Res. Microbiol. 151 (2000) 97–106

S0923250800000123-6/REV

Functional inferences from reconstructed evolutionary biology involving rectified databases – an evolutionarily grounded approach to functional genomics

Steven A. Benner[a]*, Stephen G. Chamberlin[b], David A. Liberles[a], Sridhar Govindarajan[b], Lukas Knecht[b]

[a] *Departments of Chemistry and Anatomy and Cell Biology, University of Florida, Gainesville, FL, USA*
[b] *EraGen Biosciences, 12085 Research Drive, Alachua, FL 32615, USA*

Abstract — If bioinformatics tools are constructed to reproduce the natural, evolutionary history of the biosphere, they offer powerful approaches to some of the most difficult tasks in genomics, including the organization and retrieval of sequence data, the updating of massive genomic databases, the detection of database error, the assignment of introns, the prediction of protein conformation from protein sequences, the detection of distant homologs, the assignment of function to open reading frames, the identification of biochemical pathways from genomic data, and the construction of a comprehensive model correlating the history of biomolecules with the history of planet Earth. © 2000 Éditions scientifiques et médicales Elsevier SAS

functional genomics / sequence database / Master Catalog / protein structure prediction / evolution

1. Introduction

The explosion of sequence data has generated a wealth of new problems for the bioinformaticist to solve. These begin simply with the task of curation of the data, putting the data in places where it can be found and used. Curation strategies must reflect the fact that the amount of genomic sequence data will increase still further, perhaps by two orders of magnitude, over the next decade or so.

Next comes the problem of interpretation. The sequence of a protein is written in the language of organic chemistry, which uses words such as 'atoms', 'bonds', and 'functional groups'. To be most useful to biologists, protein sequences must be translated into the language of biology. The language of biology concerns phenotype and 'function'. Under the Darwinian paradigm, function is the behavior of a molecule that confers 'fitness' on an organism, reflected in an ability to survive, select a mate, and reproduce, distinct from 'behavior', which is what is measured in the laboratory.

It is now widely appreciated that an evolutionary analysis provides one of the most powerful ways to organize genome sequences with respect to function [4, 19]. Virtually all annotation methods now being applied to sequence databases involve the search for evolutionary homologs, other proteins that are related by common ancestry. Each new sequence that is determined is used as a probe in a BLAST search to identify similar sequences in the database which (one hopes) have themselves been annotated. Once these are found, the 'homology-implies-functional-analogy' (HIFA) paradigm is followed [12]. The logic behind this paradigm holds that sequence similarity implies homology, homologous proteins have analogous conformations (or folds), analogous folds imply analogous behaviors (what is measured

* Correspondence and reprints
Tel.: +1 352 392 7773; fax: +1 352 846 2095;
benner@chem.ufl.edu

experimentally), and analogous behaviors imply analogous functions [14].

Difficulties lie at the very beginning of the functional interpretation of a genome database. Most genes are interrupted by introns; finding these is difficult, and error-prone. Tools are needed to identify errors in the database. Further, even in the pregenomic era, the HIFA logic was known to be fallible [3]. At the root of its fallibility is the biological phenomenon of 'recruitment'. At many times in the history of the biosphere, new function has been generated. Frequently, proteins required by new function do not emerge de novo. Rather, an existing protein performing a primitive function suffers gene duplication, and one of the duplicates undergoes an episode of rapid evolution where amino acid substitutions are introduced to tweak the behavior of the protein to perform new functions.

Recruitment is the rule, not the exception, in metazoan biology. For example, kinases that phosphorylate tyrosine on proteins, protein phosphatases that remove phosphate from proteins, and the src homology 2 (SH2) domains that bind phosphotyrosine form three families of proteins, recognizable as homologs, all catalyzing analogous reactions at the level of organic chemistry. They perform quite different functions in the eyes of the biologist. And recruitment is widespread even in microorganisms. For example, eubacterial adenylosuccinate lyase, aspartase, and fumarase are homologs recognizable by sequence analysis; the first is involved in nucleic acid biosynthesis, the second in amino acid metabolism, and the third in the citric acid cycle. Annotating using the HIFA paradigm would misrepresent function at a very fundamental level.

