will be recognizably built from ca. 10^5 independently evolving protein sequence “modules” (Riley and Labedan, 1997). For each of these, an evolutionary history can be built to include (a) a multiple alignment of the sequences of the proteins and genes in the module themselves, (b) an evolutionary tree, and (c) a reconstructed ancestral DNA and protein sequence for each branch point in the tree. Given a detailed model of biomolecular evolution, these histories can be used to connect sequence, structure, chemical reactivity, and biological function.

Over the past decade, we have used *Advances in Enzyme Regulation* and its associated conference to lay out tools that exploit evolutionary analysis of sequence data to solve problems in biological chemistry. These have included methods to identify functional regions of protein structure (Benner, 1989), methods that predict the conformation of proteins from a family of homologous sequences, methods that analyze evolutionary covariation at residues distant in the polypeptide chain, and methods that use protein structure prediction to detect distant homologs (Benner and Gerloff, 1991). More recently, we used the *Advances* venue to lay out tools that exploit evolutionary histories to extract information concerning protein structure, behavior, and function from a detailed understanding of how protein sequences divergently evolve under functional constraints (Benner et al., 1998). In a post genomic world, with volumes of sequence data from an unlimited number of organisms, these tools will be used to learn more from sequence data about living systems, their chemistry and their diseases.

We return again to this venue to describe the next generation of tools needed for interpretive proteomics. The first purpose of this lecture is to outline the nature of the tools and resources that are required to support an evolutionary analysis of genomic sequence data. Its second purpose is to provide the strategies and tactics that are needed to apply these tools. Its third purpose is to provide examples of where these tools, strategies, and tactics have been used to solve problems in biomedicine.

The time available for this lecture will limit the detail of the discussion. Fortunately, *Advances in Enzyme Regulation* has no page limits. For the written version of this lecture, therefore, we are limited only by the time that we have until the submission deadline that has been presented by the publisher. Even this has proven shorter than desirable. Therefore, we have been able to provide only part of the background of interest to the biomedical researcher. Much of the details must be found in references that we make to our previous work in the published and patent literature (Benner et al., 1998, 2002).

Tools and resources

The Role for Statistical and Mathematical Models in Interpretive Genomics and Proteomics

In its October 26, 2000 article entitled "Cool Jobs and How to Get Them" (Seth Stevenson, p. 104), the magazine *Rolling Stone* noted that a bioinformatics major was perfect for those who "want to cash in on the biotech gravy train without having to learn too much heavy science."

In this single sentence, *Rolling Stone* offered another of its penetrating insights into modern popular culture, capturing a paradox at the heart of contemporary bioinformatics. By "science," *Rolling Stone* was referring to the large body of empirical data describing how cells, tissues, organs, and organisms meet the challenges presented by their environments as individuals struggle to survive and reproduce. Another part of "science" is chemical, and concerns the molecules that help living systems meet these challenges. For two centuries in organic chemistry, and much longer in biology, the science has developed in its modern form by observation, theory, intuition, hypothesis, and experiment.

If the past can be used to anticipate the future, almost all of the tools that will be used to exploit genomic sequence data will come from "science." Few of the interesting features of the chemical structures found in biology will be captured by the statistical models that form the core of the studies of the student of bioinformatics. In part, this outlook reflects the well-recognized weakness of statistical methods relative to the complexity of the chemical and biological datasets. It also recognizes that the connection between chemical structure (a.k.a. "sequence") and molecular conformation, behavior, is far beyond the capabilities of any formal mathematical model at the present time.

The behavior of molecules (including protein molecules) is not correctly called "chaotic." All chemists (perhaps by faith) believe that the rationale connecting chemical structure with chemical behavior ultimately will be formally describable. But that formalism is far away. Further, as with all organic molecules, small changes in the structure of proteins can have either no detectable impact on behavior, or dramatic impact. The impact of changes at one site can be determined entirely by changes at other sites, or not at all.

This reality carries a clear message to students wishing to exploit genomic sequence data today. From a practical perspective, it is both desirable and necessary to use less rigorous methods with a bit of formal mathematical models. Statistics and mathematical formalisms are simply the handmaiden of humans as they do the activities that humans excel in (pattern recognition, intuition, conjecture, hypothesis, insight, and experiment).

We do not mean by these comments to disparage the contributions of statisticians or statistical models in bioinformatics. Some of our best friends are statisticians. We ourselves use statistical formalisms, develop them, and appreciate their power. We do mean, however, that statistical approaches must be used appropriately. Further, non-statistical methods should not be criticized because they do not meet the standards of rigor that are typical in mathematics. To do so would slow the application of genomic sequence data to biology and biomedicine.

One of the best ways to use formal mathematical models in bioinformatics is to recognize that they (generally) treat gene and protein sequences as if they were linear strings of letters, lacking either form (fold) or the functional behavior that comes with the fold (Benner, 2002). Such formalisms can serve as null hypotheses, models for how genes and proteins would divergently evolve *if they were* formless, functionless strings of letters. By observing how genes and proteins actually divergently evolve, and comparing the observations with the null hypothesis, a signal concerning form and function can be extracted. Thus, formal mathematical models for sequence evolution serve as invaluable starting points for functional analysis of gene sequences.

This insight laid the foundation a decade ago for the first set of tools that convincingly predicted the folding of proteins from sequence data alone (the "protein structure prediction problem") (Benner and Gerloff, 1991). Given a set of sequences of homologous proteins diverging under functional constraints, patterns of variation and replacement that are different from those anticipated by simple statistical models allow the identification of sites containing amino acids whose side chains protrude from the surface of the protein, those that are found inside the folded structure, and those that are near the active site. Correlation of patterns of replacement between positions nearby in the sequence makes it possible to assign elements of secondary structure, alpha helices and beta strands, to segments of the protein. Correlation of patterns of replacement at sites distant in the linear sequence provides information about how the protein folds (Benner and Gerloff, 1991).

The power of evolutionary-based tools that use formal mathematical models as a null hypothesis has been demonstrated in a large number of bona fide predictions of protein structure, those made and announced before an experimental structure was known. The first of these, for protein kinase, was reported on these pages (Benner and Gerloff, 1991). The evolution-based strategy has been applied to a large number of additional proteins, including the hemorrhagic metalloproteinase (Gerloff et al., 1993), ribonucleotide reductase (Tauer and Benner, 1997), heat shock protein 70 (Gerloff et al., 1997), phospho-beta-galactosidase (Gerloff and Benner, 1995), and synaptotagmin (Benner et al., 1995), the last three within the context of the project known as a Critical Assessment of Structure Prediction (CASP). In each case, these predictions have been sufficiently accurate to solve specific biological problems associated with the protein family. These are reviewed elsewhere (Benner et al., 1997).

A directly analogous type of evolutionary analysis allows biological scientists to infer functional information from a set of aligned homologous sequences. Indeed, in many cases, structure prediction is the first step in this chain of inference. Many of these inferences are at the level of hypothesis, of course.

In each case, however, a formal description of evolutionary processes is not valuable because it does capture the details of proteins sequence evolution, but rather because it does not. It therefore serves as a null hypotheses. Comparing how proteins actually divergently evolve with this hypothesis generates a signal that provides information about function.

In the discussion below, we use formal models in the context of insight, intuition, and non-formal models to build a broad suite of tools, tactics, and strategies to interpret genomic sequences. This suite is known by the rubric FIREBIRD (Functional Inference from Reconstructed Evolutionary Biology). The FIREBIRD suite of tools offers a powerful framework for analyzing function in proteins, identifying targets of biomedical interest, and guiding pre-clinical drug development in animal models, inter alia. When applied to whole genomes, the framework identifies metabolic pathways and regulatory networks, permits the correlation of the life history of a lineage with its historical past, and captures interconnections that will move the biomedical researcher and biological chemist from the genome to the planet.

The only metric for evaluating this combination is whether it is useful to the biological and biomedical researcher. This, in turn, is done through applied use. There is no metric that a statistician would approve for measuring success. But for the biological and biomedical research communities, this is how success must be measured.

Aligning Two Protein Sequences. Simple Formalisms

We assume that a fraction of our readers is not familiar with the basic models that describe the divergent evolution of protein sequences. We therefore outline these here. The more advanced reader might wish to skip this section.

Even in the 1960s, when very few sequences were available, tools were sought to compare two protein sequences with the purpose of asking the question: Are these two sequences related by common ancestry? To answer this question, the two sequences needed to be aligned, where the correct alignment matches amino acids in the two sequences so that the matched amino acids were true descendants of a single amino acid in the ancestral protein.[1]

[1] An entertaining collection of examples where statisticians, attempting to impose statistical formalisms upon science, have slowed progress in biomedical research can be found in peer review settings. For example, one statistician suggested that tools to develop hypotheses should not be funded unless he could first assess "the reliability of the hypotheses" that would emerge. This request, of course, is risible for an experimental scientist, who knows that the value of a hypothesis in science relates more to its experimental testability than to its reliability. Indeed, some of the most valuable hypotheses in science have been wrong.

As the historical past was considered to be unknowable, the alignment task came, in practice, to be the task of aligning two sequences with a scoring matrix in a way that identifies the alignment giving the best score. Algorithms that produced such an alignment were adapted for protein sequences by Needleman and Wunsch (1970), and by Smith and Waterman (1981).

The availability of these algorithms encouraged the development of 20 × 20 matrices that give the probability of two amino acids being matched in a sequence alignment by reason of common ancestry, divided by the probability that the two will be matched by random chance. This effort was pioneered by Margaret Dayhoff and her group at the National Biomedical Research Foundation.

Key to this scoring was the concept of a "distance" between two protein sequences. The distance was a measure of the number of amino acid replacements per site. Its calculation required the development of a model of sequence evolution, which we will call here the Standard Model.

Consider a protein sequence exactly 100 amino acids in length. Let us assume that each of the 20 amino acids is equally likely to be at each of the 100 positions in the sequence. This means that on average, each of the amino acids will appear in the sequence five times.[2]

Now, let us assume that the protein suffers duplication. The duplicates are, by definition, 100% identical. We now allow one of the duplicates to suffer replacements. Let us replace one amino acid in one sequence (we will call it the "diverging sequence") by another of the 20 standard amino acids. We will chose the site of the replacement at random (meaning, in this case, that the probability of any individual site suffering a replacement is 0.01). We will also assume that each of the 19 amino acids is equally likely to make the replacement.

Following this single replacement in one of the duplicates, the two proteins are no longer identical in sequence. In fact, the two sequences are now 99% identical. More conveniently, we will refer to the "fractional identity", and represent this by a number that can range from zero to unity. Here, fractional identity is 0.99. We will call the distance between the two proteins as being 1 PAM unit, which means that they have suffered one Point Accepted Mutation in the time since they diverged.[3]

Now, allow the diverging sequence to suffer another replacement. Again, the site suffering the replacement is chosen at random; each of the sites has a 0.01 likelihood of changing in the second cycle of replacement, with the site that suffered a change in the first cycle no more likely to suffer change in the second. Again, the replacing amino acid is chosen at random; each of the 19 other amino acids is equally likely to appear at the site after the second round of replacement. What is the percentage identity between the two diverging sequences after two cycles of replacement?

[2] Technically, proteins $p1$ and $p2$ cannot not correctly be said to be descendants of p; more correct is to say that the genes that encode proteins $p1$ and $p2$ are descendants of the gene that encoded p. The expression creates no confusion, however, and we will use it with the meaning stated here..

[3] For any particular sequence, of course, the number of amino acids is distributed around this 5% expectation value.

Naively, one might say that the two sequences will be 98% identical, as one sequence has diverged by two PAM units. In fact, this is not exactly the case, if a standard stochastic model for sequence evolution is followed. There is a 99% chance that the second site in the diverging sequence to suffer a replacement, chosen randomly, will be different from the first site that suffered a replacement. In this case, the percentage identity of the two sequences following the second round of replacement will be 98%. But there is a 1% chance that the second site chosen will be the *same* as the site that suffered a replacement in the first round of replacement. And if the same site suffers a second replacement in the second round, then there is a 1 in 19 chance ($P = 0.05263$) that the replacement will restore the original amino acid to that site. As a consequence, after two replacements, the diverging protein will have an expectation value of more than 98% sequence identity; the precise value is 98.0105% identity.

This number 98.0105% is not very different from 98.0000%, of course. But as the sequences diverge through further rounds of replacement, the odds increase that the next replacement will occur at a site that has not already suffered a replacement, and therefore not further differentiate the derived sequence from the original sequence. Likewise, the odds increase as replacement continues that the round of replacement will restore the amino acid found at that position in the original sequence.

For this reason, the percent identity between two diverging sequences is not a linear function of the number of replacements. Indeed, after suffering a large number of replacements, the sequence will reach an "equilibrium." At the equilibrium, the likelihood of a replacement restoring the amino acid found in the original sequence becomes equal to the likelihood that a replacement will make the descendent still more dissimilar from the parent. If the 20 amino acids are introduced/lost with equal frequency, this equilibrium will be achieved when the sequences are 5% identical. This level of identity is shared, of course, between two random sequences that are composed of 20 amino acids, chosen with equal frequency.

This process is well understood as an "approach to equilibrium" problem. After an indefinitely large number of replacements, the two sequences will be 5% identical. Under these circumstances, an additional replacement has a 95% chance of affecting a site that distinguishes the two sequences, with a 1 in 19 chance of increasing similarity. At equilibrium, the probability of a replacement increasing similarity is the same as the probability of a replacement decreasing similarity. The result is, in a word, equilibrium.

Fig. 1 shows a correlation between percentage identity and PAM distance. In practice, one can use the curve to determine the PAM distance between any two sequences. One simply determines the percentage identity, goes to the figure, finds the percentage identity on the *y*-axis, and reads off the PAM distance on the *x*-axis.

The assumption that all amino acids occur with equal frequency, and that all replacements were equally likely, is clearly incompatible with reality in natural protein sequences. In natural protein sequences, some amino acids are more abundant than others, and some replacements evidently occur with higher probability than others. The relative abundance of different amino acids (and the

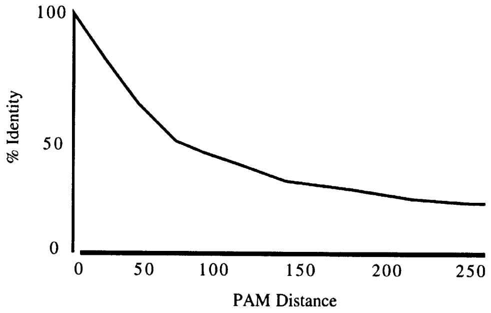


Fig. 1. Percent identity between two sequences falls off asymptotically with increasing number of mutations, which are measured by the PAM (point accepted mutations) distance between two sequences. Given a simple stochastic model for amino acid replacement (all sites suffer replacement independently, patterns of replacement are time-invariant, all 20 amino acids equally represented in the sequence), after infinite time, the two sequences will have reached an equilibrium identity of 5%.

Table 1
Frequencies of 20 standard amino acids in protein sequences[a]

9.14%	Leu	7.20 (average)	6.23%	Glu	4.12 (average)
7.23%	Ser	6-fold	5.80%	Lys	2-fold
5.22%	Arg		5.19%	Asp	
			4.39%	Asn	
7.58%	Ala	6.47 (average)	4.17%	Gln	
7.18%	Gly	4-fold	3.93%	Phe	
6.48%	Val		3.24%	Tyr	
5.94%	Thr		2.25%	His	
5.18%	Pro		1.85%	Cys	
5.35%	Ile	5.35 (average)	2.30%	Met	1.83 (average)
		3-fold	1.35%	Trp	1-fold

[a] Organized by number of codons.

redundancies in their coding systems) is shown in Table 1, extracted from a recent version of GenBank.

Dayhoff therefore enhanced the model to reflect different likelihoods that different amino acids would occupy any particular site, assuming that occupancy at any site mirrored the occupancy of the average site. She collected a set of aligned protein sequence pairs, for proteins whose sequences had only slightly diverged (let us say for sake of discussion, by one PAM unit). Dayhoff tabulated the probabilities of all of the A_i/A_j matches in her set of aligned protein sequence pairs. From these she built a 20×20 matrix showing the probability of each A_i being matched to every other in her dataset. The matrix terms $M_{i,i}$ reflected the probability that the match was an identity; the matrix terms $M_{i,j}$ reflected the probability that the match was a non-identity.

To obtain a scoring matrix that could be used in a dynamic programming algorithm, Dayhoff then modelled the probability that A_i would be matched to A_j by reason of random chance. The likelihood of an A_i/A_j match, under her probabilistic model, would depend on $\{A_i\}$ and $\{A_j\}$, the frequencies of A_i and A_j in the database as a whole, f_i and f_j. She therefore tabulated these from her database. She then divided every A_i/A_j term by f_i and f_j, and normalized the data to reflect the changes estimated for two proteins that had diverged by only one PAM unit. Last, she constructed a new "log odds" matrix, a 20 × 20 matrix that contained the logarithms of the normalized ratios, and multiplied these by 10.

We can do the same thing for aligned pairs of sequences where the partners have diverged by more than one PAM unit. We can collect a set of pairwise alignments where the pairs are ca. 91% identical (having diverged by ca. 10 PAM units) and repeat the process, generating a replacement matrix that reflects greater divergence, normalize it for the frequencies of the amino acids in the database as a whole, and compare it with the theoretical matrix.

This process is indeed feasible, but only to a point. For a replacement matrix to make biological sense, one needs to be certain that the pairwise alignments that generate the replacement data match amino acids that are descendants from a single amino acid in the ancestral protein. This means, in turn, that one needs to be certain that one has the correct alignment in the pairs of alignments that are used to generate the replacement matrix.

This is simple enough to ensure when the two proteins are 90% identical, and the pairwise alignment has no gaps. But once the sequences have diverged further, and gaps are introduced, the correct alignment is not so clear. At some point, one begins to worry about whether the sites paired in the alignment are truly homologous. If they are not, the replacement matrix is tabulating evolutionary non-sense. This, it turns out from empirical data, occurs in typical proteins as ca. PAM 100–120 (Benner et al., 1993).

Dayhoff took a different approach. She assumed that the patterns of amino acid substitution are the same in a pair of proteins separated by one PAM and in protein pairs separated by 10 PAM units. If this is the case, we should be able to simply multiply the first 1 PAM replacement matrix by itself 10 times to obtain a matrix that described amino acid matching in two proteins that are 10 PAM units distant. This is equivalent to raising the PAM 1 matrix to its 10th power.

This was the logic that was used by Dayhoff to construct what is now known as the Dayhoff matrix for scoring sequence alignments. Dayhoff collected data from pairs of sequences 5–10 PAM units distant. She then took the replacement matrices that she obtained, normalized them for amino acid frequencies, and powered them to get the equivalent of a 250 PAM matrix. She chose this distance because she felt that this was the most distant sequence that anyone would ever be productively able to align.

She then noted that this matrix contained a large number of fractional terms. Recognizing that logarithms are easy ways of representing small numbers (and anticipating the use of the matrix as a scoring matrix, see next section), she replaced the terms in the matrix by their logarithms (base ten). She then multiplied these logarithms by 10, for no good reason except to get a majority of the matrix elements to lie between 1 and 10. A typical mutation matrix is shown in Fig. 2.

C	12.1																			
S	0.9	2.1																		
T	-1.5	1.5	2.4																	
P	-2.7	1.4	0.6	6.5																
A	-1.7	1.4	1.7	1.1	2.5															
G	-1.3	0.8	-0.5	-1.7	0.8	5.8														
N	-1.6	1.2	0.5	-1.1	0.0	-0.1	3.6													
D	-3.7	-0.4	-1.2	-2.8	-0.6	0.8	2.5	5.2												
E	-4.7	-1.2	-1.6	-2.6	-0.7	0.5	1.1	4.4	5.2											
Q	-3.2	-1.4	-1.7	0.1	-1.7	-1.6	0.1	0.6	2.1	5.3										
H	-1.2	-0.9	-1.7	-0.4	-2.1	-2.1	1.4	0.1	-0.2	3.2	6.1									
R	-0.4	-0.9	-1.3	-1.3	-1.7	-0.1	-0.1	-1.5	-0.4	2.5	1.8	5.1								
K	-2.8	-1.2	-1.1	-2.3	-1.9	-1.4	1.0	-0.2	0.9	2.5	0.9	4.3	5.6							
M	-3.7	-1.3	0.6	-1.8	-0.2	-3.7	-2.5	-4.3	-4.1	-3.1	-3.4	-3.0	-2.9	4.8						
I	-3.6	-1.2	0.7	-2.0	0.1	-3.4	-2.5	-4.2	-4.1	-3.8	-3.7	-3.8	-3.8	4.0	4.4					
L	-3.8	-1.5	-0.4	-0.2	-1.3	-4.6	-3.4	-5.3	-5.0	-2.4	-2.2	-3.2	-4.1	2.9	2.4	4.8				
V	-3.1	-0.9	0.6	-1.6	0.7	-2.3	-2.4	-3.3	-3.0	-3.5	-3.8	-3.7	-3.8	3.3	3.9	1.9	4.0			
F	-0.1	-1.8	-2.4	-3.2	-3.2	-5.7	-3.5	-5.7	-6.7	-4.4	0.1	-4.9	-6.3	-0.1	0.0	2.4	-0.5	8.3		
Y	2.6	-1.8	-3.4	-3.8	-4.0	-4.9	-0.9	-2.3	-4.1	-1.4	4.4	-2.6	-4.0	-3.6	-3.3	-1.6	-3.8	5.6	9.5	
W	1.6	-2.9	-2.6	-4.8	-4.3	-1.7	-4.4	-6.3	-5.6	-2.6	-2.8	2.0	-1.4	-4.4	-5.0	-3.0	-4.8	-1.6	-0.3	14.7
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

Fig. 2. A Dayhoff matrix. The numbers are 10 times the logarithm of the probability that two index amino acids will be aligned by reason of common ancestry, divided by the probability that the two amino acids will be aligned by random chance, for two proteins that are separated by 250 PAM units (that is, have suffered 250 amino acid substitutions per 100 amino acid residues).

The *i,j* element of the Dayhoff matrix arises from an empirical measure of the probability that amino acids *i* and *j* will be matched in an alignment of a pair of proteins related by common ancestry. Because the terms are normalized for the frequencies of *i* and *j* in the database, they are normalized for the probability that amino acids *i* and *j* will be paired by random chance. Thus, the elements of a Dayhoff matrix can be viewed as the logarithm of the probability that two amino acids will be matched by reason of divergent evolution (for the number of PAM units specified in the matrix) divided by the probability that they would be matched by random chance.

A Dayhoff matrix can be used to score to a pairwise alignment. Consider a sequence alignment exactly one position in length, with an arbitrary PAM distance between them. The Dayhoff matrix element represents the probability that the two amino acids arose by divergence from a common ancestor at that PAM distance, divided by the probability that they arose by random chance. Simple enough.

Now consider an alignment of two proteins. If we treat each site as independently evolving, the probability of the two dipeptides arising from common ancestry, divided by the probability that they each arose by random chance, is the product of that probability at the first position in the dipeptide times that probability at the second position in the dipeptide.

As the Dayhoff matrix records the logarithms of these probabilities, and as the logarithm of the product of two probabilities is the same as the sum of the logarithms of the two probabilities. That is, to score the alignment, we add the term of the Dayhoff matrix for the first position to the term of the Dayhoff matrix for the second position. This is the logarithm of the probability that the two dipeptides arose by divergence, divided by the probability that they arose by random chance.

This process can be extended indefinitely. Fig. 3 shows, for example, the alignment of two segments, both 17 sites long, of the alcohol dehydrogenases from the horse and the human. Between each is written the number taken from the appropriate *i,j* matrix of the Dayhoff matrix in the figure.

The Dayhoff matrix also offers an alternative way to determine the PAM distance between two sequences. Consider a pair of aligned protein sequences having a PAM distance of 10. Consider also a series of matrices constructed for proteins separated by 0 PAM units, 1 PAM unit, by 2 PAM units, by 3 PAM units, and so on, up to PAM 100. As the PAM distance increases, the scores given to on-diagonal terms grow smaller, and the scores given to off-diagonal terms grow larger. The score given to any particular alignment will differ depending on the PAM matrix used. Thus, for

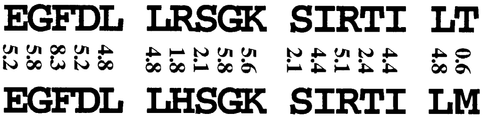


Fig. 3. Scoring an alignment between two segments, each 17 sites long, of the alcohol dehydrogenases from human and horse. The sum of the terms is 73.2, meaning that the probability of this pairing occurring by common ancestry, divided by the probability of it occurring by random chance, is equal to $10^{7.32}$ = ca. 2×10^7.

two proteins that are largely identical in sequence, a low PAM matrix will give a higher score than a high PAM matrix, simply because the low PAM matrix scores identities higher. Conversely, for two proteins that have few sequence identities, a high PAM matrix will give a higher score than a low PAM matrix, simply because the high PAM matrix scores non-identities higher. The consequence of this is that the PAM matrix that gives the highest score for a pairwise alignment is the one scaled to the PAM distance that actually separates the two sequences. Thus, one can determine the PAM distance by the process of comparing the scores given to the alignment using different PAM matrices. The PAM of the matrix that gives the highest score is the PAM distance separating the two sequences. Indeed, from this process one can also get a variance on the PAM distance, something that is not possible by simply inspecting the curve in Fig. 1.

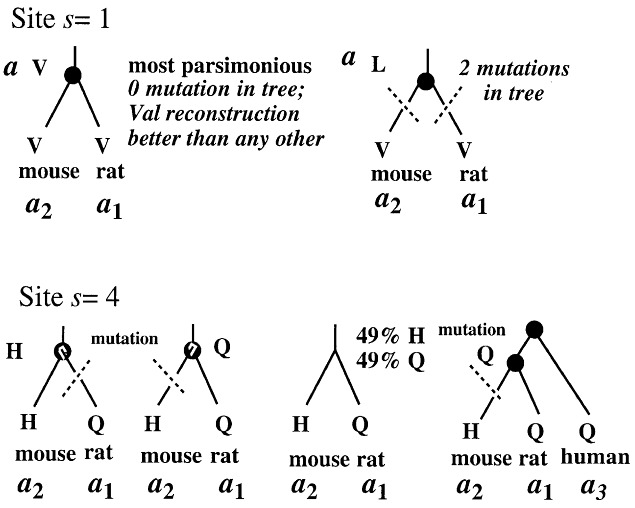
The Dayhoff matrix has also been misused. For example, the 250 PAM matrix was offered as a default on many sequence alignment programs, and was used to score the alignments of pairs of protein that had diverged much less. This was incorrect, although the consequences of using a PAM matrix that does not match the distance that separates two protein sequences are not usually severe.

Other authors have treated the log odds values for the alignment of amino acids A_j and A_i as the equivalent of a transition probability for converting A_j to A_i. This creates a paradox. When constructing a pairwise alignment, the score of A_j matched to A_i equals the score of A_i matched to A_j. But if these scores are used as transition probabilities, this equivalency implies that the rate constants $r_{i\rightarrow j}$ and $r_{i\rightarrow j}$ are equal, which implies that the ratio of amino acids in the database at equilibrium $\{\mathrm{A_j}\}/\{\mathrm{A_i}\}=1$. This is, of course, not the case, as is captured in another part of the Dayhoff formalism, where the probability of ratio of A_i being matched to A_j by random chance is normalized based on an empirical measurement of $\{A_j\}$ and $\{A_i\}$ in the database as a whole.

Simple Models for Comparing Homologous Protein Sequences

Tools to align two protein sequences to assess the likelihood that the two sequences are homologous must be different from tools that infer the evolutionary relationships between sequences, or the structure of ancestral proteins. Linus Pauling and Emil Zuckerkandl pointed out 40 years ago that the sequences of proteins from modern organisms might be used to construct phylogenetic trees and reconstruct the sequences of ancestral proteins from extinct organisms (Pauling and Zuckerkandl, 1962). Among their other insights, Pauling and Zuckerkandl pointed out that it might be possible to resurrect, through chemical synthesis, ancient proteins from extinct organisms. This task was not achieved for another 25 years, first in our laboratory (Stackhouse et al., 1990) and independently in the laboratory of the late Allan Wilson at Berkeley (Malcolm et al., 1990). This represented the start of a new discipline, known today as experimental paleobiochemistry (Chandrasekharan et al., 1996).

Pauling and Zuckerkandl had simple models in mind when they proposed the construction of trees and reconstruction of ancestral sequences, which emerged as a



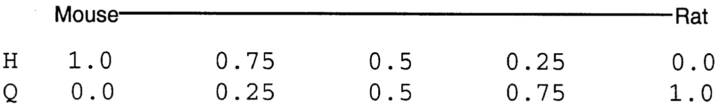
rat (p_1) VPIHKVQDDTKTLIKTIVTRINDISHTQSVSARQ
mouse (p_2) VPIQKVQDDTKTLIKTIVTRINDISHTQSVSAKQ
ancestor (p) VPIQKVQDDTKTLIKTIVTRINDISHTQSVSAKQ
human (p_3) VPIQKVQDDTKTLIKTIVTRINDISHTQSVSSKQ

Fig. 4. Reconstruction of ancestral states of leptin in the ancestor of mouse and rat. The ambiguous reconstrution at site $s = 4$ is resolved with the human sequence as the outgroup.

procedure of "maximum parsimony". According to the rule of parsimony, the best tree relating protein sequences to reconstructed ancestral sequences is the one that obtains the derived sequences from ancestral sequences with the smallest number of independent evolutionary events. This is illustrated in Fig. 4, where site s in ancestral protein p holds amino acid residue a if the amino acid residues a_1 and a_2 at the aligned positions in the two proteins derived from a, p_1 and p_2, are the same. If a_1 and a_2 are different, then no most parsimonious reconstruction exists for site s in the ancestor; the parsimony algorithm formally fails to provide a reconstruction at that site.

A variety of software packages provide parsimonious reconstructions of ancestral states given a set of homologous sequences. One of these is the MacClade software package (Maddison and Maddison, 1992). Given a set of previously aligned multiple sequences, and a user-selected tree, MacClade will return (in a colorful form) models the residues at each site in the multiple sequence alignment at each node.

