The exhaustive genome comparison effort. A quarter-century later

Steven A. Benner, Kevin M. Bradley, Stephan G. Chamberlain

A quarter century after the Benner and Gonnet groups began their collaboration in evolutionary bioinformatics, evolution-based functional genomics is a field with considerable scope. Even with the remarkable advances in computing power over this period, the explosion of data derived from genomic and protein sources have required more and more sophisticated approaches be developed and utilized. We describe here new software combined with data organization techniques and illustrate how we are harnessing these to place physiological function of protein sequence data using natural history.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

The exhaustive genome comparison effort. A quarter-century later

Steven A. Benner,* Kevin M. Bradley, Stephen G. Chamberlain
Foundation for Applied Molecular Evolution
Gainesville Florida
*Correspondence author: Steven A. Benner
720 SW 2nd Avenue, Suite 201
Gainesville, FL 32601
sbenner@ffame.org
352-219-3570

Abstract

A quarter century after the Benner and Gonnet groups began their collaboration in evolutionary bioinformatics, evolution-based functional genomics is a field with considerable scope. Even with the remarkable advances in computing power over this period, the explosion of data derived from genomic and protein sources have required more and more sophisticated approaches be developed and utilized. We describe here new software combined with data organization techniques and illustrate how we are harnessing these to place physiological function of protein sequence data using natural history.

Introduction

It has now exactly a quarter-century since the Benner and Gonnet groups began their collaboration in evolutionary bioinformatics [Gonnet and Benner 1991], a collaboration made possible when my wife (Beverly Sanders) directed me to attend a seminar that Prof. Gonnet (then from Waterloo, Canada) was giving on his work with the Oxford Unabridged English Dictionary. That collaboration had at first only a modest goal: to update the (by then) more than 20 year old amino acid substitution matrix that had been introduced in the 1960s by Margaret Dayhoff and her colleagues at the National Bureau of Standards [Dayhoff et al. 1972]. However, even though the age of genomic sequence had not been begun, it was clear that it would soon get underway, and that it would deliver a large number of whole genome sequences that could service the platform for this new field. We wanted to be prepared for this revolution in biology.

Fortunately, the tools that the Gonnet group developed to organize the Oxford Unabridged Dictionary were applicable to manage protein sequence databases. When applied to SwissProt

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

[Bairoch and Apwiler 2000] and other early databases, this yielded the first exhaustive matching of a protein genome sequence database and was published in *Science* in 1992 [Gonnet et al. 1992]. It provided not only new Dayhoff matrices, but also a clear understanding of how patterns of amino acid substitution [Gonnet et al. 1994] and gapping [Benner et al. 1993] differed as two protein sequences diverged. In the three dozen papers that were to follow, interpretation of those patterns formed the basis for the then-emerging field of "evolutionary-based functional genomics", including the resurrection of ancestral genes and proteins [Jermann et al. 1995], the use of evolutionary analyses to predict the folded structures of proteins [Benner et al. 1997], and the analysis of the natural history of two families to understand adept, drift, functional change, and pathway interactions [Benner et al. 1998] [Benner 2003]. Further, the exhaustive matching supported of some of the earliest efforts to infer the metabolism of very ancient organisms [Benner et al. 1989], including organisms standing at the branch points of the major three kingdoms, and organisms that invented protein translation [Benner et al. 1993].

This work was well underway, of course, before complete genome sequences were available for any individual organism. As these emerged for microorganisms, the Benner group, in collaboration with EraGen Biosciences, introduced a naturally organized genome database [Benner et al. 2000]. Called the MasterCatalog, the database organized protein sequences by evolutionary families, much as had been done by more primitive databases dating back to Dayhoff herself, but also in earlier versions of computerized database such as Hovergen [Duret et al. 1994]. However, the MasterCatalog also included pre-computed trees, multiple sequence alignments, and probabilistic ancestral sequences at nodes of the trees. A commercial version of the MasterCatalog was bundled with several dozen complete genome sequences from various microorganisms that had been assembled in a commercial effort at the company Genome Therapeutics. This product generated approximately $3.4 million in sales during its lifetime.

In a later version, secondary structure assignments determined by protein crystallography were added to these evolutionary models for individual protein families to create the Magnum database [Bradley and Benner 2006]. These supported a range of tools to extract functional information from evolutionary comparisons between different species.

The Magnum database was announced just as whole genome sequences of vertebrates are becoming available. This opened an entirely new direction for the assembly of an evolution-base, naturally organized database, if the families of orthologous proteins from advanced organisms

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

could be reliably inferred with few errors. Complete genomic sequences were proposed to offer, as one of their outputs, the prospect of knowing what genes and proteins are *not* present in a biological organism. This prospect has driven, now for 20 years, the technology to determine every last nucleotide in a chromosome, close all of the chromosomes in a genome, and provide a complete list of genetic components in a complex organism.

The advance of deep sequencing, of course, created a crisis in data management. These crises were associated with a series of problems briefly outlined below.

Data volume

Even in 1998, when MasterCatalog was conceived, sequence data resources were large and their size was growing almost exponentially. Early development with GenBank involved a dataset with some redundancy and about 700,000 sequences, which grew to well over 2.5 million sequences in the space of 3 years.

Today, the number of bacterial whole genomes in RefSeq (a service provided by the NCBI) is in the thousands, with is combined with the dozens of vertebrate organisms that have been sequenced. It would be easy to collect 20 million sequences from whole genome sources alone. This is such a rapid growth in computational demands that even technological increases in computer power have been unable to keep up. In particular, a naïve all-against-all comparison scales with the square of the database size. While indexing and other algorithmic tools can be used to cause the scaling factor to be smaller, even Moore's law would be unable to manage the size of the database. Therefore, any practical computational approach to organizing this data by clustering needs to balance the time and space demands.

Fortunately, this problem could be mitigated simply by exploiting the realities of natural history. The entire protein sequence space has not been explored during that natural history, not the least of which and certainly not by vertebrates. Accordingly, focusing on chordate, once genomes representing each of the major branches in the chordate tree are available, it is no longer necessary to do and “all-against-all” comparison of an entire database.

MasterCatalog was designed to manage computational challenge of the rapidly growing volume of data as efficiently as possible. Minimizing the number of sequence pairs for which redundancy must be computed is the first step; performed using a BLAST comparison tuned for nearly identical sequences, with pairwise comparisons only for sequences of the same species.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

One sequence (the longest one) is used as a representative for subsequence comparisons. A following BLAST "all-against-all" pairwise comparison considers only non-redundant sequences, storing all significant matches. If more sophisticated estimates of sequence similarity are required (true for some clustering algorithms), such as optimal local or semi-global alignment, these are performed last and only for those sequence pairs that are significant non-redundant matches.

With this, individual families within complete chordate genomes that are deemed it to be especially reliable (in this case, we initially used 18) can be exhaustively matched, the families identified, and evolutionary models (multiple sequence alignments, trees, and inferred probabilistic ancestral sequences) constructed for use family. Further, an ancestral sequence standing at the top of each nuclear family can be inferred for each of these families. Then, as new genome sequences become available, or even as individual sequences become available separate from whole genome sequencing efforts, a new exhaustive matching is not required.

Rather, the family to which the new sequence(s) belong can be identified by a search against the founder sequences for each of the previously identified families. Then, the new sequence(s) can be "tucked into" the multiple sequence alignment for the pre-computed family, and a branch within that pre-computed and rectified family can be added to indicate the point of divergence of the new sequence(s). While the pre-computed sequence would presumably have a fixed (and evolutionarily correct) species tree topology, the new sequence(s) might cause minor adjustments of the pre-computed evolutionary model for the. Of course, as the number of members within each family increases, the impact of each addition becomes smaller. Over the long-term, one can expect those models to become more or less stationary, with little change occurring further as additional genomic sequences are added. This strategy, therefore, brings to an end the computational challenge.

We report here our most recent efforts constructing evolutionarily organized databases following the strategy. We also report an outline of the use of the database to characterize the publicly available repertoire of whole genomic sequences.

Data quality

Even with confrontational challenge in hand, further problems emerge. Of particular importance is the quality of the data contained in existing publicly available genome sequence

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

databases. Data quality is a problem of broad scope caused by problems at several levels in the data collection process.

First, the sequence data delivered to genomic database can be unrepresentative of the actual sequence of the providing organism. Sometimes, this is the result of low coverage shotgun sequencing. Parts of the assembly may have coverage only from a single clone. Even worse, circumstances exist when the shotgun sequences have been assembled against a template from another source – often a different organism. Where short gaps exist arising from no coverage, the template sequence is used instead of "N", yielding a complete false impression of the quality of the data. While such problems can be easily identified when examining the primary assembly, it is impractical to do this on a genomic scale, and software working with genome-wide comparisons accepts the bulk annotation without any effort to assess its validity.

