Accessing and applying molecular history

Joshua G. Stern[1] and Eric A. Gaucher[2]

[1]Machine Learning Research, Silver Spring, MD 20910; joshstern6@gmail.com
[2]School of Biology, School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, GA 30332; General Genomics, Atlanta, GA 30332

ABSTRACT

Studying the evolutionary history of life's molecules - DNA, RNA, and protein - reveals nature-based solutions to real-world problems. We discuss an approach to applied molecular evolution that is well-known within the field but may be unfamiliar to a wider audience. Using a case study at the intersection of molecular evolution and medicine, we introduce the fundamental concepts of *orthology* and *paralogy*. We also explain a practical entry point to molecular evolution named STORI: Selectable Taxon Ortholog Retrieval Iteratively. STORI is a machine learning algorithm designed to clear a bottleneck that researchers encounter when studying evolution.

Availability. Existing source code is available for download from GitHub (https://github.com/jgstern/STORI_singlenode).

Keywords: molecular evolution, machine learning, synthetic biology

INTRODUCTION

In 2013, sickle-cell anemia killed more than 56,000 people [4]. This genetic disorder occurs when someone has a mutation in their beta-hemoglobin gene [6]. This gene is the DNA blueprint for actual beta-hemoglobin *proteins*: subcellular, nanometer-scale molecular machines made of yet smaller building blocks known as *amino acids*. Like a jeweler making necklaces from 20 different types of bead, life uses 20 different types of amino acid to build proteins. A single cell contains millions of proteins [38], although many of them share the same sequence of amino acids and thus have the same function. Protein functions include maintaining DNA, sending signals across the nervous system, and in the case of hemoglobin, transporting oxygen from lungs to tissues.

Using the evolution and biochemistry of hemoglobin as a case study, we introduce *orthology* and *paralogy* as principles that are scientifically interesting and practically relevant. We also explain a practical entry point to molecular evolution named STORI: Selectable Taxon Ortholog Retrieval Iteratively. STORI is a machine learning algorithm designed to clear a bottleneck that researchers encounter when studying evolution.

THE BIOCHEMISTRY OF SICKLE-CELL ANEMIA

Below are the amino acid sequences comprising normal human beta-hemoglobin protein, and its sickling variant. Protein sequences such as these use a 20-letter alphabet to represent which of the 20 possible amino acids presents at a position along the protein chain. We can describe every protein with a sequence of letters.

```
   >gi|229752| Human hemoglobin subunit beta
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVA
NALAHKYH
   >gi|40889142| Human sickle-cell hemoglobin S
VHLTPVEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVA
NALAHKYH
```

Naturally-occurring hemoglobin is not a straight chain of amino acids. Amino acids interact with each other and their environment, causing the chain to fold. Some amino acids attract water, and others repel water. Some carry a positive charge; others a negative charge. The colossal range of possible protein behaviors results from the variety of available amino acids and their myriad possible orderings.