This review outlines the tools that we have implemented at the University of Florida and EraGen Biosciences to address these problems. The tools exploit the fact that biomolecular sequence data reflect four billion years of biological evolution on Earth. It is possible to construct bioinformatics tools that reflect this fact. While the details of the historical past cannot be known with certainty, it turns out that tools that mimic the historical past provide powerful approaches to some of the most difficult tasks in genomics.

2. The Master Catalog™

We began working on the data management problem in genetics in 1986, when it became inevitable that genomic projects would sooner or later create this problem. First alone and later in collaboration with Professor Gaston Gonnet at the Swiss Federal Institute of Technology, we assembled what was known about sequence and behavior in biological macromolecules (nucleic acids and proteins) in context of then available sequence data [2–4, 6–8]. The goal of the first phase of our bioinformatics work was to learn what we could about molecular structure and behavior on one hand, and biological adaptation and natural selection on the other.

At the same time, the programming language DARWIN (data analysis and retrieval with indexed nucleic acid-peptide sequences) was developed as a computational workbench for manipulating and analyzing the genomic sequence data. Much of the work discussed below depends on DARWIN as a bioinformatic tool, and many of the key features of DARWIN have been made available to the public via the Web (cbrg.inf.ethz.ch).

The power of the DARWIN tool comes from several techniques drawn from computer science. One of these is a 'patricia tree' data structure, a structure that indexes sequence data. The indexing makes possible not only the easy retrieval of genetic data (much as an index allows the retrieval of data from the Oxford English Dictionary), but also the comparison of genetic data within the database. The tools are so powerful that DARWIN was able to complete the first 'exhaustive matching' of a modern sequence database, the systematic comparison of every subsequence in the database with every other [14]. Even with the 'small' (by modern standards) genetic database in 1992, the exhaustive matching would have been prohibitively expensive if undertaken without indexing.

The exhaustive matching was not simply a librarian exercise. It produced 1.7 million pairs of aligned homologous protein sequences. For the most part, these were sequences for functional proteins that contribute to the survival and reproduction of their host organisms. Thus, each pair provided some empirical information describing how the sequences of a particular biological macromolecule had changed during divergent evolution under functional constraints. Taking 1.7 million of these pairs together, the exhaustive matching proved to be a rich source of information concerning the structure and function of genetic molecules.

One conclusion from the exhaustive matching and its various updates was an estimate that when all of the genomes of all of the organisms on planet Earth are completed, all protein sequences will be easily recognizable as members of one of ca. 10 000 nuclear families, protein sequence modules 50–500 amino acids long that are related by common ancestry. This conclusion reflects the well-known fact that all organisms on the planet are descendants of a single ancestor. In the course of producing the diversity of organisms now on Earth, divergent evolution also produced the diversity of molecular genetic sequences within nuclear families. And, as Linus Pauling and Emil Zuckerkandl noted some time ago, the sequence data contained within them the evolutionary history of each of these nuclear families of proteins [20].

The evolutionary histories of nuclear families are represented by three elements illustrated in *figure 1*: a) an evolutionary tree (which shows the pedigree of each member of the family); b) a multiple sequence alignment (which shows the genetic relationship of every amino acid in the protein sequence); and c) 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. Each element is obtained entirely automatically by DARWIN from sequence data alone.

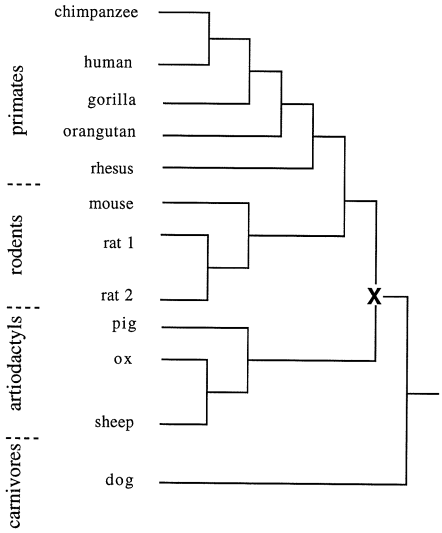
Because genetic sequences have arisen by divergent evolution, a natural organization is a more efficient way to organize a genomic database than a relational database (such as GenBank). In its natural organization, all genetic information on the planet is represented by the 10 000 founder sequences (approximately 10^7 bits). To search the database, one need search the founder sequences only. Some 90% of these are already reconstructable, meaning that the search of the naturally organized genome database will not cost significantly more when all of the genomes of all organisms on Earth are sequenced than it does today. In contrast, relational databases must be searched one sequence at a time, meaning that the time for a search is growing exponentially with the size of the database.