A second problem with data quality is simply that the gene calling is inaccurate. Occasionally this caused by gaps in the genome assembly, where whole exons (and introns) are missing from the primary sequence data. More commonly, the gene calling shows variations in the start or end point of genes that appear reasonably justified when the genome is examined in isolation, but are obviously wrong when orthologous genes are examined together. This is usually limited to errors in the end of one exon and/or the start of another, but can occasionally be much more pronounced, with genes that would be expected to be orthologs (as judged by an alignment made by MUMmer) being reported with wildly different transcripts that share few common exons.

As a third fact that creates problems when high quality alignments are required, alternative transcripts are typically not consistently reported from one whole genome project to another. Sometimes this may be because of differences in reporting criteria by the authors, but the quality of EST data upon which gene predictions are based must also be a factor. Although we must consider that EST data to be more reliable then raw DNA sequence data (it is, after all, experimental evidence for the expression of particular transcripts), it is rare to find transcripts in one organism that are supported by EST data to be disallowed due to mutation of alternative splice sites in closely related orthologs (say amongst mammals); one infers that absence of an alternative splice variant prediction is very weak evidence that the transcript is *not* present in the second organism.

A less important problem is that functional annotations (the linguistic statement reporting to describe the contribution of the protein to the fitness of the host organism) of genes in current

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

genomic databases can be quite wrong. While this makes functional annotation provided as linguistic statements of limited use, it also clearly illustrates how much the process of annotation (gene and functional) draws information from an arbitrary, but previously annotated homolog to make predictions. It is sometimes possible to see which species has been used as a template for another in the annotation process, simply by looking at the functional annotations and the gene calls.

Sequence clustering starting with 18 selected chordate families

Clustering can be an expensive process, not merely because of the number of sequences that are typically being organized, but because some clustering methods seek to cluster sequence regions instead of whole sequences. The primary challenge in building clusters is how to use (and weight) the similarities between sequences to generate desirable clusters. Many schemes have been explored over the last 20 years. Two methods have been developed internally during the development of MasterCatalog, while the GUI framework is capable of allowing the examination of arbitrary clustering (called Catalogs).

We illustrate here one early run that examined the Ensembl65 sequences for 18 chordate species (**Table 1**). The resulting catalog contained 8,199 individual families (with only clusters with four or more sequences being considered a family); on average, each family had 44 members.

Table 1. The 18 "Whole" Chordate Genomes Examined Here

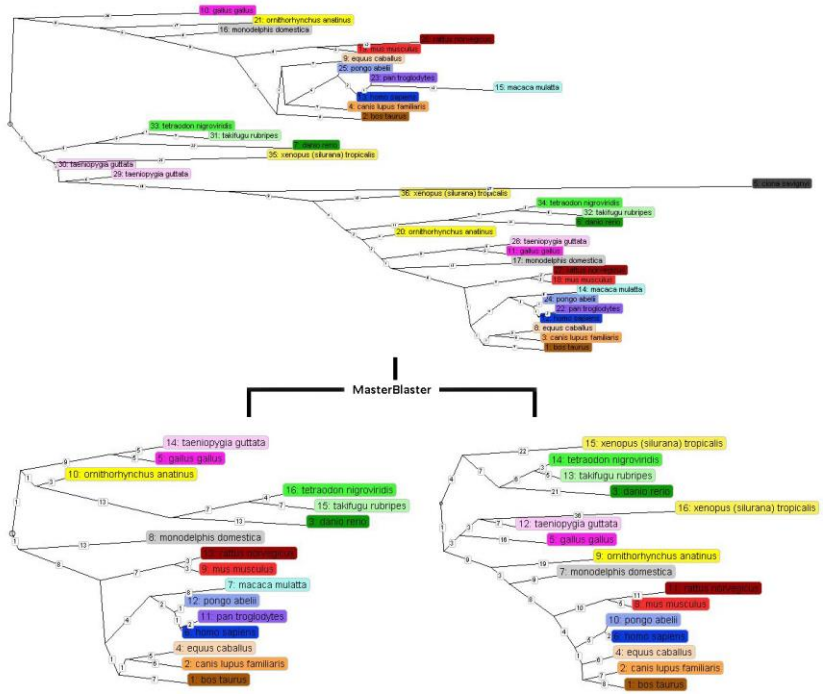
Species	**Common Name**
Bos taurus	Cow
Canis lupus familiaris	Dog
Ciona savignyi	Ciona
Danio rerio	Zebrafish
Equus caballus	Horse
Gallus gallus	Chicken
Homo sapiens	Human
Macaca mulatta	Rhesus monkey
Monodelphis domestica	Opossum
Mus musculus	Mouse
Ornithorhynchus anatinus	Platypus
Pan troglodytes	Chimpanzee
Pongo abelii	Orangutan
Rattus norvegicus	Rat
Taeniopygia guttata	Finch

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

Takifugu rubripes	Fugu
Tetraodon nigroviridis	Pufferfish
Xenopus (Silurana) tropicalis	Frog

Upon use of this catalog for individual family analyses, two problems became quickly apparent. First, protein sequences that had sufficient similarity to cluster in one family often had very different lengths. This caused problems in the creation of the MSA, which needed multiple gaps, often substantial in length, to accommodate sequences of very different lengths. This, in turn, corrupted distance metrics, which never score gaps correctly.

Second, when the clustering threshold was too low, very large families were created. In this case, 183 families contained more than 200 sequences. To address this problem, we created a tool to recluster families, internally referred to as MasterBlaster. This filter works by first examining the average sequence length with a family and removing any sequences that are of significantly different length from the average. Then, families were reclustered using BlastClust with more stringent criteria to allow larger families to be broken into smaller, more "natural" units, as well as further removing sequences that are distant enough to cause issues with the MSA. Application of MasterBlaster to the 18-genome database resulted in the creation of 14,058 individual families with an average of 20 sequences per family and only 6 families containing more than 200 sequences.



PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

Figure 1. Application of MasterBlaster to Family 1923 (HGNC symbol *LETM1*) derived from the 18-genome comparison. Note that the root is placed internal to the tree. The MasterBlaster splits a cumbersome family into two separate trees, as well as removing troublesome sequences such as *Ciona savignyi* (colored above as black). While the *Ciona* sequence is conveniently used to root the tree as the most primitive chordate outgroup, the time since its divergence from vertebrates (~1 billion years in both branches) causes sufficient sequence divergence and gapping to render distance metrics imprecise.

Missing Data Analysis

For an overwhelming majority of the families, one or more of the 18 "complete" genomes examined did not have a representative within the tree. In fact, 13,146 of 14,058 families were missing a sequence for at least one of the 18 species (93.5%). For example, **Figure 2** shows Family 5041 (cysteine dioxygenase type1, *CDO1*). This family had an apparent ortholog in 17 of the 18 species; however, no ortholog was found in the zebrafish. This could, of course, mean that the gene was lost during the episode of natural history following the divergence of zebrafish from other fish species. Alternatively, it could mean that (i) the whole genome sequence was less "whole" than desired, with the DNA segment encoding that gene missed in the sequencing effort, or (ii) the DNA segment encoding the *CDO1* gene was actually sequenced and present in the database, but the bioinformatic gene finding tool failed to find it.

To assess the relative likelihood of these alternatives, we constructed a tool that would search various sequence databases, including genomic, mRNA, and protein databases, in the event that an ortholog was missing or truncated. Both mRNA and protein sequences from species present in the tree were queried in hopes of finding missing sequences.

When this search tool was applied, we discovered a few alternative transcripts/translations that were missed. For example, for Family 5041, we found that Ensembl protein ENSDARP00000124052 from zebrafish would have clustered within this family and provided the missing ortholog. However, an alternate sequence from zebrafish, ENSDARP00000085212, was marked as the canonical protein sequence by the MasterCatalog; its sequence was too divergent to have made it into the family.

Only a minority of the families could be completed using this strategy. Specifically, we searched for 5292 sequences that were missing or truncated (defined as being < 50% of the

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

average length of sequences within the family) within 2843 families that had up to 3 of the 18 species without representative sequences. In all, we found alternate transcripts accounted for 17.1% of the missing sequences. We could further find 9.6% of the missing sequences in the genomic databases, and 7.7% within Ensembl's protein sequence database. However, even when missing sequences were found in the genomic or protein databases, these most often contained enough truncation or internal deletions to explain why they were not annotated by automated processes. For the remaining 65.6% of the sequences, our tool failed to find any reasonable trace that could complete the families. Thus an alternate approach was used to complete the families.

