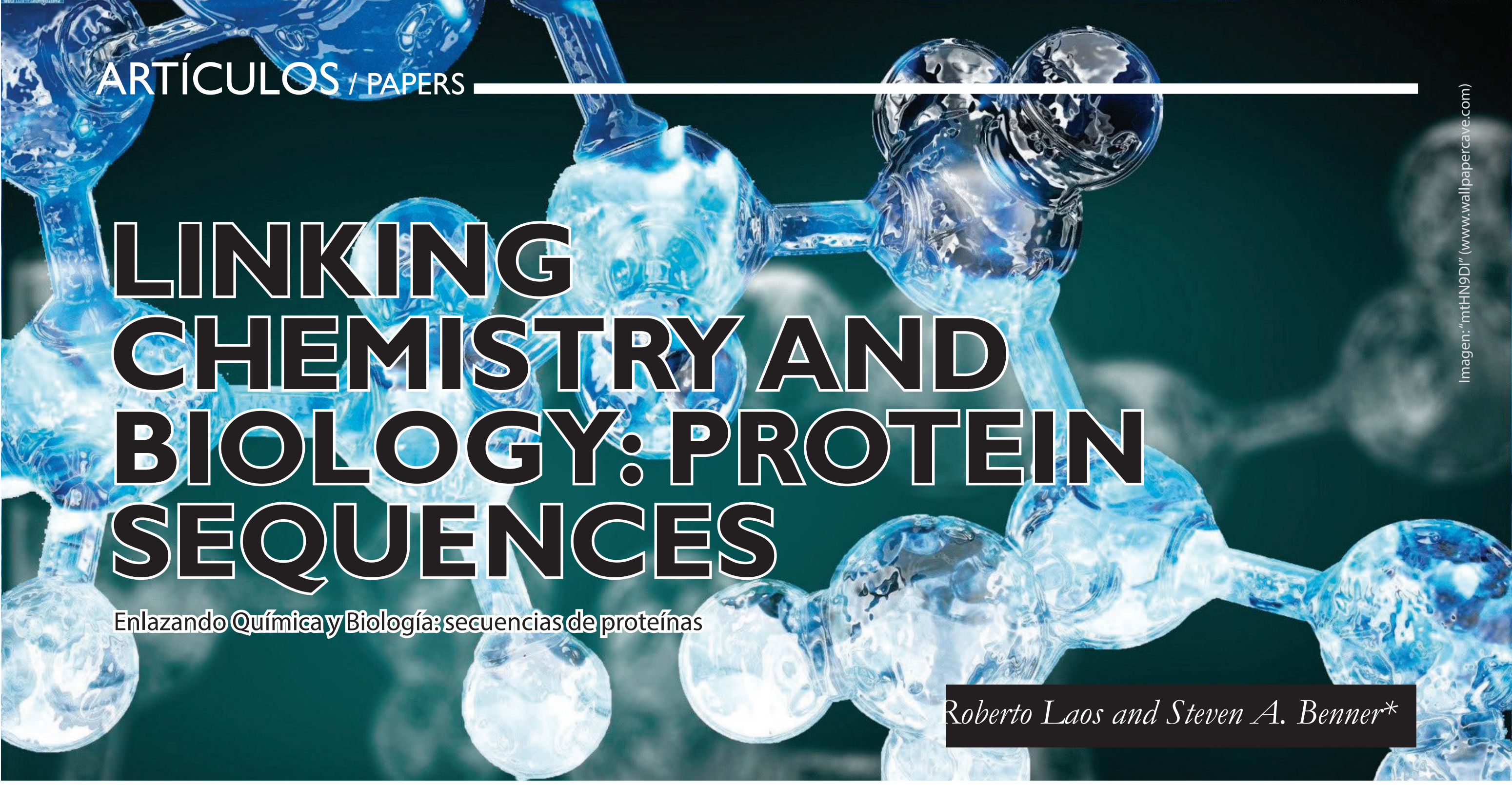

LINKING CHEMISTRY AND BIOLOGY: PROTEIN SEQUENCES

Enlazando Química y Biología: secuencias de proteínas

*Roberto Laos and Steven A. Benner**

En los últimos veinte años el número de genomas completos que han sido secuenciados y depositados en bancos de datos ha crecido dramáticamente. Esta abundancia de información de secuencias ha servido de base para la creación de una disciplina llamada paleogenética. En este artículo, sin ahondar en algoritmos complejos, presentamos algunos conceptos clave para comprender cómo las proteínas han evolucionado con el tiempo. Luego ilustraremos como la paleogenética es utilizada en biotecnología. Estos ejemplos resaltan la conexión entre la química y la biología, dos disciplinas que quizás veinte años atrás parecían ser mucho más distintas que lo que parecen ser hoy.