3. Functional genomics. Behavior by homology

The first insights to emerge from a well-organized database concerned structural biology. It had long been known that proteins diverging from a common ancestor retain their core conformation (or 'fold') [11]. This implies that proteins within a nuclear family have the same fold. This empirical generalization, combined with chemical insights from the exhaustive matching, proved to be the starting point to a solution to one of the most confounding problems in structural biology: how can the fold of a protein be deduced from sequence data alone?

Virtually all tools for comparing the sequences of homologous proteins assume a model for divergent evolution that is stochastic in outcome. This model treats a protein sequence as a linear string of letters, one letter for each amino acid. According to the model, each letter in the string changes (the gene and its corresponding protein mutates) at a rate that is independent of its position. According to the stochastic model, future and past mutations are independent. Mutations at one position are independent of mutations elsewhere.

(a) The tree for leptin, the obesity gene protein.



(b) Part of the alignment and the reconstructed ancestral leptin sequence.

```
         080       090       100       110       120
     .     |     .     |     .     |     .     |     .     |
RNVIQISNDLENLRDLLHVLAFSKSCHLPWASGLETLDSLGGVLEASGYS  human
RNMIQISNDLENLRDLLHVLAFSKSCHLPWASGLETLDSLGGVLEASGYS  chimp
RNMIQISNDLENLRDLLHVLAFSKSCHLPWASGLETLDSLGGVLEASGYS  gorilla
RNVIQISNDLENLRDLLHVLAFSKSCHLPWASGLETLDRLGGVLEASGYS  orangutan
RNVIQISNDLENLRDLLHLLAFSKSCHLPLASGLETLESLGDVLEASLYS  rhesus

QNVLQIAHDLENLRDLLHLLAFSKSCSLPQTRGLQKPESLDGVLEASLYS  rat
QNVLQIAHDLENLRDLLHLLAFSKSCSLPQTRGLQKPESLDGVLEASLYS  rat
QNVLQIANDLENLRDLLHLLAFSKSCSLPQTSGLQKPESLDGVLEASLYS  mouse

RNVIQISNDLENLRDLLHLLASSKSCPLPQARGLETLESLGGVLEASLYS  ancestor X

RNVIQISNDLENLRDLLHLLASSKSCPLPQARALETLESLGGVLEASLYS  pig
RNVIQISNDLENLRDLLHLLAASKSCPLPQVRALESLESLGVVLEASLYS  sheep
RNVVQISNDLENLRDLLHLLAASKSCPLPQVRALESLESLGVVLEASLYS  ox

RNVVQISNDLENLRDLLHLLASSKSCPLPRARGLETFESLGGVLEASLYS  dog
```

Figure 1. The evolutionary history of the leptins, proteins from the 'obesity gene' identified by genetics experiments in mice. Homologs are found in other mammals (including human). (a) An evolutionary tree showing the pedigree of each leptin family member. (b) A part of the multiple alignment, showing the genetic relationship of amino acids in the protein sequence. The reconstructed ancestral sequence from the (now extinct) ancestor of humans, rodents, and ruminants (marked 'X') is shown in the alignment.

As all structural biologists know, such a model is at best an approximation for the reality of protein evolution. In reality, proteins are not linear strings of letters. Rather, they are organic molecules that fold in three dimensions. In the folded form, some positions in a protein sequence are more easily mutatable (without destroying function) than others. Amino acids distant in the sequence but close in the fold frequently undergo correlated mutation. Thus, real proteins divergently evolving under functional constraints behave differently than expected based on the stochastic model.

The difference between the reality of divergent evolution of proteins that fold and expectation based on the stochastic model proves to be important. By comparing the patterns of substitution within a set of folded proteins undergoing divergent evolution with expectations for those patterns based on the stochastic model, one can extract information about the fold. This makes the nuclear family more than a database organizational feature. Because the nuclear family holds a history of the pattern of divergent evolution under functional constraints in the protein, it holds information about the fold of the protein. From the sequences of proteins in the nuclear family alone, one can decide which amino acids lie on the surface of the folded structure, which lie inside, and which lie near the active site. Elements of secondary structure, the helices, strands, and loops can be identified. A model of tertiary structure can be built as well, all from the evolutionary history embodied in the nuclear family.