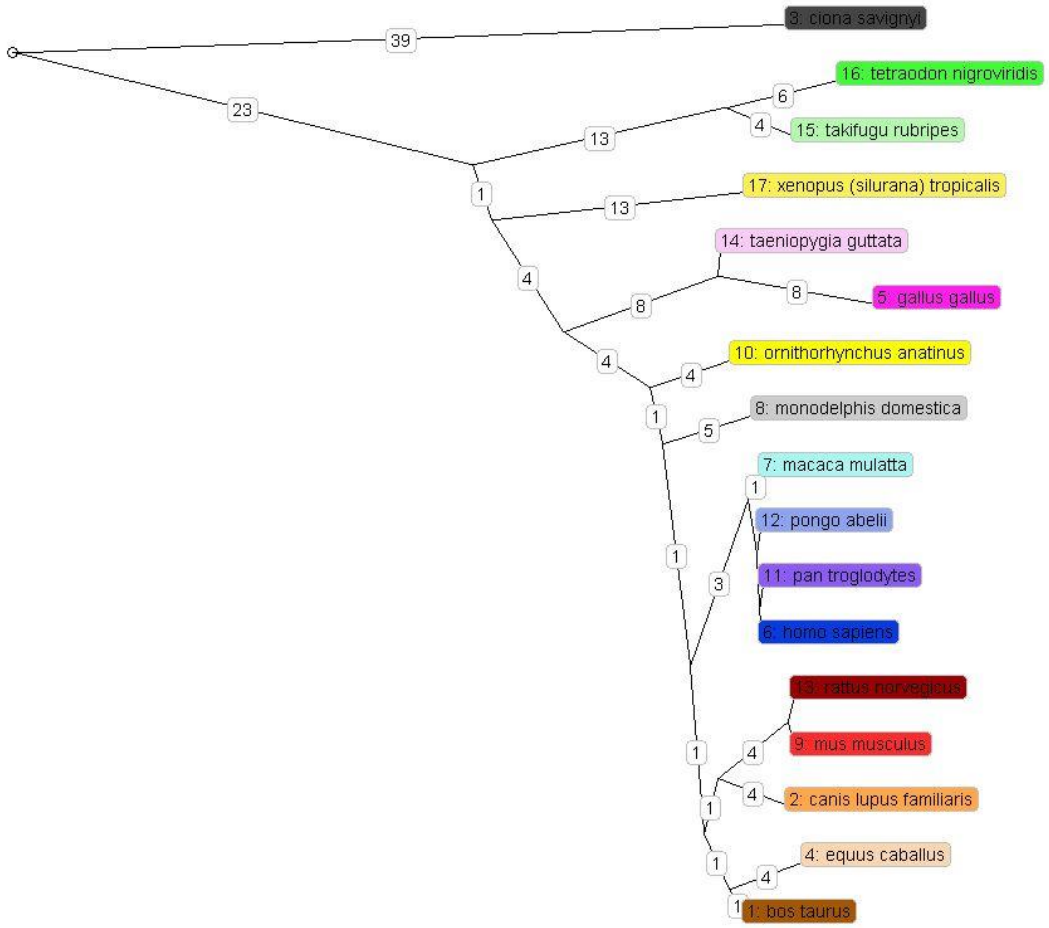


Figure 2. Family 5041, representing the *CDO1* gene, contains sequences for 17 of the 18 species within MasterCatalog. Zebrafish (danio rerio) was not present in the initial clustering of this family.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

Inclusion of Ensembl Compara Data

In 2008, Ensembl began to adopt several of the MasterCatalog innovations within a public database, the Ensembl Compara database [Vilella et al. 2009]. With their extensive computing resources, the Ensembl75 release delivered clusters and phlyogenetic trees for proteins from 66 "whole" genome species. Accordingly, we exploit this data to bypass the protein clustering, by far the most compute-expensive tasks, to allow this data to be analyzed within MasterCatalog. Our use of the Ensembl75 clustering data helped reduce issues associated with gene finding, including problems arising from alternate transcript/translation, and the missing sequence problem, both by having the capability to do more exhaustive analysis of sequences and by providing enough redundancy of species within tree branches to make an occasional missing sequence acceptable.

The core databases, containing DNA and protein sequences, for each of these 66 species were downloaded from Ensembl and loaded into the MasterCatalog database using built-in import functionality. The Compara database was then downloaded and mirrored as a local database.

Using custom scripts, two catalogs were created using the Ensembl data. The first used the protein clustering provided by Ensembl, but allowed the MasterCatalog to produce the MSAs and trees. The second used the clustering, MSAs, and reconciled trees produced by Ensembl. These two catalogs allow families to be viewed both with and without trees reconciled with the expected species tree, allowing a better overall view of the data and the error contained within.

For both catalogs, MasterCatalog is able to calculate K_a/K_s values for each ancestral node within a tree (**Figure 3**), a MasterCatalog innovation that has not yet been copied by the Ensembl Compara database. It was, of course, implemented in the TAED database [Liberles et al. 2001], which has been maintained and updated by the Liberles group. So far, the Ensembl Compara database simply has these values indicated by leaf-leaf comparisons between contemporary sequence pairs.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

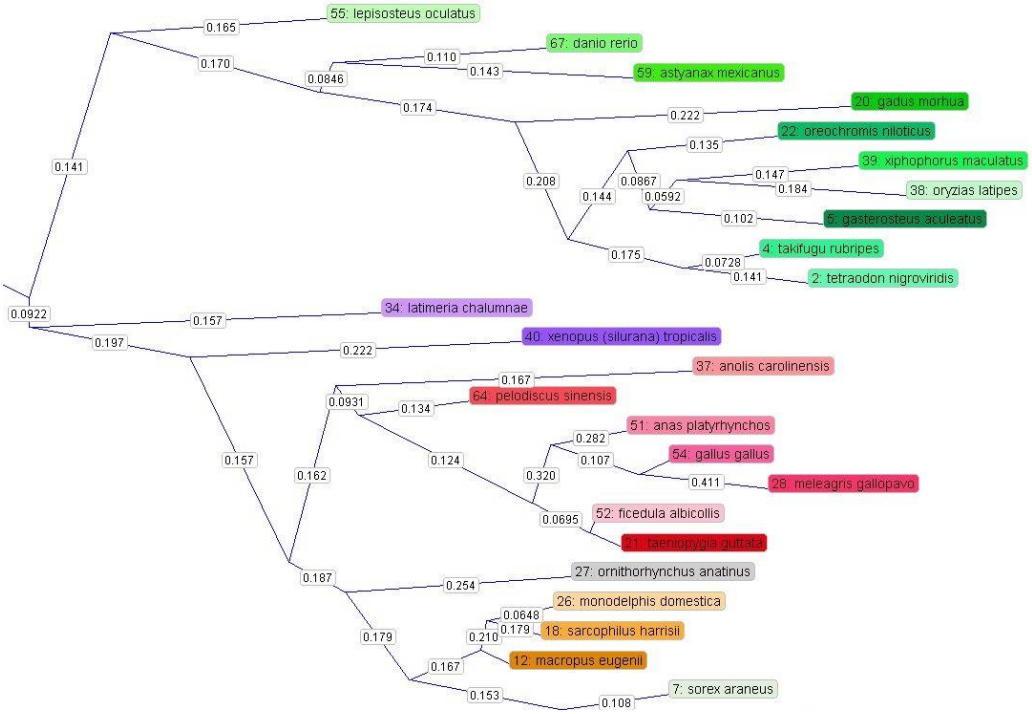


Figure 3: Computed K_a/K_s values are shown on ancestral branches for a portion of Family 4000 (HGNC symbol *FAM206A)*

Using K_a/K_s values computed for ancestral nodes [Messier and Stewart 1997], we can apply many evolutionary functional genomic tools, including identifying groups of proteins that appear to have undergone active change at given times within evolutionary history. This is a powerful tool for finding groups of proteins that emerged or significantly changed function at the time that significant evolutionary changes occurred, such as the development of the breast and prostate (see below).

The Graphical User Interface

Evolutionary functional genomics is greatly facilitated by graphical user interfaces that allow scientists to "surf the genome", examining the pre-computed evolutionary models for individual protein families quickly. Accordingly, a major advantage of MasterCatalog is its high-level graphical user interface. This interface has been used to generate all the figures presented so far in this paper.

The interface is written in Java, which allows it to be compatible with all major operating systems. Although a full description of this interface would be outside the scope of this writing, we briefly describe some of the features here.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

First, to view the underlying data, the scientist must be able to find the genes and protein products of interest. To this end, MasterCatalog contains the ability to search by many fields, including gene name, product name, gene description, and various external IDs, including Ensembl's, OMIM, and GO (**Figure 4**). Families can also be found via sequences comparison, with either protein or DNA/RNA sequence for the query. Once a family is found, it is displayed to the user in tabular format with a graphical representation of the protein sequence (**Figure 5**). From here, the user may choose to view these sequences as a multiple sequence alignment (**Figure 6**) or the phylogenetic tree can be explored. The tree viewer has many features to help the user explore the data. The entire tree can be fit to the window to gain a sense of the structure of large families (**Figure 7**), while the scale can be quickly customized in both the X and Y direction to focus on a specific region of the tree (**Figure 8**).