Palabras clave: Bioinformática, proteínas ancestrales, biología sintética, ingeniería de proteínas.

In the last twenty years, the number of complete genomes that have been sequenced and deposited in data banks has grown dramatically. This abundance in sequence information has supported the creation of the discipline known as paleogenetics. In this article, without going into complex algorithms, we present some key concepts for understanding how proteins have evolved in time. We then illustrate how paleogenetic analysis can be used in biotechnology. These examples highlight the connection between chemistry and biology, two disciplines that twenty years ago seemed to be more different than what they seem to be today.

Keywords: Bioinformatics, ancestral proteins, synthetic biology, protein engineering.

Although chemistry and biology have developed separately for a long while, advances in both disciplines are erasing the divisions between them and forcing chemists to learn about evolutionary biology and biologist to learn about synthetic chemistry. Not long ago, chemistry was a very different discipline than the one we know today, as remembered by Frank Westheimer[1], who is known as one of the first chemist that combined biology, organic chemistry and bioinformatics: when he received his PhD in 1935, instruments like UV, IR, NMR had not been developed; nor had TLC or column chromatography. A chemistry student in the present day can only imagine such a laboratory in a museum.

However, the rapid scientific advance made the situation change rapidly and in a few years NMR would be available. In fact, the first person to take a nuclear magnetic resonance of a protein in 1957[2] (Figure 1), the chemist Martin Saunders, is today the longest serving professor at Yale University. It seems that the jump from a museum-like laboratory to a modern laboratory happened in about 50 years.

DNA Sequencing and DNA synthesis

The genetic information, of all known life forms on Earth, flows between three biomolecules: DNA, RNA and proteins. DNA is the ultimate repository of genetic information (except for some viruses that use RNA for this purpose). The

*The authors (both PhD in Chemistry) are active researchers at the Foundation for Applied Molecular Evolution, FAME. (💻)
Contact: 13709 Progress Blvd. No. 7, Alachua, FL 32615, USA. (Email: rlaos@ffame.org)

[1] Westheimer, F. H., Musings. *Journal of Biological Chemistry* **2003**, *278* (14), 11729-11730. (💻)

[2] Saunders, M.; Wishnia, A.; Kirkwood, J. G., The nuclear magnetic resonance spectrum of ribonuclease. *Journal of the American Chemical Society* **1957**, *79* (12), 3289-3290. (💻)

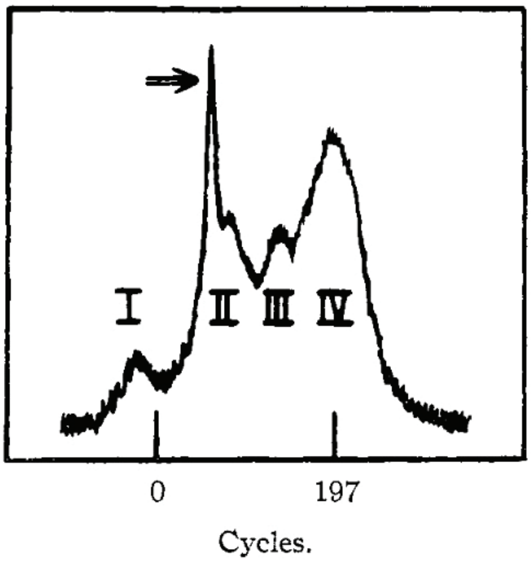

Figure 1. Nuclear magnetic spectrum of ribonuclease. Reprinted with permission from Martin Saunders, Arnold Wishnia, John G. Kirkwood The Nuclear Magnetic Resonance Spectrum of Ribonuclease. . *J. Am. Chem. Soc.*, **1957**, *79* (12), pp 3289–3290. Copyright (1957) American Chemical Society. Reproduced with permission

DNA is transcribed as RNA and this is translated into protein. Proteins serve many functions, some proteins provide structure for the organism, others catalyze reactions and are called enzymes.