A small modification of parsimony procedures creates probabilistic models for reconstructed ancestral sequences. For example, when a_1 and a_2 in the two sequences are different, residue a at site s might be reconstructed as a_1 with a 50% probability, and as a_2 with a 50% probability. Alternatively, the reconstructed residue can be defined for each point along the line connecting a_1 and a_2. Here the, probability of a_1 being present at site s in ancestor a drops smoothly from 1.0 to 0.0, and the



H	1.0	0.75	0.5	0.25	0.0
Q	0.0	0.25	0.5	0.75	1.0

Fig. 5. The probabilistic values for the amino acid residue at position 4 of the rat–mouse leptin morph smoothly along an evolutionary tree from 1.0 (for H) at the mouse leaf to 1.0 (for Q at the rat leaf).

corresponding probability that a_2 being present at site s in ancestor a increases smoothly from 0.0 to 1.0, as one proceeds from p_1 to p_2.

In the post-genomic world, it is more than likely that a sequence of a third protein, p_3, is available that is homologous to p_1 and p_2. This sequence can "root" the tree represented by the line between p_1 and p_2, by identifying the point on the line where the sequence of the reconstructed ancestral protein is most like p_3. It might also resolve the reconstruction of the sequence of a at the root. Thus, if a_3 a_3 is the same as a_1, and a_1 and a_3 are both different from a_2, then $a = a_1 = a_3 \neq a_2$. Only when $a_1 \neq a_2 \neq a_3$ does parsimony not yield a reconstruction for a. This is illustrated in Figs. 4 and 5.

More advanced statistical models represent evolution as a dynamic process. In 1989, Brian Seed and his coworkers at the Harvard Medical School applied a generator to a sequence to describe amino acid replacement within it (Stamenkovic et al., 1989). The generator has $19 \times 19 = 361$ parameters, describing the rate constants for the conversion of each of the 20 amino acids to each of the other 20 amino acids $r_{i\to j}$. Assuming a stationary amino acid composition, where "stationary" means that the composition is unchanged over the period of evolutionary time captured within the tree, the ratio of the rate constants $r_{i\to j}/_{j\to i}$ is equal to the ratio of amino acids in the database $\{A_j\}/\{A_i\}$.

Advanced Statistical Models Needed for Evolutionary Analyses of Protein Sequences

The Standard Model outlined above stands behind most analysis of divergent evolution in protein sequences. In this model, each site in a protein sequence suffers replacement independent of every other site. Patterns of replacement are presumed to be identical at each site, depending only on the specific amino acids involved in the replacement event. Future replacements are presumed to be independent of previous replacements.

The Standard Model proves to be only a poor approximation for the reality of protein sequence divergence. The first exhaustive matching of a modern sequence database (Gonnet et al., 1992) permitted a large number of empirical studies that showed how different real protein divergent evolution is from that expected by the Standard Model.

Various modifications of the Standard Model have been proposed to capture features of protein sequence evolution that are not captured within the Standard Model. For example, three decades ago, Fitch and Markowitz pointed out that different sites in a real protein sequence might be under different selective constraints

(Fitch and Markowitz, 1970). At some sites, a replacement might not change the behavior of the protein detectably, or create a change in behavior that has no impact on the fitness of the host organism. Natural selection should tolerate replacements at these sites. At other sites, however, a replacement might destroy the catalytic activity of the protein (for example). Natural selection should not tolerate replacements at these sites.

These considerations led to the "covarion" model for protein sequence evolution. The covarion model recognizes that different sites in a sequence will suffer replacement at different rates. A statistical model can help capture this feature of the behavior of real proteins. A single parameter gamma distribution is commonly used to describe the distribution of mutability at various sites in a protein sequence. This single parameter is generally written as alpha (α), and describes the shape of the distribution. This distribution can accommodate a wide range of rapidly and slowly evolving sites.

The covarion model is extremely valuable when attempting to estimate a distance between two sequences. If a second replacement is more likely to occur at a site that has already suffered a replacement than it is at a site that has not already suffered a replacement, then simple reference to Fig. 1 cannot be used to estimate the number of replacements that has occurred in the time separating two sequences. Systematically, a second and third replacement will not lower sequence identity as much as expected for a second and third replacement given the Standard Model. Reference to Fig. 1 will systematically underestimate the number of replacements that have actually occurred relative to the percent identity.

The Standard Model also assumes that each position in a protein sequence suffers replacement independent of all other positions. Examination of real proteins shows, that changes at one position are frequently correlated with change at others in the sequence. A decade ago, for example, we showed that replacements at adjacent sites were strongly correlated, both in frequency and in kind (Cohen et al., 1994). Indeed, simple inspection of any multiple sequence alignment shows that replacements are not randomly distributed along its length (Fig. 6).

Further, the Standard Model assumes that the transformation matrix that describes the second PAM unit of replacement (or, for that matter, the *n*th unit of replacement) is the same as the transformation matrix that describes the first round of mutation. This assumption is inherent in the notion that one can power the PAM 1 matrix *n* times to get the PAM *n* matrix. A decade ago, we showed empirically that patterns of amino acid replacement at longer distances are not the same as those patterns at shorter distances (Gonnet et al., 1992). Because the classical Dayhoff matrix is calculated from pairwise alignments of very similar sequences, the classical Dayhoff matrix records a pattern of amino acid replacement that is quite close to that expected from an analysis of the genetic code, and quite different from that expected from a need to conserve the functional chemistry of amino acid side chains.

Last, real protein sequences can suffer insertion and deletion events, instances where segments of a protein sequence are added (inserted) or lost (deleted). Collectively, insertions and deletions are known as "indels", since pairwise sequence analysis generally does not tell us which process generated a gap. To accommodate

```
                      0         0         0         0         0         0         0
                      1         2         3         4         5         6         7
                      0         0         0         0         0         0         0
                 .    |    .    |    .    |    .    |    .    |    .    |    .    |
                 .    |    .    |    .    |    .    |    .    |    .    |    .    |
Mouse        VPIQKVQDDTKTLIKTIVTRINDISHTQSVSAKQRVTGLDFIPGLHPILSLSKMDQTLAVYQQVLTSLPSQNV
Rat          VPIHKVQDDTKTLIKTIVTRINDISHTQSVSARQRVTGLDFIPGLHPILSLSKMDQTLAVYQQILTSLPSQNV
Macaca       VPIQKVQSDTKTLIKTIVTRINDISHTQSVSSKQRVTGLDFIPGLHPVLTLSQMDQTLAIYQQILINLPSRNV
Pongo        VPIQKVQDDTKTLIKTVITRINDISHTQSVSSKQKVTGLDFIPGLHPILTLSKMDQTLAVYQQILTSMPSRNV
Pan          VPIQKVQDDTKTLIKTIVTRINDISHTQSVSSKQKVTGLDFIPGLHPILTLSKMDQTLAVYQQILTSMPSRNM
Gorilla      VPIQKVQDDTKTLIKTIVTRISDISHTQSVSSKQKVTGLDFIPGLHPILTLSKMDQTLAVYQQILTSMPSRNM
Homo         VPIQKVQDDTKTLIKTIVTRINDISHTQSVSSKQKVTGLDFIPGLHPILTLSKMDQTLAVYQQILTSMPSRNV
# AmAcids    1112111211111111221112111111111221211111111111121111211111121112122211212
Anc-Rhesus          x                                       x    x      x     xx
Anc-ProH                                       x                                x
Hominid                       xx   x                                                  x
secondary       hhhhhhhhhhhhhhhhhhhh                           hhhhhhhhhhhhhhhh    hhh
structure
                   0         0         1         1         1         1         1
                   8         9         0         1         2         3         4
                   0         0         0         0         0         0         0
              .    |    .    |    .    |    .    |    .    |    .    |    .    |
Mouse        LQIANDLENLRDLLHLLAFSKSCSLPQTSGLQKPESLDGVLEASLYSTEVVALSRLQGSLQDILQQLDVSPEC
Rat          LQIAHDLENLRDLLHLLAFSKSCSLPQTRGLQKPESLDGVLEASLYSTEVVALSRLQGSLQDILQQLDLSPEC
Macaca       IQISNDLENLRDLLHLLAFSKSCHLPLASGLETLESLGDVLEASLYSTEVVALSRLQGSLQDMLWQLDLSPGC
Pongo        IQISNDLENLRDLLHVLAFSKSCHLPWASGLETLDRLGGVLEASGYSTEVVALSRLQRSLQDMLWQLDLSPGC
Pan          IQISNDLENLRDLLHVLAFSKSCHLPWASGLETLDSLGGVLEASGYSTEVVALSRLQGSLQDMLWQLDLSPGC
Gorilla      IQISNDLENLRDLLHVLAFSKSCHLPWASGLETLDSLGGVLEASGYSTEVVALSRLQGSLQDMLWQLDLSPGC
Homo         IQISNDLENLRDLLHVLAFSKSCHLPWASGLETLDSLGGVLEASGYSTEVVALSRLQGSLQDMLWQLDLSPGC
# AmAcids    2112211111111112111111111211322111222221221111211111111111221111212111111121
Anc-Rhesus                             x           x
Anc-ProH                    x          x       x         x
Hominid                                    x                     x
secondary    hhhhhhhhhhhhhhhhhhhh                 hhhh      hhhhhhhhhhhhhhhhhhhggg
structure                   -   binding site -
                                  SCSLPQTSGLQKPES
```

Fig. 6. Alignment of a number of leptin protein sequences. The numbers beneath the multiple sequence alignment indicate the number of positions found at each site. The short peptide written below the alignment binds at the active site of the leptin receptor. For the secondary structure, "h" = alpha helix. Even by eye, the variation in the leptin family does not appear to be randomly distributed.

indels, the Standard Model assigns a probability to indel events, in particular, a cost for their introduction, followed by an incremental cost as they become longer. This approach conveniently fits dynamic programming algorithms for aligning sequences, which is presumably why it is so frequently used. A decade ago, we developed a different empirical model for gapping (Benner et al., 1993). Most sequence analyses tools do not use it, however.

Statistical Tools Best for Constructing a Historical Model are Different from Best Tools to Analyze Function

Many features of real protein sequences diverging under functional constraints are not captured by the Standard Model. Accordingly, many have attempted to upgrade the mathematical formalism for divergent sequence evolution to "inch towards reality" (Thorne et al., 1992). While it is impossible to list all of the heroes of this effort, work by Yang (Yang et al., 2000), Nei (Suzuki et al, 2001), Felsenstein (2001), Gu (1999), Huelsenbeck (Huelsenbeck and Rannala, 1997), and Swofford (Swofford et al., 1996) is useful in many of the applications described below. These are reviewed by Felsenstein (2001), and captured in several excellent books, beginning with the one by Nei (1987).

At one level, advanced tools to capture statistical measures of actual protein sequence evolution are very useful for reconstructing improved evolutionary histories for protein families. When distance-based metrics are used to construct trees, for example, it is helpful if the distances are measured using the most realistic models for amino acid replacement. Distances calculated with gamma models that permit different sites to suffer replacements at different rates are therefore likely to yield better trees (when calculating trees with distance metrics) than those calculated by simpler models.

As another level, this effort is an attempt to do the impossible: to capture in a useful way within a statistical framework the individuality of organic molecules. When interpreting the function of proteins, the individuality of individual amino acid residues within individual protein molecules is what is interesting. To the extent that the individuality of a protein, or amino acid in a protein sequence, is captured within a statistical formalism, that formalism loses its ability to provide information about the individual contribution of that residue, or that protein, to fitness.

This creates an apparent paradox. The more models abstract within a statistical formalism the individuality of individual proteins and sites within protein sequences, the worse they are as a definition of the null hypothesis, the model that describes how proteins would divergently evolve if they were formless, functionless strings of letters. Thus, the more enhanced the model is, the more likely it is to obscure the signal indicating function.

Therefore, the more the statistical model advances, the more valuable it is for constructing the core evolutionary model, and the less valuable it is in doing interpretive proteomics. This is, of course, not a problem, as one can chose the model for the purpose. It is possible even to iterate a cycle, building models for the global proteome using an inexpensive tool, determining the characteristics of sequence evolution from the resulting families, refining the model, and then iterating.

The model that is most valuable for generating the null hypothesis is remarkably close to what Margaret Dayhoff generated three decades ago. This is because it focused on early steps in evolution, which proved to be code-driven (Gonnet et al., 1992).

Our task is now to develop tools that analyze protein function by examining non-stochastic behavior of individual proteins. As with structure prediction, we begin with a null hypothesis that models the divergent evolution of protein sequences using statistical models having various levels of sophistication, but in all cases model replacement rates at individual sites as being independent of replacements at other sites. They then extract information about function in individual members in individual protein families by observing how their behavior during divergent evolution is different from that predicted by the statistical model.

Bioinformatics Workbenches and Databases

Starting a decade ago, we and our collaborators developed a variety of databases and bioinformatics workbenches to support evolutionary analysis. Perhaps most important of these is the bioinformatics workbench that we developed in 1990 in collaboration with Prof. Gaston Gonnet (ETH Zurich) (Gonnet and Benner, 1991). Termed Darwin (Data Analysis and Retrieval with Indexed Nucleotide and protein sequences), this workbench evolved from the symbolic computation platform known as Maple (http://www.maplesoft.com/), a platform used to organize and search the Oxford Unabridged English Dictionary (http://bluebox.uwaterloo.ca/OED/index.html), and a series of databases of biomolecular structure and function that were managed on personal computers within the Benner group.

Darwin supports the analysis of protein sequences from an evolutionary perspective. It has now been used in several laboratories, and is available through the Computational Biochemistry Research Group at the ETH (http://cbrg.inf.ethz.ch/). Darwin has the distinction of being the first sequence analysis workbench to be accessible on line via server. Many of its details are describe elsewhere (Gonnet and Benner, 1991), including in an on-line resource (http://cbrg.inf.ethz.ch/Darwin/index.html).

Results of evolutionary analysis of genomes

The Exhaustive Matching

An exhaustive matching of a sequence database requires comparing every substring in the database with every other, or its equivalent. The first exhaustive matching of a modern sequence database was achieved in 1991 by Gonnet et al., (1992) using the Darwin platform. This provided over 1.7 million matched pairs of homologous sequences, which has served as a valuable resource for understanding how proteins divergently evolve under functional constraints.

From the exhaustive matching came a comprehensive empirical model for how amino acids are replaced in proteins during divergent evolution under functional constraints (Cohen et al., 1994), models describing how segments of polypeptide chains are inserted and deleted during divergent evolution (Benner et al., 1993), and the first models describing how amino acid replacement at different sites in the protein sequence is correlated (Cohen et al., 1994). These, in turn, provided the database to generate solutions to some of the more perplexing problems in protein biochemistry, including how to predict the folded structure of proteins from sequence data, how to detect changing functional behavior in protein families, and how to detect distant homologs. These have been reviewed previously (Benner, 1998; Benner et al., 1997), and form the basis for the functional proteomics tools discussed below.

First Generation Naturally Organized Databases

The exhaustive matching and its various updates provided an estimate that when all of the genomes of all organisms on Earth are completed, all protein sequences will be easily recognizable as composed of peptide segments, or modules, that come from ca. 100,000 *nuclear families*. A module is defined as a segment of amino acid

Table 2
Definitions of family types

Nuclear family:	Collection of proteins that generates a reliable multiple sequence alignment using available tools
Sequence family:	Collection of proteins where the scores of all interfamily sequence pairwise comparisons is greater than a cut-off chosen to be a significant indicator of homology
Extended family:	Collection of proteins where all interfamily sequence pairs are connected by a path of pairwise comparisons that score sufficiently to be significantly homologous
Superfamily:	Collection of homologous proteins where non-sequence based attributes (e.g., nature of the fold) are needed to establish homology
Independent innovations:	Collection of proteins where all members are descendants of a common ancestor, which represents an innovation independent of the innovation of all others.

sequence that evolves as a unit (Riley and Labedan, 1997). Typical models are 50–500 amino acids long, and are recognized by comparison of the sequences in the database itself.

A nuclear family of these is defined operationally as a collection of protein sequences, all related by common ancestry, that have not diverged beyond the point where conventional tools fail to provide convincing multiple sequence alignments (see Table 2).

The limited number of families of proteins on Earth reflects several facts of natural history. First, all organisms on Earth are descendants of a common ancestor. Further, the number of possible protein sequences is astronomically large, and considerably larger than the number of protein sequences that could possibly have been formed in the limited time (ca. 4 billion years) and space (ca. 10^{21} l of aqueous effective volume) in the history of Earth. The limited space–time available to the biosphere means that only a small part of “sequence space” could have been explored since the Earth was formed.

Last, over much of the Earth’s history, and over the past 500 million years in particular, innovation in protein function and behavior has most frequently been achieved by recruiting an existing protein and changing its behavior by point mutation, insertion and deletion, and (occasionally) contextual rearrangement, rather than by de novo creation of new polypeptide chains. This effectively limits the number of protein module families in the terrean biosphere, just as the Periodic Table limits the number of elements in the terrean biosphere.

The number of nuclear families, operationally defined, is larger than the number of families of proteins (Table 2), operationally defined as those where significant homology can be detected by sequence analysis alone. The number of families is, in turn, larger than the number of superfamilies, defined to include proteins that share common ancestry, but where sequence similarity alone is insufficient to make a convincing case for homology. Here, arguments for/against distant homology are generally based on analogy between the folded structures of proteins, although other analogies are conceivable. The number of recognizable superfamilies is presumably larger than the number of events in the history of the Earth where a gene encoding a polypeptide sequence arose de novo. For other early discussions counting possible numbers of families, superfamilies, and independently innovated proteins, the work by Dorit et al. (1990), Chothia (1992), and Gonnet et al. (1992) serves as a starting point.

Given the limited number of protein families in the terrean biosphere, and the fact that we will repeatedly analyze the evolutionary features of these families of proteins as we dissect function in the global proteome, it made sense some time ago to identify all of the nuclear families encoded by the protein sequence database, to pre-compute evolutionary models for each of these families, and to store them in a ''naturally organized database'' (naturally organized database).

The idea of a naturally organized database of protein families is not new. In its ''first generation'' form, such a database was introduced by Dayhoff in her famous *Atlas of Protein Sequence* (Dayhoff et al., 1978). This *Atlas* collected proteins by families, and presented these with evolutionary trees and multiple sequence alignments. Other implementations of this data structure have emerged since, exploiting more advanced computer platforms and web access. They include the Hovergen (Duret et al., 1994), Pfam (Bateman *et al.*, 2000), DOMO (Gracy and Argos, 1998), SCOP (Lo Conte et al., 2000), Prodom (Corpet et al., 2000) and TIGRfam (http://www.tigr.org/TIGRFAMs/) databases. These databases are not distinct in concept from the original Dayhoff concept. Indeed, some of them are worse than the original paper-bound Dayhoff naturally organized database, in that they do not offer precomputed evolutionary trees to the user. All of them, however, are accessible by computer.

The MasterCatalog advances the concept of a natural organization in several ways that enhance the value of a naturally organized database for biological and biomedical researchers. First, the MasterCatalog contains all of the elements expected within a first generation naturally organized database. For all of nuclear families of modules derived from GenBank, the MasterCatalog contains a pre-computed collection of evolutionary models that each contains:

(a) A collection of homologous protein sequences, obtained from the most recently indexed version of GenBank.
(b) Top line annotation for each of the family members, together with database ID numbers (such as gi numbers) that give the user the option of accessing the full record from GenBank.
(c) An evolutionary tree of reasonable quality, calculated by an advanced distance matrix where the distances between sequences is calculated with a variance. The tree shows the family relationship between the protein sequences. Each leaf is labelled with the species from which the corresponding sequence is derived.
(d) A multiple sequence alignment, again of reasonable quality, which shows the evolutionary relationship between individual amino acids in the sequences of the proteins in the nuclear family.
(e) Bridges, which identify other families in the database that might be distant homologs for each nuclear family.

Next, the MasterCatalog uses the nuclear family as the organizational feature. Other family databases have attempted to capture within a single family as many distant homologs as possible. Thus, the sequences collected within the extended family do not lend themselves easily to the construction of a reliable multiple sequence alignment. Indeed, some of the databases identify only "motifs", short segments of protein sequence that are extremely conserved, as their distinctive features. As we illustrate below, high quality multiple sequence alignments are keys to the functional interpretation of sequences within a protein family.

Further, in constructing the MasterCatalog, separate evolutionary models are built for independently evolving units of protein sequence. The polypeptide chain is not necessarily the unit of sequence evolution. For example, the src homology 1 (SH1), src homology 2 (SH2), and src homology 3 (SH3) domains are homologous among themselves, but are often moved, cut, added, swapped, or rearranged with each other, within a single genome, to give different polypeptide chains. Any attempt to construct a history of these chains without recognizing their composite evolutionary nature will fail. Evolutionary models must be constructed for each of these module families. The MasterCatalog does this.

Also, during the building of the MasterCatalog, ancestral sequences of genes and encoded proteins are reconstructed at nodes throughout the tree. The ancestral sequences are represented at each site in the protein sequence by a vector in 20 dimensions, where the components of the vector sum to unity, and at each site in the DNA sequence by a vector in four dimensions, where the components of the vector again sum to unity. Details of amino acid replacement throughout the tree are captured within the MasterCatalog, where they stand available for use by the biomedical researcher.

Second Generation Naturally Organized Databases

The number of innovations built within the MasterCatalog is sufficient as to earn it the designation as a "second generation" naturally organized database. A particularly inventive feature of the MasterCatalog is its use of explicitly reconstructed ancestral states throughout the tree. These add a dimension of interpretive value to an evolutionary model that is not captured by first generation databases. With reconstructed ancestral sequences come statements, in probabilistic form, about every event that occurred along every branch of every tree. Our Darwin server generates an report that provides the user with a list of the nucleotide substitutions and amino acid replacements that have occurred along each branch.

Starting with ancestral sequences and information about events, each branch in an evolutionary tree can be characterized. The first feature of a branch is simply its length. Metrics for length include the number of mutations in the DNA sequence that occurred along the branch, the number of silent mutations that occurred along the branch, the number of non-silent mutations that occurred along the branch, the number of silent transitions that occur along the branch, the PAM length of the branch, or the number of amino acid replacements that occurred along the branch.

Other features of the branch can involve ratios of these. For example, the ratio of non-synonymous to synonymous substitutions can be calculated for each branch. The MasterCatalog exploits a heuristic used by Pamilo and Bianchi (1993) for performing this calculation. The output is a K_a/K_s ratio for every branch in the tree. The significance of this ratio is discussed below.

The features of the tree overall can be drawn from these reconstructed ancestral sequences. We can ask, for example, what the average K_a/K_s is per branch across the tree. We can weight this average by branch length. We can ask what statistical model best represents the features of the evolutionary model that describes the tree as a whole. We can identify individual amino acids that are replaced in a branch with a high K_a/K_s ratio. The value of these metrics in interpretive proteomics will be discussed below.

More importantly, a second generation database proves to be an excellent starting point to identify points in the protein and in the history of its family where divergent behavior does not follow simple stochastic models. These are, of course, the focus of this lecture. Before we explore these, however, we need to apply another novel enhancement of the evolutionary model, one involving dating.

Enhanced second generation naturally organized databases

Adding Dates

A key strategy in the analysis of genomic and proteomic sequence database involves *temporal correlation*. Following this strategy, the scientist seeks events in the history of the biosphere that occurred near the same time. When a correlation is observed, it suggests (as a hypothesis) a functionally significant relationship between the correlated events. Many of these events are recorded in the geological and paleontological record.

Temporal correlation is a staple of interpretive paleontology. Ph.D. dissertations are written in paleontology analyzing hypotheses that imply causal relationships between historical events based on their near simultaneity (with the scale in millions of years) of events recorded in the paleontological record.

With second generation naturally organized databases such as the MasterCatalog, it becomes possible to also ask about dates of events captured in the molecular record. To make a correlation between the geological and paleontological records on one hand, and the molecular record on the other, however, we need a tool to date events in the molecular record.

Each of these records has different tools for performing dating. The tools have different accuracies. For example, the dates of crystallization within igneous rocks are determined by examining the amounts of radioisotopes and their decay products within the rocks. Radioactive isotopes are useful for dating events in the geological record because of the first order nature of nuclear decay, and the remarkable extent to which the associated rate constants are independent of environmental factors. Its first order nature means that the decay can be modelled using a simple exponential rate law, with the fraction of initial atoms remaining f after time t defined by the expression $f = 1 - \exp(-kt)$. Here, k is the rate constant for the decay, which gives the half-life $\tau = \ln 2/k$. The independence of k of environmental factors means that one need know nothing about the history surrounding the sample to calculate a date from this process. Simply by measuring the amounts of decay products from two isotopes of uranium in a zircon crystal, for example, precision to better than a million years is nearly routine when dating an igneous rock 500 million years old (Bowring et al., 1993).

The paleontological record is dated by the association of specific fossils with specific radiochemically dated rocks. Unfortunately, fossils are found in sedimentary rocks. Crystallization of a rock from molten rock is needed to set the radiochemical clock, making it radiochemical dating possible only for igneous rocks. In some cases, volcanic strata or igneous rocks are closely associated with sedimentary rocks, enabling the transfer of a date from one to another. More frequently, dates of igneous rocks constrain the dates of fossils, without establishing their age precisely. Correlating igneous rocks with sedimentary rocks, and correlating sedimentary strata with the fossils that they contain, is an ongoing exercise in geobiology.

No known chemical process has rate properties that are comparable to those displayed by radioactive decay. Many chemical processes display first order (or pseudo-first order) kinetics, of course. But the rate constants for nearly all of these are influenced dramatically by environmental factors, including temperature, salt concentration, and pH (for example). How unsuitable chemical processes are as a metric for age is well illustrated by examples where dating tools based on chemical reactions were sought. Amino acid racemization is perhaps the most widely used of these. But the rates of amino acid racemization vary dramatically depending on conditions, making this a “second choice” dating tool, at best.

Given this, it may appear hopeless to try to identify a chemical process in living systems that has sufficient first order character to be useful to date biological events, especially one reflected in DNA or protein sequences (Fitch, 1976). We do not know the microscopic chemical processes that are responsible for natural mutations in natural populations. Indeed, it is conceivable that many microscopic chemical processes contribute, including deamination, oxidative damage, polymerase error, and failure of repair. Further, natural selection can play a major role in determining what mutations are fixed in a population. When DNA mutations result in the replacement of an amino acid in an encoded protein (a non-synonymous mutation), the behavior of the protein can change. Protein behavior can be intimately connected to function and natural selection. Therefore, encoding DNA sequences are not expected to diverge with a time-invariant rate constant whenever the demands of selective pressure are changing, even if the microscopic chemical processes that create a pool of mutations occurs with a time-invariant rate constant (Ayala, 1999).

Nevertheless, one can hope that some parts of a DNA sequence will diverge in a process that might display first order kinetics approximately. Synonymous sites in a gene, sites where nucleotide substitution does not change the encoded amino acid, are frequently examined for this purpose (Li et al., 1985). Because these cannot alter the behavior of a protein, synonymous substitutions are likely to be free of selective pressure than substitutions at non-silent sites. Thus, these are candidates for mutations that diverge with (pseudo) first order kinetics.

Recent studies that examine synonymous substitutions do not, however, use an approach-to-equilibrium kinetic processes to model these. Rather, most approaches attempt to enumerate substitutions at synonymous sites by comparing two extant sequences, counting the silent differences, and using a correction to estimate the number of times multiple substitutions have occurred at the synonymous site (Tiffin and Hahn, 2002). From this is extracted a number for the synonymous substitutions per site. Further, these metrics attempt to count all synonymous mutations that occur at each site, including those within two-fold redundant codon systems, within four-fold redundant coding systems, and within codons that have also suffered non-synonymous mutations well. Most treat transitions (which replace pyrimidines by pyrimidines, or purines by purines) and transversions (where pyrimidines and purines are interconverted) together, even though these are known to occur at different rates (Gojobori et al, 1982).

Some time ago, we introduced into the patent literature a new tool to date sequences based on the approach-to-equilibrium formalism taken from chemical kinetics. While we recognize that no DNA mutation process will ever be described as a first-order process to high accuracy, the value of these approach-to-equilibrium models in sorting out networks and pathways in whole genome analysis has proven to be so valuable in ad hoc cases that we believe it is timely to report details of the approach-to-equilibrium model in correct form.

It is well known that a two state system interconverting species A and G in the kinetic scheme:

$$A \underset{k_{G\to A}}{\overset{k_{A\to G}}{\leftrightarrows}} G \tag{1}$$

approaches equilibrium via an exponential process, where the observed rate constant $\mathrm{k_{obs}}$ is equal to the forward rate constant for the conversion of A to G, *plus* the reverse rate constant for the conversion of G to A, that is, $k_{\mathrm{obs}}=k_{A\to G}+k_{G\to A}$. Further, at equilibrium, the ratio of G to A is equal to the ratio of the forward and reverse rate constants, that is, $G_{\mathrm{eq}}/A_{\mathrm{eq}}=(k_{A\to G})/(k_{G\to A})$, where G_{eq} and A_{eq} are the respective concentrations of G and A at equilibrium. In the general case, the concentration of A as a function of time is

$$\frac{A(t)}{A_0}=f_{G_{\mathrm{eq}}}\ \exp-(k_{A\to G}+k_{G\to A})t+f_{A_{\mathrm{eq}}}, \tag{2}$$

where $f_{G_{\mathrm{eq}}}$ and $f_{A_{\mathrm{eq}}}$ are the fractions of G and A at equilibrium (that is $G_{\mathrm{eq}}/(G_{\mathrm{eq}}+A_{\mathrm{eq}})$ and $A_{\mathrm{eq}}/(G_{\mathrm{eq}}+A_{\mathrm{eq}})$). These two fractions, expressed in terms of the microscopic rate constants, are $k_{A\to G}/(k_{A\to G}+k_{G\to A})$ and $k_{G\to A}/(k_{A\to G}+k_{G\to A})$. The analogous expression can be written for the concentration of G as a function of time:

$$\frac{G(t)}{G_0}=f_{A_{\mathrm{eq}}}\ \exp-(k_{A\to G}+k_{G\to A})t+f_{G_{\mathrm{eq}}}. \tag{3}$$

The two fractions of G and A at time t always sum to unity.