When we introduced evolution-based structure prediction (ESP) methods in 1990 for predicting the conformation of protein families from a set of homologous protein sequences, we recognized that the experimental genetics community was skeptical of all methods to predict protein folds from sequence data. An expedient was therefore adopted, whereby fold predictions were made and published before an experimental structure was known. These were termed bona fide predictions. Protein kinase was the first bona fide prediction made using ESP methods [5].

Published in 1990, the prediction for protein kinase was evaluated using a crystal structure of a member of the kinase nuclear family published 1 year later. The crystallographers noted that the prediction was "remarkably accurate" [16], because it identified correctly the secondary structural elements and the antiparallel sheet at the core of the first domain. Thornton and her colleagues reviewed both the prediction and the experimental structure, and noted that the prediction was "much better than expected from standard methods" [26]. Lesk and Boswell foresaw that this prediction would come to be regarded as a "major breakthrough" [17].

ESP methods have now accumulated a track record of over two dozen predictions made and announced before an experimental structure was known [9]. The antiparallel azurin-type fold of synaptotagmin, the eight-fold alpha-beta barrel of phosphogalactosidase, the helix bundle of the obesity gene protein (leptin), and the core fold of heat shock protein 90 (Hsp90) are four examples that illustrate the variety of predictions that have been made using ESP methods. A set of predictions has focused on proteins involved in signal transduction in higher organisms: the src homology 2 (SH2) and src homology 3 (SH3) domains, protein tyrosine phosphatase, protein serine phosphatase, and the pleckstrin homology domain are examples of these.

The success of ESP predictions made and announced before an experimental structure was known had another impact: it caused a re-emergence of bona fide predictions as a tool for testing structure prediction methods. Although this tool was controversial in 1990, the structural biochemistry community now comes together every second year in Asilomar to test their skills in prediction contests against experimental structures that crystallographers keep secret until after predictions are announced. In 1996, these contests had attracted the attention of both the scientific and lay press.

This work with evolutionary genetic databases has generated a solution to one of the oldest and most frustrating problems in structural biology, predicting protein folds. The resulting optimism in the structural biology community today is a strong contrast to the pessimism that existed just 5 years ago, when no less an authority than the editors of *Trends in Biochemistry* declared protein fold prediction "more a matter for soothsayers than scientists" [15]. If a database organized using evolutionary strategies and analyzed with organic chemistry can change the outlook in a field in such a short time, imagine what other problems might be solved should our organic, organizational, and evolutionary tools become more sophisticated.

4. Detecting long distance homologs by structure prediction

Prediction contests are marvelous sport. A predicted model of a protein structure becomes interesting to a biologist, however, when it is used to solve a biological problem. The power of ESP methods makes them useful to the geneticist for precisely this reason. Let us begin by showing how predicted structures can be used to solve a specific problem in genomics: how to identify distantly related genes.

Let us suspect that two nuclear families of proteins (and their genes) are in fact homologous, descendent from a still more ancient ancestor and belonging to one large extended family. The suspicion might be based on statistically marginal sequence similarity between members of the family; perhaps the two protein families share a sequence 'motif', a few amino acid residues presumed to be important for folding or function.

Homologous proteins have analogous folds. Conversely, nonanalogous folds in two protein families indicate that the two families are not homologous. Thus, if two protein families are predicted to have the same fold, they are more likely to share common ancestry. If two protein families are predicted to have different folds, the two families do not share common ancestry.

To illustrate how this works, consider again the protein kinase family. Protein kinases all contain the sequence motif Gly-Xxx-Gly-Xxx-Xxx-Gly (where Xxx is any amino acid). A similar motif is found in adenylate kinase, whose crystal structure was known in 1990. Therefore, many researchers proposed that protein kinases were homologs of adenylate kinase, and inferred from this proposal that protein kinases adopt the same fold as adenylate kinase. Several groups built models for the conformation of protein kinase based on this inference [22, 24, 25, 27]. All of these models turned out to be wrong.

An ESP model for the protein kinase nuclear family [5] predicted that the Gly-Xxx-Gly-Xxx-Xxx-Gly motif was flanked by two beta strands embedded in an antiparallel beta sheet at the core of the protein fold. In adenylate kinase, this motif is flanked by a strand and a helix, and is embedded in a parallel beta sheet. Thus, the fold predicted for protein kinase was not analogous to the fold known for adenylate kinase, and this led to the prediction that protein kinases were not homologous to adenylate kinases. This prediction was shown to be correct. For the first time, a predicted structure had been used to infer the absence of homology between two families catalyzing analogous chemical reactions.