The abundant genetic information we can easily access nowadays comes largely from obtaining the sequences of the DNA from different organisms. The sequences are entered in data banks and from them the encoded protein sequences can be inferred.

DNA molecules can also be synthesized by automatic procedures. The prices for both sequencing and synthesizing DNA have been reduced drastically. This accessibility to large amounts of data combined with affordable prices has made feasible many projects that previously would have been too costly to be considered.

Bioinformatics: How to study protein sequences using a computer

Proteins are chains of amino acids joint covalently. Twenty amino acids are encoded by genes, and each of them has a conventional abbreviation of three letters (Ala for alanine, Cys for cysteine, for example) and another abbreviation of one letter (A and C); bioinformaticians generally use the single letter abbreviation. These chains of letters can be compared with other chains of letters in homologous proteins that share a common ancestor.

To illustrate what we will see with protein sequences, let us consider an analogy for the reconstruction of words meaning “snow” in different Indoeuropean languages (Figure 2). The reconstruction shows similarities between words in languages that are related by having a common ancestor. This analysis can lead to the discovery of words used thousands of years ago in ancient civilizations.

What can we learn from this kind of reconstruction? The reconstruction of a language can tell us a lot about a civilization that existed thousands of years ago. Obviously they lived in a place that had snow. Further analysis of the descendent Indoeuropean languages shows that the ancestral language had no words for gold or silver. It did, however, have words for vehicles with wheels, which were used to transport grains. A similar reconstruction with protein sequences can show us how the animal’s diets have changed with time or how the capacity to metabolize certain substances has changed with the weather and with it the optimal temperature for the proteins activity.

Proteins, once placed in certain format to be used by a computer program, can be compared between them. Inevitably, you will find words or hidden messages in these 20-letter-strings, see **Figure 3**.

“Nothing in biology makes sense except in the light of evolution”[3]

The gene (DNA) that encodes for the sequence of a particular protein can experience random mutations, these sometimes, turns into changes in the sequence of the protein encoded. Some of these changes confer an advantage to the organism and are retained for the next generations. Other changes are lethal, these disappear quickly from the population. There are also neutral changes that do not confer neither an advantage nor a disadvantage and these can be retained or not.

For 3,500 million years, since life on Earth has left a trace, the genes of the last common ancestor, have been copied and passed from generation to generation, occasionally two organisms carrying a different mutation can be isolated from each other and this can lead to new species. In this way, new families and subfamilies of the same protein are created.

To fully understand biology we need to understand the evolution of the biomolecules of life. The evolutionary history combines three descriptions. First an evolutionary tree shows the relationship of proteins within a family. Secondly, an alignment of protein sequences shows how amino acids of a protein are related with the amino acids of another protein. Finally, the sequences of ancestral proteins can be inferred for ancestral proteins represented by points in the tree that join the branches. These ancestors are intermediates in the evolution of that protein family. The tree, the alignment and the reconstructed sequences are interdependent.

[3] Dobzhans.T, Nothing in biology makes sense except in light of evolution. *American Biology Teacher* **1973**, *35* (3), 125-129.