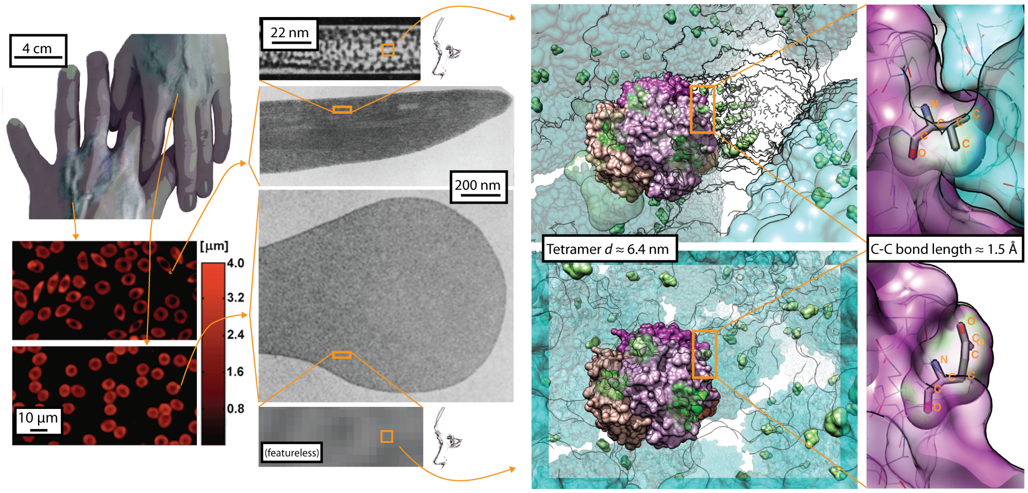


Figure 1. Physical differences between normal blood and that of someone experiencing a sickle-cell crisis. The right-side person's hand has normal blood, so even when oxygen is low, their red blood cells (RBCs) are flexible and shaped like shrink-wrapped, vacuum-sealed, stale doughnuts. The other hand belongs to someone having a sickle-cell crisis, caused by RBCs with sharp corners that rigidified in response to low-oxygen conditions. Wide-field digital interferometry thickness profiles (first column, bottom half) and transmission electron micrographs (second column) of single RBCs show the pointy shape and internal structure that RBCs develop when they sickle. More detailed electron micrographs of the internal structure (second column, top image) suggest that when sickling occurs, hemoglobin tetramers organize into double-stranded, helical multimers (third column, top). These double-strands seem to assemble into groups of seven, resulting in 14-stranded fibers [53] (third column, top). X-ray crystallography confirms that double-strands are the keystone of hemoglobin fiber formation [23]. Models of the atomic coordinates responsible for the crystallography data show double-strand assembly depends on a valine amino acid from a "donor" tetramer fitting between two other amino acids from a "receiver" tetramer [42] (fourth column, top). Normal hemoglobin has a glutamate instead of a valine at this position, which is too big and hydrophilic to bind any receiver (fourth column, bottom). The 3D molecular visualizations in this figure show the sixth residue of beta subunits in light green, hemes of the subject tetramer in dark green, and surrounding tetramers in cyan. The four subunits of the subject tetramer are dark and light purple for the two betas and dark and light copper for the two alphas. We generated these 3D visualizations using the UCSF Chimera software [49], Adobe Photoshop, and Protein Databank (PDB) files 2HBS and 2HHB. These PDB files resulted, respectively, from [23] and [14]. *Image credits:* Hands modified from "friends" (© SuperFantastic, 2007) under CC BY 2.0. Wide-field digital interferometry thickness profiles of RBCs (© Shaked et al. 2011) are reprinted under CC BY 3.0 from their original publication in the Journal of Biomedical Optics [57] (doi:10.1117/1.3556717). Sickled and normal RBCs (© Döbler and Bertles 1968) are reprinted under CC BY-NC-SA 3.0 from their original publication in The Journal of Experimental Medicine [12] (doi:10.1084/jem.127.4.711). Detailed hemoglobin fiber micrograph (© Rodgers et al. 1987) reprinted with permission from the authors from their original publication in Proceedings of the National Academy of Sciences [53] (doi:10.1073/pnas.84.17.6157). Profile of a person's eyes (© Judith Gallant 2015) used under CC BY-NC-SA 3.0.

Little changes can make a big difference. With the exception of the sixth residue, highlighted in yellow, the above two sequences are identical. In normal human beta-hemoglobin, the sixth amino acid is glutamate (E). In human sickle-cell beta, the sixth amino acid is valine (V). Glutamate attracts water, and valine repels water. Valine is also smaller than glutamate. Although these two protein sequences are over 99% identical, a single substitution has significant downstream consequences.

Sickling hemoglobins polymerize (aggregate, or stick together) when the concentration of oxygen in red blood cells falls below the normal range - during intense exercise, for instance [15]. Hundreds of individual hemoglobins, sticking together, form unwieldy hemoglobin fibers that cause their red blood cell containers to rigidify and develop sharp corners [15]. Sickle-shaped red blood cells clog blood vessels and cause serious health problems (Fig. 1) [51].