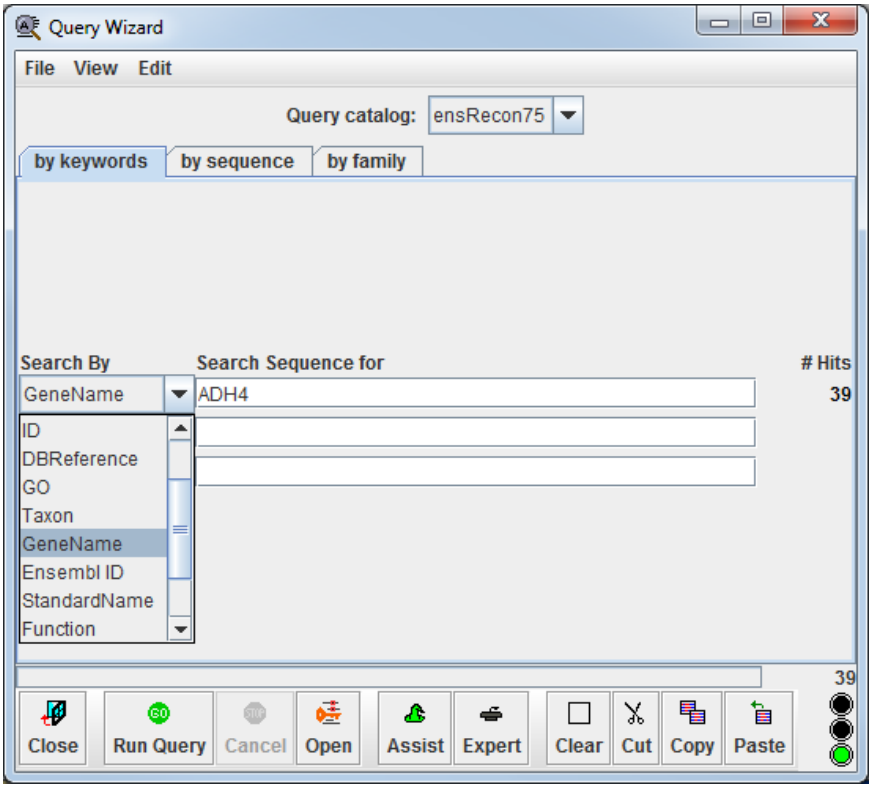


Figure 4. The query window for MasterCatalog allows the user to search by multiple fields to assist in finding the correct sequence family.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

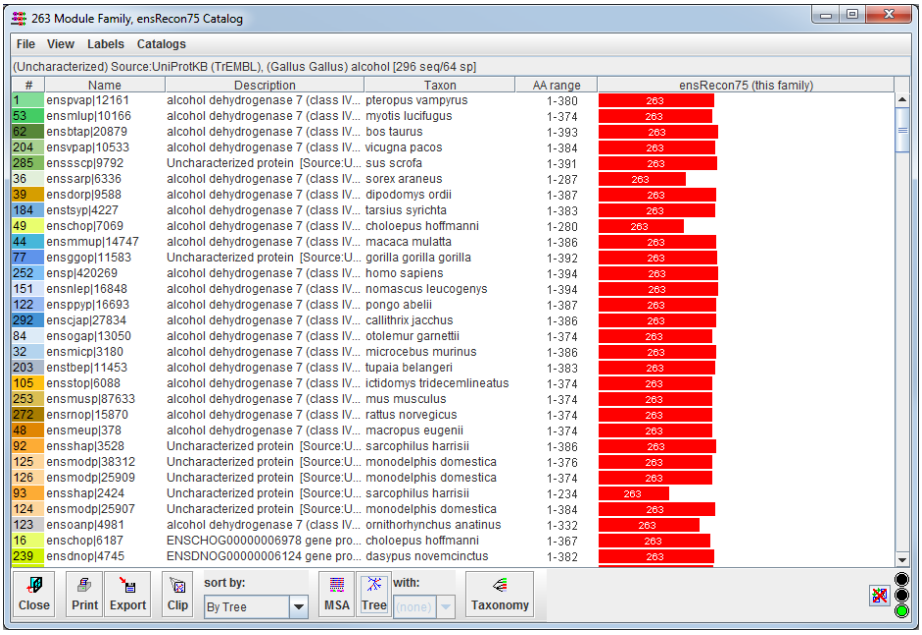


Figure 5. A tabular view of the sequences in a MasterCatalog family with graphical representation of sequences.

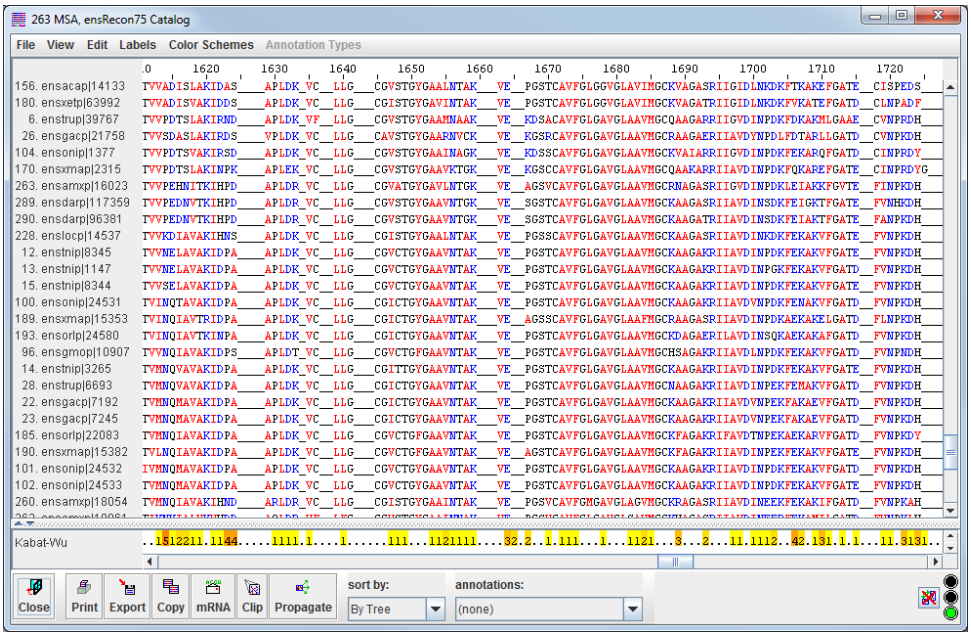


Figure 6. A view of MasterCatalog's MSA display window.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

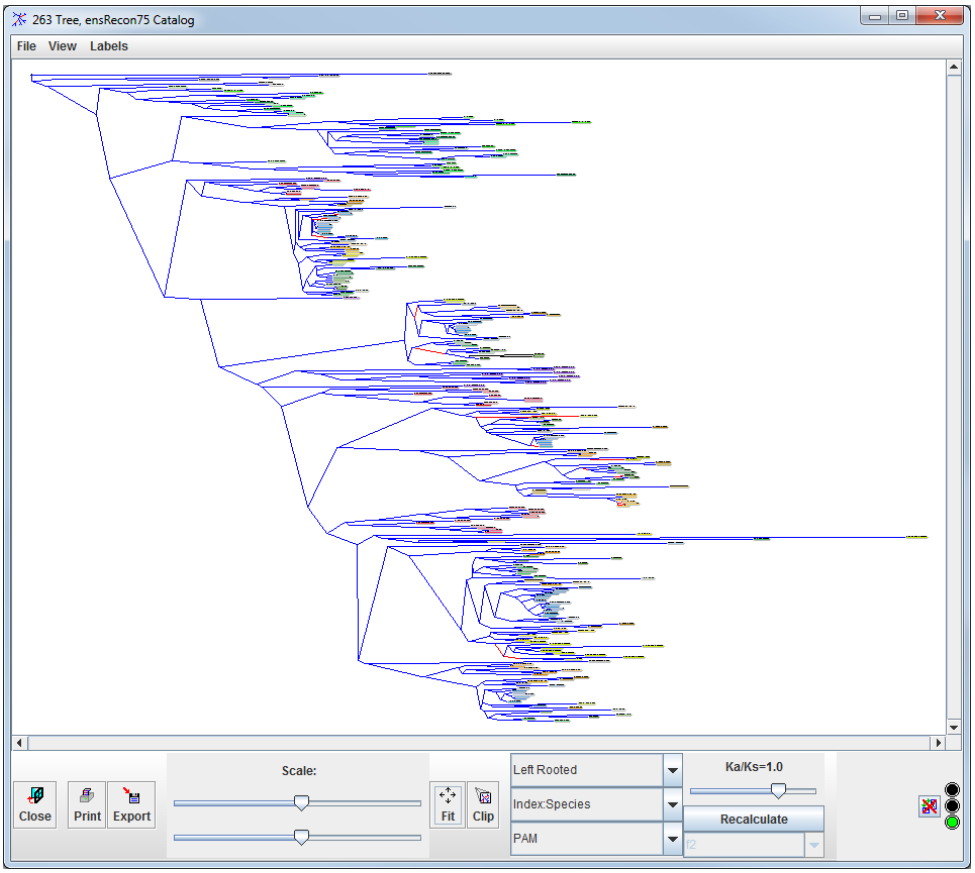


Figure 7. A top level view of a large phylogenetic tree in MasterCatalog (here displaying the sequence for alcohol dehydrogenase.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