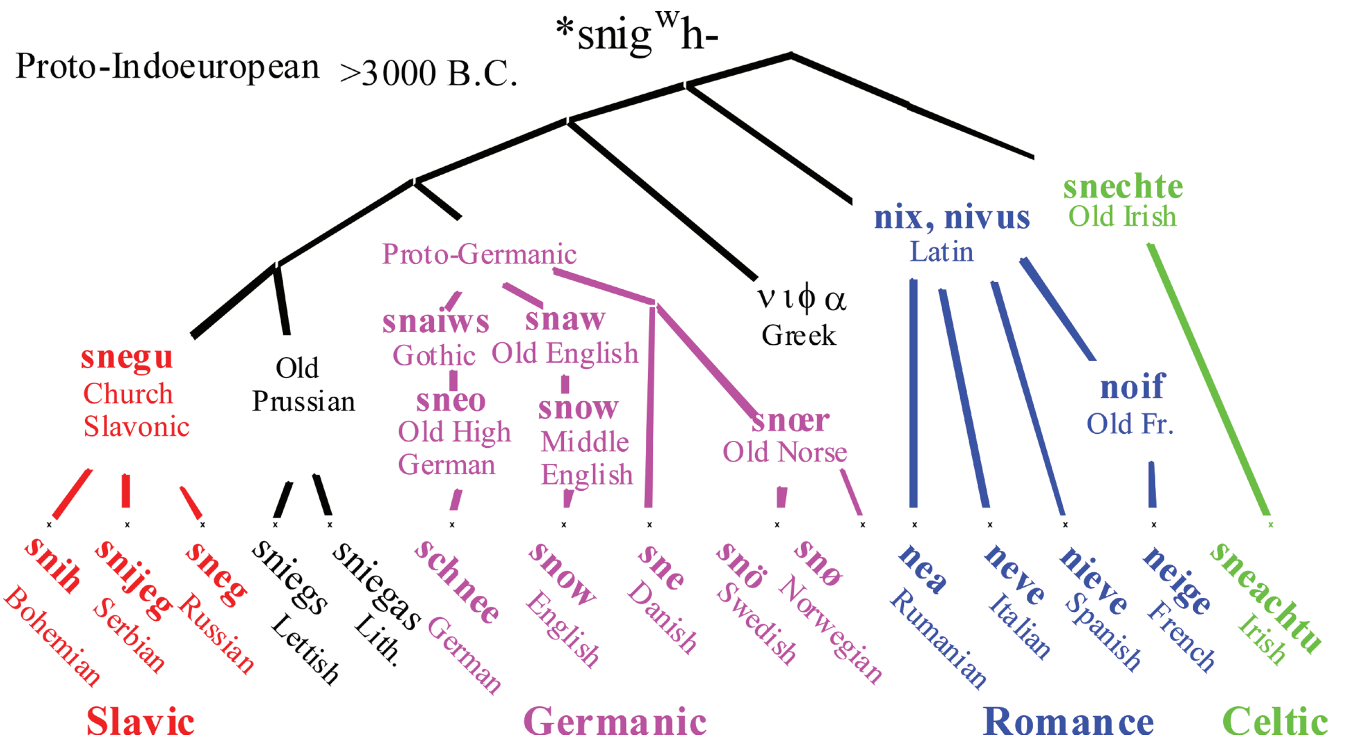

Figure 2. Most European languages descend from an ancestral language called Proto-Indoeuropean. All these languages have a word for snow, which are related by common ancestry. All the words for snow in modern languages are descendants from a word with the same meaning in the ancestral language.

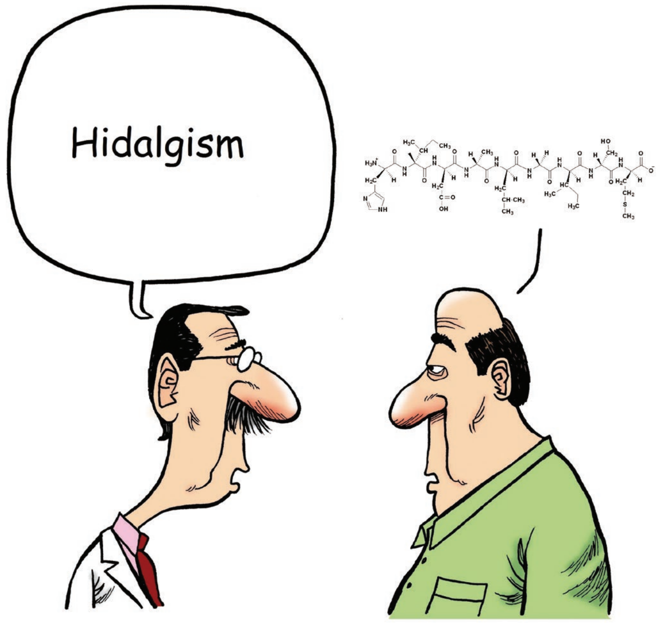

Figure 3. A chemist and a biologist talking about proteins. The sequence of amino acids shown: HIDALGISM[&] was found the year 1993 at the same time in the Unabridged Oxford English dictionary and in the protein data bank Swiss-Prot[#]. This was one of the first exercises in searching two databases. The word "hidalgism" (as well as the word "ensilsits") was, at the time, the longest English word found in that search. Illustration adapted from *Life, the Universe and the Scientific Method* by Steven A. Benner © Steven Benner, 2009.