Sometimes protein molecules stick together in unhelpful ways, as above, and sometimes they stick together in helpful ways. Whether or not they are part of a sickle fiber, hemoglobin molecules form four-member packs. In adults without sickling beta-hemoglobin, two of the pack members are beta-hemoglobins and two of the pack members are *alpha-hemoglobins* [41, 59]. We call each pack a *tetramer*.

Human hemoglobin efficiently transports oxygen from areas of surplus (lungs) to areas of deficit (tissues). The transport is efficient because the hemoglobin tetramer has a high oxygen affinity in our lungs and a low oxygen affinity in our tissues. If the oxygen concentration in a red blood cell decreases, then the hemoglobin tetramer changes the physical arrangement of its four subunits, which causes the tetramer to decrease its oxygen affinity, and potentially let go of its oxygen. Conversely, increasing oxygen concentration makes hemoglobin tetramers oxygen-greedy. Sequence differences between the alpha and beta subunits are what make these two types of subunit stick together and change into the appropriate shape for the circumstances [41, 59, 26].

For people with sickle-cell anemia, the tetramers usually do their job, however as we said above, the E to V substitution in sickling beta-hemoglobins causes hundreds of hemoglobin tetramers to form fibers when oxygen gets particularly scarce. In healthy individuals, low oxygen does not cause hemoglobin tetramers to polymerize.

One treatment for sickle-cell anemia is to stimulate production of gamma-hemoglobin, because it can pinch-hit for defective beta-hemoglobin and reduce polymerization [51]. Human fetuses produce gamma-hemoglobin, which is only slightly different than the beta-hemoglobin that adults produce. Fetal blood consists almost entirely of fetal hemoglobin (two alpha and two *gamma* subunits); blood in people older than 10 months is almost entirely adult hemoglobin (two alpha and two *beta* subunits) [52, 43].

The human genome (our ”instruction manual” written in DNA) contains genes for both gamma and beta, but human red blood cells *express* these genes as proteins at different stages of a human's development. At birth, gamma-hemoglobin production starts a 10-month wind down. However, in a pinch, doctors prescribe non-infants hydroxyurea. For reasons still being studied [63], hydroxyurea enhances production of gamma-hemoglobin in adults and children, thereby improving sickle-cell anemia symptoms.

Like protein, DNA is a chain of smaller constituents (nucleotides). Similar to how we can describe a protein with a sequence of characters chosen from a 20-letter alphabet, we use sequences containing four possible letters to represent the four possible nucleotides in DNA. When subcellular machines synthesize a protein, they follow instructions encoded by genes, which are made out of DNA. The properties of nucleotides make DNA suited to store information, and the properties of amino acids make proteins suited to perform tasks.

MEDICINE IN AN EVOLUTIONARY CONTEXT

Gamma-hemoglobin is a *paralog* of beta-hemoglobin, which means that historically, beta and gamma hemoglobin were exactly the same. These two types of protein, which we'll refer to as *subfamilies*, were once only one type of protein, coded by one gene. According to the latest research, two duplication events led to contemporary gamma-hemoglobin. First, a proto beta-hemoglobin gene duplicated 220 million years ago, producing a new beta and a new epsilon gene. Next, this ancestral epsilon gene duplicated 130 million years ago, producing what would become *contemporary* gamma-hemoglobin and beta-hemoglobin [45, 44].