Consider now the case where A and G are nucleotides at a site constrained to accept only purines, A or G. The rate constants $k_{A\rightarrow G}$ and $k_{G\rightarrow A}$ now correspond to pseudo-first order rate constants for two transition processes, the mutation of A to give G, and the mutation of G to give A. Again, the k_{obsR} (R for purines) is equal to $k_{A\rightarrow G}+k_{G\rightarrow A}$. The fraction of sites occupied by A and G reflect the A/G bias at such sites at equilibrium (which we assume holds throughout).

Let us now consider two identical sequences that are given the opportunity to diverge. We assume that the initial proportion of A and G at these sites is equal to the bias, that is, that the fractions of A and G represent the fractions expected at equilibrium. We also assume that each site suffers mutation independent of other sites, and that the forward and reverse transition rate constants are the same for all sites. How will the identity at purine-constrained sites diverge?

Let us consider separately the sites that are occupied by A at $t=0$ and the sites that are occupied by G at $t=0$. For those that are occupied by A, the sites that are considered to be "conserved" at time t are those that retain A at time t. As a fraction of the total sites originally A, Eq. (2) can be deconvoluted as follows:

$$\begin{array}{l}\text{conserved sites}\\ \text{arising from } A\end{array} [f_{G_{\mathrm{eq}}}\exp(-k_{\mathrm{obsR}}t)+f_{A_{\mathrm{eq}}}]\,f_{A_{\mathrm{eq}}}, \tag{4a}$$

$$\begin{array}{l}\text{conserved sites}\\ \text{arising from } G\end{array} [f_{A_{\mathrm{eq}}}\ \exp(-k_{\mathrm{obsR}}t)+f_{G_{\mathrm{eq}}}]\,f_{G_{\mathrm{eq}}}. \tag{4b}$$

The fraction of all sites conserved as a function of time is the sum of these two:

$$\begin{aligned} f_2 &= f_{A_{\mathrm{eq}}}\, f_{G\mathrm{eq}} \exp(-k_{\mathrm{obsR}}t) + f_{A_{\mathrm{eq}}} f_{A_{eq}} + f_{G_{\mathrm{eq}}} f_{A\mathrm{eq}}\ \exp(-k_{\mathrm{obsR}}t) + f_{G_{\mathrm{eq}}} f_{G_{\mathrm{eq}}} \\ &= 2f_{A_{\mathrm{eq}}}\, f_{G_{\mathrm{eq}}}\ \exp(-k_{\mathrm{obsR}}t) + f^2_{A_{\mathrm{eq}}} + f^2_{G_{\mathrm{eq}}} = P_R\exp(-k_{\mathrm{obsR}}t) + E_R, \end{aligned} \tag{5}$$

where P_{R} is the pre-exponential term $(=2\{f^2_{A_{\mathrm{eq}}}+f^2_{G_{\mathrm{eq}}}\})$ and E_{R} is the f_2 reached at equilibrium, and is equal to $f^2_{A_{\mathrm{eq}}}+f^2_{G_{\mathrm{eq}}}$.

Thus, f_2 as a function of time follows a first order exponential decay from unity to an end point defined by the expression $(f^2_{A_{\mathrm{eq}}}+f^2_{G_{\mathrm{eq}}})$. These two terms, in turn, are defined by $\{k_{G\rightarrow A}/(k_{A\rightarrow G}+k_{G\rightarrow A})\}^2$ and $\{k_{A\rightarrow G}/(k_{A\rightarrow G}+k_{G\rightarrow A})\}^2$. If A and G appear with equal frequency, then the end point $f_2=0.5$. If, however, A and G appear with a relative frequency of 0.6 and 0.4, then the end point is 0.52.

If the rate constants are assumed to be time-invariant, we can treat f_2 as a molecular clock. It is a very special one, in that it involves only two specific rate constants from the 12 that are possible with the four letters in the genetic alphabet. Further, it considers only those sites where the amino acid has not diverged, constraining the site to accept only a transition. As the rate constants for transitions and transversions are known to be different, this particular clock should (from first principles) generate better dates than one that aggregates the 12 different processes.

To implement this clock, we need only to identify sites in natural DNA sequences that are constrained to mutate between A and G only. Codons for three amino acids (Glu, Gln, and Lys, or E, Q, and K in the one letter code) are so constrained if the amino acids are not replaced. In practice, we can examine a pair of aligned protein sequences for positions where Glu, Gln, and Lys are conserved between the two. Making only the approximation that homoplasy at these sites has not occurred, we can use the synonymous (third position) sites in these codons as candidate sites that fit the purine-constrained criterion.

An analogous kinetic expression can be written for pyrimidine–pyrimidine transitions. The third positions of six amino acids (Cys, Asp, Phe, His, Asn, and Tyr, or C, D, F, H, N, and Y in the one letter code) are constrained to have only T or C, if they are not replaced in the protein coding sequence. Again, inspection of a pair of aligned protein sequences for positions where these amino acids are conserved identifies synonymous sites as candidates that fit a pyrimidine-constrained kinetic behavior.

The TREX (transition redundant exchange) clock does not require that silent transitions be absolutely neutral. The equilibrium fractions of nucleotides at silent sites need not be equal (codon bias), and this corrects for any selective pressure that causes a time-invariant bias at the silent sites. It does require, however, that this selection pressure be time-invariant, that is, that it not change in the time separating the sequences whose divergence is being dated. As we note below, a comprehensive analysis of a genome permits one to assess the extent to which codon bias and transition rate constants have changed in the historical past of a lineage. Absence of time-invariant bias means something too, for example, that the evolutionary processes that lead to natural mutation are changing or that the properties of tRNA molecules in the system are changing. One of the key purposes of whole genome analyses (see below) is to model these processes and properties over time.

The precisions of TREX distances depend on the number of characters used to derive them. All dates contain uncertainty. Uncertainties in geological dates based on exponential radiochemical decay are small, often less than 0.1% for dual isotope chronology on well-preserved igneous rocks. Paleontological dates of divergence (from fossils) have larger uncertainties, primarily because of the incomplete fossil record. Fossils near branch points in a phylogenetic tree are rarely found, and those that are need not be associated with isotopically datable igneous formations.

Work with yeast, fly, and vertebrates suggests that the main sources of variance in dating using standard silent substitution metrics (Li et al., 1985; Pamilo and Bianchi, 1993, Lynch and Conery, 2000) are the approximations made by the underlying model. These create imprecision much greater than the uncertainties due to fluctuations, and the uncertainty in paleontological dates. The approximations embedded into the TREX tool, however, are not as severe (although some still exist, of course). Further, TREX distances can be calculated from reconstructed ancestral sequences, allowing a correction for generation times of ancestral organisms (see below). Therefore, the variance in N^2ED dates arises primarily from fluctuation (a typical TREX value is calculated from 100 characters); fluctuation accounts for $>90\%$ of the variance observed in a data set from mammals (Caraco, 2002). We need not invoke differential rates of silent substitutions in different genes (“hot spots”), different codon biases in different genes, or other non-first order processes to account for the variance. Further, the error in a TREX date is less than the typical errors in dating branch points from the fossil record. For the purpose of planetary biology and genome annotation, this is as good a precision as is useful.

The TREX tool can be used to add dates to nodes in the tree captured within a second generation naturally organized database back in time to a point where the two-fold redundant sites become equilibrated. The greater the number of sequences descendent from a particular ancestor present in the database, the more precisely the sequences of the ancestral proteins can be defined. Because node–node distances are shorter than leaf–leaf distances in a tree, the process of reconstruction can permit TREX dates farther back in time with greater precision than would be possible for leaf–leaf dating alone.

Perhaps the most direct use of the TREX tool, however, is to distinguish orthologs from paralogs. Orthologs are two homologs found in different taxa where the most recent common ancestor of the two proteins was found in the most recent common ancestor of the two taxa. Orthologous proteins are generated by gene duplication, of a sorts. But the gene duplication generating orthologs is the same as the duplication that is associated with speciation. The fate of one duplicate is associated with the fate of the organism(s) that eventually founded taxon 1; the fate of the other duplicate is associated with the fate of the organism(s) that eventually founded taxon 2.

In contrast, paralogs are homologs found in a single genome. They arose by a true gene duplication, meaning a genetic event that creates two loci on two different positions of a chromosome, or possibly on two different chromosomes.

The "ortholog–paralog problem" arises from the fact that a homolog in taxon A need not have diverged from its counterpart in taxon B at the same time as the two taxa diverged. Gene duplication prior to the divergence of the two taxa, and possible gene loss (or incomplete genome sequencing), can generate paralogs whose true evolutionary relationship is not recognized by analysis of a tree alone. Fig. 7 illustrates how the perception regarding the point(s) on an evolutionary tree that represent the last common ancestor can be altered by discovering new sequences.

Adding Structural Biology

Proteins are organic molecules. Therefore, the three concepts of structure (constitution, configuration, and conformation) that apply to all organic molecules apply to proteins as well. Many of the interpretive proteomics tools that we use involve manipulation of those strings as strings. Therefore, interpretive proteomics tools that incorporate conformational analysis add a new "dimension" to the analysis.

In interpreting the functional significance of sequence change, a particularly powerful tool involves correlations that identify specific amino acids that are being replaced during an episode of sequence evolution, and view those specific amino acids in relation to the three-dimensional structure of the protein. For example, if a change occurs in the active site of an enzyme, this fact alone suggests functional hypotheses that are different than if the change occurs distant from the active site (Benner and Gerloff, 1991).

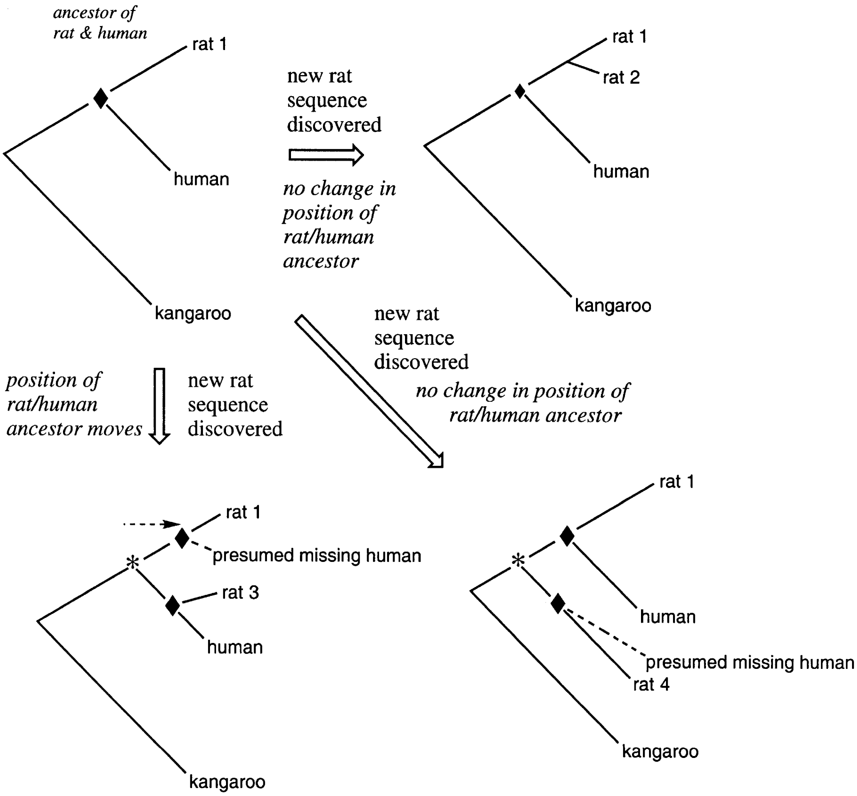

ancestor of rat & human
rat 1
human
kangaroo
new rat sequence discovered
no change in position of rat/human ancestor
rat 1
rat 2
human
kangaroo
new rat sequence discovered
no change in position of rat/human ancestor
position of rat/human ancestor moves
new rat sequence discovered
rat 1
presumed missing human
rat 3
human
kangaroo
rat 1
human
presumed missing human
rat 4
kangaroo

Fig. 7. The point(s) on an evolutionary tree of a protein family that represent the last common ancestor can be altered by discovering new sequences. Note that in the bottom trees, *two* points represent the most recent common ancestor of rat and human. This is because the protein family suffered a duplication prior to the divergence of rat and human, meaning that the last common ancestor of rat and human had two members of this protein family. TREX dates permit us to distinguish between these without having to find the missing sequences. Thus, if the diamond in the upper left diagram is associated with a TREX date that reflects the known time of divergence of human and rat, then the rat 1 and human sequences are true orthologs. In contrast, if the diamond in the upper left diagram is associated with a TREX date that is larger than the known time of divergence of human and rat, then the rat 1 and human sequences must be paralogs, and the true orthologs need to be sought (if the genome is not complete) or can be presumed to have been lost (if they are not there in the completed genomes).

The most direct way to associate a conformation with a protein family is to access an experimentally determined conformation for one of the proteins in the family. This is done either by X-ray crystallography or nuclear magnetic resonance spectrometry. The MasterCatalog records the experimental structure for every protein whose conformation has been determined experimentally.

It is also possible to predict secondary structure for proteins from a set of sequences of homologous proteins undergoing divergent evolution under functional constraint. The tools implemented on the Darwin server for predicting secondary structure are based on a general, site-by-site analysis of mutability. These tools have been reviewed elsewhere (Benner et al., 1997). The MasterCatalog contains a predicted secondary structure for each protein family.

Information about the conformation of a protein can be used to rectify the evolutionary model for the protein family. In particular, knowledge of an experimental structure can help in the placement of gaps within a multiple sequence alignment, in the alignment of key residues, and in selecting the preferred tree.

The last is especially useful. When protein sequences divergently evolve under functional constraints, some individual amino acid replacements that reverse the charge (lysine to aspartate, for example) may be compensated by a replacement at a second position that reverses the charge in the opposite direction (glutamate to arginine, for example). When these side chains are near in space (proximal), such double replacements might be driven by natural selection, if either individually is selectively disadvantageous, but both together restore fully the ability of the protein to contribute to fitness (are together "neutral").

This type of behavior is called compensatory substitution. It represents a higher order behavior of protein sequences that is not captured by the Standard Model. Some time ago, we noted that a modest signal could be obtained by searching for compensatory changes on branches of trees that lie between two nodes. The signal is most evident when a crystal structure is available, as it can be determined whether the amino acids that are suffering complementary replacement at the same time are actually close in space.

The strength of the compensatory covariation signal undoubtedly depends on the degree to which the trees and the reconstructed ancestral sequences accurately reflect the history of the family. If the branching of the tree or the reconstructed sequences themselves are not correct, a pair of charge compensatory replacements that are coincident, in fact, may not be assigned to the same branch of a tree. In this case, the signal from this pair will be lost.

Getting the branching correct in an evolutionary tree is a difficult problem. Part of the difficulty arises because of the trade-off between the accuracy of the tree and the cost of generating it. For example, the ClustalW (Thompson et al., 1994) and Fitch parsimony tools are relatively inexpensive methods for reconstructing trees and ancestral sequences. ClustalW uses a neighbor joining tool based on estimates of the distances between sequence pairs derived from the Kimura empirical formula (Kimura, 1983). Ancestral sequences reconstructed by parsimony are well known to be sensitive to incorrect branching topology. This may be the principal error associated with the choice of this inexpensive reconstruction tool.

Even more expensive tools do not guarantee a correct tree, of course. In practice, the approximations made in the model may create systematic error larger than fluctuation error. To date, the only way to benchmark a tree requires knowledge of the evolutionary history of the sequences in question (Hillis et al., 1994) or a reconstruction of a simulated evolutionary process (Takahashi and Nei, 2000). The first is difficult to get for sequences emerging from natural history. The second requires a mathematical model for evolution, which is often the same one that is used to construct the tree in the first place.

Here, the compensatory covariation signal, extracted from reconstructed ancestral sequences, may provide a metric for the quality of a tree based on organic chemistry, independent of any mathematical model for evolution. Hypothetically, the best tree should be the tree that places compensatory replacements truly driven by natural selection on the same branch. This requires the construction of a tree that reflects the actual evolutionary history. This, in turn, implies that the tree has the most compensatory covariation is the tree that is most likely to reflect the actual history.

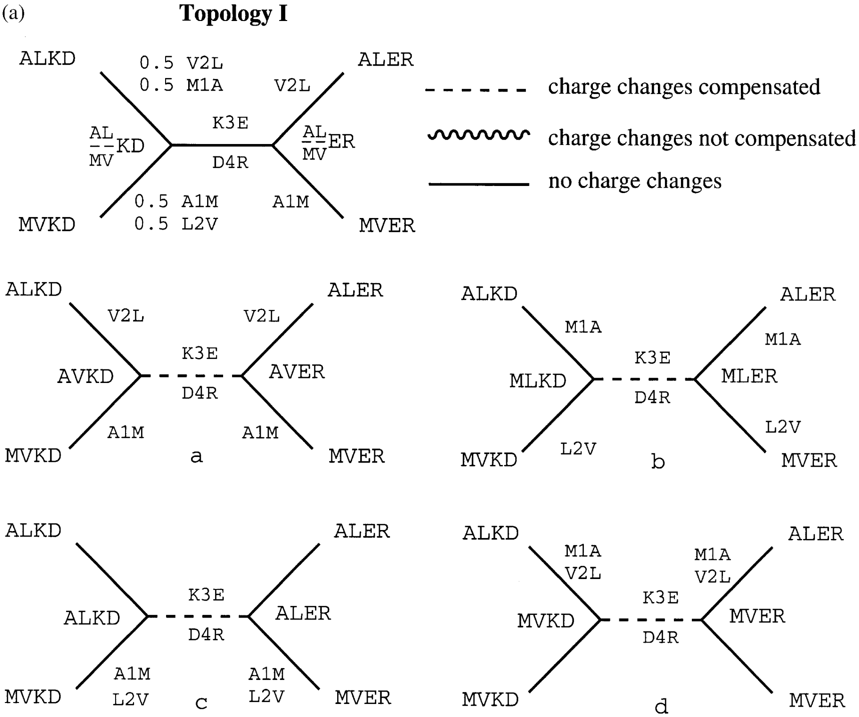


Fig. 8. A schematic illustration of the use of compensatory covariation to select a preferred tree from two equally parsimonious trees. The two tree topologies relating the four sequences (ALKD, MVKD, ALER, and MVER) each require six changes. The changes are marked on individual branches, with fractional changes arising from the ambiguity in the ancestral sequences. The ancestral sequences are placed at the nodes in the tree, with ambiguous sites (by parsimony) noted by placing the two possible residues above and below a horizontal line. For each topology, identical trees holding all four possible ancestral sequences are shown. Each, by parsimony, has equal likelihood (0.25 for each). In Topology I, the ancestral sequences are ambiguous at the first two positions. In Topology II, these are ambiguous at the last two positions. Both trees require the same amount of homoplasy (convergence). Classical parsimony analysis is indifferent with respect to the two topologies. In Topology I, however, the likelihood that a charge reversal is compensated is unity. In Topology II, the likelihood that a charge altering replacement is compensated is only 0.5. Thus, Topology I is preferred if compensatory covariation is maximized. This criterion is independent of mathematical formalisms used to construct the tree. Further, the metric weights changes at position *i* depending on events at position *j*, making this metric for evaluating a tree fundamentally different from any metric based on a first order stochastic analysis of protein sequences.

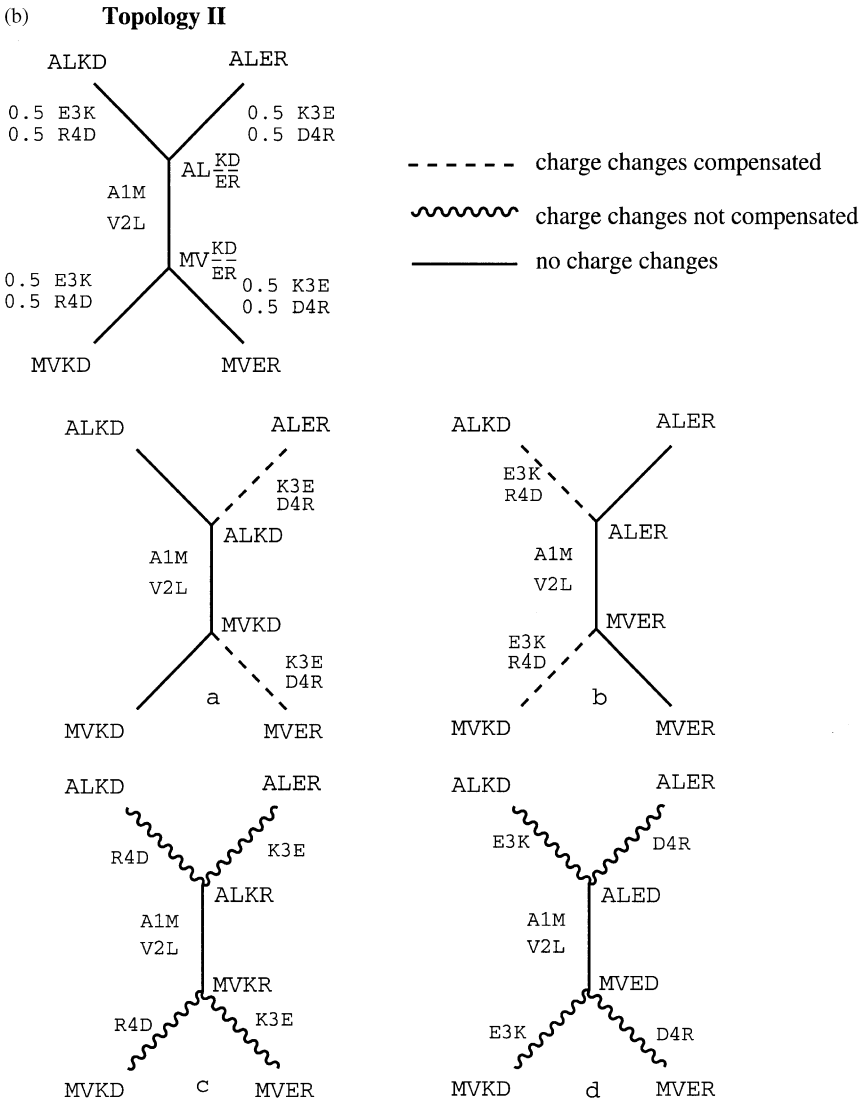


Fig. 8 (*continued*).

To illustrate this application, consider four hypothetical proteins, just 4 amino acids in length, having the sequences ALKD, MVKD, ALER, and MVER. Exactly two topologies exist for unrooted trees that relate these four sequences (Fig. 8). Both reconstructions have two ambiguous sites in both ancestors. In Topology I, the first two positions are ambiguous; in Topology II, the last two positions are ambiguous. Both trees require four "homoplastic" events (independent mutations that cause sequence convergence). Both trees require exactly six changes. Classical parsimony therefore ranks these two topologies as equally likely.

The two topologies are different, however, with respect to the extent to which charge changes are compensated. In Topology I, a charge altering replacement is 100% likely to be compensated. In Topology II, however, a charge altering replacement is only 50% likely to be compensated. This is illustrated in Fig. 8 by writing out four trees, each equally likely, that carry reconstructions that the ambiguities require. If we postulate that compensatory covariation is maximized, then Topology I is preferred over Topology II.

Conversely, an analogous logic can be used to assign preferred ancestral states involving charged residues. For the tree on the left, the ancestral states involving charged residues are fixed. For the tree on the right, the preferred ancestral sequences are in reconstructions IIa and IIb.

This metric can be applied even if no crystal structure is available for a protein family. If, however, a crystal structure is available, then (as a practical matter) one would maximize the number of charge compensatory changes that are physically near in space when identifying the preferred tree.

This approach is the first to identify the correct tree by seeking a physical organic property of the molecular evolution. Again, a statistician will find no numerical metric to assess the approach's reliability. This is chemical science, not mathematics. Nevertheless, the tool is useful, if only because it generates a preference for one tree as a hypothesis.

Identifying the Superfamily

The assembly of a nuclear family stops once the sequences being added fail to meet a cut-off that is selected to ensure high quality MSAs. The cut-off is, to a degree, arbitrary. Therefore, more distant homologs are retained within the MasterCatalog in a list of "bridges," connections to other nuclear families that reflect indisputable inter-family homology, but where the extent of sequence divergence is too great to permit a single nuclear family to be constructed from the two. As noted above, families connected by bridges can be used to root the individual nuclear families. Again, over time, the MasterCatalog will evolve to incorporate these roots, and the table of bridges is a resource.

Ultimately, we wish to identify superfamilies, collections of protein sequences that may share common ancestry even though the similarities in the sequences of their most distant members is insufficient to support (with an acceptable level of significance) the conclusion of homology. Today, the only reasonably validated to tool for inferring homology at this distance is by noting analogies in the conformation, or fold, of the families. Analogous folds between two protein families may help align sequence motifs, sequence strings that are too short to support significantly any conclusions of homology, but might be regarded as being suggestive. Alternative tools, including mechanistic analogy (when enzymes are involved) are too susceptible to convergence to be reliable, although they can support a conclusion of homology based on weak sequence similarity and analogous conformation.

Table 3
Steps used to identify distant homologs

Add obvious bridges to the family; these are bridges that meet test of statistical significance
Refine the placement and sequence of the founder based on a root
Use the founder sequence to confirm/deny speculative bridges based on sub-statistical sequence similarity
Use experimental secondary structures to confirm/deny speculative bridges based on sub-statistical sequence similarity
Refine the secondary structure prediction
Use the secondary structure prediction to confirm/deny speculative bridges based on sub-statistical sequence similarity

Both predicted and experimental structures have been shown to be useful to identify superfamilies. Especially valuable are tools introduced by Benner and Gerloff (1991), which are able to both confirm and deny distant homology. Using these tools, for example, protein kinase was correctly predicted *not* to belong to the superfamily that also contains adenylate kinase, even though motif analysis suggested that it did (Benner and Gerloff, 1991).

Table 3 summarizes the steps used to assemble a superfamily. Superfamily connections aid interpretive proteomics most significantly when no cultural annotation, defined as the linguistic construct that indicates function, is available for any members of the nuclear family, or for any members of the extended family linked by bridges. Failing any experimental evidence for function within the nuclear and extended families, the experimentalist is delighted if any broader homology indicators identify possible homologs that have an assigned function. This being said, it is important to note that function can change dramatically within a nuclear family, and it certainly changed within extended families. Therefore, annotation transfer (see below) between members of a superfamily may be only conjectural.

Modifying the Family as Delivered

The MasterCatalog saves time. Without the MasterCatalog, every time an evolutionary analysis is desired, we are faced with the grim tasks of submitting searches to a BLAST server, downloading the sequences that are identified, and building an evolutionary model for the sequences that have been identified. We must make decisions about what proteins to include and exclude from the family, the significance of the scores, and the suitability of trees and multiple sequence alignments.

All of these are pre-computed in the MasterCatalog, for all of the families in the global proteome currently identified. Given the MasterCatalog, therefore, we can start the day possessing a naturally organized database as a resource, having in hand all of the elements of a first generation naturally organized database, and many of a second generation naturally organized database. The models within the MasterCatalog are sufficiently advanced that we can move immediately to the next

Table 4
Modifying the contents of a nuclear family

Adding sequences, perhaps proprietary sequences
Removing obvious duplication
Removing sequences without DNA (for analyses that require DNA)
Removing "defective sequences" (fragments, dubious intron assignments)
Superfluous sequences as a problem to presentation

phase: devising interpretive strategies and tools assuming that the task is already complete. In this regard, the 100,000 families within the MasterCatalog, collected by family, are a deliverable. With the MasterCatalog, we can immediately begin thinking about biology.

This does not mean that one must accept the evolutionary model delivered by the MasterCatalog, as it is delivered, of course . Before we begin an analysis of a family of proteins starting with a deliverable provided by the MasterCatalog, we may wish to modify its contents. Some possible modifications are listed in Table 4.

Most commonly, the user may possess sequences that is not in the public domain. In this case, the sequence data may be added to the family as it is delivered by the MasterCatalog.

Alternatively, we may wish to remove sequences from the family as it is delivered by the MasterCatalog. GenBank has a bias towards redundancy; virtually every sequence variant that is submitted to GenBank ends up in the database. Often, exact duplicates or near duplicates contain interesting information, and should be retained. But for other purposes, a family with too many sequences may be difficult to visualize, or may slow subsequent computation. In these cases, the delivered family may be culled, to remove exact or near duplicates, or to remove sequences from exceptionally bushy branches.

In either case, the original tree and MSA may be recomputed. The Darwin server supports recomputation of trees and MSA when presented with an SGML file that contains the desired set of sequences. In practice, all interpretive proteomics efforts begins by downloading the MasterCatalog family as a deliverable, adding or removing sequences as appropriate, generating an SGML file from the final set of sequences, and submitting the final set to the Darwin server (or to any other utility that the user desires) to recompute the tree and the MSA.

The result is the "rectified" family that is the starting point for a functional analysis. We can now make some general observations about the family. First, we can note the overall PAM distance of the family, a measure of the number of point accepted mutations per 100 amino acids between the most distant sequences, captured within the family. A divergent family is better than a family composed from only highly similar sequences. The former contains more information, while the latter resembles (in terms of information) many copies of the same newspaper, therefore not containing much more information than the first copy.

Further Rectification of the Evolutionary Model for the Family

Rectification is a complex concept. Generally, an evolutionary model is rectified whenever the user is dissatisfied with some feature of the evolutionary model for the family of proteins as delivered by the naturally organized database.

For example, different practitioners are advocates of different tools to construct trees and MSAs. Some prefer parsimony methods, such as those implemented by PAUP (Swofford, 1998). Others prefer maximum likelihood methods as implemented by PHYLIP (http://evolution.genetics.washington.edu/phylip.html). Preferences vary concerning scoring matrices. Table 5 contains some alternative approaches that can be used to construct a tree and/or a MSA from a set of protein sequences derived from the MasterCatalog. No effort has been made to be exhaustive.