Comparing protein sequences
Reconstructing proteins from extinct organisms

As in the analogy shown in **Figure 2**, knowing the sequences of homologous proteins, we can reconstruct the sequence of the ancestral protein from a common, now extinct, ancestor. Below are shown the partial sequences of the enzyme alcohol dehydrogenase from human and horse to illustrate this.*

```
    5    10    15
EGFDL LRSGK SIRTI LT   horse
EGFDL LHSGK SIRTI LM   human
```

It is evident that this is the best way of aligning these two proteins. Their similarities, as well as the differences, are evident. We can add more sequences of alcohol dehydrogenases from other species to produce a multiple alignment. Below is shown the multiple alignment that includes the sequence of a dehydrogenase of a plant (*Arabidopsis thaliana*).

```
    5    10    15
EGFDL LRSGK SIRTI LT   horse
EGFDL LHSGK SIRTI LM   human
KAFDY MLKGE SIRCI IT   plant
^ **^ ^  *^ *** * ^
```

* To reproduce this in your computer use Courier New font.

[&] Gonnet, G. H.; Benner, S. A.: *Nature* **1993**, *361* (6408), 121.

[#] Bairoch, A.; Boeckmann, B.: *Nucleic Acids Research* **1992**, *20*, 2019-2022.

Under each column, a mark indicates if the amino acids are identical (*) or different but conserved (^). A conserved change means that although the amino acids are not the same, they are structurally similar. For instance, leucine (L) and isoleucine (I) are structurally similar; aspartic acid (D) and glutamic acid (E) have both a carboxylic acid as a side chain (Figure 4).

The sequence of the protein that comes from the plant is more different from the two proteins from the animals than the two animal sequences themselves. It is still evident that these sequences are homologs. There are three possibilities for which two proteins are similar: (1) By accident: the similarities appear by chance. (2) For having a common ancestor: it is possible that both enzymes are descendants from the same ancestral enzyme and they are similar for the same reason that two siblings look alike. The differences between them represent divergent evolution. (3) Both enzymes catalyze the oxidation of alcohol to aldehyde and it is possible that a protein that catalyzes this reaction must have this particular sequence. This would imply that these sequences appeared by convergent evolution.

Let us see now an example for the reconstruction of an ancestral protein, defined as the sequence with the highest probability of being the original sequence from which the modern sequences have derived by the minimum number of independent amino acid replacements. The simplest case is when both ancestors have the same amino acid in certain position. In that case the ancestor is inferred to have had the same amino acid, as the amino acids in the two derived sequences can be obtained without any amino acid replacements at all.

```
    5    10    15
EGFDL LRSGK SIRTI LT  horse
EGFDL LHSGK SIRTI LM  human
EGFDL L?SGK SIRTI L?  ancestor
```

Let us next consider now the position 17 in the alignment shown above. The human enzyme has an M (methionine) at that position; the horse's enzyme has a T (threonine). If we only had this information, the ancestor would have the same possibility of having M or T in this position. Both reconstructions require a single mutation to explain the modern sequences (see **Figure 5A and 5B**).

If, however, we add the sequence from the plant (which has a T in this position) then the reconstruction with T gives a phylogenetic tree that requires a single mutation, while the reconstruction with M gives a phylogenetic tree that would require two mutations (see Figure 5C and 5D).