Predicted conformations can be used to confirm the presence of long distance homology as well. For example, ribonucleotide reductases from different organisms use different cofactors, including vitamin B12, iron, and manganese. The reductases share no sequence similarities that show that they are related by common ancestry [18]. In an exercise in 'postgenomic' biochemistry, we isolated, cloned, sequenced, and expressed a ribonucleotide reductase from the archaebacterium *Thermoplasma acidophilum* [23]. Prediction tools applied to the protein family showed that all of the ribonucleotide reductases were descendants of a single ancestral sequence present in the most recent common ancestor of archaebacteria, eubacteria, and eukaryotes. This conclusion had implications not only for the chemical mechanisms by which ribonucleotide reductases biosynthesize genetic molecules, but also for how genetic molecules evolved on the planet.

Successes such as these underlie the current and future work in our group to organize a 'Master Catalog' covering all genetic data from all organisms, including human. The Master Catalog contains the evolutionary histories of ca. 10 000 nuclear families described using a multiple sequence alignment, an evolutionary tree, and reconstructed ancestral sequences. To this is added a predicted secondary structural model for the protein family (or an experimental structure, if available). Using a combination of reconstructed ancestral sequences and predicted conformations, bridges are built between the nuclear families, joining them to give extended families and superfamilies. As these

bridges are built, the number of 'genesis events', instances when an independent fold of a protein is presumed to have emerged, required to explain existing genetic diversity on Earth decreases.

Geneticists do not often speak of genesis. But it is clear that genetic information, organized with sophisticated bioinformatics tools, analyzed with organic chemical insight, and supplemented by evolutionary models that predict folds, has relevance to the problems of 'origins'. This, in turn, is one of the great intellectual problems in science. The Master Catalog distils the diversity found in the modern world of molecular biology into fewer than 1 000 events of biomolecular genesis. This is not yet 'origins'. But it is a step in this direction.

5. Deducing biological function from genetic sequence data

Assigning long distance homologs within genetic data is directly relevant to the evolution of genetic systems on the planet. But such assignments gain additional value if they can be used to attribute biological function to genetic elements known by their DNA and protein sequences. A plausible (if tenuous) logic suggests that this might be possible. If homologous proteins have analogous conformations, perhaps they generally have analogous behaviors and analogous functions as well.

The possibility that homologous proteins have analogous functions makes the detection of homology through structure prediction central to the emerging discipline of functional genomics. The heat shock protein 90 (Hsp90) family provides an illustration. As with many sequences in modern genetics, Hsp90 was known only as an open reading frame expressed under specific conditions (increased temperature). This provides no particular insight into fold, catalytic behavior, or biological function. Indeed, in 1996, the biological function of Hsp90 was disputed. Earlier experiments suggested that Hsp90 had ATPase activity, but these were later ascribed to contaminating kinases.

To resolve this problem, we made an ESP prediction for the structure of Hsp90 [13]. The predicted secondary structural elements were assembled to yield a model for the tertiary fold. From this model, the predicted fold of Hsp90 was recognized to resemble the fold found in the ATP-binding fragment of DNA gyrase B. From this observation, Hsp90 was predicted to be a distant homolog of gyrase and to bind ATP. After this prediction was announced publicly, an experimental structure of Hsp90 bound to ATP was solved. The experimental structure showed not only that the predicted model for the conformation of the protein was largely correct, but that the predictions concerning function were correct as well. In the words of the crystallographers who solved the structure [21]:

"The tertiary fold of Hsp90 N-domain has a remarkable and totally unexpected similarity to the N-terminal ATP-binding fragment of... DNA gyrase B protein. This similarity was not initially recognized by the authors of either the human or yeast structures but was determined within the CASP2 structure prediction competition. The observation of specific ADP/ATP binding to Hsp90 completely contradicts... [earlier] and widely accepted biochemical analysis."

Using predicted structural models to detect or deny long distance homology and suggest function are especially important for human genetics. Manipulative experimentation that has been so powerful in organisms from *Escherichia coli* to mouse are denied to those who study human genetics. No matter how technology improves, humans will never be experimental animals. Key, therefore, to the future of human genetics are tools that permit extrapolation from nonhuman models to address problems in human genetics, and these tools provide the first step. Naturally organized databases yielding ESP fold predictions provide the human geneticist some of these tools.