We are neither advocates nor opponents of any particular tool. Each represents a particular mathematical formalism representing divergent evolution, with its own set of assumptions. Extensive discussion can be had as to which is "better". As we outlined above, we are less concerned with how accurately different tools reflect reality than we are with how useful they are, especially to biological and biomedical researchers. This determination must be made through the actual use of the tools, not by applying statistical metrics.

A plurality of models becomes valuable for this reason. Different mathematical formalisms of divergent evolution can give somewhat different trees and multiple sequence alignments. It is conceivable that these differences will lead to differences in interpretations made by an evolutionary analysis of these models. It is therefore useful to ask how robust the interpretations are with respect to plausible variation in the formalism used to construct the model. Indeed, it is useful to ask whether the interpretation is robust even with respect to implausible variations in the formalism.

For this reason, it makes sense during a program of interpretive proteomic analysis with a family to re-compute the tree and MSA using formalisms different from those used to create the deliverables within the MasterCatalog. At the same time, the user might consider recomputing the tree, MSA, and ancestral sequences using the most expensive formalism that the budget will allow, and then determine whether the interpretations are robust with respect to the resulting changes. We ourselves frequently wish to adjust the placement of gaps within the MSA using advanced gap placement heuristics (Benner et al., 1994), and to examine alternative

Table 5
Alternative approaches to constructing trees and multiple sequence alignments

Use a distance-based tree that does not incorporate variances (like the MasterCatalog tree does)
Use a tree based on DNA sequence analysis
Use a distance-based tool that incorporates a gamma model
Use paleontological information to constrain the tree
Use a silent codon metric, such as a TREX distance, to build a part of the tree
Use different metrics in different parts of the trees, such as (for example) TREXs for closely related sequences, PAM for more distant parts, and gapping for still more distant parts of the trees.
Build trees with alternative sampling of the database (robustness to sample size)

trees using our compensatory covariation tools, as discussed above (Fukami-Kobayashi et al., 2002).

Another rectification process recognizes the possibility that the sequences themselves might contain mistakes. In particular, genes that are found by hidden Markov models (HMMs) misassign introns, starts, and stops within found genes, with an unknown frequency. These yield incorrect gaps in a multiple sequence alignment. Advanced gap placement tools help remove these mistakes as part of a rectification process.

Again, over time, the pre-computed models for individual families within a naturally organized database will be rectified to remove mistakes, enhanced by statistical analysis, and refined through the introduction of non-sequence information (including paleontological information). This, over time, these families come to reflect the historical reality more accurately.

The vision is captured by the analogous development of the Periodic Table a century ago. The characteristics of the chemical elements were obtained approximately for each element when it was discovered. Over time, the description of the element was enhanced, however, and more precise values were obtained. Eventually, for each element, the model converged to an endpoint.

We expect the same to occur for each family in the MasterCatalog. There is only one true molecular history for any individual family of proteins. As more data emerge, including sequence, paleontological, and geological, the parts of this history that can be reconstructed will be reconstructed with increasing accuracy. The parts that have been irretrievably lost will also become evident in the process. Over time, the description of the evolutionary history of each module family will converge to a stationary point.

Enhancing the Evolutionary Model for a Protein Family of Interest

Given a rectified evolutionary model, we are prepared to enhance the trees and multiple sequence alignments of the nuclear families extracted from a second-generation database. Enhancements can be pre-computed at any point in the assembly of an evolutionary model, and are worth storing in a secondary database, as they are used repeatedly throughout any interpretive proteomics efforts. Like the basic model itself, the enhancements will improve as more sequences are collected. Eventually, as they come to be accepted for individual families, the enhancements will be incorporated within the primary database itself. Table 6 lists a set of the most used enhancements.

The first enhancement assigns f_2 values to nodes in the tree. Nodes having lower f_2 values represent events that are more ancient than events represented by nodes with higher f_2 values. These can be used to place TREX lengths for each branch, as discussed above.

The next enhancement places a root on the tree, where the root is the point on the tree that represents the most ancient sequence. When a root cannot be assigned to a specific point, it is useful to identify a region on the tree that contains the root.

Table 6
Enhancing the evolutionary model for a nuclear family

Assigning TREX f_2 values to nodes in the tree
Placing the root on the tree
Placing gaps on the tree
Reference K_a/K_s values for individual branches on the tree to the average K_a/K_s on the tree; then to sub-branches

Two ways are available to place a root on the tree for a nuclear family, given a naturally organized database. The first is to find an outgroup family. When one exists for a family, it is recorded in the MasterCatalog as a bridge. Conventional methods can be used to root a tree using a bridged family as an outgroup.

The second approach exploits the TREX methods to find regions of the tree where the root might lie. Here, the root is the node with the lowest f_2 value, if the two-fold redundant silent sites have not suffered so many mutations that their nucleotides have equilibrated, and if the assumption of constant codon bias holds across the tree. Alternatively, the root lies within the region of the tree where equilibration is observed. Where equilibration has not occurred, the TREX values permit us to place approximate dates for nodes on the tree.

By placing a root on the tree, one puts a direction to time for every branch that is beneath it. This is useful when correlating the molecular record with the paleontological and geological records, as noted below.

A further enhancement places insertions and deletion events (indels) on the preferred tree. Placing indels on a tree is becomes increasingly easier as the tree becomes more articulated. Obviously, the gain or loss of a chunk of peptide sequence is a dramatic change in the structure of a protein, more dramatic in most cases than a point mutation. The more precisely indel events can be mapped to the tree, the more likely they are to be interpretable in terms of natural history.

An enhanced, second generation, evolutionary model for a family of proteins provides information far beyond the information contained in the "first generation" models collected in Dayhoff's *Atlas*, or the Hovergen, Pfam, DOMO, COG, or TIGRfam databases. As the coverage of the global proteome improves, we expect these enhancements to stabilize, and eventually become incorporated within the MasterCatalog families.

Functional inference from reconstructed evolutionary biology involving rectified databases (firebird): single family analysis

While the first law of Structure Theory in Organic Chemistry holds that all behaviors of a molecule are determined by its molecular structure, and while these include biological behaviors, the complexity of interactions between the many molecules in a living cell makes direct inference of molecular behavior and molecular function impossible, at least today, with current theory. Indeed, the complexity of interaction between an organic molecule and the water molecules that dissolve it is too great to permit any currently available theory to be useful, either predictively or manipulatively.

These considerations encouraged many to believe in 1990 that the task of predicting the conformation of proteins from a set of sequence data was not likely to be solved. Indeed, number crunching approaches that attempted to directly model the protein molecule, the energetics of interactions between its pieces, and the structure of the surrounding solvent, had failed to provide convincing solutions to the problem. They even had failed to show measurable progress towards solutions.

A decade ago, we offered an alternative approach to number crunching as a way to extract information about the conformation of proteins, starting from a set of homologous sequences diverging under functional constraints. This approach avoided the problems associated with a frontal assault on the protein fold prediction problem, and generated the first convincing tools to solve the problem. In a variety of settings, including the project known as the "Critical Assessment of Structure Prediction" (CASP) project, evolution-based tools have repeated provided accurate structural models, as well as occasional statements about protein function based on these (Gerloff et al., 1997). The approach is reviewed elsewhere (Benner et al., 1997), including on these pages (Benner et al., 1998).

The same types of approaches that are used to predict the conformation of proteins can be used to extract information about the function of a family of proteins. Indeed, predicting the fold of a protein is often the key step when making statements about functional behavior in proteins. As before, a set of homologous proteins diverging under functional constraints is assembled, and a model is built that makes explicit statements about the relationship between inferences about function are extracted when homologous protein sequences are presented as family, set within the context of an explicit evolutionary model, diverging under functional constraints.

To students who might view what follows as intimidating, we offer this process as a "game". The goal of the game is to draw as many inferences as possible about a family of proteins, without concern as to how reliable these inferences might be. The best inferences include as many interconnections as possible, with as broad a scope as possible, and lead to as many experimentally testable hypotheses as possible. Any information from "common knowledge" can be used. The winner of the game is the player who has generated the most testable hypotheses, that bring together the most (previously viewed as disparate) facts about the natural and biochemical worlds, and provides the most interesting framework for future work.

We will not be concerned about the reliability of inferences as one plays the game. This notion is, of course, alien to statisticians who often seek tests of reliability. The notion is not alien, however, to the experimental biologists. Weak hypotheses can be valuable if they suggest experiments that test them. Indeed, incorrect hypotheses can be valuable if they are testable.

What is "Function"?

Examining sequence data for inferences about function is fundamentally different from examining sequences for statements about conformation, or fold. The fold is a statement about the molecule itself, part of the trinity of constitution, configuration, and conformation that define structure of an organic molecule. It is expressed in universal terms, coordinates of atoms in space, or the relative positions of amino acid side chains.

Function, in contrast, is a linguistic construct. Those who ask "What is the function of my protein?" expect (Benner and Gaucher, 2001) a sentence or two of "cultural annotation" written in the language of the biologist. The answer might take, as an example, the form: "Your protein is a leptin, which regulates the feeding behavior of mice. When the gene is mutated or deleted, the mouse becomes obese" (Zhang et al., 1994).

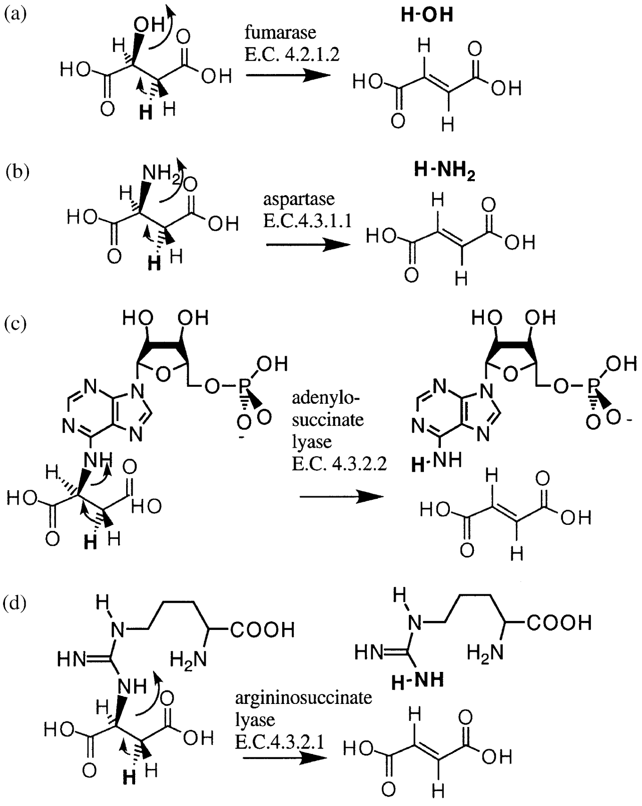
Such linguistic constructs are distinct from the underlying concept of fitness, of course. In principle, the contribution to fitness made by any specific behavior of a specific biomolecule is subject to operational measure. One must alter the structure of the biomolecule in a way that alters that behavior (and no other behavior). One must then introduce the altered biomolecule into a natural population living in a natural environment. One then must measure how the altered biomolecule is distributed in the population after an arbitrary number of generations.

As is often remarked in the literature, this experiment is difficult to do. Chemists, who do the same sort of thing when trying to connect chemical structure to behavior, fully appreciate the problem. There is no structural change in a molecule that affects only one of its behaviors, even for an isolated molecule in a test tube. The chemist's joke is that the structural change changes the behavior that you know about, as well as the behavior that you do not know about.

Despite this limit, over time, this approach in chemistry has generated a somewhat coherent (if elaborate) view of the relationship between molecular structure and molecular behavior. Doing so has required chemists to consciously *not* use the types of statistical analyses that are part of the culture of population biology, bioinformatics and molecular evolution. Instead, chemists favor a metalanguage that abstracts features of molecular structure, a cultural annotation of a sorts. Much can be learned from the development of chemical metalanguage that is applicable to the tasks presented to the modern interpretive proteomicist.

To make functional annotation, contemporary bioinformatics generally attempts to bridge chemical sequence to biological fitness using a doctrine of "functional equivalency" in the linguistic descriptions of function (for example, see Eisenberg et al., 2000). This doctrine writes a linguistic construct for a new protein sequence by expropriating the linguistic construct from another sequence having a similar chemical structure. A protein with unknown function is found in one genome. It is inferred, from its sequence similarity, to be homologous to a different protein found in a different organism. Homologous proteins are then assumed to have equivalent functions. The functional language assigned to the protein with the known function is then transferred to the new protein.

Long before the genomics revolution began, many cases were known where this doctrine failed (Benner and Ellington, 1988). Fig. 9 illustrates one example. Here, four proteins from microbial metabolism, adenylosuccinate lyase, argininosuccinate lyase, aspartase, and fumarase clearly group into homologous pairs based on sequence similarity, and are part of an evolutionary superfamily that includes all four proteins (Aimi *et al.*1990). One protein is involved in nucleic acid biosynthesis, another is involved in amino acid biosynthesis, another is involved in amino acid degradation, and the last is involved in central metabolism,



```
LPENEPGSSIMPGKVNPTQC fumarase
LPELQAGSSIMPAKVNPVVP aspartase
FEKDQIGSSAMPYKRNPMRS adenylosuccinate lyase
SDRVTSGSSLMPQKKNPDAL argininosuccinate lyase
      *** ** * **
```

Fig. 9. Homologous enzymes catalyze four reactions: (a) in central metabolism (the citric acid cycle) (b) in amino acid degradation, (c) in nucleic acid biosynthesis, and (d) in amino acid biosynthesis. The enzymes are indisputably homologous; even a simple BLAST sequence search identifies significant similarities. The catalyzed reactions are analogous from the perspective of organic chemistry. The functions of the proteins, from their roles in pathways, are quite different. An annotation strategy that assumes homologous proteins confer fitness in their host organisms in an analogous way would be defeated by this example.

however. The biologist certainly does not regard the function of these proteins as equivalent.

But should they? All of these proteins use fumarate as a substrate. They all, in the language of the chemist, add the elements of H–X to fumarate using a Michael reaction, where the carboxylic acid functional group acts as an electron sink. This type of language is very close to that used by the Enzyme Commission when it assigns "EC" numbers to enzymes. In the language of the chemist, all of these proteins have analogous function because they all catalyze an E2 addition reaction to fumarate. Evolutionary recruitment in this family presumably occurred because of this mechanistic similarity (Gerlt and Babbitt, 1998).

The point to be made here is not that one cannot infer function by homology alone. Nor do we wish to argue that the biologist's view of function is right, while the Enzyme Commission's view is wrong. Rather, the point to be taken is that the analysis of function is tied to the language used to describe it. The language used to describe the systems determines whether one sees "equivalency" or "non-equivalency".

Orthologs and Paralogs: Functional Analogs?

The organismic contexts of orthologs and paralogs are sufficiently different that the two can be analyzed separately in functional terms. Some degree of functional non-analogy is especially likely in paralogs, even when the basics of their behaviors are similar. If a gene duplications creates two genes, and if the duplicates are retained for long periods of time within a single genome following the duplication event, then it is nearly axiomatic that the duplicates have not served truly redundant functions. If they have, then one should have been lost (where "loss" includes conversion to a pseudogene) (Trabesinger-Ruef et al., 1996) without any corresponding loss of fitness.

It is commonly believed that orthologs are more likely to have analogous functions than paralogs. For this reason, databases that explicitly collect orthologous sequences have been constructed to facilitate the study of proteins with presumably analogous function. Particularly well known is the COG database (Tatusov et al., 1997) developed by Koonin and his collaborators.

Certainly, to the extent that the environments of the descendent taxa are analogous, and the demands imposed by natural selection in the two lineages are analogous, the behaviors expected from orthologous proteins in the two taxa are expected to be analogous. Indeed, to the extent that these conditions hold, all of the differences in the sequences of two orthologous proteins are expected to be attributable to neutral drift.

At the same time, it is clear that two species living in the same physical space, almost by axiom, cannot have identical strategies for survival. This, in turn, implies that two orthologous proteins may not contribute to fitness in exactly the same way in two species, nor are the behaviors demanded by the two environments exactly the same. This implies the possibility that some of the changes in the sequences of the two proteins may reflect differences in the behaviors of the two proteins that are, respectively, optimized for the two environments.

Encapsulated within these comments are many heated debates in molecular evolution. We cannot review the neutralist–selectionist dispute here. Indeed, we believe that this debate has been largely unhelpful to the science, casting issues in an "either–or" fashion about molecules in general, ignoring the evident reality that the question must be addressed using a formalism that is not "either–or", and requiring molecule-by-molecule analysis, rather than treating molecules as statistical aggregates. In these respects, the debate reminded us, as chemists, of the "non-classical carbonium ion" debate in organic chemistry (Brown 1977), and the "transition state stabilization versus "ground state destabilization" debate in mechanistic enzymology. These debates also suffered from a formalism that grouped molecules together, asked about the behavior of molecules "in general", and did not focus on individual molecular structures.

For the purpose of developing interpretive proteomics tools, the principal challenge is to avoid being paralyzed by issues that are important to the principals in the controversy.

Changing Functional Behavior

Behind annotation transfer stands the notion that proteins related by common ancestry perform analogous functions in different organisms. A useful dialectic, therefore, can be established with a tool that suggests that function is *not* analogous in homologs.[4]

In this section, we discuss tools that create inferences that functional behavior within a family has changed. We begin with models obtained from the MasterCatalog, enhanced and rectified as discussed above. For each family, we have in hand:

(a) A collection of homologous protein sequences, obtained first from the most recently indexed version of the MasterCatalog, augmented and/or culled as outlined above.
(b) Top line annotations for each of the family members, which provide an overview of how the community regards the function of the protein, including those that might be found in the extended and superfamilies.
(c) An evolutionary tree that captures the best model of the historical past, using the tools that the user prefers, augmented by alternative trees that will support robustness tests, and enhanced with the assignment of specific insertions, deletions, and amino acid replacements to specific branches of the tree, f_2 values dating nodes in the tree, a root on the tree, and normalized K_a/K_s values.
(d) A multiple sequence alignment, also adjusted to reflect the preferred tools of the user, including gap placement.

[4] The term is from Dayhoff. Modern terminology prefers the term "replacement" at the amino acid level, reserving "mutation" for an event that occurs in a DNA molecule.

(e) Bridges, which indicate other families in the database that might be more distant homologs, and an understanding of the superfamily within which the family is embedded.
(f) A set of reconstructed ancestral sequences for ancient proteins from now-extinct organisms at branch points in the tree. These include a reconstructed "founder" sequence near the root of the tree, the most ancient sequence from which all of the members of the nuclear family are descendent.
(g) Alignment of the MSA with a consensus three-dimensional structure obtained, where possible, from experimental data, together with a predicted secondary structure.
(h) A summary of the general properties of the divergent evolution within the family, including an overall view of mutability and the adaptive history.
(i) A list of all of the events that have occurred within the family (mutations at the DNA level, replacements at the protein level) assigned to specific branches on the tree, and reconstructed sequences throughout the tree.

We then ask what features of change within this evolutionary model are sufficiently indicative of changes in functional behavior that it is worthwhile considering hypotheses relating to them. Further, if we can identify these features, we will ask what can be inferred about when and where they occurred.

Sequence Change as a Surrogate for Change in Functional Behavior

The simplest tool to detect a change in functional behavior looks for a change in sequence. Certainly, without a change in sequence, the behavior of a protein cannot change. Conversely, it is axiomatic (from Structure Theory) that whenever two proteins differ in sequence, even at one site, their behaviors differ. The differences in behavior can be small, where "small" is a human perception derived from an operation that experimentally measures the difference. Indeed, the difference might be so small that the differences cannot be detected by any particular experiment. But the difference must be there, and this difference in behavior could conceivably have an impact on fitness.

The neutralist–selectionist dispute has made clear that is difficult to know a priori whether a behavioral difference, large or small, associated with changes in sequence, few or many, has an impact on function. Therefore, an indirect approach is needed to evaluate the potential that sequence changes reflected positive selection for them, as opposed to the absence of selection against them. The first, according to the model, represents adaptive change, while the second represents neutral drift.

One approach seeks to define and interpret rates of change in sequence, the number of amino acid replacements per unit time. As originally proposed, a certain rate of protein sequence change might be expected purely from neutral drift. This might accumulate with approximately clock-like behavior, where amino acid replacements accumulate with a time-invariant rate constant, number of replacements per site per unit time. If this were true, then an episode of sequence divergence that contains more numerous replacements than expected for the time elapsed would be one that holds a functional change.

From a practical perspective, it is difficult to apply this approach to reconstructed events on an evolutionary tree. Prior to the introduction of the TREX metric, there has been no particularly useful tool to date nodes on a tree. Even today, there remains no reliable method to date nodes on a tree when silent sites have equilibrated.

More seriously, however, an overwhelming body of empirical data suggests that no clock-like rate constant for amino acid replacement can be found universally in protein sequences (Ayala, 1999). The historical rate of replacement in amino acids can vary over orders of magnitude between protein families. This could mean that some protein families are suffering more adaptive change than other, of course. It could mean, however, that proteins whose sequences are rapidly diverging are simply subject to fewer functional constraints; more of their amino acids serve no "function", and then are available to drift.

One approach to resolve this problem seeks the maximum rate of drift. Let us assume that we could learn the rate by which mutations at the DNA level were generated and fixed. Let us also assume that this underlying mutation rate were time-invariant, the same for all sites, and intrinsic to a genome (and therefore the same for all proteins). Then, proteins that evolved more slowly than they could, given this intrinsic rate, must be subjects of purifying selection. Proteins evolving more rapidly must be subjects of positive selections, and those evolving at the same rate as mutations were introduced and fixed were subject to no selection pressure.

To estimate the natural rate of mutation, the redundancy of the genetic code is frequently exploited. Due to the redundancy of the genetic code, mutations at the DNA level can be either synonymous or non-synonymous. Let us assume that synonymous substitutions, which do not alter the sequence of the encoded protein, have no impact on fitness. They therefore are sites that suffer mutation without functional constraint, positive or negative. The rate at which mutations accumulate at silent sites, therefore, will reflect the rate at which mutations occur and are fixed independent of selective pressure. Mutations at silent sites were therefore proposed as approximate metrics of time.

Non-synonymous mutations, in contrast, result in changes in amino acid sequence which can alter the folding, kinetics, binding specificity, or binding affinity of the protein. They can therefore be the targets of selective pressures. Starting in 1985, Li and coworkers proposed that the ratio of non-synonymous to synonymous mutations might be an indicator of adaptive change (Li et al., 1985). They applied this first to compare sequences at the leaves of evolutionary trees (leaf-leaf comparisons), A decade later, the tool was applied to node–node comparisons (Endo et al., 1996, Trabesinger-Ruef et al., 1996, Messier and Stewart, 1997).

Let us examine the proposal with greater detail. First, let us assume that the rate of all nucleotide substitutions (transitions and transversions) is equal, and that each site in the DNA sequence suffers mutation with equal frequency. Consider an idealized gene encoding a protein that is simply a string of valines (chosen because valine is encoded by a four-fold redundant coding system). If mutations are introduced into this sequence at random, initially the ratio of non-synonymous mutations to synonymous mutations will be 2:1. Every mutation at the first and second positions will convert the encoded amino acid to something other than valine, while every mutation at the third position will be silent.

In such a simple system, it is also easy to calculate the ratio of non-synonymous and synonymous sites. Here again, it is 2:1, which the first and second sites considered to be non-synonymous sites, and the third considered to be a synonymous site. Hence, if divergence is completely without functional constraints (if, for example, our gene is a pseudogene), then the ratio of non-synonymous to synonymous substitutions (2), normalized for the number of non-synonymous and synonymous sites (also 2) will be unity.

The ratio of non-synonymous to synonymous substitutions, normalized for the number of non-synonymous and synonymous sites, is the K_a/K_s ratio. This is equal to unity for a gene randomly diverging without functional constraints, such as a pseudogene. But can the ratio be used diagnostically?

It can for some cases. If the values of the K_a/K_s ratio is significantly greater than unity, this can be explained under the Darwinian paradigm only by invoking "positive selection". Here, non-synonymous mutations must have accumulated faster relative to synonymous mutations than can be explained by random fixation of neutral mutations. In episodes represented by branches of the tree where the K_a/K_s ratio is greater than unity, some of the mutant children must have been more fit than non-mutant children.

Many cases are now known where positive selection must have occurred. We collected these a few years ago and prepared an "adaptive evolution database" for many plants and vertebrates where adaptive evolution were possible (Liberles et al., 2001).

The K_a/K_s ratio as a metric for functional change has several obvious limitations. First, the number of characters used to calculate the ratio is no greater than the number of amino acids in a protein. Therefore, "fluctuation" error, a type of sampling error, can make the measurement less precise than desirable, especially with short protein sequences. There is little that can be done to address this problem with a protein of a fixed length.

Other limitations can be diminished, although not removed. For example, the K_a/K_s calculation cannot be applied when the time separating the two sequences is so large that the nucleotides at the silent sites have equilibrated. As node-node distances are shorter than leaf–leaf distances, this problem is first addressed by performing K_a/K_s calculations between nodes. This is the calculation that is built into the MasterCatalog.

The accuracy of a K_a/K_s ratio obtained from ancestral sequences is determined by the accuracy of the ancestral sequences, of course. This can always be improved by increasing numbers of derived sequences, which permits more reliable reconstruction of ancestral gene sequences. As the most certain feature of our post-genomic future is an increase in the number of sequences, ancestral sequences will improve over time.

Even so, there is reason to believe that real world factors will prevent K_a/K_s calculations from being performed indefinitely far back into the past. The number of speciation events may be insufficient to articulate a tree over a period of time judged relative to the silent site drift rate to permit reliable reconstruction. Extinction may have erased part of the record, placing a limit on the number of derived sequences that can today be found in the biosphere to increase the articulation of a tree. While future discoveries are difficult to anticipate, it is not clear that sufficient sequences have survived in the contemporary biosphere to support the reconstructions that would be needed to apply the K_a/K_s metric back to the divergence of the three primary kingdoms of life (archaebacteria, eubacteria, and eukaryotes).

Another limitation to the K_a/K_s ratio as a practical tool arises with poorly articulated trees in general. Adaptive episodes in sequence evolution can be brief, and be surrounded by periods of adaptive stasis. Thus, when reconstructing ancestral episodes of sequence evolution, it is possible (and perhaps likely) that individual branches of the evolutionary tree will contain both episodes of adaptive evolution and episodes of conservative evolution. Only if speciation events occurred right before and right after the episode of adaptive evolution, and if the relevant proteins from all of the derived taxa have been sequenced, will a branch isolate the episode of adaptive evolution, and a K_a/K_s ratio for that episode be calculable without dilution from other episodes. Otherwise, the high K_a/K_s ratios characteristic of adaptive change in a protein will be averaged with the low K_a/K_s ratios characteristic of no adaptive change. This implies that high K_a/K_s ratios that might alert the functional genomicist to a change in function in a protein can easily be diluted to the point where they cannot be recognized within a background of change.

Other limitations are more fundamental to the method. The first center around the assumptions of neutrality associated with synonymous mutations in coding regions. Codon selection bias implies that the mutation of a silent nucleotide need not be exactly neutral. While a modest codon bias does not have a large impact on the metric, a changing bias can. Codon bias can change between taxa that are "closely" (by human standards) related, especially in plants (Tiffin and Hahn, 2002).

This limitation will also be ameliorated as more whole genomes are sequenced. Complete genome sequences will enable us to reconstruct general features of ancestral genomes, such as codon bias, and how these features have changed historically. Within plants, for example, we may soon be able to identify branches within the taxonomical tree where codon bias changed. This will provide the information needed to correct the tools that apply the K_a/K_s metric. Whole genome analyses are discussed in detail below.

Another class of limitations in the K_a/K_s ratio as a metric for functional change arise from the chemical reality of proteins, and how their sequences are related to their behaviors. Behavior in a protein can radically change even if only a few amino acids are replaced, as long as the replacements are near an active site. As the K_a/K_s metric encompasses the entire protein, a few sites that suffer functionally significant replacement may be lost among a larger number of sites that need not change to produce new function.

This might not be problematic if these sites drifted, as this would contribute to the K_a term. Simple models of functional recruitment suggest that the remainder of the sequence may be constrained from drifting, however, even during an episode of functional change. At the very least, some residues must undoubtedly be conserved to retain the "core" behaviors required for both the old and the new functions. A well-recognized core property is the folding of the polypeptide chain itself. In general, recruitment retains the folded protein scaffold.

This means that a high K_a/K_s ratio may characterize only a few of the sites in the polypeptide sequence, with a low K_a/K_s ratio characterizing the remaining sites. The K_a/K_s ratio for the sequence as a whole can therefore be significantly less than unity, even though the sequence evolution includes changes are driven by selective forces.

We have suggested some crude "fixes" for this problem. These are often intuitively reasonable, and are very useful within the context of a biomedical research program, even though they are not supported by any statistical formalism. For example, it may be possible to identify adaptive episodes by comparing K_a/K_s ratios for different evolutionary episodes within a single protein family. If we assume that the number of positions that must be conserved in a protein family to conserve its core behaviors (e.g., folding) is constant throughout its evolutionary history, and if punctuated equilibrium (Gould and Eldredge, 1993) at the molecular level is the rule, then one might expect some episodes to have a K_a/K_s ratio to be higher than the rest in a biphasic (or multiphasic) distribution of K_a/K_s ratios. The episodes displaying higher K_a/K_s ratios would be associated with functional change, while those with low ratios would be associated with conservation of function, regardless of the absolute value of the ratio of their respective distribution maxima.

Again, the MasterCatalog model serves as a starting point, with its explicit reconstruction of the sequences of ancestral genes in the tree and pre-calculation of K_a/K_s values for every branch. We can now ask about the distribution of K_a/K_s values across the tree for that particular protein family. For families where functional behavior is conserved over most of the tree, then the typical K_a/K_s value for a typical branch might be used.