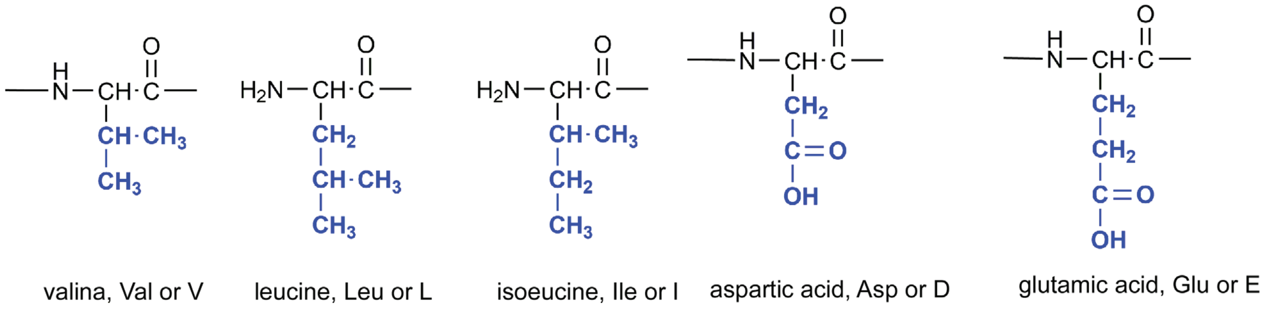

Figure 4. Amino acids with similar chemical side chains (shown in blue). Valine, leucine and isoleucine have a non-polar side chain. Aspartic acid and glutamic acid have both a hydrophilic carboxylic acid side chain. When an amino acid is replaced by another of similar structure the change is considered to be conservative.

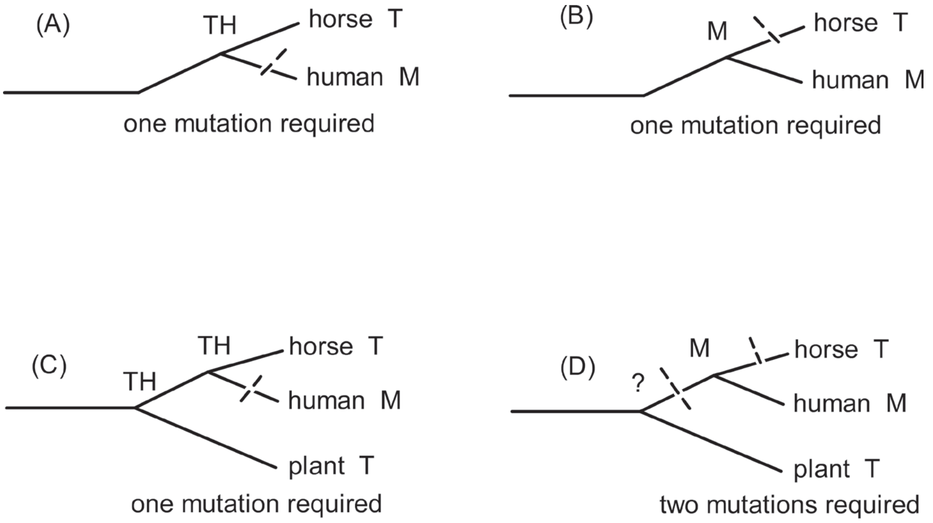

Figure 5. Reconstruction of ancestral proteins in the position 17 of the alignment shown in the text. (A) and (B) show two possible reconstructions for the ancestor of horse and human in position 17. Both reconstructions have the same probabilty (the ancestor could have a T or an M in this position), both reconstructions require a single mutation indicated by the dashed line. (C) The ambiguity for the reconstruction disappears when the sequence corresponding to the plant is introduced. The common ancestor for humans and horse would be T since this reconstruction requires a single mutation. (C) and (D) show other two possible reconstructions. The reconstruction for the common ancestor for humans and horse with an M requires two mutations indicated by the dashed lines. The reconstruction shown in (C) is better than the one shown in (D).

We might extend this example by introducing the sequence corresponding to the dehydrogenase from yeast then the multiple alignment would be:

```
    5    10    15
EGFDL LRSGK SIRTI LT  horse
EGFDL LHSGK SIRTI LM  human
KAFDY MLKGE SIRCI IT  plant
EIYEK MEKGQIVGRYV VD  yeast
```

Next, we can try to reconstruct the ancestor on position 8. We have two possibilities and one is more probable than the other (**Figure 6**).

The reconstruction of proteins from extinct organisms has created a discipline called paleogenetics. The Benner research group has done several studies in which sequences of proteins from extinct organism have been reconstructed. Once the desired sequences are reconstructed (as proteins) then a sequence of DNA can be produced that codes for the protein of interest. The DNA sequence can be synthesized using automated DNA synthesis, inserted in a vector and expressed in a host, typically *E. coli*. The purified protein can be manipulated and studied in the laboratory[4].