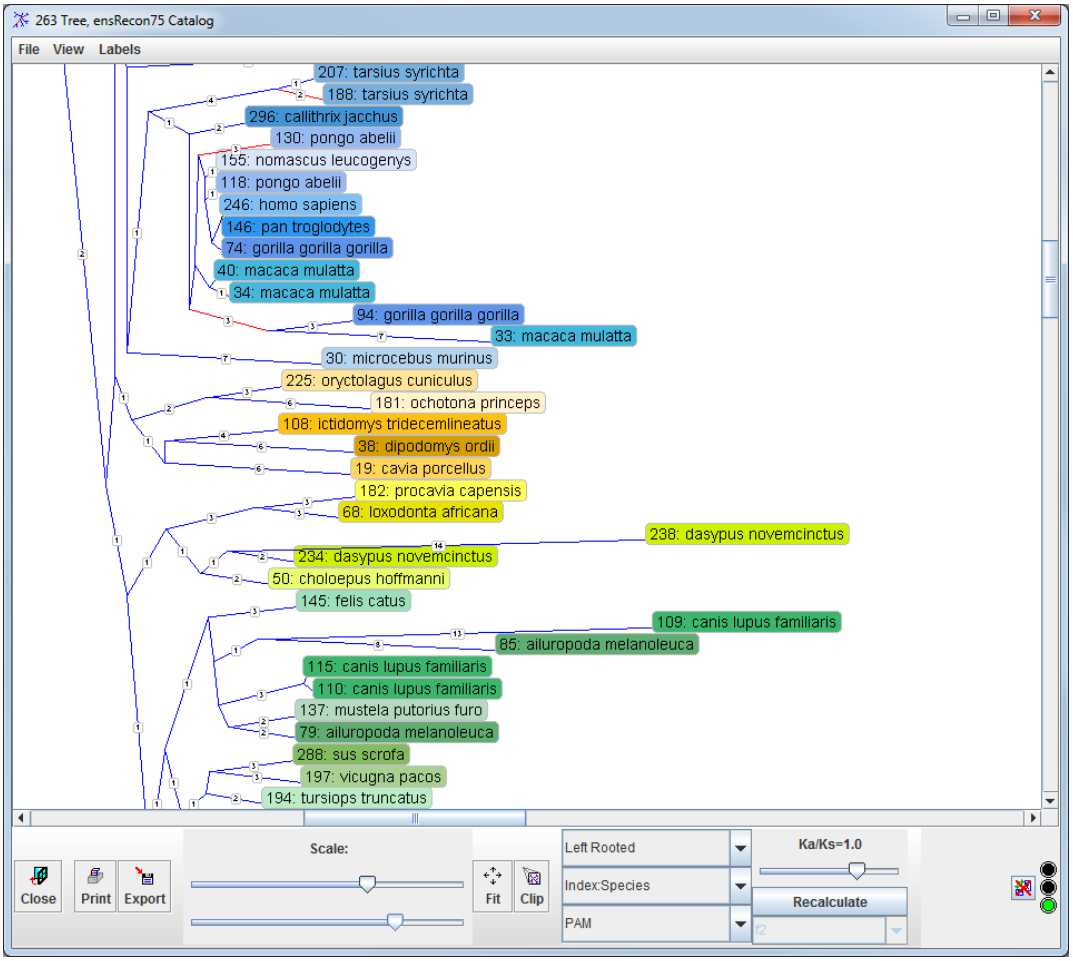


Figure 8. A zoomed in view of alcohol dehydrogenase sequences.

The user may also change tree structures (left rooted, top rooted, and unrooted), change the description attached to leaves (species name, sequence name, sequence description, etc.), and select the data displayed on the branches (PAM distance or K_a/K_s values). One advanced feature within MasterCatalog is the ability to "clip" data, which allows for subsets of sequences within a family to be displayed in either MSA or tree form. This clipped data can also have K_a/K_s values recalculated, which can be helpful when removing error-filled sequences that are obviously interfering with such calculations. Another advanced feature is the ability to view sequences in multiple catalogs at the same time. This allows comparison between different clustering criteria. Such comparisons can include sequence clustering vs. structural clustering, internal clustering vs. Ensembl clustering, and comparisons of trees reconciled against the species trees with non-reconciled trees. Such comparisons are unique to the MasterCatalog and allows the user a much greater insight into the data. With this comes the ability to select groups of sequences in one

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

window and have those sequences select across all open datasets, helping the user quickly navigate through a wealth of information.

Examples of use

The database allows us to explore “high level” questions in biology using genome-supported evolutionary analyses. For example, breast cancers display an intriguing mixture of characteristics, each having an associated diagnostic/prognostic/therapeutic problem. For example, improvements in screening (including ductal lavage) have allowed ductal carcinoma *in situ* to be detected quite often. However, most ductal carcinoma do not become invasive, and it is not understood why. Understanding of the "why” could minimize unnecessary interventions, providing immediate improvements in the management of breast cancer.

Each of these characteristics presumably has one or more genetic/epigenetic correlates, suggesting that if associated genes could be found, they might be sequenced in individual patients to identify markers that would help diagnose the primary cancer, prognosis its course, and choose a preferred therapy. Already this is done for estrogen-sensitive and insensitive cancers. However, to date, only about half of breast cancers are explained by the broadest set of risk factors; the most commonly used risk marker (BRCA1) covers perhaps only a fifth of total cancer incidence. Likewise, early exposure to radiation changes the spectrum of cancer risk, presumably by mutating genes, some perhaps not yet identified. Last, even if a breast cancer patient has survived for five years, a good chance remains of recurrence, again with uncertain etiology.

Thus, the ultimate overarching challenge (the elimination of mortality associated with breast cancer) is associated with a challenge: Can we identify a spectrum of genes that, if sequence-analyzed, can guide the genetic counselor, diagnostician, and physician in understanding the risk of breast cancer in individual patients, identify consequential cancers at an early stage, distinguish between early-stage aggressive from indolent cancers, and choose therapies that are neither too much nor too little? Analyses of these genes in combination will, we hope, cover all breast cancers, much as genetic analyses of BRCA or estrogen receptors offer similar guidance for a fraction of those cancers.

Following an evolution-based functional genomics strategy, we begin by recognizing that the breast (as a tissue) emerged only recently in the history of Earth, approximately 300 million

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

years ago (**Figure 9**). This episode is indicated in the tree in **Figure 9** by a blue line, the episode when suckling vertebrates emerged via divergence from other amniotes, most notably non-suckling birds and reptiles, which diverged still earlier from amphibians and, even earlier, from fish. The assignment of this time in natural history as the time when the breast emerged is, of course, identical to the statement from biological systematics that mammals form a true vertebrate class. The episode is recent enough in history to avoid much of the ambiguity that arises when bioinformatics tools model more ancient events.

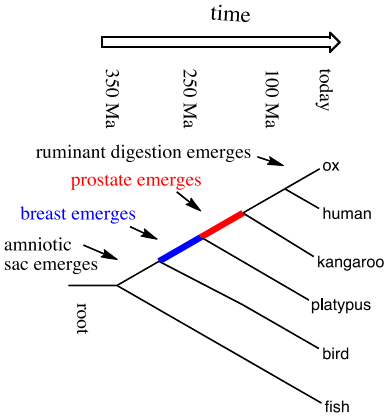


Figure 9. A schematic outlining the evolution of vertebrate tissues. Time is in million years. The red and blue lines indicate the episodes for the emergence of the prostate and breast (respectively).

A second tissue new in mammals emerged in the episode immediately following (the red line): the prostate. This is indicated by the lack of a prostate in the platypus, but its presence in marsupials. The prostate is also a tissue that appears to generate cancer without obvious "insult" (although environmental factors can increase the incidence of prostate cancer, as they can breast cancer). Again, the red episode is sufficiently recent to avoid many ambiguities that make difficult bioinformatics analysis of more ancient events. Further, it is convenient to have two tissues from opposite genders equally susceptible to cancer and equally accessible to evolutionary analysis, as they can serve as controls (of a sort) for each other.