Systematic examination of the proteins with high K_a/K_s ratios may be one approach for identifying those biomolecules important for the distinctive properties of different organisms. For example, for 2820 orthologous rodent and human sequences, the average K_a/K_s ratio (calculated leaf-to-leaf) is approximately 0.2 (Makalowski and Boguski, 1998). Naively as this number is far below the value of unity that is the hallmark (for example) of pseudogene neutral drift, the K_a/K_s test might suggest that most mammalian genes have experienced purifying selection during recent evolution (the last 80 million years).

The K_a/K_s ratio does not lead to a useful hypothesis about functional change when its value lies between ca. 0.5 and ca. 1.0. At the high end of this region, pseudogene drift is always a possible interpretation, of course. If no other indicators of pseudogene status are present (such as stop codons), it is compelling to conclude that the protein is suffering an episode of adaptive evolution, where the K_a/K_s value has not convincingly surpassed unity because sites that are rapidly changing are being hidden beneath sites that must be conserved to retain core function.

The K_a/K_s value is one of the most widely recognized tools for detecting functional change. It is often incorrectly used, however. For example, in their recent analysis of the mosquito and Drosophila genomes, Bork and his colleagues used values of K_a/K_s as an indicator that individual genes are pseudogenes, overlooking the aspects of the metric outlined above that suggest, as a more likely interpretation, that these are proteins suffering adaptive evolution (Zdobnov et al., 2002).

Further, efforts to calculate a K_a/K_s value for a single site seem misplaced. It is useful to calculate the number of non-synonymous mutations at each site. These can indicate that natural selection returns again and again to that site to create adaptive changes. But the normalization can remain the number of synonymous substitutions per synonymous site over the entire sequence, as long as the entire sequence has evolved as a unit over the period of time of interest. Conversely, the number of non-synonymous substitutions per TREX unit is an equally sensible metric of functional adaptation at a site.

Example: Leptin

Even with these limitations, the K_a/K_s ratio is useful metric to draw inferences, if only at the level of hypothesis, that relate to change in function. These can be extremely valuable to a biomedical researcher, even if the hypothesis has an undefinable reliability, and is unsupported by any metric that a statistician might recognize.

The protein leptin, for example, is known from genetics to be related to the obesity phenotype in the mouse. Deletion of the gene from mouse led to overeating and obesity (Zhang et al., 1994). Following the discovery of the leptin gene in mouse, a human homolog was sought. This is almost certainly the ortholog, as judged by the TREX distances. Almost immediately, both academic and industrial biomedical researchers began research programs using leptin as a potential target for managing or treating human obesity. As of this year, over 165 grants funded by the National Institutes of Health make reference to leptin.

Some details of the molecular history of the leptin protein family, however, suggested that leptin might not be a clear target for drug development as an obesity gene in humans. A reconstruction of the evolutionary history of the leptin family (Figs. 10 and 11) found that as primates emerged from the cenancestor of mouse and human, the leptin gene underwent an episode of rapid sequence evolution involving many non-synonymous substitutions in the leptin gene (Benner et al., 1998). Indeed, the reconstructed evolutionary history of the gene family shows that the number of non-synonymous changes that accumulated in the gene during this episode, divided by the number of synonymous changes, normalized for the number of non-synonymous and synonymous sites is remarkably high. The K_a/K_s ratio in this episode is ca. 2.1 fold higher than that displayed by a pseudogene.

The only explanation consistent with Darwinian theory for this episode is that leptin was under positive selection pressure (Yang and Bielawski, 2000) as it entered the lineage leading to hominoid apes, perhaps 40 million years ago (MVA). Mutant forms of the primitive primate leptin evidently contributed more to the fitness of the primate descendants than non-mutant forms of the protein. This led us to suggest on these pages four years ago (Benner et al., 1998), that human leptin may not play a role in humans analogous to the role it plays in mice. At the very least, a primate model is recommended for pharmacological analysis of compounds targeted towards

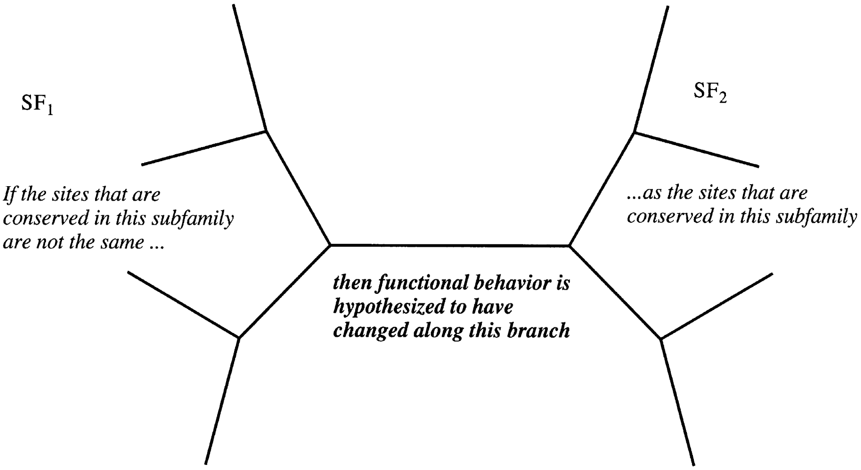

SF$_1$
SF$_2$
If the sites that are conserved in this subfamily are not the same ...
...as the sites that are conserved in this subfamily
then functional behavior is hypothesized to have changed along this branch

Fig. 10. Non-stationary behavior in the details of sequence evolution, in particular, if more conserved sites in one subfamily are not the same as those in another, then functionally significant change in behavior is implied along the branch that connects the two subfamilies. See Benner (1989).

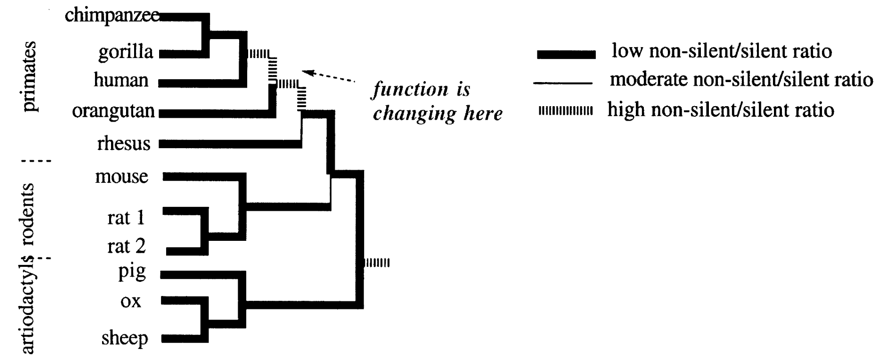

primates
rodents
artiodactyls
chimpanzee
gorilla
human
orangutan
rhesus
mouse
rat 1
rat 2
pig
ox
sheep
function is changing here
low non-silent/silent ratio
moderate non-silent/silent ratio
high non-silent/silent ratio

Fig. 11. Evolutionary analysis of the leptin family shows episodes of adaptive evolution separating primate and rodent leptins, indicated by high K_a/K_s values. This alerts the scientist to the possibility that the leptin homolog in humans need not have the same function as the leptin in mouse, as was concluded using the homology-implied-analogous function paradigm for database annotation. For the researcher, this implies that the mouse model may not have predictive value for humans (Benner et al., 1998).

this system. And now, articles are appearing with titles such as "Whatever happened to leptin?" (Chircurel, 2000), noting that "the hormone's precise physical role seems to vary from species to species." This was anticipated by the evolutionary analysis.

The leptin example shows how a second generation naturally organized database can support functional analysis in proteins, identifying targets of biomedical interest, and guiding pre-clinical drug development in animal models. When developing a drug targeted against leptin, the tree in Fig. 11 strongly suggests that one use a primate animal model, not a rodent animal model. As is known to all medicinal chemists, selection of a correct animal model is one of the most important things that determines the success or failure of pre-clinical research. The MASTERCATALOG helps make this decision correctly.

Eric Gaucher and I have submitted an analysis of the leptin family in light of its three-dimensional crystal structure. Since the paper is not yet in print, let me summarize only its broadest conclusions. We first identified the sites that were suffering amino acid replacements along the branch having a high K_a/K_s value. These are, of course, more likely to be the residue changes that are important for the change in function. These are not distributed randomly on the structure. Rather, they cluster, and cluster suggestively of changing interactions between leptin and other proteins. This leads to further hypotheses having biomedical significance, and illustrate how the FIREBIRD analysis is useful practically.

Correlating with Events in the Historical Past

With the MASTERCATALOG model for an evolutionary family as a starting point, and using its pre-computed K_a/K_s values, we can immediately identify segments of a tree where functional change might have occurred. This is at the level of hypothesis, which can be strong (if K_a/K_s is very significantly greater than unity), or weaker (if, for example, the case is based on the fact that the K_a/K_s for a branch is less than unity, but greater than the typical K_a/K_s branch in the protein family).

The next phase of interpretive analysis seeks temporal correlation. For this purpose, we need to extract dates for the tree. Classically, dates for nodes on trees have been assigned by noting the taxa that provided the derived sequences. We then refer to paleontological information to constrain the geological dates when the taxa might have diverged. This requires that the sequences within the family be true orthologs.

Once paleontological information is extracted, we can ask whether the molecular data are compatible with events in changing physiology at the time when the molecular changes occurred. This is easily illustrated using the leptin family of proteins. Whenever a mouse is foraging, he/she is just as likely to be food as to find food. Hominoid apes, in contrast, occupy a very different position in the food chain, and have a different feeding behavior. For mice, the instinct to forage must be under tight control, with over 90% of any mouse's offspring not surviving (on average) to themselves reproduce. Foraging mice take greater risks in the autumn than in the spring, balance opportunity with cost, and the corresponding behavioral instinct must be under strong selective pressure. In contrast, hominoid apes have more opportunity to learn.

The next step in the cycle of hypothesis generation asks: What did the ancestor of hominoid apes and rodents look like? Here, we must turn to the paleontological record.

The first lesson taught to paleontologists is that no fossil corresponds to an ancestor at the node from which two taxa branch. But as the paleontological record becomes more complete, it constrains with narrower and narrower bands the date that two taxa diverged. Further, a fossil from the paleontological and historical vicinity of the taxa that represents a last common ancestor can define very well the physiology of the true ancestor.

For example, the ancestor of hominoid apes and rodents lived in the mid-Cretaceous (Table 7). In this particular case, the fossil record has improved dramatically in its ability to describe the animal that was near the divergence of mouse and humans. A complete skeleton of Eomaia ("dawn mother") is preserved as a pressing, complete with fur imprint (we know how long the animal's hair was) from the very early Cretaceous in China (Ji et al., 2002).

Eomaia was more similar in many features of its physiology to mouse than hominoid apes. The implication is that it was not at the top of the food chain, like mouse, but not like human. Indeed, the episodes of rapid sequence evolution that is found on the leptin tree is associated with the increase in size of hominoid apes, a change that presumably is associated with a change in position in the food chain. It is not surprising that a protein like leptin, presumed to be involved in managing feeding behavior, would have an episode of sequence evolution at the time.

It is important to recognize that many of these discussions do not address the details that are of hot debate among people who specialize in these questions. For example, we do not know the relative sequence in which the mammal orders containing rodents (Rodentia), rabbits (Lagomorpha) and humans (Primata) diverged. The issue has been contentious in the past, is unresolved at present, and is likely to be both until the rabbit genome is completely sequenced, and perhaps even after that.

But the reason for the uncertainty in the tree is because the short lengths of the branches around which alternative trees differ.[5] These alternatives are of interest to specialists in the field. But, as outlined above, they interest us only if the biological conclusions that we draw are not drawn robustly with respect to small changes. Therefore, in constructing a evolution-based biological hypothesis, it is worth re-running the analysis with all possible tree topologies swapped around short branches, just to see if the biological hypothesis survives these swappings.

Alternative Measurements of Function: Non-Stationary Gamma Models

As noted above, different sites in a real protein sequence are under different selective constraints. As a consequence, natural selection tolerates replacements at some sites better than replacements at others. This distribution of mutability can be captured by a single parameter (alpha) in a gamma distribution.

The gamma model fits the statistician's culture, and several dozen papers have now appeared discussing and applying it. The statistical treatment, however, loses most of the information contained in the sequences themselves. Most specifically,

[5] This is directly analogous to the ability of evolution-based structure prediction tools to deny homology. Homology modelling and threading, for example, can only suggest that two proteins might be homologs. These tools cannot offer a statement that two proteins might *not* be homologs. This is one reason why the evolution-based tools for predicting the folded structure of proteins are so valuable.

Table 7
An approximate time scale for the paleontological record, and a layman's view of major features in the historical record near this time (with apologies)[a]

Million years before present	Name of the era; prominent features
0.0	Pleistocene (Cenozoic)
1.6	Pliocene (Cenozoic)
5	Miocene (Cenozoic)
24.5	Oligocene; (Cenozoic) at the beginning, have the massive cooling of the Earth; grasslands emerge; this is the radiation of the artiodactyl families deer/antelope/camels
38	Eocene (Cenozoic) this is the garden of Eden, warm weather.
54	Paleocene; (Cenozoic) warm weather, mammals take over from dinosaurs; the secondary orders of placentals diverge here (whales, artiodactyls) this date corresponds to a mass extinction
65	Cretaceous (Mesozoic) the principal orders of plancentals diverge here (primate, rodent, elephant, carnivora, ungulates); angiosperms become dominant
146	Jurassic, (Mesozoic) first angiosperms, according to Dilcher
208	Triassic (Mesozoic) by this point, mammals, birds (dinosaurs) and reptiles are diverged); this date corresponds to a mass extinction
250	Permian (Paleozoic)
280	Pennsylvanian (Paleozoic) coal beds
320	Mississippian (Paleozoic) coal beds, plants heavy on land without very successful land animals to eat them; animals are starting to go on to the land Hedges speaks of stem amphibians 338 MYA
345	Devonian (Paleozoic) 370/360 lobe finned fish become tetrapods ready to go on to land
395	Silurian (Paleozoic) bony fishes
438	Ordovician (Paleozoic) fishes
510	Cambrian (Paleozoic) tunicate versus other chordates probably diverge by end
543	(Paleozoic starts, the start of the Phanerozoic) This is recognized as the last date for the divergence of the major metazoan phyla, such as worm, fly, chordate Precambrian
1000	Major lineages probably established, probable eukaryotic fossils
2200	Oxygenic photosynthesis clearly established; certain microbial fossils
3800	First fossils (?)
4500	Earth forms

[a] The rodent–primate divergence, for example, was clearly not later than 70 MYA, probably not earlier than 150 MYA. The marsupial–placental divergence was certainly not later than 150 MYA; Hedges, using a protein-based molecular clock, suggests 173 MYA, while fossil evidence says 178–143, but is poorly attested. The mammal–archosaur divergence was certainly not after 310, and probably not before MYA. Hedges says that it occurred a bit more than 310. The land–fish divergence was certainly not after 338 (Hedges' date for amphibians), and probably not before 370 Hedges suggests 360 MYA. The bony fish-cartilaginous fish (shark) divergence was probably around 400 MYA

aggregation into one parameter (alpha) loses all of the information that is contained by knowledge of which amino acids are suffering replacement at which sites.

When functional behavior is changing, it is likely that the particular sites where replacements are tolerated will change too. Some sites that were not critical to the previous function (and therefore freer to drift) may perhaps become critical to the new function.

It is not clear that statistical models will help. Indeed, statisticians as notable as Felsenstein have despaired at the possibility of ever capturing non-stationary, time variant behavior within a statistical model (Felsenstein, 2001). Like most statisticians, Felsenstein proposes another hidden Markov model treatment. Such treatments, of course, bury useful information still deeper within a mathematical formalism.

More preferable is to return to a chemical analysis, one that treats the molecule and its individual sites individually. Consider a family of proteins divided into two subfamilies, SF_1 and SF_2, each with its own set of functional behaviors, where the two sets are not equal (Fig. 10). Let us also define a set of sites in each subfamily, C_1 and C_2, at which natural selection does not tolerate replacement, and a set of sites in each subfamily, V_1 and V_2, at which natural selection does tolerate replacement. Let us further assume that the differences in the sets of functional behaviors results in two inequalities: $C_1 \neq C_2$, and $V_1 \neq V_2$. This means that sites exist where replacement is not tolerated in SF_1, but is in SF_2, and where replacement is not tolerated in SF_2, but is in SF_1.

Given a sufficient articulation of the trees in the two subfamilies, the two inequalities will be apparent above fluctuation. This then provides a test for change in functional behavior independent of the test involving K_a/K_s ratios. In some senses, it is superior to the K_a/K_s ratio test. Changing specifics in the distribution of more and less replaceable sites in a protein sequence can be observed even after synonymous sites have suffered so many mutations that their occupancy has equilibrated.

Further, the analysis retains much more information, about which sites are involved. Knowing which sites are involved, we can apply other tests supported by the MasterCatalog. First, we can examine where the sites with changing mutability are located within the three-dimensional crystal structure. This process can bring to bear the insight and intuition of the organic chemist to bear on the problem.

The first case where this tool was applied, together with a crystallographic analysis, was reported on these pages in 1989 (Benner, 1989). The alcohol dehydrogenases from yeast and mammalian livers are homologous. They perform different functions, however. In different yeasts, the enzyme has the same, narrow substrate specificity, interconverting only acetaldehyde and ethanol, and this substrate specificity has clear physiological significance, as the catalytic process that recycles NADH to regenerate NAD^+ in the glycolytic pathway. One expects the amino acids lining the pocket in the enzyme where the substrate binds to be highly conserved to maintain this substrate specificity.

In contrast, the enzyme from mammalian liver plays (according to the best hypothesis) a role in the detoxification of foreign organic compounds, which themselves have varying (and not necessarily anticipatable) structures. Many mammals have paralogs of the liver alcohol dehydrogenase, having different substrate specificities. One expects that sites near the substrate binding site of mammalian ADH to be highly variable.

Benner (1989) presented a three-dimensional crystal structure highlighting sites that were variable in mammalian ADH subfamily, but conserved in the yeast ADH subfamily. The entire substrate binding region of the active site was highlighted. This is a graphic illustration of how a three-dimensional model can be used to make a compelling case that non-stationary behavior in the replacability at different sites indeed indicates change in function.

Gaucher et al., (2001a) made another compelling case by observing non-stationary, time-variant gamma distributions in the family of elongation factors related to EF-Tu. These proteins are involved in the translation of mRNA in protein synthesis, and serve to present charged tRNA molecules to the ribosome. They are among the most highly conserved proteins on Earth, and no one suspected (from a first generation evolutionary analysis) that they would display functional diversity. Indeed, they would seem to be archetypal examples of a protein that performs the "same" function in all three kingdoms of life. If transfer of the linguistic construct describing function from one member of a protein family to another is ever secure, it would seem to be secure with elongation factors.

This study began with a statistical perplexity. The alpha parameter for the subfamily of eukaryotic elongation factors, and the alpha parameter for the subfamily of bacterial elongation factors were comparable, but not comparable to the alpha value calculated to the family as a whole. Thirty EF-Tu/EF-1α protein sequences were aligned over 380 sites using the alignment program Darwin. Replacement rates per site for bacterial and eukaryotic EFs were estimated using a gamma-based, maximum likelihood (ML) model for protein sequences (JTT + Γ) and the phylogeny of Baldauf et al. (1996) for EF-Tu and EF-1α. An α of 0.78 was calculated for the entire tree, with a standard deviation (SD) of 0.05 using parametric bootstrapping (evolutionary simulations) (Swofford et al., 1996). The α values for the bacterial and eukaryotic subtrees were significantly different from that for the entire tree (0.46 and 0.38, respectively). These reductions in α for bacteria and eukaryotes alone are expected of a non-stationary process.

Thirty seven percent of the sites had essentially the same rate in the two groups (rate difference of ~0), as expected under a stationary gamma process. However, 18 and 21 sites had evidently evolved >2 standard deviations faster in bacteria than eukaryotes, and vice versa, respectively. These 10% of the sites are most responsible for the covarion characteristics of EF-Tu and EF-1α.

Residues displaying abnormal evolutionary behavior were then mapped to a three-dimensional model of the protein based on a crystal structure of ET-Tu. These were used to generate structural hypotheses for the different behavioral differences that were known. For example, bacterial EF-Tu binds GDP ~100-fold tighter than GTP. Eukaryotic EF-1α, in contrast, binds both with similar affinities. EF-Tu regenerates its active form by binding to the single-subunit nucleotide exchange factor EF-Ts. EF-1α requires the multi-subunit nucleotide exchange factor EF-1βγδ. EF-1α in eukaryotes also interacts with the cytoskeleton as it moves from the nucleus to the cytoplasm. EF-Tu, in bacteria, have no nucleus to move from.

Non-stationary behavior in this case indicated very subtle changes in functional behavior, for sure. But they are meaningful. The Firebird analysis led to predictions about the roles of specific residues that are functionally important, for different functions, in the two subfamilies of elongation factors. Several of these predictions have subsequently been validated (Gaucher et al., 2001b). It is interesting to note that virtually every specialist in the field of "elongation factors" had overlooked the phenomenon that was caught by the FIREBIRD analysis.

Homoplasy

So far, we have emphasized features of divergent evolution that indicate a change of function. High K_a/K_s ratios and non-stationary time-variant features of amino acid replacement can both point to a branch on a tree that represents the episode when functional behavior changed. Addition of information from secondary, tertiary, or quaternary structure can provide a confidence that can come only when chemical analysis is done to the chemical system.

These tools can also be metrics for the conservation of function. That is, a low K_a/K_s value, or a stationary gamma model, indicate that annotation transfer across the portion of the tree where it holds is likely not to be extremely deceptive.

Many features of the divergent evolution within a protein sequence family can provide indication of functional conservation. Let us consider just two. The first is homoplasy.

Homoplasy is defined as a character similarity that arose independently in different subfamilies of an evolutionary tree (Strickberger, 2000). Molecular homoplasy is best illustrated by an example (Fig. 12). Homoplasy so defined is the observed phenomenon; no statement is made as to the mechanism by which homoplasy arises. It may reflect selection pressures. The MASTERCATALOG gives us the opportunity to systematically search for molecular homoplasy in the database as a whole.

At one level, homoplasy is simply the statement that selective pressures are forcing the protein to select from a subset of the 20 standard amino acids. Thus, it is similar to the bias that is seen in membrane proteins, for example (where residues are chosen more frequently from a subset of hydrophobic amino acids than in the database as a whole). Homoplasy is more. Not only (in the example in Fig. 12) is position 30

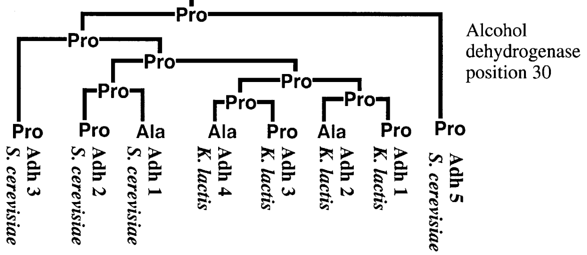


Fig. 12. An example of homoplasy taken from the evolution of alcohol dehydrogenase from yeast (position 30). No matter what the reconstructed ancestral sequences are, at at least three points in the tree, a P→A substitution occurred independently.

limited to A and P, but the selection pressures have toggled between the two more than once in the module's evolutionary history.

This is, of course, a signature that a functional constraint is conserved in the various branches of the tree across which homoplasy is observed. For this reason, molecular homoplasy is expected to be a contrarian signature to high K_a/K_s or non-stationary covarion behavior in a protein. We expect it to occur more frequently with proteins that are *not* undergoing functional recruitment.

Some informative features are already evident from preliminary work. For example, a preliminary search of 38 protein families with high resolution crystal structures identified over 2000 examples of molecular homoplasy. These were characterized first by the nature of the amino acids identified. A number of very obvious patterns emerged. First, the majority of the examples involve the interchange of hydrophobic side chains of nearly identical volume. The homoplasy involving I and V was the most frequent. It occurred 230 times in the dataset. The I/V molecular homoplasy was far more abundant than the next most popular hydrophobic/hydrophobic homoplasy, F/Y, which was found 68 times, and the I/L hydrophobic/hydrophobic homoplasy, which was found 44 times. As might be expected, the majority of these were buried in the three-dimensional structure of the protein.

The most interesting homoplasies are those that involve multiple steps. For example, the Pro/Gly homoplasy (at the codon level, CCN to GGN) requires two substitutions. Either of these alone creates a change in the encoded amino acid (CGN, Arg, or GCN, Ala). Observing examples of these without observing the intermediates anywhere else in the tree suggests that selection pressure is remarkably strong at this position, even though two amino acids appear to be nearly equally suited to perform function.

Molecular homoplasy indicates a constraint on structure that implies a constant behavior, which in turn implies a constant function. If this is true, it should correlate negatively with K_a/K_s ratios. That is, homoplasy should be found less frequently in branches separated by a branch with a high K_a/K_s ratio than in branches not separated by such a branch. Case studies developed under this project will develop ways to exploit such a correlation.

Compensatory Changes

The conservation of a fold after extensive divergences raises the possibility that amino acid substitutions at one position in a polypeptide chain might be compensated by substitutions elsewhere in a protein. For example, if a Gly at one position inside the folded protein core is replaced by a Trp, it might be necessary to substitute a Trp by a Gly at a position distant in the sequence but near in space to conserve the overall volume of the core, and therefore the overall folded structure. These assume that if a substitution is not compensated, the organism hosting the protein is less fit.

Individual examples of compensatory changes in proteins have been proposed (Oosawa and Simon, 1986), both by analysis of families of natural proteins with known structures (Lesk & Chothia, 1980, 1982; Chothia and Lesk, 1982; Altschuh et al., 1987; Bordo and Argos, 1990). In these examples, amino acid residues distant in the sequence but near in three-dimensional space in the folded structure have been observed to undergo simultaneous compensatory variation to conserve overall volume, charge, or hydrophobicity.

Compensatory covariation has been used in the prediction of the tertiary folds. For protein kinase (Benner and Gerloff, 1991), for example, an antiparallel beta sheet was predicted for the core of the first domain because of two specific compensatory changes identified in consecutive strands in the predicted secondary structural model. The subsequently determined crystal structure (Knighton et al., 1991) showed not only that antiparallel beta sheet existed, but that the side chains of the two residues undergoing compensatory covariation were indeed in contact.

Systematic studies have suggested, however, that the compensatory covariation generates only a small signal. The early work by Lesk and Chothia with the globin family found that replacements of hydrophobic residues in the core of the protein fold are usually accommodated by small shifts of secondary structural elements rather than by size complementary amino acid substitutions (Lesk and Chothia, 1980, 1982; Chothia and Lesk, 1982). More recent studies have suggested that a weak compensatory covariation signal might exist (Taylor and Hatrick, 1994; Shindyalov et al., 1994; Göbel et al., 1994; Neher, 1994). Some authors have doubted, however, that the signal is adequate to be useful in structure prediction (Taylor and Hatrick, 1994). Others have been more optimistic (Neher, 1994; Shindyalov et al., 1994). More recently, Chelvanayagam et al. pointed out that the signal might be improved if examples of compensatory covariation were sought within explicit evolutionary context (Chelvanayagam et al., 1997, 1998).

In the literature, compensatory changes have been sought by comparing the sequences of two extant proteins from contemporary organisms. In principle, any position where an amino acid residue had undergone substitution at any point in the time separating the two proteins via the common ancestor might be paired with any other position that had also suffered substitution in this time. Such an approach is problematic because the evolutionary time separating two contemporary protein sequences can be long; in years, it is twice the time since the most recent common ancestor of the two proteins.

With Kaoru Fukami from the National Institute of Genetics in Japan (Fukami-Kobayashi et al., 2002), we examined 71 families of proteins from the MasterCatalog to learn whether reconstructed ancestral sequences will generate a more useful signal for compensatory covariation than can be obtained by examining extant sequences. We noticed anecdotally that covariation was more likely to occur along branches with low K_a/K_s values. This makes sense, as compensation is necessary only if function is conserved. Case studies developed under this project will test this.

Frequently, the charge compensatory signal is weak, perhaps even weaker "than expected." We might be disappointed in this fact, because it limits the technological value of the signal (in predicting the three-dimensional fold of a protein, for example). Balancing this disappointment, however, may be the significance of the scientific implications of this observation.

Charge compensatory covariation might be weak because the coulombic interactions being sought may themselves be largely unimportant to the selective fitness of proteins. Gaining or losing them, in this view, has insufficient impact on fitness to ensure that natural selection will prevent uncompensated charge reversals from entering the sequence database. This implies a limit to the tool generally, one imposed by the physical organic chemistry of the protein sequences.

An alternative explanation should be considered, however. Observation of a compensatory pair of substitutions implies that natural selection preserved some global feature of a protein during the episode represented by the branch between two nodes. This, in turn, implies some degree of constancy in the behavior of the protein before and after the episode where compensatory change has occurred. In this view, compensatory substitution patterns should be observed only in protein families whose behavior must remain largely constant during this branch. This, in turn, implies that compensatory covariation should be observed only during episodes where "function," defined as the behavior that contributes to fitness, is largely conserved.

Conversely, when functional behavior is changing, there may be no need to compensate individual replacements in a sequence. Indeed, an uncompensated change is more likely to generate a protein with different behaviors, whose (now) different behaviors contribute most to the (now different) requirements for fitness. In this view, compensatory covariation should not be observed, or should be observed less frequently, whenever functional behavior is changing.

In this view, compensatory covariation is scarce (at least when compared to perhaps naive expectations) because branches of an evolutionary tree where functional behavior is rigorously conserved are scarce. This is, of course, a controversial suggestion, again relating to the neutralist–selectionist dispute.

Given this observation, compensatory substitutions may become a powerful tool in functional genomics, complementary to K_a/K_s values that are widely used to detect change in functional behavior (Li et al., 1985). Here, compensatory changes would indicate functional constancy, while uncompensated changes would indicate functional change. Because compensatory analysis rests on protein sequences, while the K_a/K_s value requires measurement of silent substitution rates, and because silent substitution rates are frequently rather high, this metric for functional recruitment may ultimately prove to be more valuable than K_a/K_s ratios.

A Combination of These

This discussion makes evident the power of second generation tools to analyze function within a single protein family. Unanticipated, however, is the power of these when combined. In this combination, reinforcing and contradicting metrics support with varying degrees the emergence of hypotheses. These are summarized in. Tables 8,9 and 10.