We have seen in a simplified way how multiple sequence alignments (MSAs) can be used to infer more information about the sequences of ancestral, now extinct, proteins. By examining more than just a small fragment of the protein, we can decide if one alignment is better than another, or one tree is better than another. For most protein families, MSAs can have the sequences of hundreds of proteins. This is when computers become useful, where the MSA, the tree, and the best ancestral sequences must be inferred by algorithms implemented by computer programs. These programs assign scores to MSAs and to the phylogenetic trees. The pioneer in scoring protein alignments is Margaret Dayhoff.[5] The reader interested in more detail in this field should consult her seminal work. Another good starting point is Russell Doolittle's book "*Of URFS and ORFS*".

Patterns of evolution

Phylogenetic trees and their associated alignments can show patterns of interest. Some especially interesting patterns are shown in Figure 7.

The pattern known as *heterotachy* (from Greek, *heteros*: different and *tachy*: speed) has been extremely useful in our research**. By introducing amino acid changes in sites that display heterotachy we have been able to produce variants of a DNA polymerase that accept new "alphabet letters" of an Artificially Expanded Genetic Alphabet (AEGIS)[6], developed in our laboratories. The advantage of introducing changes in a site that displays heterotachy is that nature has already test these sites. These sites are relevant enough to influence the activity of the enzyme so we see conservation in a sub-family but can be changed without losing the activity as we see variation in another sub-family.

Protein engineers (those tasked to modify an enzyme so that it has a new or improved activity) are often faced with a monu-

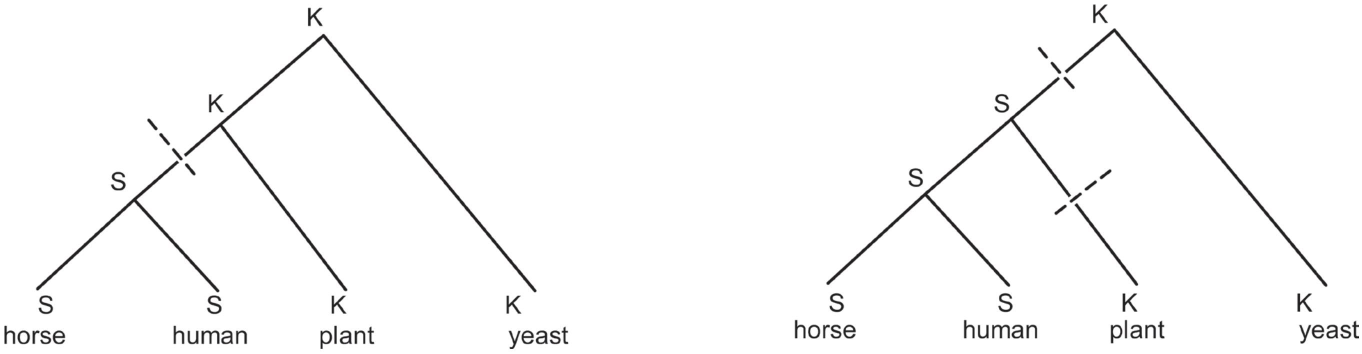

Figure 6. Two phylogenetic trees shown two possible reconstructions. The tree at the right requires two mutations while the one on the left requires a single mutation (indicated by the dashed lines). The tree on the left is a better reconstruction.

[4] (a) Gaucher, E. A.; Thomson, J. M.; Burgan, M. F.; Benner, S. A., Experimental paleogenomics as a tool to analyze protein function and predict environmental temperature during early life. *Astrobiology* **2002**, *2* (4), 501; (b) Sassi, S. O.; Benner, S. A., The resurrection of ribonucleases from mammals: from ecology to medicine. 2007; p 208-224; (c) Gaucher, E. A.; Govindarajan, S.; Ganesh, O. K., Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* **2008**, *451* (7179), 704-U2; (d) Carrigan, M. A.; Uryasev, O.; Frye, C. B.; Eckman, B. L.; Myers, C. R.; Hurley, T. D.; Benner, S. A., Hominids adapted to metabolize ethanol long before human-directed fermentation. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (2), 458-463.