It is axiomatic in evolutionary developmental biology that the emergences of the breast and prostate in the blue and red episodes were associated with genetic changes. Further, Bayesian analyses are well known to be able to infer genetic events in the historical past through the analysis of modern gene sequences [Yang 1997]. Thus, we (and others) have long inferred the

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

sequences of ancestral genes and proteins from ancient genomes by analyzing the sequences of their descendants. In a field invented in the Benner laboratory [Benner 2007], paleogenetics can go still further, resurrecting inferred ancestral sequences from extinct animals by recombinant DNA technology, making ancient proteins available for study in the laboratory. Since maximum likelihood DNA and protein sequences at nodes in an evolutionary tree can be inferred using Bayesian analysis, probabilistic changes can be assigned to individual branches in a tree like that shown in **Figure 9**. Therefore, when applied to entire genomes, protein family by protein family, we can say what amino acids were replaced, inserted, or deleted during the episode when the breast emerged, or when the prostate emerged. While we agree that non-coding regions are also important to an "evo-devo" analysis, these are not addressed here because of the greater difficulty in inferring their histories. Further, numerous examples suggest that when a gene is recruited to perform a new role, a signature of recruitment and its associated adaptive evolution can be inferred by examining what amino acids are replaced, inserted, or deleted.

More sophisticated analyses can be applied across whole families and whole pathways. Thus, we can suggest a central hypothesis: *To identify genes and proteins involved in the emergence of the breast and/or prostate, we might go stepwise, family by family, through the genomic history of vertebrates to find those that carry signatures of functional adaptation at the time when the breast emerged and/or when the prostate emerged*. This work will deliver this family-by-family analysis.

Various hypotheses give such analyses medical relevance. First, we hypothesize that genes involved in the emergence of the breast (and, as a control, prostate) are likely candidates for genes involved in regulating the growth, development, and functioning of these tissues in modern mammals. Further, we hypothesize that mutations in these genes create susceptibility to these cancers, determine the types of cancers that result, control the likelihood that those cancers will progress and metastasize, and govern susceptibility of the resulting cancers to different therapies. In this view, our evolutionary analysis will complement "classical" approaches to the same goal, such as "deep sequencing" of multiple specimens of breast cancer tissue in search of mutations with etiological significance, in established cancer-linked proteins, or the use of large-scale typing of genetic markers in a case/control study format.

A naturally organized database can help, especially if provided with semi-automated tools that will address error in genome annotation, heuristic development, and expert analysis [Benner et

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

al. 1995]. This supports a multi-metric approach to identify individual nuclear families that have signatures of adaptive change during the episodes represented by the blue and red lines, including the obvious gene duplications/loss, gene shuffling, and K_a/K_s metrics [Liberles et al. 2001], but also more subtle changes, such as homoplasy, heterotachy, and clustering of adaptive sites on 3D-structures (if known). The ultimate goal is to identify individual families whose paleogenetic history suggests underwent adaptive change during the historical episode when the breast emerged. Parallel studies in the immediately following episode when the prostate emerged provide a "control" with similar bioinformatics constraints.

In developing this theme, it is useful to review one of many cases where the Benner group applied multi-metric tools to generate hypotheses of adaptive change in vertebrate protein families. Most relevant to breast cancer is our study of aromatase (**Figure 10**) [Gaucher et al. 2004]. This protein oxidatively transforms androgens to create estrogens using heme, nicotinamide, and a coreductant. During the divergent evolution of artiodactyls (cloven hooved ungulates), the aromatase gene family suffered two duplications to generate three paralogous aromatases in the lineage that led to modern pigs (**Figure 10a**).

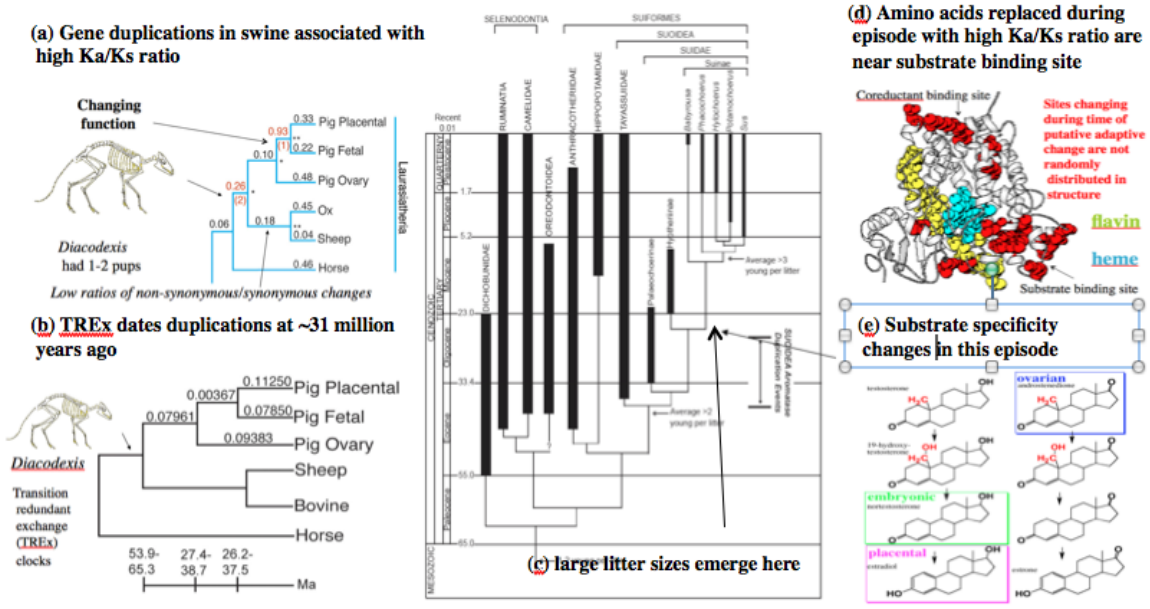


Figure 10. A multi-metric analysis of the molecular evolution of the vertebrate aromatase gene family. The goal of this project is to do similar analyses for vertebrate protein families with focus on identifying those important in the historical emergence of the breast and (as a control) the

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

prostate. In addition the work will integrate and exploit the wealth of available new data from "modern biology" (e.g. functional genomics)

These are dated using the TREx clock [Li et al. 2006] to have occurred 31 ± 5 million years ago (**Figure 10b**). One of the duplications is associated with a relatively high (0.93) K_a/K_s ratio (**Figure 10a**). While this ratio is not greater than unity, it is large compared to ratios in other branches of the tree (which are typically ~0.2) (**Figure 10a**). Therefore, this high *relative* ratio suggests that this family is undergoing functional change. The very survival of paralogs, of course, also suggests adaptation.

Had we stopped here, this case might have been just one many disputed examples in a literature containing many of these. However, the next step in the multi-metric approach noted that the amino acids replaced during the episode with a high K_a/K_s ratio were not randomly distributed across the structure of the protein (**Figure 10d**). Rather, they were clustered near the substrate binding site and the co-reduction binding site (**Figure 10d**). This suggests that during this episode, details of the structure of the substrate had changed, as did the co-reductant. This led to experimental work that showed that the different paralogs had different substrate specificities (**Figure 10e**) [Corbin et al. 1999].

But what does it mean functionally? A cladogram based on fossil records (**Figure 10c**) suggested that this episode of adaptive change occurs near the time in which pig litters went from one piglet (with occasional twinning) to five or more piglets. This generated the hypothesis that this gene triplication emerged to manage a new reproductive physiology in pigs (large litter size). This drove an analysis of the molecular physiology, which confirmed this inference Corbin et al. 2004] [Kao et al. 2000] [Conley et al. 2001].

To date, multi-metric analyses have been generally developed case-by-case. From the Benner group, these include analyses of dehydrogenases [Benner 1989], ribonucleases [Sassi and Benner 2007], leptins [Gaucher et at 2003], sulfotransferases [Bradley and Benner 2005], inflammatory proteins [Benner 2002], hypertension [Johnson et al. 2008], SARS [Benner et al. 2003], cystic fibrosis [Gaucher et al. 2006], uterin serpins [Peltier et al. 2000], ribonucleotide reductase [Tauer and Benenr 1997], and elongation factors [Gaucher et al. 2001], among others [Benner et al. 2002]. In some cases, we have been interested in developing statistical heuristics that assess the number of free variables that should be used to model adaptive divergence [Sassi et al. 2007]. In

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

other cases, we have explored the use of heterotachy (see below) to identify episodes of functional adaptation [Gaucher et al. 2002]. In other cases, we have asked how codon models [Benner 2012], scoring tools [Gonnet et al. 2000], and gapping models [Benner et al. 1993] [Chang and Benner 2004] improve multiple sequence alignments, or the impact of homoplasy in corrupting gene trees with short branch lengths [Carrigan et al. 2012]. In other cases, we have used the approach to improve the models upon which Bayesian inference relies [Gonnet et al. 1994]. In others, the analysis has been the start of paleogenetics experiments, where ancestral proteins from now-extinct organisms are resurrected in the laboratory for study [Stackhouse et al. 1990] [Jermann et al. 1995] [Ciglic et al. 1998] [Opitz et al. 1998] [Gaucher et al. 2003] [Thomson et al. 2005].