The leptin family of proteins (*figure 1*) arising from the 'obesity gene' in mouse provide an interesting illustration of this. Mutation in leptin in mice is correlated with obesity. An ESP structure prediction suggested that the protein

would have a fold similar to that found in a family of cytokines [10]. The predicted similarity in the folds of leptin and the cytokines implied in turn that leptin would have a receptor that would be homologous to the receptors known for the cytokines. The leptin receptor was subsequently identified and found indeed to belong to the cytokine family of receptors [1].

Proceeding from genetic sequence data to biological function promises to be one of the most significant ways in which genome sequencing projects will be put to work. The applications are first and foremost technological: the detection of new mechanisms for metabolic regulation, the identification of pharmaceutical targets, and the design of therapeutic agents are the most obvious. The reader should keep in mind that these 'functional bioinformatics' tools are still in their nascent stages. If problems such as these can be solved using such simple tools, we can expect many more complex biological problems to be solved once the tools become sophisticated.

6. When homology does not imply analogous biological function

As powerful as homology is as a tool for assigning function to open reading frames from genome projects, the underlying logic remains problematic. Homologous proteins need not have either analogous behaviors or analogous functions. As has been reviewed in detail elsewhere [3], old protein folds are frequently recruited by evolutionary processes to perform new functions. For example, fumarase (functioning in the citric acid cycle), adenylosuccinate lyase (functioning in nucleotide biosynthesis) and aspartate ammonia lyase (functioning in amino acid metabolism) are all identified (correctly) as homologs by a BLAST search. Yet their behaviors are analogous only at the level of organic reaction mechanism, and there only at the most abstract level. Their functions are quite different.

In proteins involved in 'advanced' functions (in development, for example) in more complex organisms, difficulties with the 'homology-implies-analogous-structure/behavior/function' paradigm become confounding. For example, protein serine kinases and protein tyrosine kinases are clearly homologous, the latter having been recruited from the former ca. 600 million years ago. The chemist would say that both classes of enzyme operate via analogous reaction mechanisms, differing only in the source of the oxygen nucleophile in the phosphoryl transfer reaction. The biologist would note, however, that the physiological functions of the two classes of proteins are greatly different. For any biomedical application, the biologist would be correct. The physiologically relevant differences in behavior, central to the understanding of biological function (phosphorylation on tyrosine versus phosphorylation on serine) cannot be inferred for one family from the other using the conventional logic.

The deeper the chemistry of developmental biology is probed in metazoa (multicellular animals), the more apparent it becomes that function in the Darwinian sense can change with very little change in sequence [3, 4]. For example, SH2 domains quite similar in sequence all bind peptide sequences containing phosphotyrosine residues. The binding specificities of the different SH2 domains are different, however, for the surrounding peptide sequences. It is these specificities that determine which protein binds to which individual SH2 domain, and from there, the physiological function. Thus, any statement of function for any particular SH2 domain must at least identify its phosphotyrosine-containing partner. To assign function at this level, the conventional evolutionary logic has little to say.

How can recruitment be detected within a nuclear family? One approach returns to the Master Catalog and exploits the degeneracy of the genetic code. More than one triplet codon encodes the same amino acid. Therefore, a mutation in a gene can be either silent (not changing the encoded amino acid) or expressed (changing the encoded amino acid). Especially in multicellular organisms, and most particularly in multicellular animals (metazoa), silent changes are not under (large) selective pressure.

In contrast, expressed changes can change the properties of the protein. This frequently places these changes under selective pressure.

Consider in a 'thought experiment' three cases of divergent evolution of a hypothetical protein whose sequence has been optimized to perform a specific biological function. In the first, the function of the protein remains constant during the episode of evolution that follows. Changes in the gene that change the sequence of the encoded protein (expressed changes) will diminish the survival value of the protein and will be removed by natural selection. Silent changes will not be removed by natural selection. Thus, the ratio of expressed to silent changes will be low during an episode of evolution where the ancestral and derived proteins share a common function.

In the second, the protein acquires a new (derived) function during the episode of evolution. This, almost by definition, requires a change in the behavior of the protein, which requires a change in its amino acid sequence. Expressed changes will have a chance of improving the behavior of the protein vis à vis its new biological function; these will be selected for. The ratio of expressed to silent substitutions at the DNA level will be high.