Table 8
A summary of tools used to analyze change in functional behavior

I. Tools that detect change in functional behavior along a branch
   A. High rates of amino acid replacement per unit time along a branch
   B. High ratios of silent to non-silent substitution along specific branches of an evolutionary tree including tools that address normalization issues
   C. Non-stationary gamma models in subfamilies connected by a branch
   D. Low amounts of compensatory covariation
   E. Low amounts of homoplasy across the branch

II. Tools that indicate conservation of functional behavior along a branch
   A. Compensatory changes
   B. Homoplasy across the branch
   C. Low rates of amino acid replacement per unit time along a branch
   D. Low ratios of silent to non-silent substitution along specific branches of an evolutionary tree including tools that address normalization issues

III. Tools that identify individual sites involved in changes in functionally significant behavior.
   A. Sites changing along branches with high rates of replacement.
   B. Sites changing in episodes with high K_a/K_s values, minus sites changing in episodes with low K_a/K_s values.
   C. Sites causing non-stationary gamma behavior
   D. Sites suffering replacement not randomly scattered on the folded protein
   E. Sites that suffer replacements repeatedly
   F. Replacement on the surface of the folded protein clustered in space and time

IV. Tools that identify individual sites involved in conserved of functionally significant behavior
   A. Sites suffering compensatory changes
   B. Sites displaying homoplasy
   C. Sites that do suffer replacement are scattered on the fold, generally on the surface

V. Tools that involve correlation between the evolutionary histories of two families of proteins.
   A. Correlating the topology of evolutionary trees in two families of proteins.
   B. Correlating the connectivity of proteins in a gene family.
   C. Dating events in the molecular history.
   D. Correlating evolutionary events in two protein families occurring at approximately the same time.
   E. Correlating evolutionary events in two protein families that are associated with analogous behavior involving expressed/silent ratios.

VI. Tools that involve correlation between the evolutionary history of a family of proteins and the evolutionary history of the organism as known from some source other than genomic sequence data, including paleontology, geology, ecology, ontogeny, phylogeny, or systematics (collectively known as the "non-genomic record").
   A. Correlating the topology of an evolutionary trees and the non-genomic record.
   B. Correlating features of patterns of evolution in specific branches in the evolutionary tree with the non-genomic record.
   C. Correlating evolutionary events in several protein families occurring at approximately the same time with the non-genomic record.

Testing Hypotheses with Experimental Paleobiochemistry

As we have noted elsewhere, the hypothesis, generated in silico using these tools, can be tested by an experiment in resurrective paleobiochemistry. In this experiment, the proteins at the nodes on each end of the branch suspected of holding a discontinuity in functional behavior are resurrected and studied in the laboratory.

Table 9
Signatures of a branch having discontinuities in functional behavior

A high K_a/K_s
A K_a/K_s value higher than the norm for the rest of the protein family
A change in replaceable sites in different sub-branches joined by the branch
Different patterns of homoplasy on different sides of the branch
A branch with abnormally low compensatory covariation, compared with other branches in the tree
Non-canonical placement of mutable residues along this branch in the three dimensional structure.

Table 10
Tools to display residues with special evolutionary properties

Displays based on its sampling of the 20 amino acids (a profile)
Displays based on the mutability of the position across the tree
Displays of positions that have non-stationary patterns of mutability across the tree
Displays of positions that have "accelerated evolution"
Displays of positions that have a homoplasy history, locally and globally
Displays of positions that suffer mutation on branches with high K_a/K_s
Displays of positions that suffer mutation on branches with low homoplasy

The Firebird *Recipe*

For students who wish to play the game, we provide here a step-by-step recipe for performing a Firebird analysis of the single family.

1. Find in the MasterCatalog the families of modules from which the target protein is built.
    1.1. Download these
    1.2. Assemble full length sequences from these
2. Complete the inventory of homologs (optional)
    2.1. Add sequences of your own
    2.2. Identify genes that have been entered since the last MasterCatalog was built.
    2.3. Go to the current whole genomes, and get a complete inventory of the homologs in these.
3. Rectify the multiple sequence alignment and tree
    3.1. Apply alternative tools to construct the multiple sequence alignment and tree
    3.2. Apply alternative non-classical strategies to build the tree
        3.2.1. DNA instead protein-based analysis
        3.2.2. Distance-based tool using gamma models, or other refined distance metrics
        3.2.3. Incorporate paleontological information to constrain trees
        3.2.4. Use TREX distances to construct trees
        3.2.5. Hybrid constructions, applying different tools to different branches of the tree

- 3.2.6. Build trees with alternative sampling of the database (robustness to sample size)
- 3.3. Refine gap placement
  - 3.3.1. Identify gaps introduced by gene finding mistakes[6]
  - 3.3.2. Place indel events on specific branches of the tree
  - 3.3.3. Refine MSA if crystal structures available
- 3.4. Retain alternative alignments and trees for use to test robustness of biological conclusions

4. Correlate the tree with the paleontological and geological record
   - 4.1. Assigning TREX f_2 values to nodes in the tree
   - 4.2. Assign TREX distances to the tree
   - 4.3. Placing the root on the tree.
   - 4.4. Obtain a rate constant for silent transitions on branches of the tree
     - 4.4.1. Using datable orthologs from the tree itself
     - 4.4.2. From whole genome analysis (see below)
5. Perform a Firebird analysis
   - 5.1. Determine typical global characteristics of the tree
     - 5.1.1. PAM width
     - 5.1.2. Typical K_a/K_s ratio for a typical branch
     - 5.1.3. Calculate parameters of the gamma model for the tree overall
   - 5.2. Dissect the tree into subtrees
     - 5.2.1. PAM width of subtree
     - 5.2.2. Typical K_a/K_s ratio for a typical branch in subtree
     - 5.2.3. Calculate parameters of the gamma model for subtree
   - 5.3. Identify branches where function might be changing
     - 5.3.1. Identify all branches that have high rate of amino acid replacement per unit time
     - 5.3.2. Identify all branches that have high K_a/K_s ratios
     - 5.3.3. Identify all branches that have high K_a/K_s ratio relative to the typical ratio in the subfamily
     - 5.3.4. Identify all branches that have high K_a/K_s ratio relative to the typical ratio in the family
     - 5.3.5. Identify subtrees with different gamma model parameters
   - 5.4. Identify branches where function might be conserved
     - 5.4.1. Identify all branches that have low rate of amino acid replacement per unit time
     - 5.4.2. Identify all branches that have low K_a/K_s ratios
     - 5.4.3. Identify all branches that have low K_a/K_s ratio relative to the typical ratio in the subfamily

[6] We do not wish to deny the importance of determining the precise order of branching of phylogenetic trees around short branches. The fact remains, however, that alternative branchings do not, as a rule, alter the biomedically relevant conclusions that are drawn from an evolutionary analysis. Therefore, those interested in practical applications of genome sequences using evolutionary models need not concern themselves with controversies of this type.

5.4.4. Identify all branches that have low K_a/K_s ratio relative to the typical ratio in the family
5.4.5. Identify subtrees with uniform gamma model parameters
5.4.6. Identify branches with large amounts of compensatory covariation
5.4.7. Identify subfamilies large amounts of homoplasy

6. Residue by residue analysis
   6.1. Establish a correlation between the MSA and a representative crystal structure
   6.2. Identify sites potentially involved in adaptive change
      6.2.1. Sites changing along branches with high rates of replacement
      6.2.2. Sites changing in episodes with high K_a/K_s ratio
      6.2.3. Sites causing non-stationary gamma behavior
      6.2.4. Sites that suffer replacements repeatedly
   6.3. Map sites potentially involved in adaptive change on the crystal structure
      6.3.1. Identify such sites that are on the surface
      6.3.2. Identify such sites that are near the active site
      6.3.3. Identify such sites that are interior to the fold.
      6.3.4. Analyze spatial relation of multiple sites.
   6.4. Identify sites potentially involved in adaptive stasis
      6.4.1. Sites that display homoplasy
      6.4.2. Sites that are highly conserved
      6.4.3. Sites that display compensatory replacement
   6.5. Map sites potentially involved in adaptive stasis on the crystal structure
      6.5.1. Identify such sites that are on the surface
      6.5.2. Identify such sites that are near the active site
      6.5.3. Identify such sites that are interior to the fold.
      6.5.4. Analyze spatial relation of multiple sites.
7. Consider correlations outside of the family
   7.1. With other protein families
   7.2. With non-sequence records, including records from paleontology, geology, ecology, ontogeny, phylogeny, or systematics (collectively known as the "non-genomic record").

Example. Why Do Pigs Have Three Paralogous Genes for Aromatase?

Logan Graddy, a masters degree candidate working with Rosie and Frank Simmen, presented a simple question: Why do pigs (*Sus scrofa*) have three genes encoding aromatase? Aromatases are enzymes, dependent on cytochrome P450, that catalyze a three step reaction that converts an androgenic steroid to an estrogenic steroid. The paralog structure of the aromatase gene family in vertebrates is complex. Two aromatase genes are known in goldfish, for example (Callard and Tchoudakova, 1997). In contrast, only a single gene is known in the horse (Boerboom et al., 1997), the rat (Hickey et al., 1990), and the mouse (Terashima et al., 1991). Oxen have both a functional gene and a pseudogene built from homologs of exons 2, 3, 5, 8, and 9 interspersed with a bovine repeat element (Fürbaß and Vanselow, 1995). In several mammalian species, including humans and rabbits, a single gene (Harada, 1988; Delarue et al, 1996) yields multiple forms of the mRNA for aromatase in different tissues via alternative splicing (Simpson et al., 1997; Delarue et al., 1998).

Logan expected to find two paralogous aromatases in pigs because two are found in goldfish. This expectation itself captures an evolutionary concept. In this concept, the last common ancestor of pigs and fish had two genes for aromatase, and the number of aromatase genes is conserved since pigs and goldfish diverged. If this model were true, we would ask what we know about that ancestor. It lived perhaps in an Ordovician ocean. It certainly laid eggs. It probably resembled a bony fish more than a shark or an amphibian, and certainly more than a pig.

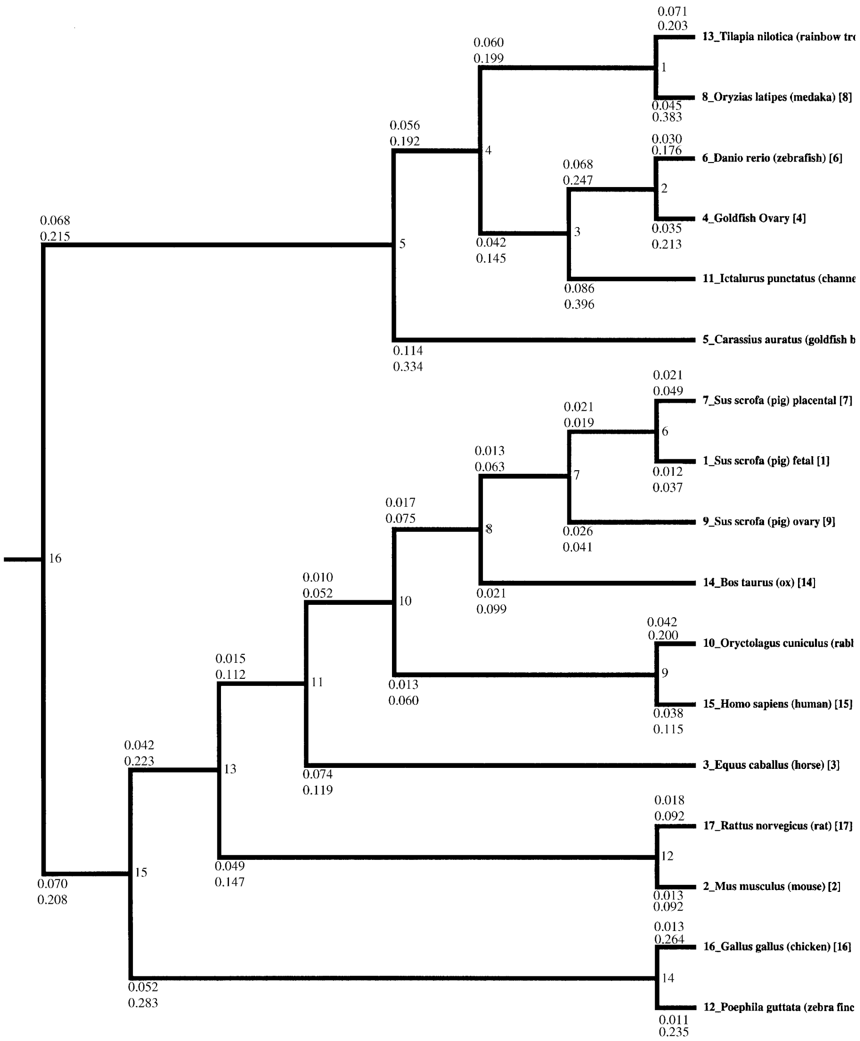

0.071
0.203
13_Tilapia nilotica (rainbow tr
0.060
0.199
1
8_Oryzias latipes (medaka) [8]
0.045
0.383
0.056
0.192
0.030
0.176
4
6_Danio rerio (zebrafish) [6]
0.068
0.247
2
0.068
0.215
4_Goldfish Ovary [4]
3
0.035
0.213
5
0.042
0.145
11_Ictalurus punctatus (channe
0.086
0.396
5_Carassius auratus (goldfish b
0.114
0.334
0.021
0.049
0.021
0.019
7_Sus scrofa (pig) placental [7]
6
0.013
0.063
1_Sus scrofa (pig) fetal [1]
7
0.012
0.037
0.017
0.075
9_Sus scrofa (pig) ovary [9]
8
0.026
0.041
16
14_Bos taurus (ox) [14]
0.010
0.052
0.021
0.099
10
0.042
0.200
10_Oryctolagus cuniculus (rabl
0.015
0.112
9
11
0.013
0.060
15_Homo sapiens (human) [15]
0.038
0.115
0.042
0.223
3_Equus caballus (horse) [3]
13
0.074
0.119
0.018
0.092
17_Rattus norvegicus (rat) [17]
12
15
0.049
0.147
0.070
0.208
2_Mus musculus (mouse) [2]
0.013
0.092
0.013
0.264
16_Gallus gallus (chicken) [16]
14
0.052
0.283
12_Poephila guttata (zebra finc
0.011
0.235

Fig. 13. A tree showing that the pig aromatase paralogs diverged after the divergence of pigs from oxen.

These thoughts would lead to the question: Why would an egg-laying fish-like creature living in an Ordovician ocean need two proteins that catalyze analogous reactions for the synthesis of estrogens? Why would these two functions be conserved in the subsequent 350 million years?

A simple click on a MasterCatalog family shows that this is not the story with the pig (Fig. 13), Here, the pig paralogs arose near the time when pigs diverged from oxen, perhaps 60 MYA in the Eocene. The average branch in the aromatase evolutionary tree has a value of K_a/K_s of 0.35. Inspection of the tree shows that the highest K_a/K_s values anywhere in the mammalian aromatase family (0.85 and 0.66) are found within the divergent evolution of the pig aromatases, in the branch leading to the embryonic and placental paralogs.

The evolutionary history of the aromatase family was then analyzed using the TREX analysis. Using a fixed single lineage first order rate constant of 3×10^{-9} changes per base per year, the TREX analysis indicated that fish and land vertebrates diverged 340 MYA, birds and mammals diverged 250 MYA, primates and ungulates diverged 73 MYA, horse and artiodactyls diverged 71 MYA, and pigs and ruminants diverged 62 MYA. Each of these dates is close to the date suggested by the paleontological record (Carroll, 1988).

The TREX dating was used to assess two alternative models to explain the triplication of aromatase gene family in pigs. The first, advanced by Callard and Tchoudakova (1997), holds that the physiological specialization of aromatases through the formation of paralogs occurred early in vertebrate divergence, perhaps 350 MYA, before fish and mammals diverged. If this were the case, then a functional explanation for the aromatase genes must be sought in fundamental features of vertebrate developmental biology, those that emerged early in vertebrate evolution. Conversely, the triplication of aromatase may occur in response to the domestication of pigs.[7] In this case, a functional explanation for the aromatase genes would be found in the selective pressures applied by breeding programs.

[7] This is a new invention. When an intron is missed, the protein sequence in which it was missed has an insertion relative to other homologs in a multiple sequence alignment. When a segment is removed under a mistaken impression that it is an intron, it leaves a gap relative to other homologs in a multiple sequence alignment. We could in principle identify incorrectly removed/missed introns by looking for gaps. The difficulty is that indel processes occur naturally. Therefore, the problem of intron assignment rectification based on MSA alignment analysis comes down to trying to detect which gaps in an alignment arise through true insertion/deletion events in the history of the protein family, and which arise through mistakes in gene finding/intron finding. The strategy is to recognize that when a gap is created through a true indel event, the segment inserted/deleted is not random. Rather, true indels occur in parsing regions, as defined by the Benner parent patent. Therefore, a gap that does not occur in a parsing region defined by the sequences of the other proteins in the MSA has a higher probability of arising from a misassigned intron (over or under). The amino acids found in positions just before a gap in the gapped sequence (A), just after a gap in the gapped sequence (B), just before the position of the gap in the aligned sequence (P), just after the position of the gap in the aligned sequence (R), and in the insertion in the ungapped sequence (Q), do not have the same distribution as the amino acids in the database as a whole, when the gap is derived from a true, historical event:

XXXA———BXXXX
XXXPQQQQQQQQRXXXX

The TREX distances separating the three pig isoforms range from 0.154 (corresponding to a distance of 51 million years between the proteins) to 0.199 (corresponding to a distance of 66 million years). Recognizing that the total distances between two proteins are twice the distance along a single lineage from the point of divergence to the modern protein (half of the distance occurs along one lineage after divergence, and half of the distance occurs along the other lineage), the TREX dates suggest that the first duplication led to the three porcine aromatase genes occurred ca. 33 MYA, and the second occurred ca. 25 MYA.

An evolutionary tree constructed from these TREX distances is consistent with these conclusions, showing that the porcine aromatases branched after the lineage leading to pig diverged from the lineage leading to ox (Fig. 13). This tree shows a different branching order for the three porcine paralogs than the tree based on amino acid sequences, something not uncommon in the presence of substantial adaptive evolution. Nevertheless, the data are consistent with an evolutionary model that holds that the ancestor of pig and oxen (approximated in the fossil record most closely by the now extinct *Diacodexis*, which lived perhaps 55 MYA) contained a single aromatase gene, and that the paralogous genes in pig arose ca. 25 million years later.[8,9] Thus, the paralogs in pig can be explained neither in terms of the fundamentals of vertebrate reproductive endocrinology (established in the

(*footnote continued*)

An empirically derived set of parameters can distinguish more likely and less likely gap assignments. The pattern of evolution in the region designated Q is also different when it is gappable. This includes both the rate of substitution (it is higher) and the amino acid distribution in the family (it is more like a parse). One can place the putative indel event on the evolutionary tree. When a true indel occurs, the rest of the protein responds with an episode of rapid sequence evolution, a change in mutability distribution, loss of compensatory covariation signal, and other events indicative of changing function. If the putative indel is not real, these associated signals will not be found.

[8] The analogy between the evolution of proteins and the evolution of language is a subject of frequent comment, and we cannot resist making a comment here. Thus, the proto-Indoeuropean language had words for "pig" (PIE **su-*, compare Tocharian B *suwo*, Latin *sus*, Greek *us*, Sanskrit *sukara*, Church Slavic *svinija*, Old High German *swin*, and English *sow*; and PIE **porko-*, compare Latin *porcus*, Church Slavic *prase*, Old High German *farah*, etc.), indicating that the pig has been under human domestication for at least 6000 years, enough time to have suffered a significant impact on its genotype through husbandry.

[9] A comment can be made about the uncertainty in the TREX dating. The uncertainty can arise from two sources, standard error (which arises from fluctuation) and systematic error (which arises from the fact that the evolutionary model does not represent actual evolution). The first can be calculated by standard statistical approaches using standard statistical assumptions. The second cannot be calculated, as too little is known about possible systematic errors in the evolutionary model. The f_2 distances are each based on ca. 120 two-fold redundant codon systems, and variances can be directly calculated. The calculated distance from the divergence of the three porcine enzymes to the type II enzyme is 31 million years, to isoform I is 32 million years, and to isoform III is 30 million years. Thus, the average reported (31 MYA) could be as low as 30 and as high as 32 MYA. All of these dates are in the Oligocene, after the first episode of cooling. The divergence of isoform I and III ranges from 24-26 MYA. These uncertainties are less than the uncertainties associated with the dating (from the fossil record) used to set the molecular clock. Further, the uncertainties are far smaller than are needed to distinguish the three hypotheses that might be used to explain these paralogs, as arising when fish and pigs diverged, arising in the Oligocene-Miocene periods, or arising 6000 years ago as a consequence of domestication.

Oligocene), nor as a consequence of swine domestication (which occurred ca. 6000 years ago).

Instead, an understanding of why pigs have three genes for aromatase must lie in the environment of (and events that occurred during) a time on Earth 25–33 MYA. For this we turn to the paleontological, paleogeographical, and paleoclimatological records of that period, which is near the boundary between the Oligocene (38–25 MYA) and the Miocene (25–5 MYA), two epochs in the Cenozoic "Age of Mammals" (Prothero, 1994). This period is an unusual one in the history of the Earth. When characterized globally, the Earth during the Eocene (54–38 MYA) was warm and tropical, evidently free of ice over the entire planet. By the end of the Eocene, however, the Earth had begun to suffer a dramatic cooling that was to lower the mean annual temperature by as much as 15°C (Wolfe, 1978). Areas of the planet became covered with ice. And the impact of the cooling on the biosphere was dramatic. For example, perhaps 80% of the North American faunal genera became extinct (Prothero, 1994, pp. 113–114; Stucky, 1990). By the end of the Oligocene and into the Miocene 25 MYA, however, the global cooling abated, the climate turned warmer, and the biosphere became more tropical (Azanza, 1993).

Did this climate change occur in the environment where the ancestors of modern pigs were living just before the Oligocene–Miocene boundary? At this time, the North American and Eurasian fauna were geographically isolated. Modern peccaries (*Tayassuidae*), not pigs, emerged in the New World from ancestral suids that immigrated from Asia. North America cannot be the site for the triplication of the aromatase genes in pig, therefore, and its climate 25–33 MYA is irrelevant to an explanation for the triplication of the aromatase genes in pigs.

Instead, modern pigs most likely emerged in Europe near the end of the Oligocene ((Cooke and Wilkinson, 1978), but see also (Pilgrim, 1941)) from more primitive entelodonts such as *Archaeotherium*. During the Oligocene, the Dichobunids (the most probable ancestral stock) were most abundant in Europe. Likewise, the first true pig, *Propalaeochoerus*, from the late Oligocene, was common only in Europe (Cooke and Wilkinson, 1978; Carroll, 1988). This makes the paleoenvironment of Europe near the Oligocene–Miocene boundary relevant to the functional implications of the aromatase gene triplication in pigs.

Various paleobiological evidence suggests that the climate in Europe also deteriorated in the Oligocene and warmed in the Miocene. A study of amphibian distribution in the Oligocene of Europe, for example, is consistent with a significant drop of mean annual temperatures in the European Oligocene. In the Miocene, amphibians populations rebounded, corresponding to an improvement in the climate (Rocek, 1996). Likewise, analysis of the deer population suggested a subtropical climate returning to Europe in the early Miocene (Anzanza, 1993). The Iberian peninsula in the early Miocene had an intertropical to subtropical climate (Murelaga et al., 1999). Crocodiles also returned to Europe at the Oligocene–Miocene boundary (Antunes and Cahuzac, 1999). The presence of arboreal primates in the European Miocene also suggests a forested environment (Qi and Beard, 1998). Each of these facts (and many others) suggests that the second duplication of the aromatase gene in pigs occurred at the same time as the return of subtropical and warm temperate forests and woodlands to Europe, the type of environment for which suids are best adapted (Fortelius et al., 1996).

Immediately thereafter, the suids underwent a significant radiative divergence, and came to occupy all of the Old World. By the early Miocene, the two basal members that were to lead to all modern pigs, *Hyotherium* and *Xenochoerus*, were widespread in Europe, Asia, and Africa. The amelioration of the climate evidently assisted in this spread. For example, the pigs now in Africa apparently came from southwest Asia in the Early Miocene. A fossil of this date of a tetraconodontine pig has been reported from the Levant (van der Made and Tuna, 1999), through which the pigs would have migrated to get from Eurasia to Africa, and which was a tropical environment at the beginning of the Miocene (Tchernov, 1992). In the middle and late Miocene, modern suids had diversified in Europe in further response to the change in the paleoclimate (Fortelius et al., 1996).

Why might a change in climate with a return of forested (and perhaps tropical) ecosystems have led to a selection of pigs that had three different aromatase genes? We turned to porcine reproductive physiology for insight. We recently found that the type III aromatase was expressed by the embryo between day 11 and day 13 following fertilization, during the late pre-implantation period (Choi et al., 1997a,b). The estrogen generated by the type III isoform causes uterine undulation. This undulation, in turn, is expected to cause the spacing of the ca. 30 eggs that are fertilized in a typical conception, which eventually yield the 8–12 piglets that are normally birthed. In pigs, if the litter does not contain at least 5 individuals, the entire conception is aborted. Thus, the embryonic form of aromatase may have a role in spacing the embryos uniformly around the uterus, and preventing abortion. These are useful adaptations if one wants to have an increased litter size.

Evidence in the paleontological record suggests that the size of the litter in pigs increased dramatically 25–30 MYA, at the same time as isoform III of aromatase was generated by triplication, the local paleoclimate warmed, and the pigs began a major radiative divergence. The ancestral suid *Archaeotherium*, disappearing from the fossil record at the end of the Oligocene, may have given birth to a single pup. All of the contemporary forms of pigs arising from the divergence of Hyotherium and Xenochoerus, known from the Early Miocene, have large litter sizes. Further, *Archaeomeryx*, the early Eocene artiodactyl that is presumed to be the ancestral ruminant, resembles the contemporary chevrotain, which also births a single pup.

The biogeography of the suids was again consulted to test the hypothesis that litter size increased in the suids near the time that the climate changed and the aromatase gene triplicated. As noted above, peccaries were isolated in the New World in the Early Oligocene, before the TREX-derived date for the triplication of the aromatase gene in the Old World pigs. Consistent with the model, the peccary has only 1–2 offspring. The model predicts as well that the peccary should have only a single aromatase gene.

The molecular biological, fossil, paleoecological, and physiological evidence are all consistent with a model that proposes that climate changes in Europe at the end of the Oligocene selected for pigs that had larger litter sizes. The successful lineage generated a new embryo aromatase by gene duplication, and expressed it at the time of implantation, forming the molecular basis of the physiology that enabled large litter sizes. It is possible to speculate on why a conversion from an open, savannah like environment to a forested environment might enable larger litter sizes. Contemporary Savannah babies are large and born with the ability to run, presumably because hiding is no alternative. In contrast, in a forested environment, pups are easier to hide, permitting them to be smaller and less precocious at birth, permitting in turn a larger number of pups for the same total birth weight. Indeed, the contemporary *Sus scrofa* sow hides her piglets in earthen hollows covered with leaves (Eisenberg, 1981).

Implantation is one of the least well understood steps in mammalian reproductive biology, including human reproductive biology. Implantation is, of course, found only in mammal reproductive physiology, and is itself therefore a relatively recent innovation in physiology, emerging perhaps 200 million years ago. This analysis emphasizes the degree of innovation and experimentation that is continuing in mammalian reproductive physiology. Further, the analysis is a combination of computational informatics, geology, paleontology, physiology, molecular biology and chemistry. Analogous analyses should be applicable in functional genomics throughout the biological, biomedical and biochemical sciences, especially as genome projects are completed and as new tools become available to analyze genomic databases.

But what about the high K_a/K_s values? With Eric Gaucher's help, we retrieved the sites that were suffering replacement at the time of the high K_a/K_s values, and mapped them on to the three-dimensional structure of a homologous P450 enzyme whose structure had been done. The sites were not distributed randomly in the structure. Instead, they were found in two regions, the first near the active site, the second near the site where the P450 enzyme docks to its co-protein. These results suggested that the substrate specificity of the aromatase was changing during this episode of evolution. This may be consistent with recent reports that the substrate specificity of aromatases is indeed different in the different isoforms in pigs (Kao et al. 2001).

Analysis of the entire genome of a single species

Analysis of the function of individual families is artificial in a very fundamental way. A protein does not act in a vacuum. If the protein is an enzyme, then its substrate generally arises from another enzyme, and its product is generally consumed by another enzyme. As part of a regulatory network, proteins directly interact with other proteins, as substrates, as their substrates, and without an associated chemical reactions. Further, through regulatory and metabolic networks, the performance of a protein can influence proteins that are not in direct physical contact.

These networks are a key part of the definition of function for a protein. For this reason, we distinguish between "functional behavior," something that concerns (and can be measured for) an individual protein, and "function" itself.

The goal of an analysis of a single family is to generate as much information as possible about the interrelationship between a protein family, its members, their structure, and their behavior. This information then generates inferences and hypotheses about functional behavior.

Networks and pathways, however, can be identified only through a horizontal analysis that involves many, and perhaps all, families of proteins represented in the genome of an organism. As always, hypothesis generation involves correlation. A completed genome offers a particularly interesting environment within which to make correlations, however. Here, an evolutionary analysis looks back in time along the entire lineage that led to the organism.

For the purpose of this discussion, we shall assume that the starting point is a curated database of the proteome encoded by an organismic genome. The ideal is not frequently met, especially for higher organisms, and in particular for the human genome (where the word "draft" clearly applies). Ideally, the curated genome contains a list of all of the open reading frame that matches the sequence of the proteins that the genes encode with the sequence of the encoding gene. An evolutionary analysis of a genome is itself a tool for identifying open reading frames, of course, so it is possible to iterate the cycle of gene finding, evolutionary analysis, and then further gene finding.

Paralog Identification

The first step for analyzing a genome begins with a comprehensive identification of paralogs. This is, in fact, the only thing that can be done from an evolutionary perspective starting with the genome alone.