Just as gamma-hemoglobin came from ancestral beta- and epsilon-hemoglobin, ancestral beta-hemoglobin also came from somewhere. Jawed vertebrates have beta hemoglobin, but jawless vertebrates don't. The jawless vertebrates have proteins with sequences similar to beta, epsilon and gamma hemoglobin, but not *that* similar. The model for this discrepancy is that when ancestral vertebrates diversified (speciated) into the jawed type and the jawless type, 475 million years ago, DNA coding an ancestral hemoglobin precursor duplicated along the jawed branch of life's tree but *not* along the jawless branch [24]. Where there was previously only one copy of instructions for the ancestral hemoglobin, a duplication event yielded two copies, but only in jawed vertebrates. One of these genes became alpha-hemoglobin, and the other became beta-hemoglobin. As a consequence of the historical gene duplication that birthed

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1293v1 | CC-BY 4.0 Open Access | rec: 11 Aug 2015, publ: 11 Aug 2015

alpha and beta-hemoglobin, in humans these subfamilies must cooperate to carry oxygen through the blood.

Occasionally, cells make mistakes when replicating their DNA, causing life's genetic code to accumulate mutations. As time passes, dissimilarity (aka diversity) usually grows between DNA in two different species, or two duplicated genes in the same species [66]. After the initial gene duplication that birthed ancestral alpha and beta hemoglobin in jawed vertebrates, this particular lineage of vertebrates diversified. Natural selection and genetic drift evolved [22, 35] the sequences of ancestral beta and alpha hemoglobin to what they are now: collaborative proteins with complementary functions and similar - but nonidentical - sequences. Today's alpha and beta hemoglobins are palimpsests of whatever ancestral gene preceded their birth.

As we said earlier, after alpha hemoglobin and beta-hemoglobin were born, but before the present day, another two duplication events occurred. The gene for ancestral beta-hemoglobin duplicated, and then its daughter epsilon-hemoglobin duplicated again. These events led to today's gamma-hemoglobin and beta-hemoglobin. One explanation of these gene duplication events is that they enabled the hemoglobin family to become more efficient at specific tasks. Alpha/gamma tetramers have a higher oxygen affinity than alpha/beta tetramers in a situation where a pregnant mother is breathing for two [59].

Because of how evolution unfolded, in humans the gamma sequence and the beta sequence are more related than the gamma sequence and the alpha sequence. Put another way, gamma and beta diverged more recently than gamma and alpha. In light of this model, for humans it makes sense that gamma-hemoglobin can pinch-hit for beta-hemoglobin but alpha-hemoglobin cannot. Tetramers of human alpha-hemoglobins simply do not assemble [1, 56]. Without either alpha/beta tetramers or alpha/gamma tetramers, a person cannot live. Using hydroxyurea to artificially boost gamma-hemoglobin in people with sickle-cell anemia is an example of using a defective protein's paralog as a pinch-hitter. Studying evolutionary history reveals nature-based solutions to real-world problems (16).

SEARCHING FOR EVOLUTIONARY CONTEXT WITH ORTHOLOGY RECONSTRUCTION

Our review of hemoglobin evolution and biochemistry relies on published research. However, biologists have yet to tell the stories of thousands of subfamilies besides hemoglobin. Here is where *ortholog retrieval* comes in.

Sequences can be similar to each other in different ways and to different degrees. If two sequences, each in a different species, are similar because they "derived from a single ancestral gene in the last common ancestor of the compared species [29]," then the sequences are *orthologs*. Alpha-hemoglobin in monkeys is orthologous to alpha-hemoglobin in humans. If two proteins are similar because a gene duplication event allowed their sequences to diverge, then they are *paralogs*. Alpha-hemoglobin in monkeys is paralogous to beta-hemoglobin in monkeys. Duplication causes paralogs and speciation causes orthologs.

With some exceptions [7, 64, 28] it is rare to witness speciations and gene duplications in real time. More often than not, the only evidence of these historical events is from present day organisms that have too much in common for coincidence and too many differences for biologists to give them the same name. Replaying the tape of life means finding the most probable historical model given features of contemporary organisms [18].