[5] Dayhoff, M. O., Origin and evolution of protein superfamilies. *Federation Proceedings* **1976**, *35* (10), 2132-2138.

[6] Geyer, C. R.; Battersby, T. R.; Benner, S. A., Nucleobase pairing in Watson-Crick-like genetic expanded information systems. *Structure* **2003**, *11* (12), 1485-1498.

** Laos *et al*., **2016**, submmitted.

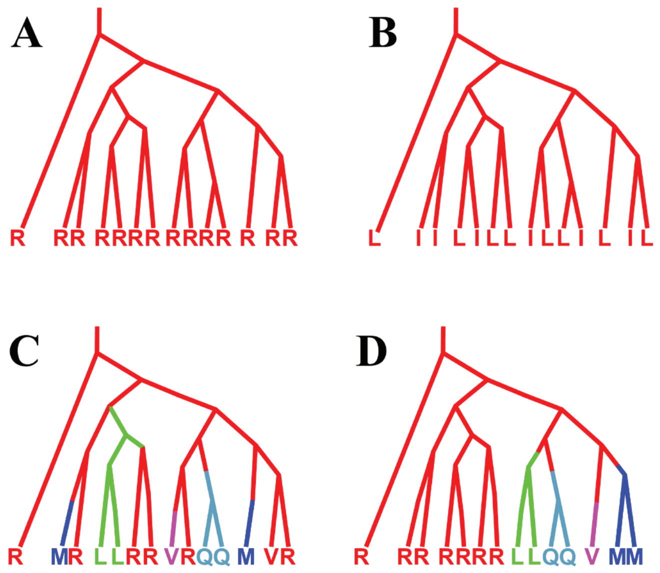

Figure 7. Different patterns of evolution. This figure shows four phylogenetic trees, with the letters at the bottom indicating the amino acid of each protein at a specific position. (**A**) The amino acid is conserved, in this case an arginine (R) is seen at this position in all the species present on the tree. Ancestors that tried a different amino acid in this position are dead. (**B**) Similar to A the site is conserved, it has either a leucine (L) or an isoleucine (I) which are very similar. (**C**) The site is highly variable, this site was probably irrelevant since it tolerates a wide range of amino acids. (**D**) One sub-family has the site highly conserved while another sub-family has high variability on this site. This is consequence of the two subfamilies being exposed to different evolutionary pressures. This pattern is called *heterotachy*.

mental task, given the immensity of the sequence space[7]. With 20 amino acids available for each position, there are 20100 possibilities to modify a protein of 100 amino acids. This number is colossal, you cannot test each possible enzyme variant. The evolutionary guidance can help the biotechnologist to overcome the "numbers problem" associated with this huge sequence space.

Outlook

The differences between chemistry and biology come from the time when chemists synthetized small biomolecules to confirm their structure while biologist studied biological systems by removing parts of the system and by perturbing the system infer some understanding of it. Today, synthetic biology, a good example of the merge of chemistry and biology, has a higher bar which is to include and/or replace natural biomolecules by artificial analogs. The recent interest on synthetic biology can be illustrated by the American Chemical Society starting a journal dedicated to synthetic biology in 2012.

In this article we intend to introduce young chemistry students to see biology from the point of view of evolutionary biology. The take home message to the chemistry student is "*nothing in biology makes sense except in the light of evolution*[3]".

Acknowledgments

This project/publication was made possible through the support of a grant from Templeton World Charity Foundation, Inc. 0092/AB57, the Defense Threat Reduction Agency (HDTRA1- 13-1-0004), and by NASA under awards: NNX-14AK37G and under award NNX15AF46G. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of Templeton World Charity Foundation, Inc., or the National Aeronautics and Space Administration, or the Department of Defense. R. Laos also would like to than Eliana Esparza for her suggestions during the preparation for this article.

Received: July 1st, 2016
Acepted in final format: september12th, 2016

The spanish version of this paper is available online here Descargar

ESSENTIAL LITERATURE

Doolittle, R: "*Of URFS and ORFS. A Primer on How to Analyze Derived Amino Acid Sequences*", University Science Books: Mill Valley, 1986.

[7] Smith, J. M., Natural selection and concept of a protein space. *Nature* **1970**, *225* (5232), 563-564.