When using the MasterCatalog, however, the families have already been found. Therefore, it is possible to leap over the first step, and to generate directly a set of nuclear families containing all paralogs, together with all of the homologs that were identified in other species at the time of the last MasterCatalog build. The process occurs as follows:

1. Constructing the paralog families

1.1. We first identify in the MasterCatalog all families that contain at least one member that has the name of the target species as the species descriptor. The output is a list of MasterCatalog family numbers for sequence modules that contribute at least one polypeptide segment to one protein in the species proteome list. We print this list and generate a universally accessible electronic file containing it. A paralog analysis obviously needs to identify those families that have two or more members from the target species. We identify all of those that have exactly one as well, simply because later, we might wish to recover these if the families that they are in are bridged to other families that contain representatives of the target species.

1.2. We then recover from the database the full length sequences that correspond to the proteins in the MasterCatalog family, including those from the target genome and the non-target genome.

1.3. We then rectify the family of full length proteins

1.3.1. We remove duplicates from the target genome by comparing the entries in the MasterCatalog family with entries in the curated species sequence database (which is treated as the gold standard).

1.3.2. We remove entries from the MasterCatalog families that lack DNA sequences. Presumably, these do not include any sequences from the target species genome, but may include protein sequences from other species.

1.3.3. (Optional) Cull the size of all of the families (if the family contains more than ca. 80 sequences)

1.3.3.1. We may fragment the MasterCatalog family into subfamilies, one that contains all of the target genome, and one that contains no representatives of the target genome.

1.3.3.2. If there are fewer than ca. 80 members in the subfamily that contains all of the members of the target organism, then we use this subfamily.

1.3.3.3. If there are more than 80 members in the subfamily that contains all of the members of the target organism, then we divide this subfamily into subfamilies with fewer than 80 sequences, and proceed with separate subfamilies.

1.4. For each family of full length sequences, we create an SGML file.

1.5. We then submit the family of full length sequences to the Darwin server for an all-against-all comparison. The output, for every family that contains one or more proteins from the target genome is:

1.5.1. Output: A PAM distance matrix for pairs of proteins in the family.

1.5.2. Output: An f_2 matrix for or pairs of proteins in the family.

1.5.3. Output: A multiple sequence alignment, together with a NEXUS file, for every family.

1.5.4. Output: An evolutionary tree for every family

For the biologist browser, these are printed out in a loose leaf notebook, *Book of the Species*. This will become the index to a paper reference resource for use by those who prefer to browse in hand rather than on the screen (there are reasons to prefer this, even in the age of computers). The data are also recorded in an electronic version that is generally accessible as a resource. At the end of the age of the proteome, we expect these resources to be available for hundreds of organisms.

Inspection of the Modularization

We next must address the modularization question within the species proteome. Modularization becomes an issue whenever units of protein sequence are shuffled in the course of evolutionary history. This happens frequently in the sequences of higher organisms, as noted above. Therefore, the MasterCatalog will, in some cases, divide a full length protein sequence into pieces, and place those pieces into separate MasterCatalog families. Sometimes, in the analysis of a single genome, we want to keep this division. Sometimes, we do not. The rule is: We want to keep the division if it reflects actual gene shuffling within the datable history of the target genome. If, however, it reflects shuffling events that occur before the datable history, we are not as interested in it.

In practice, we handle this by looking at the family of proteins from the target genome. The process conceptually requires us to construct a rooted tree of the sequences in the MasterCatalog family, and note where the sequences from the target genome lie. Then we ask, is the sequence that caused the MasterCatalog to fragment the sequences from the target genome an "in-group" with respect to the target sequences? Or an outgroup? It is conceivable that the sequence that causes the modularization of sequences within the target genome lies within the target genome itself. In this case, the modularization within the MasterCatalog is retained. On the other hand, if the fragmentation–modularization of the full length sequence in the target genome is due to a protein that is an outgroup, then we do not include the modularization.

In this analysis, the power of a second generation naturally organized database is clear. Modularization is a difficult tool to implement, and cannot be implemented by any rationale that can be assessed by standard statistical methods. In practice, modularization must be iterative, and iteration can occur over years as the true relationship between proteins and their segments is revealed and appreciated. MasterCatalog is an enormously valuable resource because a first pass modularization has been completed. It will undoubtedly be revised as civilization completes more genomes. We may wish to revise the modularization as we continue with our own analysis of a specific protein or a specific genome. But the MasterCatalog allows us to begin the day doing biology, not worrying about a first pass modularization.

Lineage-Specific Resources Created from a Whole Genome Analysis

We next prepare a series of lineage-specific resources from the genome. These are secondary databases that are used repeatedly in the future analysis of the genome, making it sensible to pre-compute the information that they contain, and store it.

Lineage-specific Resource 1: A database of rectified pairs of paralogs within the target proteome. With each is associated a pairwise alignment of both DNA and protein sequences, a PAM distance (with a variance), an f_2 value, and a TREX distance, and the top line annotation of a set of proteins that are found within the family within the naturally organized database.

Lineage-specific Resource 2: The pairwise alignment rank ordered in decreasing f_2 value.

Lineage-specific Resource 3: A histogram that records the collects of paralogization events within the genome in clusters based on the f_2 values of the paralog pairs.

The next lineage specific resource involves dating. We have introduced the TREX dating tool above, and it forms the core of any effort to correlate the molecular record of a genome with the paleontological and geological record.

Any effort to correlate events in a genomic record with the geological and paleontological histories requires that we assign dates to points of gene duplication. This requires some reference to geological dates, which in turn arises from a combination of radioisotope dating of geological strata, and the association of fossils with dated strata, with the hope of constraining dates when specific taxa diverged.

Because paralogs are created within a single lineage unassociated with speciation, it is impossible to use paralogs to calibrate clock. Instead, a clock can be calibrated only by finding orthologous pairs of protein, where the date of divergence of the two taxa that contain them can be constrained by the fossil record. In the MasterCatalog, these orthologs are already identified; we need only to extract them to do the analysis. In principle, the TREX clock can be calibrated at points along the slice back in time captured in the genome by comparing f_2 values for orthologs that diverged, as far back in time as possible before the silent sites have equilibrated.

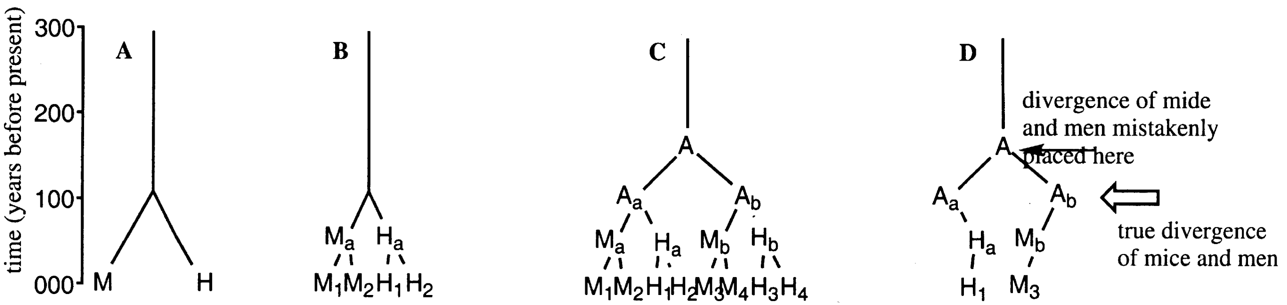
This simple approach is complicated by gene duplication prior to speciation, and possible gene loss (or incomplete genome sequencing). Together, these processes can generate paralogs whose true evolutionary relationship is not recognized by analysis of a tree alone (Fig. 14).

As genomes become completed, the paralog problem can be addressed by plotting a histogram that collects all interspecies pairs, and fitted a Gaussian curve based on the number of characters used to calculate the TREX tool. Knowing the number of characters used to determine f_2, we can calculate the shape of the distribution assuming fluctuation as its only cause. Where this has been done for mammalian genes, the shape of the distribution is fit well by a binomial distribution for the most recently diverging genes. This suggested that we need not invoke "hot spots" or other higher order behaviors to explain the plot.

From these histograms, and a divergence date of the two taxa, a time-invariant transition rate constant can be estimated. For example, the human and mouse taxa diverged perhaps 80 MYA. From the average f_2 value of the presumed orthologs, a rate constant for transitions at two fold redundant sites can be estimated to be ca. 3×10^{-9} transitions/site/year. This corresponds to a single lineage half-life ($\tau=\ln2/k$)$\approx$200 million years. Typical sequences generate sufficient characters to get reasonably accurate TREX dates back three or four half-lives, which correspond to divergence dates of 300–400 MYA (note that one must halve a single lineage time to get a date of divergence; half of the time is along one descendent branch, half is along the other).

The transition rate constant need not be time-invariant, of course. Because the second generation MasterCatalog reconstruct the sequences of nodes within the tree, however, we can count the number of substitutions that occurred along any individual branch of the tree in any kind of site. To the extent that the speciation can be dated from the fossil record, rate constants for any process (including transitions) can be calculated for any episode of interest.

Obviously, both the reconstructions, the trees, and the error due to statistical fluctuations will be improved as additional sequences become available. The first phase of a lineage specific model is complete when we can place on each branch of a genome tree the rate constants for pyrimidine–pyrimidine transitions and purine–purine transitions on individual branches in the target lineage extending back to the point where so many mutations have occurred that it is no longer possible to reconstruct events along branches with acceptable levels of uncertainty.


time (years before present)
300
200
100
000
A
M
H
B
M_a
H_a
$M_1M_2H_1H_2$
C
A
A_a
A_b
M_a
H_a
M_b
H_b
$M_1M_2H_1H_2M_3M_4H_3H_4$
D
A
A_a
A_b
H_a
M_b
H_1
M_3
divergence of mide and men mistakenly placed here
true divergence of mice and men

No duplication in family before speciation or after speciation. Pair of genes are orthologs in human & mouse; cannot be confused as anything else.

No duplication before speciation; in both lineages after. M_1 & M_2 are paralogs; H_1 & H_2 are paralogs. Tree from protein sequences detects true relation.

Duplication before & after speciation, but no gene loss. Tree based on protein sequences should detect true relation. Ancestor of human & mouse had two paralogous members of the A family, A_a and A_b.

Duplication before & after speciation, with gene loss. Ancestor of human & mouse had two paralogous members of the A family, A_a & A_b. H_1 & M_3 mistakenly identified as orthologs. Speciation mistakenly placed at A.

Fig. 14. Complications in evolutionary history of protein families creating difficulties in assigning ortholog–paralog relationships between proteins from different species. All genes "duplicate" when two species diverge.

Here, the value of reconstructed ancestral sequences is evident. An f_2 value can be calculated for any two extant sequences. A third sequence, which is an outgroup of the first two sequences, roots the tree holding the first two sequences. The DNA and protein sequences (probabilistic) at this root are then reconstructed. Now, an f_2 value (and the corresponding TREX distance) can be calculated between the ancestor and the third sequence.

This approach permits calculation of 12 rate constants for all 12 processes at silent sites (Fig. 15). This calculation can be supplemented with calculations at non-coding sites other than silent sites, including sites in introns and putative pseudogenes.

The model assumes that the transition rate constant is independent of location on the gene. This has proven to be the case (within statistical error) in the vertebrate genomes that we have examined so far. It may or may not be true in the history of the target species genome. Part of the assessment of the lineage-specific model for molecular evolution includes a genome-wide assessment of the variation of rate constants within the genome (Morozov et al., 2000). For each family, the behavior of silent sites in coding regions will be compared with the norm and the associated variance.

A similar analysis will permit us to understand the characteristics of processes that insert and delete gene segments during divergent evolution in the specific lineage that contains the target. We showed nearly a decade ago that a "penalty plus increment" formula does not accurately describe accepted indels in the general protein (Benner et al., 1993). These processes are sequence-dependent, and taxon-dependent. Again, the reconstructed ancestral sequences permit us to trace the history of indel events going back in time.

Lineage-specific Resource 4: A set of parameters for fundamental rates of events in the target genome.

(a) A composite rate constant k_Y for pyrimidine–pyrimidine transitions in the target lineage, assuming time invariance.
(b) A composite rate constant k_R for purine–purine transitions in the target lineage, assuming time invariance.
(c) A set of 12 rate constants describing all silent point mutation in the target genome, again assuming time invariance.

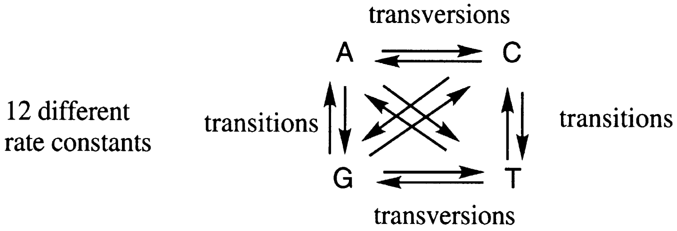


Fig. 15. Transitions and transversions in nucleotide replacement. No two arrows need have the same rate constant. See Gojobori et al. (1982). One goal of a whole genome analysis seeks to determine the history of all of these rate constants through the analysis of the planetary genome.

(d) Empirical parameters describing the rate insertion, deletion, gene duplication, and gene loss in the target lineage.
(e) The same as (a)–(d), but estimated for individual branches of the evolutionary tree, based on reconstructed evolutionary sequences. The degree of detail in this lineage-specific model depends on the availability of orthologs.
(f) Rate ratio tested parameters as described above.
(g) The set of trees, MSAs, and ancestral sequences for the families of proteins that are represented in the target genome, from which these parameters are calculated.

Together, these constitute a model for microscopic processes at the DNA level for the history of the target lineage. With whole genome analysis, therefore, we no longer must calibrate clocks on a single gene family, and hope that the variance is acceptable. We need not blindly assume time invariance. We can test whether time invariance holds and, if it does not, adjust the models to compensate for this.

The results of this analysis will be estimates for chronological dates for all nodes in all families containing target representatives, based on models that reflect rate constants (including variation in these) for transitions, supported by models for all other change processes at the level of the DNA molecule, with estimation for how these processes changed in specific lineages over specific periods of time.

At this time, this particular resource cannot be found in any existing public database. As the age of the genome progresses, databases capturing these features of the history of the basic processes of evolution, insertion, deletion, and mutation, will accumulate for hundreds of slices back in time. These will be a key to interpretive proteomics research for the next century.

Searching for Temporal Correlation in the Historical Record of a Lineage. Pathway and Network Hypotheses

Defining the interaction between proteins is very much part of defining the function of a protein. It is possible to give an enzyme a name based on the reaction that it catalyzes. This is the strategy behind the Enzyme Commission nomenclature for proteins, for example. But while most of mechanistic enzymology is based on this simple characterization of behavior of a protein, an understanding of how this reaction contributes to fitness (which is, from a Darwinian perspective, the only correct definition of "function") requires much more. This includes identifying other enzymes that provide the substrate(s) and consume the product(s) of the enzyme of interest, of course, as well as the enzymes that produce *their* substrates and consume *their* products.

Chemical theory is far from being able to calculate from first principles what substrate an enzyme binds, what reactions it catalyzes, and what product it produces. This is true even if a crystal structure of the enzyme, at an ultra-high resolution, is available. Computational biologists, therefore, cannot begin to identify metabolic pathways.

An evolutionary analysis based on the lineage-specific resources outlined above suggests a pass around the limitations of conventional chemical analysis. Specifically, the rank ordering of duplications in the history of a genome permits us to say which duplications occurred at approximately the same time.

When two duplications occurred at the same time, the time correlation of the duplications is consistent with the notion that the two duplications events are related functionally, and this in turn implies that at least one of the duplicates from one family interacts with at least one of the duplicates from the other.

These are, of course, "soft hypotheses". Time correlation of events is not expected in the history of two proteins that do not interact as they function. It may occur in these nevertheless, by random chance, of course. Therefore, the existence of time correlation of two events in the historical record of a genome is not sufficient reason to conclude that there is a functional relationship. This is, of course, a statement that applies equally throughout human inference, including in areas outside of genetics. Post hoc need not imply propter hoc, as every student of logic learns.

This does not mean, however, that temporal correlation is useful as a tool of inference, either in genetics or in general. A genome that contains n genes has on the order of n^2 possible pairwise interactions. It holds an astronomically larger number of higher order interactions. Any tool that identifies, even at the level of a soft hypothesis, a limited number of these, can focus the experimentalist on a set of these. This is extraordinarily useful in post-genomic science.

Lineage-specific Resources 2 and 3 are remarkably useful when analyzing the origin of new pathways that involve gene duplication. In principle, when two proteins suffer duplication near the same time, this observation immediately suggests the hypothesis that these two proteins interact.

Example: Identifying Pathways and Networks Within the Yeast Genome

The most compelling demonstration of second generation tools and databases over first generation tools comes by comparing the results of their implementation. Nowhere is this demonstration more dramatic than in the analysis of the yeast genome. The yeast genome has been the subject of numerous analyses ever since it was released (Lynch and Conery, 2000). These all used first generation tools. These were capable of identifying crude features of the molecular history of the yeast lineage, including past large-scale duplication events. But none of them captured the second level signals, all of which held biological insights.

The resources detailed above were created for the *S. cerevisiae* genome, which encodes ca. 6000 proteins. Fig. 16a shows a histogram that emerges when the gene duplications are clustered based on their f_2 value. Fig. 16b shows the histogram created by Lynch and Conery (2000) attempting the same cluster using first-generation tools, here, a "statistician approved" method for counting the number of mutations per silent site.

The first obvious feature in both histograms is the large number of duplications that occurred recently, represented by the bars on the right side of our histogram

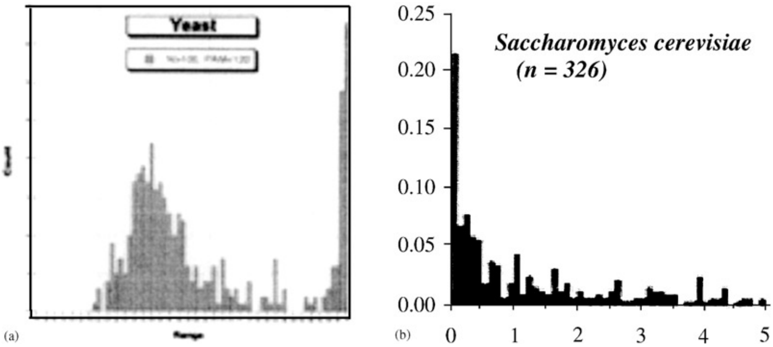


Fig. 16. (a) Histogram showing distribution of duplications in *S. cerevisiae* using the f_2/TREX metric and (b) the Li (1985) silent substitution metric (from Lynch and Conery 2000)). In (a), the most recent duplications are at the right. In (b) histogram, the most recent duplications are at the left. The f_2 metric (a) shows an episode of recent gene duplication, and an episode at $f_2 \approx 0.84$ (first major island of duplication to the left) that corresponds to events occurring ca. 80 MYA. The time correlation between approximately simultaneous events occurring in a single genome permits the assignment of pathways and networks from genome sequences (see Fig. 17).

(Fig. 16a) and the classical histogram (Fig. 16b). Differing in the two histograms is the evident presentation in ours of a large number of duplications occurring in the past, where silent substitutions have equilibrated (hump near $f_2 \approx 0.5$). The classical histogram does not transparently indicate regions where stochastic models provide uncertain answers, although this is clearly the case after 5 mutations have occurred in a single site.

Also present in the second generation histogram, but missing in the classical histogram, is a prominent episode of gene duplication at f_2 near 0.84. This corresponded to duplication events that occurred ~80 Ma, based on a calibration of the clock using fungal fossils (Berbee and Taylor, 1993). This generated a hypothesis, that protein families generating these duplications interact functionally.

Because so much is known about the yeast genome, it was possible to evaluate this hypothesis. These particular duplications created several new sugar transporters, two new glyceraldehyde-3-phosphate dehydrogenases, the non-oxidative pyruvate decarboxylase that generates acetaldehyde from pyruvate, a transporter for the thiamine vitamin that is used by this enzyme, and two alcohol dehydrogenases that interconvert acetaldehyde and alcohol.

This is not a random collection of proteins. Rather these proteins all belong to the pathway that yeast uses to ferment glucose to alcohol (Fig. 17). Correlating the times of duplication of genes in the yeast genome using the TREX method has identified a pathway.

Dating allows us to add geological and paleontological records to the analysis. By doing so, these pathways assume additional biological meaning. Fossils suggest that fermentable fruits also became prominent ~80 Ma, in the Cretaceous, during the age of the dinosaurs (Dilcher, 2000). Indeed, over-grazing by dinosaurs may explain why flowering plants flourished (Bakker, 1978; Barrett, 2001). Other genomes also

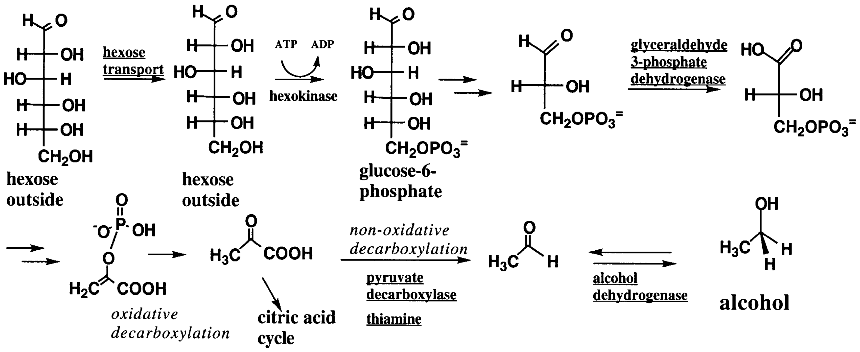


Fig. 17. The metabolic pathway identified by contemporaneous events in the history of yeast, as found using second generation tools for dating paralogization events in the genome. Genes in red (underlined) are duplicated in the historical event represented by the peak at $f_2 = 0.84$ in the histogram in Fig. 16.

record episodes of duplication near this time, including those of angiosperms (which create the fruit) and fruit flies (whose larvae eat the yeast growing in fermenting fruit) (Ashburner, 1998; Pereira, 1995).

Thus, time-correlation between the three records connected by approach-to-equilibrium dates generates a planetary hypothesis about function of individual proteins in yeast, one that goes beyond a statement about a behavior ("this protein oxidizes alcohol...") and a pathway ("...acting with pyruvate decarboxylase...") to a statement about planetary function ("...allowing yeast to exploit a resource, fruits, that became available ~80 Ma"). This level of sophistication in the annotation of a gene sequence is difficult to create in any other way.

We know of no other approach that can generate this level of functional insight, or capture pathways and regulatory networks as effectively. In particular, the approach-to-equilibrium dating tools can be more effective at inferring possible pathways from sequence data than approaches developed within other programs (Marcotte et al., 1999a, 1999b; Pazos and Valencia, 2001). Neither of these alternative tools captures dates, chemical, or paleontological information as effectively as the TRATE tool supported by a second generation naturally organized database such as the MasterCatalog.

Two yeast genome contains other illustration of the power of second generation strategies, especially when compared with conventional approaches. Consider the conventional histogram in Fig. 16b (from Lynch and Conery, 2000). Here, duplications in the yeast genome were dated using the conventional K_s metric (Li et al., 1985; Li, 1993). The conventional metric is adequate only to note that duplications do indeed occur, and that many are recent, and to suggest a rate for duplicate loss. Lynch and Conery (2000) interpreted this as random duplications that created redundancies that had not yet been removed by random loss. These conclusions remain controversial, in part because of criticism of the silent substitution metric to rank-order events in the genome (Long and Thornton, 2001; Zhang et al., 2001).

Second generation analyses suggested an alternative interpretation. All of the recent duplication events in the yeast genome fall into three metabolic categories: (a) genes that allow yeast to divide more rapidly, (b) genes that allow yeast to synthesize proteins more rapidly, and (c) genes that allow yeast to ferment malt (Benner et al., 2002). This is not a signature of random gene duplication, with the randomly created duplicates present in the yeast genome only because insufficient time has passed since they were created for them to be lost as functionless redundancies.

More plausible is the hypothesis that contact with humans has offered yeast a relatively rich environment to grow, far richer than the environment encountered by yeast in the wild (where few feasts are interspersed with long famines). The hypothesis is therefore more compelling that we are observing in the genome of yeast the record of its interaction with humans in the most recent episode of gene duplication, just as we are observing the record of yeast's acquaintance with angiosperms in the episode of gene duplication where $f_2 = 0.84$.

Example: Identifying Pathways and Networks within Mammalian Genomes

These results show how second generation dating tools, evolutionary models, and interpretive strategies address problems that are not addressed with first generation tools. Within yeast, so much is known that hypotheses are rapidly validated.

Within mammalian genomes, hypotheses drawn using a Firebird analysis can remain hypotheses longer, guiding biomedical researchers in the selection of Targets. Fig. 18 illustrates one example. Here, inspection of the STAT family within the MasterCatalog resource identifies a gene duplication in the mouse genome occurring since the divergence of mouse and rat. Because of the power of the MasterCatalog as a second generation naturally organized database, this inspection is possible by a browser with one click of a mouse button; the biologist need not to first make a commitment to investigate the STAT family, suffer through BLAST searches, and build his/her own evolutionary models before the first biological information returns as feedback.

This approach is useful for non-directed discovery. For example, once one notices the duplication in the STAT family, one can search the mouse genome for duplications occurring near the same time. Fig. 18 shows that this is in fact the case in the JAK family. As JAK and STAT interact in a regulatory networks generally, one generates the hypothesis that this particular JAK and this particular STAT are involved in the same regulatory pathway.

But which JAK is involved with which STAT? Again, the Firebird strategy generates a working hypothesis. Inspection of the pre-computed trees within the MasterCatalog for each of the branches leading from the ancestral JAKs and STATs, one notices that one JAK and one STAT in mouse lie at the ends of branches with particularly high K_a/K_s ratios. The working hypothesis is that this particular JAK and this particular STAT work together in a new pathway that emerged in the last 10 million years. This hypothesis is exactly the type of hypothesis that biological scientists would like to extract from a contemporary genome database.

Example: Paralogization within the Human Genome during the Oligocene

The human genome is still in draft form, and it is clear that all of the human proteome has not yet been identified. Yet the human paralog histogram, Resource 3 from above, offers functional hypotheses. With many more genes, no isolated episodes of duplication were observed in the human histogram. We were nevertheless able to select interesting episodes based on what we know about the paleontology and paleoecology of mammals (Eisenberg, 1981), and the transition rate constant calculated from the inter-species comparisons discussed above. In particular, we knew that ca. 40–35 MYA, the Earth suffered an episode of global cooling (the average temperature dropped perhaps 15°C) (Wolfe, 1978; Prothero, 1994).

We knew that this dramatic cooling had repercussions throughout the biosphere. Grasses emerged for the first time, adapted to survive in the newly formed savannahs

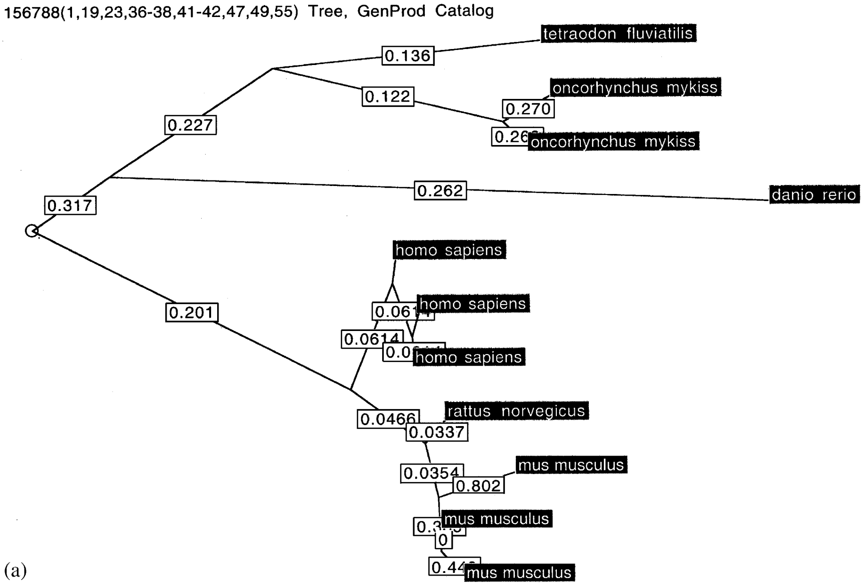


Fig. 18. (a) The tree showing divergence of two mouse paralogs within the STAT family. The numbers on the branches are K_a/K_s ratios. Note the duplication at the bottom of the tree of mouse paralogs, where one of the paralogs has a K_a/K_s ratio of 0.802. This is insufficient to compel the conclusion that adaptive evolution has occurred along the branch, but is suggestive, as this ratio is higher than ratios elsewhere in the tree. This screenshot from the MasterCatalog shows color, not captured here. (b) The same tree, but showing GenBank gi numbers (instead of species names) as labels on the leaves of the trees, and PAM distances as the numbers on the trees. (c) A portion of the JAK tree showing paralogization (duplication) in the mouse lineage after the divergence in rat, where the JAK duplication is associated with the same TREX date as the STAT duplication. The numbers on the branches are K_a/K_s ratios. The hypothesis emerging from this Firebird analysis correlates specific JAK kinases with specific STATs, suggesting a regulatory network that is now open to experimental test.