Searching for the most plausible model given the data is the keystone of evolutionary reconstruction. We reconstruct orthology to facilitate other types of reconstruction. Modeling a set of sequences as orthologous does not specify the order of speciation events leading to the observed sequence diversity. Rather, orthology conveys that the sequences are different *because of speciation*, not duplication. Modeling two subfamilies as paralogous does not specify where in evolutionary history the duplication occurred. Rather, paralogy conveys that sequences in subfamily A are different from those in subfamily B *because of duplication*, not speciation. Having a model for which sequences are orthologs and which sequences are paralogs makes it easier to reconstruct speciations, ancestral sequences, gene gain, gene loss, and gene transfer. Researchers have developed several methods for orthology reconstruction [10, 32, 60, 65, 69, 67] because this step facilitates detailed reconstructions downstream.

We write historical fiction – retrodictions of unwitnessed past events - to explain present-day observations (10) [40]. Often we use algorithms to build millions of alternative reconstructions, holding on to the ones under which the likelihood of the data improves, and letting go of the ones under which the likelihood worsens [13]. As we iterate through each model, we traverse a rugged terrain of plausibility in search of the highest ”mountain”.

Finding the highest ”plausibility mountain” (aka the global optimum) can be tricky. However, we improve our odds of getting there by strategically choosing which data we wish to explain and by sampling different areas of model space with the right balance between small steps and giant leaps [54].

When we say, “these sequences are orthologs, but *those* are paralogs” we are proposing a plausible history given the contemporary protein (or DNA, or RNA) sequence data. Researchers have created many definitions of plausibility. One popular definition of a plausible orthology is *best-hit symmetry*.

USING SYMMETRIC BEST-HITS TO REVEAL CLUSTERS OF ORTHOLOGOUS GROUPS

A single *E. coli* bacterium contains about 1000 different protein subfamilies, as does a single *V. cholera* bacterium. (Furthermore, each cell contains about one million protein molecules, since a cell produces each protein subfamily many times.) In other words, each of these two example cells contains roughly 1000 nonredundant sequences. Suppose we choose one sequence from the *E. coli*, and search through the *V. cholera* sequences for the one sequence that is most similar to the *E. coli* query. This search result is the *best hit*, similar to if we entered a query on the Google page and clicked “I’m Feeling Lucky.” Now, take the *V. cholera* best hit, and use it to query the *E. coli* sequences. Is the best hit the same sequence that we initially chose from *E. coli*? If yes, then these two sequences are symmetric best hits, and we infer that they are orthologs. The National Center for Biotechnology Information (NCBI) pioneered this technique in the 1990s, and used it to build the *Clusters of Orthologous Groups (COG)* paradigm [60].

COGs help us understand how subfamilies were born. A COG is a set of sequences from different species. Every sequence in a particular COG must be the symmetric best hit of at least two other sequences in that COG. When a set of sequences satisfies this condition, most phylogeneticists feel reasonable saying that the sequences are orthologs (descended from a single gene in the species’ last common ancestor). Examining COGs for a set of species reveals a *phyletic pattern* [39] for each subfamily. The phyletic pattern is the presence or absence of a subfamily (ortholog) in a particular species (lineage). If an ortholog is present in one lineage but not another, one explanation is that a gene duplication occurred when those lineages diverged.

To build COGs we first need to know all pairs of symmetric best hits, which means running many best-hit searches (sometimes hundreds of thousands). We perform the best-hit searches using the Basic Local Alignment Search Tool (BLAST) software [2]. We choose a sequence, say from the panda, and we find its’ best hit in the cow database. Then we find its best hit in the human database. Then we find its best hit in the mouse database. We repeat until we’ve queried each of our organisms (taxa). Then we choose another sequence from panda and repeat the process. For each sequence in each taxon’s database, we BLAST it against the remaining databases and record the best hit (1). If we have n taxa represented by n protein databases, and S sequences per database, then each taxon requires $S*(n-1)$ BLAST searches. The total number of BLAST searches conducted is $S*(n-1)*n$. Finding all pairs of symmetric best hits requires time proportional to the square of the number of taxa (2).