Again a principal challenge doing such work by automated methods arises from annotation and gene finding errors in whole genome sequence databases. These are illustrated in **Figure 11** for the primate BRCA1 gene family. Purely automated assembly of the family encounters situations where (in this case) it appears as if gene duplications created two paralogs in *Macaca* and *Pongo* (the rhesus monkey and orangutan, respectively). Of course, it is conceivable that this gene actually did suffer duplication independently in the two lineages leading to *Pongo* and *Macaca*. If so, this duplication would (like in the aromatase case) indicate functional adaptation, especially as the half-life for survival of nonfunctional duplicates that have not acquired a new function is about 11 million years [Trabesinger-Ruer et al. 1996]. In fact, this is not the case; the apparent paralogs in *Pongo* and *Macaca* are evidently the result of mistaken gene finding.

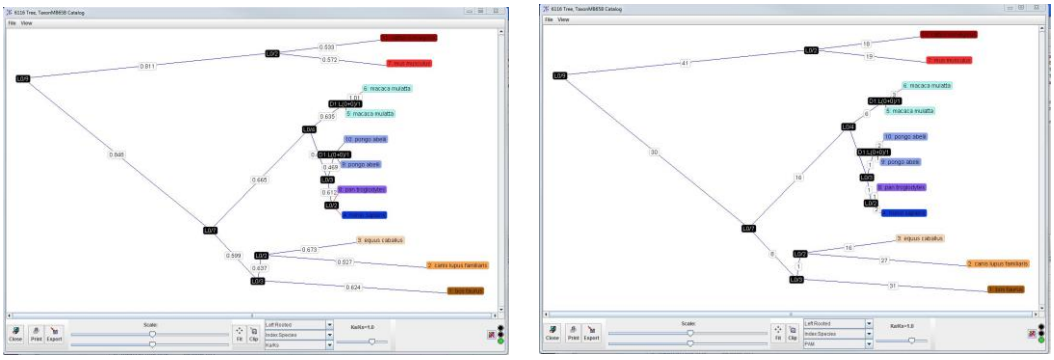


Figure 11. The 18 vertebrate whole genomes within the MasterCatalog are illustrated here for the BRCA1 nuclear family for primates. This example illustrates “false paralogization”, where the *Pongo* and *Macaca* genomes have extra sequences that suggest duplicates where none exist.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

These families will be rectified to remove such errors, when they corrupt the interpretation about historical events at the time where the breast and prostate emerged.

We can further apply advanced metrics to detect functional change that complement the K_a/K_s ratio and duplications [Benner 2004]. For example, heterotachy recognizes that two different branches of a tree whose respective members have different functions also have different site-by-site constraints on functional divergence (**Figure 12**). In lay language, that means that in proteins having different functions, different sites evolve more rapidly while other sites evolve more slowly. Likewise, homoplasy can indicate specific sites having specifically changing functional roles. Other tools include an analysis of compensatory changes [Fukami-Kobayashi et al. 2002] and crystallographic clustering, both of which bring crystallographic data to bear on an analysis of functional divergence [Benner et al. 1997]. For example, amino acids being replaced during an episode of relatively high non-synonymous substitution are often not distributed randomly across the three-dimensional structure, but rather are clustered, perhaps near a substrate binding or regulatory site. This crystallographic clustering is strong evidence for adaptive change and, as in the aromatase example, can guide specific experiments to confirm/deny a hypothesis of changing function.

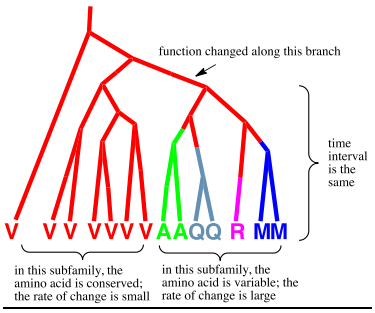


Figure 12. "Heterotachy" is a change in the *rate* of amino acid substitution at a site that indicates a change in function. It requires whole family analysis to detect. Shown the amino acids reside at a site in a hypothetical protein. Purifying selection retained a valine at this position in the left branch, but did not retain any amino acid in the right branch. The change in functional constraints at this site indicates that the function in the protein changed in the episode indicated by the arrow.

Summary

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

A quarter century has passed since Gaston Gonnet began to help us use evolutionary analysis to extract function in ever-growing sequence databases. The results of this collaboration are now having impact throughout biomedical research, much no longer acknowledged by (or even known to) today's beneficiaries of a research program that began so long ago. However, the hour spent 25 years ago in a seminar that Gaston gave on the Oxford Unabridged English Dictionary was more than well spend. Thanks again to Beverly for making me aware of, and encouraging me to attend, it.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

Gonnet, G. H., Benner, S. A. (1991) Computational Biochemistry Research at ETH. *Technical Report 154, Departement Informatik*, March

Dayhoff, M. O., Eck, R. V., Park, C. M. (1972) A *Atlas of Protein Sequence and Structure.* Vol. 5 (ed. M. O. Dayhoff)

Bairoch, A., Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45-48

Gonnet, G. H., Cohen, M. A., Benner, S. A. (1992) Exhaustive matching of the entire protein sequence database. *Science* **256**, 1443-1445

Gonnet, G. H., Cohen, M. A., Benner, S. A (1994) Analysis of amino acid substitution during divergent evolution. The 400 by 400 dipeptide mutation matrix. *Biochem. Biophys. Res. Comm.* **199**, 489-496

Benner, S. A., Cohen, M. A., Gonnet, G. H. (1993) Empirical and structural models for insertions and deletions in the divergent evolution of proteins. J. Mol. Biol. 229, 1065-1082

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

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

Benner, S. A., Trabesinger-Ruef, N., Schreiber, D. R. (1998) Post-genomic science. Converting primary structure into physiological function. *Adv. Enzyme Regul*. **38**, 155-180

Benner, S. A. (2003) Interpretive proteomics. Finding biological meaning in genome and proteome databases. *Adv. Enzyme Regul*. 4**3** 271-359

Benner, S. A., Ellington, A. D., Tauer, A. (1989) Modern metabolism as a palimpsest of the RNA world. *Proc. Nat. Acad. Sci.* **86**, 7054-7058

Benner, S. A., Cohen, M. A., Gonnet, G. H., Berkowitz, D. B., Johnsson, K. (1993) Reading the palimpsest. Contemporary biochemical data and the RNA world. in *The RNA World*. Gesteland. R. F., Atkins, J. F. editors, Cold Spring Harbor Laboratory Press, 27-70

Benner, S. A., Chamberlin, S. G., Liberles, D. A., Govindarajan, S., Knecht, L. (2000) Functional inferences from reconstructed evolutionary biology involving rectified

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

databases. An evolutionarily-grounded approach to functional genomics. *Research Microbiol.* **151**, 97-106.

Duret, L., Mouchiroud, D., Gouy, M. (1994) Hovergen, a database of homologous vertebrate genes. *Nucleic Acids Res.* **22**, 2360-2365

Bradley, M. E., Benner, S. A. (2006) Integrating protein structures and precomputed genealogies in the Magnum database. Examples with cellular retinoid binding protein. *BMC Evol. Biol.* **7**, Art. No. 89

Vilella, A. J., Severink J., Ureta-Vidal, A., Heng, L., Durbin, R., Birney, E. (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* 19, 327-335

Liberles, D. A., Schreiber, D. R., Govindarajan, S., Chamberlin, S. G., Benner, S. A. (2001) The adaptive evolution database (TAED) *Genome Biol.* **2**, 0028.1-0028.6

Messier, W., Stewart, C. B. (1997) Episodic adaptive evolution of primate lysozymes. *Nature* **385**, 151-154

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

Benner, S. A. (2007) The early days of paleogenetics. Connecting molecules to the planet. Experimental Paleogenetics, D. A. Liberles, ed. Academic Press, pp. 3-19

Benner, S. A., Jermann, T. M., Opitz, J. G., Stackhouse, J., Knecht, L. J., Gonnet, G. H. (1995) Uncertainty in ancient phylogenies. *Nature* **377**, 109-110

Liberles, D. A., Schreiber, D. R., Govindarajan, S., Chamberlin, S. G., Benner, S. A. (2001) The adaptive evolution database (TAED). *Genome Biol.* 2, 0003.1-0003.18

Gaucher, E. A., Graddy, L. G., Simmen, R. C. M., Simmen, F. A., Kowalski, A. A., Schreiber, D. R., Liberles, D. A., Janis, C. M., Chamberlin, S. G., Benner, S. A. (2004) The planetary biology of cytochrome P450 aromatases. *BMC Biology*. **2**, Art. No. 19

Li, T. *et al.* (2006) Analysis of transitions at two-fold redundant sites in mammalian genomes. Transition redundant approach-to-equilibrium (TREx) distance metrics. *BMC Evolutionary Biology*, **6,** 25.241

Corbin, C. J., Trant, J. M., Walters, K. W., Conley, A. J. (1999) Changes in testosterone metabolism associated with the evolution of placental and gonadal isozymes of porcine aromatase cytochrome P450. *Endocrinology* **140**, 5202-5210.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

Corbin, C. J., Mapes, S. M., Marcos, J., Shackleton, C. H., Morrow, D., Safe, S., Wise, T., Ford, J. J., Conley, A. J. (2004) Paralogues of porcine aromatase cytochrome p450: A novel hydroxylase activity is associated with the survival of a duplicated gene. *Endocrinology* **145**, 2157-2164.