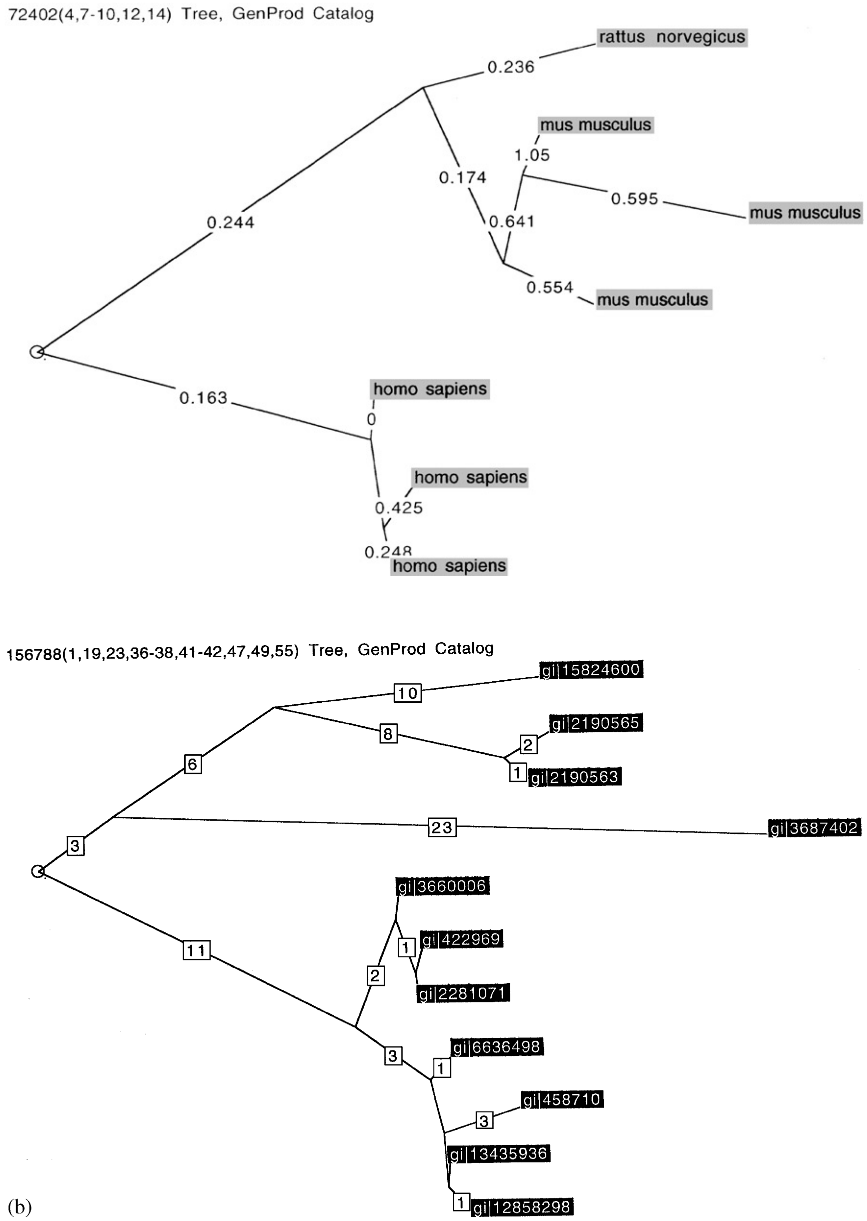


Fig. 18 (*continued*).

that replaced tropical rain forests throughout much of the temperate zone. Artiodactyls responded with a spate of gene duplication and rapid evolution at this time leading to the emergence of ruminant digestion (Rose, 1982; Jermann, 1995). Indeed, a set of experiments in paleobiochemistry, illustrated in Fig. 19, traced

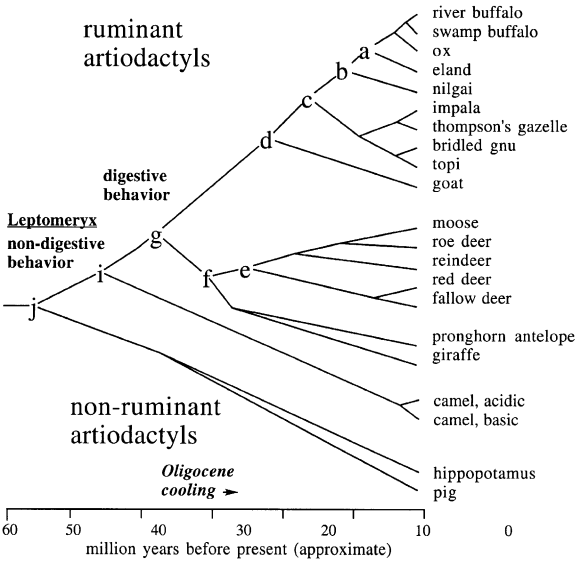


Fig. 19. An evolutionary tree relating ribonucleases responsible for the digestion of nucleic acid from bacteria fermenting grass in the first stomach of ruminants. Experimental analysis of reconstructed ancestral proteins suggests that digestive-like behavior in the protein arose near the time that its putative role in digestion in ruminants arose, near the time when grasses arose in response to the global climatic upheaval known as the Oligocene cooling.

molecular change in individual proteins in the ruminant digestive system all of the way to the planetary environment.

Explicit reconstructions of evolutionary intermediates assign specific amino acid replacements to specific episodes in the history of a protein family. In ancestral lysozyme genes, for example, rapid sequence evolution occurred as ruminant and ruminant-like digestion emerged (Messier and Stewart, 1997). Rapid change in the sequence of a lysozyme implies rapid change in the behavior of lysozyme, which in turn suggests a change in its functional behavior. This hypothesis is inferential, of course, but can be tested. Further, it makes sense in light of a historical model. New lysozymes are expected to emerge to break open bacterial cells in the new ruminant digestion.

The microbial communities within the rumen must have responded as well. It will be interesting to learn whether this response has been captured in their genomes. We expect that it should, and it should be well within the dating range, even if transitions are much faster in microbes than in their hosts.

How did primates respond to this global cooling? To answer this question, we examined duplication events that generated paralogs separated by an $f_2 \approx 0.91$. A total of 22 gene families that suffered duplication during the Oligocene cooling. Over half of these are not annotated. Remarkably, all of the genes that *are* annotated might be interpreted as being involved in neurological development (Table 11), either directly or conceivably.

Table 11
Some genes duplicated in the human genome at the time of the great Oligocene cooling

gp.24532	protocadherin 68, neuronal network patterning (Hilschmann et al., 2001)
gp.24558	protocadherin 43, neuronal network patterning (Hilschmann et al., 2001)
gp.13983	serine kinase PAK homolog, mental retardation (Blanco et al., 2000, Allen et al, 1998)
gp.16242	MNB protein kinase; Down syndrome (Kentrup et al, 2000)
gp.28010	desmolase, conceivably involved in neurosteroid biosynthesis (?)
gp.21865	butyrophilin, a possible neurosteroid receptor (?)

Especially interesting are the recently discovered protocadherins, proteins involved in the patterning of neural networks. These suffered frequent duplication in the evolution of human biology in the Oligocene. Also generating paralogs are serine kinases associated with X-linked mental retardation, and a kinase associated with Down syndrome.

Ruminants responded to the global crisis associated with the Oligocene cooling by learning to eat grass. Did primates respond to the Oligocene cooling by becoming more intelligent by altering their central nervous systems? This is clearly a hypothesis, but clearly one with significance that ranges from biomedicine to the planet. The example serves to illustrate how the FIREBIRD analysis supported by the MASTERCATALOG helps generate hypotheses that are inaccessible to any current publicly available tool, and far beyond the potential of methods constrained by statistical formalisms.

Lineage-specific Resource 5: A matrix of interconnections between protein families containing representatives of the *Arabidopsis* proteome that hypothetically interact when they function, based on the evolutionary history of the family (contemporaneous duplications, episodes of putative functional conservation, episodes of putative functional change). These networks, based on time-correlation, will indicate hypothetical pathways in which members of individual families function together.

Interspecies Genome Comparisons

Just as it is artificial to examine the evolutionary history of a single protein family, so is it artificial to examine the genome of a single species. Species interact with other species as they struggle to survive and reproduce. While the analysis is beyond the scope of this review, it is clear that the events recorded in the yeast genome as it evolves to ferment fermentable fruit are correlated with events that are recorded in the genomes of the fruit-bearing plants. At the same time, fruit flies evolved to fill this new niche in the planetary biosphere have evolved, and this response will be captured within the fly genome.

Summary

The paleontological, geological, and molecular records of life on Earth can be joined to obtain a comprehensive model for the proteins that are found throughout the life on the planet (Benner et al., 2002). Individual models for ca. 100,000 nuclear families of protein sequences have been collected in the MasterCatalog, a commercial naturally organized database developed in collaboration with EraGen Biosciences (Madison, WI). The MasterCatalog facilitates interpretive proteomics by providing the user with high quality, pre-computed, second generation models for the evolutionary history of all of the protein families in the known biosphere. These models are advanced over those offered by first generation evolutionary databases (such as those presented by Pfam, TIGRfam, and Hovergen) in several ways, including its focus on nuclear families, extensive identification of modules that undergo independent evolution, and the use throughout of explicitly reconstructed sequences from ancestral proteins that were present in now-extinct species. This second generation database is combined with a set of powerful interpretive proteomics tools that make up the Firebird (Functional Inference from Reconstructed Evolutionary Biology) strategy for moving from genomes to biology. The Firebird suite of tools offers a powerful framework for analyzing function in proteins, identifying targets of biomedical interest, and guiding pre-clinical drug development in animal models, inter alia. When applied to whole genomes, the suite identifies metabolic pathways and regulatory networks, permits the correlation of the life history of a lineage with its historical past, and captures interconnections that will move the biomedical researcher and biological chemist from the genome to organismic biology, ecosystems, and the planet.

Acknowledgements

The research described here was supported, both intellectually and financially, through the MasterCatalog, an interpretive proteomics platform offered by EraGen Biosciences. One of the authors (SAB) has a personal financial interest in EraGen Biosciences and the MasterCatalog product.

References

Aimi J, Badylak J, Williams J, Chen ZD, Zalkin H, Dixon JE. Cloning of a cDNA encoding adenylosuccinate lyase by functional complementation in *Escherichia coli*. J Biol Chem 1990;265: 9011–4.

Allen KM, Gleeson JG, Bagrodia S, Partington MW, MacMillan JC, Cerione RA, Mulley JC, Walsh CA. PAK3 mutation in nonsyndromic X-linked mental retardation. Nat Genet 1998;20:25–30.

Altschuh D, Lesk AM, Bloomer AC, Klug A. Correlation of coordinated amino acid substitutions with function in tobacco mosaic-viruses. Protein Eng 1987;1:228–36.

Antunes MT, Cahuzac B. Crocodilian faunal renewal in the Upper Oligocene of Western Europe. C R Acad Sci Ser II Fascicule A-SciTerre Planetes 1999;328:67–72.

Ashburner M. Speculations on the subject of alcohol dehydrogenase and its properties in *Drosophila* and other flies. Bioessays 1998;20:949–54.

Ayala FJ. Molecular clock mirages. Bioessays 1999;21:71–5.

Azanza B. Systematics and evolution of the genus *Procervulus* (Cervidae Artiodactyla Mammalia) of the lower Miocene of Europe. C R Acad Sci Ser II 1993;316:717–23.

Bakker RT. Dinosaur feeding-behavior and origin of flowering plants. Nature 1978;274:661–3.

Baldauf SL, Palmer JD, Doolittle WF. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. Proc Natl Acad Sci USA 1996;93:7749–54.

Barrett PM, Willis KJ. Did dinosaurs invent flowers? Dinosaur-angiosperm coevolution revisited. Biol Rev 2001;76:411–47.

Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer ELL. The Pfam protein families database. Nucleic Acids Res 2000;28:263–6.

Benner SA. Computational and interpretive genomics. In: Pardalos PM, Principe J, editors. Biocomputing. Amsterdam: Kluwer, 2002. p. 25–43.

Benner SA. Patterns of divergence in homologous proteins as indicators of tertiary and quaternary structure. Adv Enzyme Regul 1989;28:219–36.

Benner SA, Caraco MD, Thomson JM, Gaucher EA. Planetary biology. Paleontological geological and molecular histories of life. Science 2002;293:864–8.

Benner SA, Cohen MA, Gonnet GH. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. J Mol Biol 1993;229:1065–82.

Benner SA, Ellington AD. Interpreting the behavior of enzymes. Purpose or pedigree? CRC Crit Rev Biochem 1988;23:369–426.

Benner SA, Gaucher EA. Evolution language and analogy in functional genomics. Trends Genet 2001;17:414–8.

Benner SA, Gerloff DL. Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure. The catalytic domain of protein kinases. Adv Enzyme Regul 1991;31:121–81.

Benner SA, Gerloff DL, Chelvanayagam G. The phospho-β-galactosidase and synaptotagmin predictions. Proteins Struct Funct Genet 1995;23:446–53.

Benner SA, Trabesinger-Ruef N, Schreiber DR. Post-genomic science. Converting primary structure into physiological function. Adv Enzyme Regul 1998;38:155–80.

Benner SA, Turcotte M, Cannarozzi G, Gerloff DL, Chelvanayagan G. Bona fide predictions of protein secondary structure using transparent analyses of multiple sequence alignments. Chem Rev 1997;97:2725–843.

Berbee ML, Taylor JW. Dating the evolutionary radiations of true fungi. Can J Bot 1993;71: 1114–27.

Blanco P, Sargent CA, Boucher CA, Mitchell M, Affara NA. Conservation of PCDHX in mammals; expression of human X/Y genes predominantly in brain. Mamm Genome 2000;11: 906–14.

Boerboom D, Kerban A, Sirois J. Molecular characterization of the equine cytochrome P450 aromatase cDNA and its regulation in preovulatory follicles. Biol Reprod 1997;56(suppl 1):479.

Bordo D, Argos P. Evolution in protein cores. Constraints in point mutations as observed in globin tertiary structures. J Mol Biol 1990;211:975–88.

Bowring SA, Grotzinger JP, Isachsen CE, Knoll AH, Pelechaty SM, Kolosov P. Calibrating rates of early Cambrian evolution. Science 1993;261:1293–8.

Brown HC. The nonclassical ion problem. New York: Plenum Press, 1977.

Callard GV, Tchoudakova A. Evolutionary and functional significance of two CYP19 genes differentially expressed in brain and ovary of goldfish. J Steroid Biochem Mol Biol 1997;61:387–92.

Caraco MD. Nearly neutral evolutionary distance: a new dating tool and its applications. Dissertation, ETH Zurich, 2002.

Carroll RL. Vertebrate paleontology and evolution. New York: Freeman, 1988.

Chandrasekharan UM, Sanker S, Glynias MJ, Karnik SS, Husain A. Angiotensin II forming activity in a reconstructed ancestral chymase. Science 1996;271:502–5.

Chelvanayagam G, Eggenschwiler A, Knecht L, Gonnet GH, Benner SA. An analysis of simultaneous variation in protein structures. Protein Eng 1997;10:307–16.

Chelvanayagam G, Knecht L, Jenny TF, Benner SA, Gonnet GH. A combinatorial distance constraint approach to predicting protein tertiary models from known secondary structure. Fold Design 1998;3:149–60.

Chircurel M. Whatever happened to leptin? Nature 2000;404:538–40.

Choi I, Collante WR, Simmen RCM, Simmen FA. A developmental switch in expression from blastocyst to endometrial/placental–type cytochrome P450 aromatase genes in the pig and horse. Biol Reprod 1997a;56:688–96.

Choi IH, Troyer DL, Cornwell DL, Kirby-Dobbels KR, Collante WR, Simmen FA. Closely related genes encode developmental and tissue isoforms of porcine cytochrome P450 aromatase. DNA Cell Biol 1997b;16:769–77.

Chothia C. One thousand families for the molecular biologist. Nature 1992;357:543–4.

Chothia C, Lesk AM. Evolution of proteins formed by β-sheets. Plastocyanin and azurin. J Mol Biol 1982;160:309–23.

Cohen MA, Benner SA, Gonnet GH. Analysis of mutation during divergent evolution. The 400 by 400 dipeptide mutation matrix. Biochem Biophys Res Commun 1994;199:489–96.

Cooke HBS, Wilkinson AF. Suidae and tayassuidae. In: Maglio VJ, Cooke HBS, editors. Evolution of African mammals. Cambridge: Harvard University Press, 1978. p. 438–82.

Corpet F, Servant F, Gouzy J, Kahn D. ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. Nucleic Acids Res 2000;28:267–9.

Dayhoff MO, Schwartz RM, Orcott BC. In: Dayhoff MO, editor. 5 (suppl 3). Atlas of protein sequence and structure. Washington DC: Nat Biomed Res Found, 1978.

Delarue B, Breard E, Mittre H, Leymarie P. Expression of two aromatase cDNAs in various rabbit tissues. J Steroid Biochem Mol Biol 1998;64:113–9.

Delarue B, Mittre H, Feral C, Benhaim A, Leymarie P. Rapid sequencing of rabbit aromatase cDNA using RACE PCR. C R Acad Sci Ser III Sci De La Vie-Life Sci 1996;319:663–70.

Dorit RL, Schoenbach L, Gilbert W. How big is the universe of exons? Science 1990;250:1377–82.

Duret L, Mouchiroud Dand Gouy M. HOVERGEN—A database of homologous vertebrate genes. Nucleic Acids Res 1994;22:2360–5.

Eisenberg D, Marcotte EM, Xenarios I, Yeates TO. Protein function in the post-genomic era. Nature 2000;405:823–6.

Eisenberg JF. The mammalian radiations. An analysis of trends in evolution adaptation and behavior. Chicago: University Chicago Press, 1981, p. 196.

Endo T, Ikeo K, Gojobori T. Large-scale search for genes on which positive selection may operate. Mol Biol Evol 1996;13:685–90.

Felsenstein J. Taking variation of evolutionary rates between sites into account in inferring phylogenies. J Mol Evol 2001;53:447–55.

Fitch WM. An evaluation molecular evolutionary clocks. In: Ayala FJ, editor. Molecular study of biological evolution. Sunderland MA: Sinauer, 1976. p. 160–278.

Fitch WM, Markowitz E. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem Genet 1970;4:5 79–93.

Fortelius M, van der Made J, Bernor RL. Middle and late Miocene suoidea of central Europe and the eastern Mediterranea evolution biogeography and paleoecology. In: Fanas RL, Bernor V, Fahlbusch H-W, Mittmann, editors. The evolution of Western Eurasian neogene mamma. New York: Columbia University, 1996. p. 348–77.

Fukami-Kobayashi K, Benner SA. Joining structural biology with genomics using reconstructed ancestral sequences. The case for compensatory covariation. J Mol Biol 2002;319:729–43.

Fürbaβ R, Vanselow J. An aromatase pseudogene is transcribed in the bovine placenta. Gene 1995;154:287–91.

Gaucher EA, Das UK, Miyamoto MM, Benner SA. The crystal structure of eEF1a supports the functional predictions of an evolutionary analysis of rate changes among elongation factors. Mol Biol Evol 2001b;19:569–73.

Gaucher EA, Miyamoto MM, Benner SA. Function–structure analysis of proteins using covarion-based evolutionary approaches. Elongation factors. Proc Natl Acad Sci USA 2001a;98: 548–52.

Gerloff DL, Benner SA. A consensus prediction of the secondary structure for the 6-phospho-β-D-galactosidase superfamily. Proteins Struct Funct Genet 1995;21:273–81.

Gerloff DL, Cohen FE, Korostensky C, Turcotte M, Gonnet GH, Benner SA. A predicted consensus structure for the N-terminal fragment of the heat shock protein HSP90 family. Proteins: Struct Funct Genet 1997;27:450–8.

Gerloff DL, Jenny TF, Knecht LJ, Benner SA. A secondary structure prediction of the hemorrhagic metalloprotease family. Biochem Biophys Res Commun 1993;194:560–5.

Gerlt JA, Babbitt PC. Mechanistically diverse enzyme superfamilies: The importance of chemistry in the evolution of catalysis. Curr Opin Chem Biol 1998;2:607–12.

Göbel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins: Struct Funct Gen 1994;18:309–17.

Gojobori T, Li W-H, Graur D. Patterns of nucleotide substitution in pseudogenes and functional genes. J Mol Evol 1982;18:360–9.

Gonnet GH, Benner SA. Computational biochemistry research at ETH. Technical Report 154 Departement Informatik, March, 1991.

Gonnet GH, Cohen MA, Benner SA. Exhaustive matching of the entire protein sequence database. Science 1992;256:1443–5.

Gracy J, Argos P. DOMO: a new database of aligned protein domains. Trends Biochem Sci 1998;23: 495–7.

Gu X. Statistical methods for testing functional divergence after gene duplication. Mol Biol Evol 1999;16:1664–74.

Harada N. Cloning of a complete cDNA encoding human aromatase immunochemical identification and sequence analysis. Biochem Biophys Res Commun 1988;156:725–32.

Hey J. The neutralist, the fly and the selectionist. Trends Ecol Evol 1999;14:35–8.

Hickey GJ, Krasnow JS, Beattie WG, Richards JS. Aromatase cytochrome P450 in rat ovarian granulosa cells before and after luteinization. Adenosine 3′,5′-monophosphate-dependent and independent regulation. Cloning and sequencing of rat aromatase cDNA and 5′ genomic DNA. Mol Endocrinol 1990;4:3–12.

Hillis DM, Huelsenbeck JP, Cunningham CW. Application and accuracy of molecular phylogenies. Science 1994;264:671–7.

Hilschmann N, Barnikol HU, Barnikol-Watanabe S, Gotz H, Kratzin H, Thinnes FP. The immunoglobulin-like genetic predetermination of the brain: the proto-cadherins blueprint of the neuronal network. Naturwissenschaften 2001;88:2–12.

Huelsenbeck J, Rannala B. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science 1997;276:227–32.

Jermann TM, Opitz JG, Stackhouse J, Benner SA. Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. Nature 1995;374:57–9.

Ji Q, Luo ZX, Yuan CX, Wible JR, Zhang JP, Georgi JA. The earliest known eutherian mammal. Nature 2002;416:816–22.

Kentrup H, Joost HG, Heimann G, Becker W. Minibrain/DYRK1A-gene: candidate gene for mental retardation in Down syndrome? Klin Padiatr 2000;212:60–3.

Kimura M. The neutral theory of molecular evolution. New York: Cambridge University Press, 1983.

Knighton DR, Zheng J, Ten Eyck L, Ashford FVA, Xuong NH, Taylor SS, Sowadski JM. Crystal structure of the catalytic subunit of cyclic adenosine-monophosphate dependent protein-kinase. Science 1991;253:407–14.

Lesk AM, Chothia C. How different amino acid sequences determine similar protein structures. The structure and evolutionary dynamics of the globins. J Mol Biol 1980;136:225–70.

Lesk AM, Chothia C. Evolution of proteins formed by b-sheets II. The core of the immunoglobulin domains. J Mol Biol 1982;160:325–42.

Li WH. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 1993;36:96–9.

Li WH, Wu CI, Luo CC. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 1985;2:150–74.

Liberles DA, Schreiber DR, Govindarajan S Chamberlin SG, Benner SA. The adaptive evolution database (TAED). Genome Biol 2001;2:00031–000318.

Lo Conte L, Ailey B, Hubbard TJP, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res 2000;28:257–9.

Long MY, Thornton K. Gene duplication and evolution. Science 2001;293:U1.

Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science 2000;290: 1151–5.

Maddison WP, Maddison DR. Macclade analysis of phylogeny and character evolution. Sunderland, MA: Sinauer Associates, 1992.

Malcolm BA, Wilson KP, Matthews BW, Kirsch JF, Wilson AC. Ancestral lysozymes reconstructed, neutrality tested and thermostability linked to hydrocarbon packing. Nature 1990;345:86–9.

Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. A combined algorithm for genome-wide prediction of protein function. Nature 1999a;402:83–6.

Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. Detecting protein function and protein–protein interactions from genome sequences. Science 1999b;285:751–3.

Messier W, Stewart CB. Episodic adaptive evolution of primate lysozymes. Nature 1997;385:151–4.

Morozov P, Sitnikova T, Churchill G, Ayala FJ, Rzhetsky A. A new method for characterizing replacement rate variation in molecular sequences: application of the fourier and wavelet models to *Drosophila* and mammalian proteins. Genetics 2000;154:381–95.

Murelaga X, de Broin FD, Suberbiola XP, Astibia H. Two new chelonian species from the Lower Miocene of the Ebro Basin (Bardenas Reales of Navarre). C R Acad Sci Ser II Fascicule A-Sci Terre Planetes 1999;328:423–9.

Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequences of two proteins. J Mol Biol 1970;48:443–53.

Neher E. How frequent are correlated changes in families of protein sequences? Proc Natl Acad Sci USA 1994;91:98–102.

Nei M. Molecular evolutionary genetics. New York: Columbia University Press, 1987.

Oosawa K, Simon M. Analysis of mutations in the transmembrane region of the aspartate chemoreceptor in *Escherichia coli*. Proc Natl Acad Sci USA 1986;83:6930–4.

Pamilo P, Bianchi NO. Evolution of the zfx and zfy genes. Rates and interdependence between the genes. Mol Biol Evol 1993;1:271–81.

Pauling L, Zuckerkandl E. Molecular paleontology. Acta Chem Scand 1962;17:S9–16.

Pazos F, Valencia A. Similarity of phylogenetic trees as indicator of protein–protein interaction. Protein Eng 2001;14:609–14.

Pereira HS, Macdonald DE, Hilliker AJ, Sokolowski MB. Chaser (CSR) a new gene affecting larval foraging behavior in *Drosophila melanogaster*. Genetics 1995;141:263–70.

Pilgrim GE. The dispersal of the Artiodactyla. Biol Rev 1941;16:134–63.

Prothero DR. The eocene–oligocene transition paradise lost. NY: Columbia University Press, 1994.

Qi T, Beard KC. Late Eocene sivaladapid primate from Guangxi Zhuang Autonomous Region People's Republic of China. J Hum Evol 1998;35:211–20.

Riley M, Labedan B. Protein evolution viewed through *Escherichia coli* protein sequences: introducing the notion of a structural segment of homology the module. J Mol Biol 1997;268:857–68.

Rocek Z. The salamander *Brachycormus noachicus* from the Oligocene of Europe and the role of neoteny in the evolution of salamanders. Palaeontology 1996;39:477–95.

Rose KD. Skeleton of diacodexis oldest known artiodactyl. Science 1982;236:621–3.

Stamenkovic I, Clark EA, Seed B. A B-lymphocyte activation molecule related to the nerve growth-factor receptor and induced by cytokines in carcinomas. EMBO J 1989;8:1403–10.

Shindyalov IN, Kolchanov NA, Sander C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng 1994;7:349–58.

Simpson ER, Michael MD, Agarwal VR, Hinshelwood MM, Bulun SE, Zhao Y. Expression of the CYP19 (aromatase) gene. An unusual case of alternative promoter usage. FASEB J 1997; 11:29–36.

Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–7.

Stackhouse J, Presnell SR, McGeehan GM, Nambiar KP, Benner SA. The ribonuclease from an extinct bovid. FEBS Lett 1990;262:104–6.

Strickberger MW. Molecular evolution. Sudbury MA: Jones and Bartlett, 2000 (p. 644).

Stucky RK. Evolution of land mammal diversity in North America during the Cenozoic. Curr Mammal 1990;2:375–432.

Suzuki Y, Gojobori T, Nei M. ADAPTSITE: detecting natural selection at single amino acid sites. Bioinformatics 2001;17:660–1.

Swofford DL. PAUP*Phylogenetic analysis using parsimony. Version 4. Sunderland MA: Sinauer Associates, 1998.

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. In: Hillis DM, Moritz C, Mable BK, editors. Molecular systematics. 2nd ed. Sunderland MA: Sinauer Associates, 1996. p. 407–514.

Takahashi K, Nei M. Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. Mol. Biol. Evol. 2000;17:1251–8.

Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science 1997;278:631–7.

Tauer A, Benner SA. The B12-dependent ribonucleotide reductase from the archaebacterium *Thermoplasma acidophila*. An evolutionary conundrum. Proc Natl Acad Sci USA 1997;94:53–8.

Taylor WR, Hatrick K. Compensating changes in protein multiple sequence alignments. Protein Eng 1994;7:341–8.

Tchernov E. The Afro-Arabian component in the levantine mammalian fauna. A short biogeographical review. Israel J Zool 1992;38:155–92.

Terashima M, Toda K, Kawamoto T, Kuribayashi I, Ogawa Y, Maeda T, Shizuta Y. Isolation of a full-length cDNA encoding mouse aromatase P450. Arch Biochem Biophys 1991;285:231–7.

Thompson JD, Higgins DG, Gibson TJ, Clustal -W. Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–80.

Thorne JL, Kishino H, Felsenstein J. Inching toward reality. An improved likelihood model of sequence evolution. J Mol Evol 1992;34:3–16.

Tiffin P, Hahn MW. Coding sequence divergence between two closely related plant species: *Arabidopsis thaliana* and *Brassica rapa* ssp *Pekinensis*. J Mol Evol 2002;54:746–53.

Trabesinger-Ruef N, Jermann TM, Zankel TR, Durrant B, Frank G, Benner SA. Pseudogenes in ribonuclease evolution. A source of new biomacromolecular function? FEBS Lett 1996;382: 319–22.

van der Made J, Tuna V. A tetraconodontine pig from the Upper Miocene of Turkey. Trans Soc Edinburgh Earth Sci 1999;89:227–30.

Wolfe JA. A paleobotanical interpretation of Tertiary climates in the Northern Hemisphere. Am Sci 1978;66:694–703.

Yang ZH, Bielawski JP. Statistical methods for detecting molecular adaptation. Trends Ecol Evol 2000;15:496–503.

Yang ZH, Nielsen R, Goldman N, Pedersen AMK. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 2000;155:431–49.

Zdobnov EM, von Mering C, Letunic I, Torrents D, Suyama M, Copley RR, Christophides GK, Thomasova D, Holt RA, Subramanian GM, Mueller HM, Dimopoulos G, Law JH, Wells MA, Birney E, Charlab R, Halpern AL, Kokoza E, Kraft CL, Zhongwu L, Lewis S, Louis C, Barillas-Mury C, Nusskern D, Rubin GM, Salzberg SL, Sutton GG, Topalis P, Wides R, Wincker P, Yandell M, Collins FH, Ribeiro J, Gelbart WM, Kafatos FC, Bork P. Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. Science 2002;298:149–59.

Zhang LQ, Gaut BS, Vision TJ. Gene duplication and evolution. Science 2001;293:U1–2.

Zhang YY, Proenca R, Maffei M, Barone M, Leopold L, Friedman JM. Positional cloning of the mouse obese gene and its human homolog. Nature 1994;372:425–32.