In the third case, the gene becomes a pseudogene, and neutrally drifts without any function. In this case, the expressed/silent ratio will reflect random introduction of point mutations into a genetic element that was formerly encoding, but no longer. Given the genetic code and a typical distribution of amino acid codons within the gene, a ratio of expressed to silent changes will be approximately 3:1.

Within a nuclear family, the reconstructed ancestral sequences (both DNA and proteins) at branch points in the tree permit one to assign expressed and silent substitutions to different branches of the tree. This approach can be illustrated with the protein leptin (*figure 1*). In mouse, leptin is known from genetics to be associated with obesity. Accordingly, the protein has attracted interest in the pharmaceutical industry, based on the assumption that the leptin homolog in humans has an analogous

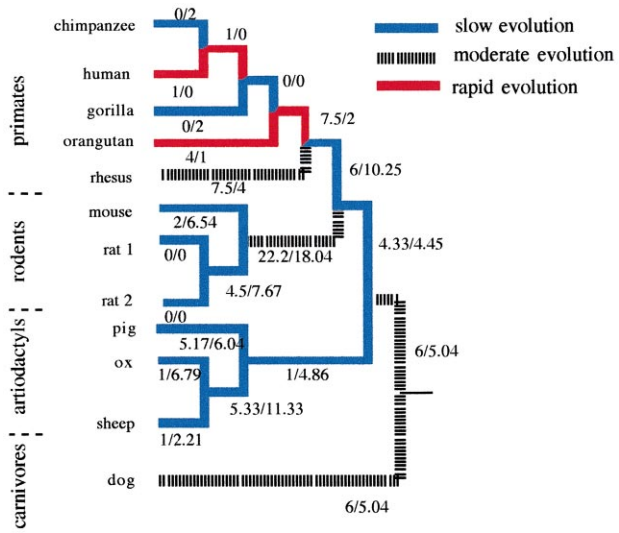


Figure 2. The evolutionary tree for leptin showing episodes of rapid (red) and slow (blue) evolution as the primate proteins diverged from the ancestor. Numbers on the tree branches indicate the number of expressed and silent changes that occurred. The analysis implies that the physiological function of leptin in primates is different from the physiological function of leptin in (for example) rodents.

function. Many pharmaceutical firms have begun to seek leptin analogs as drugs for combating human obesity.

Figure 2 reproduces the evolutionary history of the leptin protein, showing episodes of low high expressed/silent ratios indicative of change in function in red, and low high expressed/silent ratios indicative of conserved function in blue. The branches on the evolutionary tree leading to the primate leptins from their ancestors at the time that rodents and primates diverged have an extremely high ratio of expressed to silent changes. This analysis suggests that the biological function of leptins has changed in the primates relative to the function of the leptin in the common ancestor of primates and rodents.

This conclusion had practical implications for pharmaceutical companies interested in leptins as pharmaceutical targets. At the very least, it suggested that the mouse is not a good pharmacological model for analogs of leptin that might be developed to combat obesity in humans.

In future work in these laboratories, expressed/silent ratios will be placed on individual branches of the trees associated with the nuclear families in the Master Catalog. Thus, in addition to a database organization tool, the Master Catalog will contain a molecular record of the emergence of new function within each of the nuclear protein families. Again, the impact will be largely technological. The use of genetic sequence data not only to assign function, but also to identify the change in function should add power to functional bioinformatics.

7. Conclusion

The Master Catalog organizes the sequence database using a natural architecture based on the evolutionary history of individual protein module families, each with a reconstructed evolutionary history. The 20 000 histories of nonsynonymous and synonymous substitution are reconstructed comprehensively to detect episodes throughout the database where recruitment may have occurred. Functional genomics examples are given using these histories. From these emerge the concept of a 'natural annotation', one that reflects the history of a protein module in the biosphere.

Acknowledgments

This work was supported in part by NIH grants MH 55479 and HG01729.

References

[1] Baumann H., Morella K.K., White D.W., Dembski M., Bailon P.S., Kim H., Lai C.F., Tartaglia L.A., The full-length leptin receptor has signaling capabilities of interleukin 6-type cytokine receptors, Proc. Natl. Acad. Sci. USA 93 (1996) 8374–8378.