When we find symmetric best hits, we take note of them. Notice the metadata (gi|229752|... and gi|40889142|..., highlighted in green) before beta- and sickling-hemoglobin’s actual amino acid sequences. The NCBI designed those codes so that each GI number is a unique identifier for a unique record in its protein sequence database. Using a GI number enables us to refer to a particular protein sequence exactly and concisely. Using GI numbers, we make a network (graph) of protein sequences and their symmetric best hits. Nodes are GI numbers (references to protein sequences) and an edge connecting two nodes means the two sequences are symmetric best hits. If each protein sequence had a LinkedIn account, then symmetric best hits would be connections. However, sequences from the same organism cannot connect.

TELLING STORIES WITH COGS

A graph by itself does not say anything about the history of protein families. To get closer to the story we need to build the COGs, and to do this we traverse the graph using a method called *EdgeSearch* [31]. We start our first COG with an edge – any edge - and its vertices – the symmetric best-hit relationship between two orthologs. Then we ask: which vertices in the graph are adjacent to (one hop away from) both of this edge’s vertices? We add the new adjacent vertices, and their edges with the previous two vertices, to our COG. We mark the first edge as “processed”, and move on to the next unprocessed edge in the COG. We repeat the question: which vertices in the graph are adjacent to both of this edge’s vertices? We add the answer to our COG, mark the query edge processed, and continue. Eventually our COG will run out of unprocessed edges and at that point it is complete. Thousands of unbuilt COGs remain (in Bacteria, hundreds would remain, since bacteria have ~10x fewer gene families than animals). To continue building COGs, we choose an unprocessed edge and start all over. Eventually we will have examined every edge in the graph.

Choosing several known sequences for any family of interest, e.g. hemoglobin, we can run a small number of BLAST searches by hand and fish out different COGs. The COGs tell us the phyletic pattern of which orthologs are present or absent in which species, and this information contributes to the story of which genes were duplicated or lost by who and when. We integrate this model with other information about genomic structure, phenotype, and environmental history, revealing how evolution provides both causes and solutions to life’s problems.

Like EdgeSearch, other popular methods for predicting ortholog sets require the pre-computation of a large number of BLAST searches (3). In other words, if we want orthologs, we will probably use a method that takes ~n^2 time. Not everyone has access to that kind of power.

NO ONE LAB SHOULD HAVE ALL THAT POWER

STORI aims to provide an alternative to all-against-all searching, thereby reducing the time-to-orthologs. Three main characteristics of STORI distinguish it from the COG approach: “some-against-some” rather than all-against-all, “most-popular best hits” rather than reciprocal best hits, and a quasi-deterministic rather than deterministic output.

Instead of traversing a pre-computed dataset, STORI begins by accepting from the user a set of ”seed” protein sequences and an upper limit to the number of different subfamilies (sets of orthologs) it may retrieve. STORI randomly scatters the seeds into two parallel, independent iterator processes. Each iterator BLASTs the seeds against taxon databases. Within each iterator, the results of each search become queries for subsequent searches. Each iterator dynamically assigns sequences to different subfamilies as the results accumulate. By keeping track of the frequency with which queries from each subfamily choose a particular sequence as a best hit, STORI assigns each hit to a subfamily. Over the course of a run, STORI learns [9] which subfamily each sequence belongs to.

Let’s drill down into the flow of one of the two independent STORI iterators, to depict this algorithm concretely. The iterator begins with seed sequences. The crucial point here is that STORI depends on prior knowledge supplied by the user. The user must provide at least one seed sequence that is evolutionarily related to the subfamilies of interest; the closer the better. By default, the seed sequences come from a random draw from the results pool of a user-initiated keyword search of the natural language annotations (green highlighted text, above) of every taxon’s protein sequence database. If annotations were perfect, ortholog retrieval would be trivial because the subfamily name would be present in each annotation. However, annotations often do not indicate which subfamily a protein is a member of, as in the case of this frog hemoglobin:

```
>gi|213982769|ref|NP_001135556.1| uncharacterized protein LOC100216102
[Xenopus (Silurana) tropicalis]
```