Kao, Y. C., Higashiyama, T., Sun, X., Okubo, T., Yarborough, C., Choi, I., Osawa, Y., Simmen, F. A., Chen, S. (2000) Catalytic differences between porcine blastocyst and placental aromatase isozymes. *Eur, J, Biochem*. **267**, 6134-6139.

Conley, A., Mapes, S., Corbin, C. J., Greger, D., Walters, K., Trant, J., Graham, S. (2001) A comparative approach to structure-function studies of mammalian aromatases. *J Steroid Biochem*. **79**, 289-297.

Benner, S. A. (1989) Patterns of divergence in homologous proteins as indicators of tertiary and quaternary structure. *Adv. Enz. Regul.* **28**, 219-236

Sassi, S. O., Benner, S. A. (2007) The resurrection of ribonucleases from mammals. From ecology to medicine. *Experimental Paleogenetics*, D. A. Liberles, ed., NY, Academic Press, pp 208-224

Gaucher, E. A., Miyamoto, M. M., Benner, S. A. (2003) Evolutionary, structural and biochemical evidence for a new interaction site of the leptin obesity protein *Genetics* **163**, 1549-1553

Bradley, M. E., Benner, S. A. (2005) Phylogenomic approaches to common problems encountered in the analysis of low copy repeats: The sulfotransferase 1A gene family example. *BMC Evolutionary Biology* **5**, Art. No. 22

Benner, S. A. (2002) The past as the key to the present. Resurrection of ancient proteins from eosinophils. *Proc. Natl. Acad. Sci. USA* **99**, 4760-4761

Johnson, R.J., Gaucher, E. A., Sautin, Y. Y., Henderson, G. N., Angerhofer, A. J., Benner, S. A. (2008) The planetary biology of ascorbate and uric acid and their relationship with the epidemic of obesity and cardiovascular disease. *Medical Hypotheses* **71**, 22-31

Benner, S. A., Gaucher, E. A., Li, T. (2003) Post-genomic evolutionary analyses of the Severe Acute Respiratory Syndrome (SARS) virus genome using the MasterCatalog interpretive proteomics platform. *Pharmagenomics –Application Notebook* **2003**, 23

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

Gaucher, E, A. De Kee, D. W., Benner, S. A. (2006) Application of DETECTER, an evolutionary genomic tool to analyze genetic variation, to the cystic fibrosis gene family. *BMC Genomics* **7**, No. 44

Peltier, M.R., Raley, L.C., Liberles, D. A., Benner, S.A., Hansen, P.J. (2000) Evolutionary history of the uterine serpins. *J. Exp. Zool. (Mol. Devel. Evol.)* **288**, 165-174

Tauer, A., Benner, S. A. (1997) The B12-dependent ribonucleotide reductase from the archaebacterium *Thermoplasma acidophila*. An evolutionary conundrum. *Proc. Natl. Acad. Sci.* **94**, 53-58

Gaucher, E. A., Miyamoto, M. M., Benner, S. A. (2001) Function-structure analysis of proteins using covarion-based evolutionary approaches. Elongation factors. *Proc. Natl. Acad. Sci. USA* **98**, 548-552

Benner, S. A., Caraco, M. D., Thomson, J. M., Gaucher, E. A. (2002) Planetary biology. Paleontological, geological, and molecular histories of life. *Science* **293**, 864-868

Sassi, S. O., Braun, E. L., Benner, S. A. (2007) The evolution of seminal ribonuclease. Pseudogene reactivation or multiple gene inactivation events? *Mol. Biol. Evol.* **24**, 1012-1024

Gaucher, E. A., Gu, X., Miyamoto, M. M., Benner, S. A. (2002) Predicting functional divergence in protein evolution by site-specific rate shifts. *Trends Biochem. Sci.* **27**, 315-321

Benner, S. A. (2012) Use of codon models in molecular dating and functional analysis. In *Codon Evolution. Mechanisms and Models*. Oxford University Press, 133-144

Gonnet, G., Korostensky, C., Benner, S. (2000) Evaluation measures of multiple sequence alignments. *J. Comput. Biol.* **7**, 261-276

Benner, S. A., Cohen, M. A., Gonnet, G. H. (1993) Empirical and structural models for insertions and deletions in the divergent evolution of proteins. J. Mol. Biol. 229, 1065-1082

Chang, M., Benner, S. A. (2004) Empirical analysis of insertions and deletions in protein sequence evolution. *J. Mol. Biol.* **341**, 617-631

Carrigan, M. A., Uryasev, O., Davis, R. P., Zhai, L-M., Hurley, T. D., Benner, S. A. (2012) The natural history of Class I primate alcohol dehydrogenases includes gene duplication, gene loss, and gene conversion. *PLOS 1* **7** (7) e41175

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014

660 Gonnet, G. H., Cohen, M. A., Benner, S. A (1994) Analysis of amino acid substitution during
661 divergent evolution. The 400 by 400 dipeptide mutation matrix. *Biochem. Biophys. Res.*
662 *Comm.* **199**, 489-496
663 Stackhouse, J., Presnell, S. R., McGeehan, G. M., Nambiar, K. P., Benner, S. A. (1990) The
664 ribonuclease from an extinct bovid. *FEBS Lett*. **262**, 104-106
665 Jermann, T. M., Opitz, J. G., Stackhouse, J., Benner, S. A. (1995) Reconstructing the
666 evolutionary history of the artiodactyl ribonuclease superfamily. *Nature* **374**, 57-59
667 Ciglic, M. I., Jackson, P. J., Raillard-Yoon, S.-A., Haugg, M., Jermann, T. M., Opitz, J. G.,
668 Benner, S. A. (1998) Origin of dimeric structure in the ribonuclease superfamily.
669 *Biochemistry* **37**, 4008-4022
670 Opitz, J. G., Ciglic, M. I., Haugg, M., Trautwein-Fritz, K., Raillard-Yoon, S.-A., Jermann, T. M.,
671 Moore, R., Benner, S. A. (1998) Origin of the catalytic activity of bovine seminal
672 ribonuclease against double-stranded RNA. *Biochemistry* **37**, 4023-4033
673 Gaucher, E. A., Thomson, J. M., Burgan, M. F., Benner, S. A. (2003) Inferring the
674 paleoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature* **425**,
675 285-288
676 Thomson, J. M., Gaucher, E. A., Burgan, M. F., De Kee, D.W.,Li, T., Aris, J. P., Benner, S. A.
677 (2005) Resurrecting ancestral alcohol dehydrogenases from yeast. *Nature Genetics* **37**,
678 630-635
679 Trabesinger-Ruef, N., Jermann, T. M., Zankel, T. R., Durrant, B., Frank, G., Benner, S. A.
680 (1996) Pseudogenes in ribonuclease evolution. A source of new biomacromolecular
681 function? *FEBS Lett*. **382**, 319-322
682 Benner, S. A. (2004) Evolution-based functional proteomics. US Patent Application.
683 US20040204861 A1
684 Fukami-Kobayashi, K., Schreiber, D. R., Benner,S.A. (2002) Detecting compensatory
685 covariation signals in protein evolution using reconstructed ancestral sequences. *J. Mol.*
686 *Biol.* **319**, 729-743
687 Benner, S. A., Cannarozzi, G., Gerloff, D. L., Turcotte, M., Chelvanayagam, G. (1997) *Bona*
688 *fide* predictions of protein secondary structure using transparent analyses of multiple
689 sequence alignments. *Chem. Rev.* **97**, 2725-2843
690

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.424v1 | CC-BY 4.0 Open Access | received: 26 Jun 2014, published: 26 Jun 2014