[2] Benner S.A., Enzyme kinetics and molecular evolution, Chem. Rev. 89 (1989) 789–806.

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

[4] Benner S.A., Ellington A.D., Evolution and structural theory. The frontier between chemistry and biochemistry, Bioorg. Chem. Frontiers 1 (1990) 1–70.

[5] Benner S.A., Gerloff D.L., Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure. The catalytic domain of protein kinases, Adv. Enzyme Regul. 31 (1991) 121–181.

[6] Benner S.A., Allemann R.K., Ellington A.D., Ge L., Glasfeld A., Leanz G.F., Krauch T., Macpherson L.J., Moroney S.E., Piccirilli J.A., Weinhold E.G., Natural selection, protein engineering and the last riboorganism. Evolutionary model building in biochemistry, Cold Spring Harbor Symp. Quant. Biol. 52 (1987) 53–63.

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

[8] Benner S.A., Glasfeld A., Piccirilli J.A., Stereospecificity in enzymology. Its place in evolution, Top. Stereochem. 19 (1989) 127–207.

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

[10] Benner S.A., Trabesinger-Ruef N., Schreiber D.R., Exobiology and post-genomic science. Converting primary structure into physiological function, Adv. Enzyme Regul. 38 (1998) 155–180.

[11] Chothia C., Lesk A.M., The relation between the divergence of sequence and structure in proteins, EMBO J. 5 (1986) 823–826.

[12] Fetrow J.S., Skolnick J., Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T-1 ribonucleases, J. Mol. Biol. 281 (1998) 949–968.

[13] Gerloff D.L., Cohen F.E., Korostensky C., Turcotte M., Gonnet G.H., Benner S.A., A predicted consensus structure for the N-terminal fragment of the heat shock protein HSP90 family, Proteins Struct. Funct. Genet. 27 (1997) 450–458.

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

[15] Hunt T., Purton M., 200 issues of TIBS, Trends Biochem. Sci. 17 (1992) 273.

[16] Knighton D.R., Zheng J.H., Ten Eyck L.F., Ashford V.A., Xuong N.H., Taylor S.S., Sowadski J.M., Crystal structure of the catalytic subunit of cyclic adenosine-monophosphate dependent protein-kinase, Science 253 (1991) 407–14.

[17] Lesk A.M., Boswell D.R., Does protein structure determine amino acid sequence?, Bioessays 14 (1992) 407–410.

[18] Mao S.S., Holler T.P., Yu G.X., Bollinger Jr J.M., Booker S., Johnston M.I., Stubbe J., A model for the role of multiple cysteine residues involved in ribonucleotide reduction: Amazing and still confusing, Biochemistry 31 (1992) 9733–9743.

[19] Marcotte E.M., Pellegrini M., Ng H.L., Rice D.W., Yeates T.O., Eisenberg D., Detecting protein function and protein-protein interactions from genome sequences, Science 285 (1999) 751–753.

[20] Pauling L., Zuckerkandl E., Molecular paleontology, Acta Chem. Scand. 17 (1962) (Suppl. 1) S9–S16

[21] Prodromou C., Roe S.M., O'Brien R., Ladbury J.E., Piper P.W., Pearl L.H., Identification and structural characterization of the ATP/ADP binding site in the HSP90 molecular chaperone, Cell 90 (1997) 65–75.

[22] Shoji S., Parmelee D.C., Wade R.D., Kumar S., Ericsson L.H., Walsh K.A., Neurath H., Long H.L., Demaille J.G., Fischer E.H., Titani K., Complete amino acid sequence of the catalytic subunit of bovine cardiac muscle cyclic AMP-dependent protein kinase, Proc. Nat. Acad. Sci. USA 78 (1981) 848–851.

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

[24] Taylor W.R., Identification of protein sequence homology by consensus template alignment, J. Mol. Biol. 188 (1986) 233–258.

[25] Taylor W.R., Thornton J.M., Recognition of super-secondary structure in proteins, J. Mol. Biol. 173 (1984) 487–514.

[26] Thornton J.M., Flores T.P., Jones D.T., Swindells M.B., Protein structure. Prediction of progress at last, Nature 354 (1991) 105–106.

[27] Wierenga R.K., Terpstra P., Hol W.G.J., Prediction of the occurrance of the ADP-binding beta-alpha-beta fold in proteins using an amino acid sequence fingerprint, J. Mol. Biol. 187 (1986) 101–107.