A user interested in the hemoglobin family could provide the expression ”[hH]emoglobin” to STORI and the program would create a pool of sequences whose annotation contain ”hemoglobin” or ”Hemoglobin”. This pool could contain thousands of sequences with annotation matching the user query. However, the size of the pool depends on the completeness of the sequence annotations and the specificity of the user query. STORI chooses seeds by randomly sampling a fraction of the sequence pool, and passing the sample to an iterator. The random sampling happens twice; once for each iterator.

When an iterator begins, STORI assigns each seed to its own unique “quasi”-subfamily. The quasi-subfamilies are not biologically meaningful subfamilies. However, as iteration progresses these quasi-subfamilies will become plausible predictions of subfamilies (orthologous proteins). At the beginning of an iteration, STORI chooses a small, user defined number of taxa (usually 4). STORI makes this choice by sliding a window of user-defined size (usually 4) down a list of all the taxa. At iteration 1, the window is at the top of the full taxa list, and the window will advance by one list element with each subsequent iteration. After selecting the small list of taxa, STORI cycles through each quasi-subfamily and BLASTs any sequences assigned to that subfamily, within the small taxa window, against each taxon in the window.

STORI parses the BLAST results and assigns any best hits to their parent taxa within the subfamily. After using BLAST to retrieve best-hits for each quasi-subfamily, STORI identifies best hits assigned to multiple subfamilies and prunes all but the most popular. After pruning, the small taxa window slides down the full taxa list by one element.

FOLLOWING THE WHITE RABBIT

For example, suppose the small window contains four taxa: panda, cow, dog and horse (Fig 2A). The window resides at the first quasi-subfamily, which we will arbitrarily call “M” (Fig 2A, left column). The only sequence present in the window (GI 301769567) is a seed protein sequence from panda, which STORI chose based on the user’s natural-language query ”[hH]emoglobin”. The program BLASTs the panda protein sequence query against the panda, cow, dog, and horse databases. The best hit for the panda is the query sequence, and the searches against the other three taxa each return a different best hit. The score for all four sequences increases by one, because each sequence was a best hit once (Fig. 2B, left column).

Next, the small window slides over to quasi-subfamily “Z” (Fig. 2A, right column) and the equivalent steps occur. As before, the only available query is a seed from panda (a different sequence than the query used in “M”, from a different panda protein). BLAST searches yield four best hits, one for each taxon, and these “Z” sequences have best-hit scores which STORI increments by 1 (Fig. 2B, right column). If more subfamilies exist, the window slides over to each and repeats this process.

Although STORI is unaware of it, quasi-subfamilies M and Z correspond to mu- and zeta-hemoglobin, respectively. The BLAST searches that just occurred did an almost perfect job of putting the mu sequences in M and the zeta sequences in Z. However, horse alpha-hemoglobin got assigned to horse in the M subfamily, which does not make biological sense (Fig. 2B, left column). Mu and alpha are distinct sets of orthologs. The two subfamilies are paralogous and are related by a duplication event, but the current arrangement implies that speciation alone caused an ancestral mu to evolve into horse alpha.

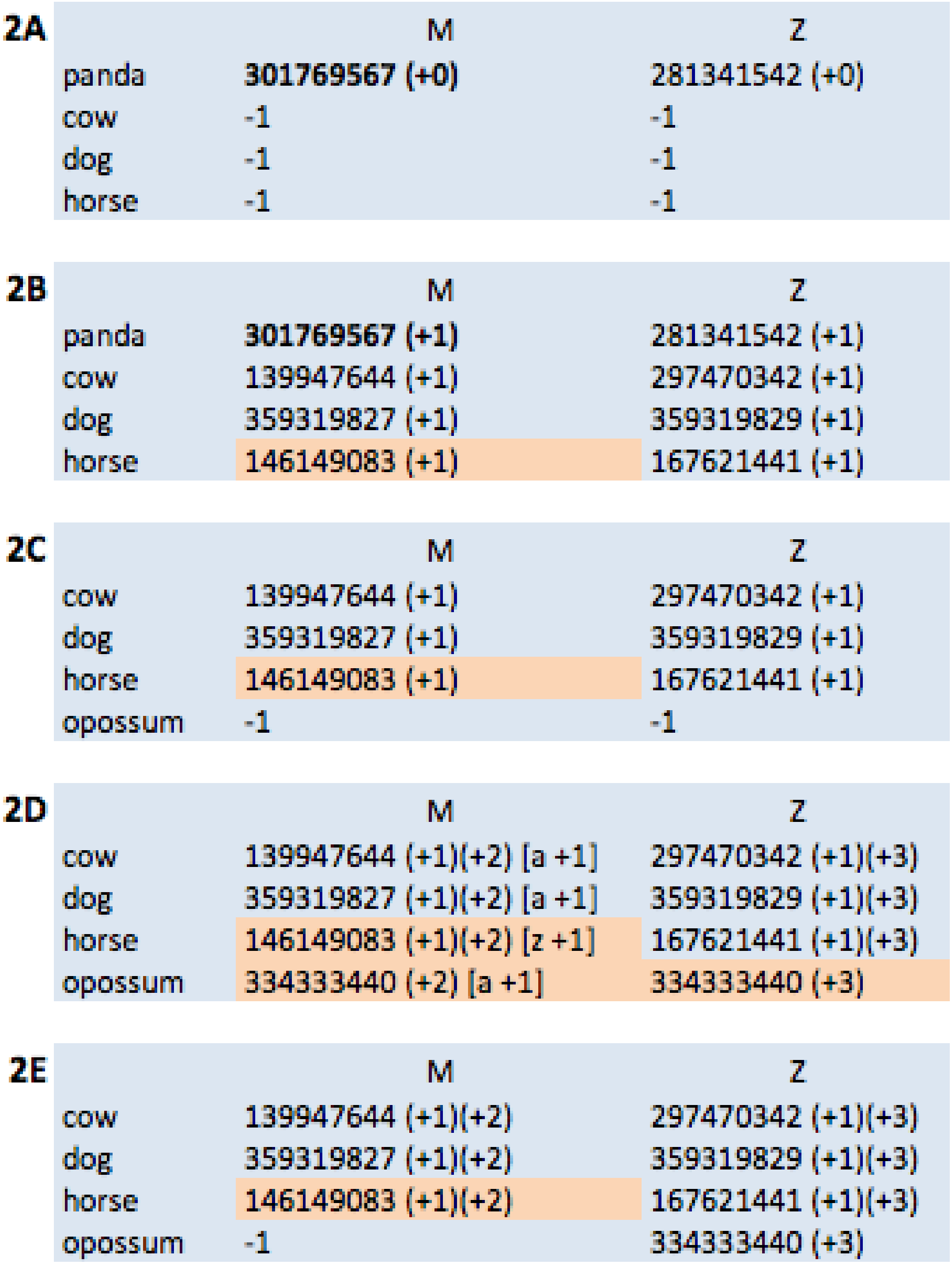
What’s going on here? Why didn’t the BLASTing return a horse mu-hemoglobin sequence? *Because it turns out that the mu sequence does not exist in the horse database.* As a result, the search returned horse alpha, since it is most similar to the query. Probably the database is incomplete, or perhaps a gene loss event occurred. We would need to do more digging to be confident.

Our sliding window, still at subfamily Z (Fig. 2B, right column), is almost ready to return to the beginning of the subfamily list and advance one taxon. However, before it does that, STORI attempts to do some pruning. It checks to make sure that no sequence exists in more than one subfamily at a time (5). Finding no redundant assignments, the 4-taxon sliding window advances (Fig. 2C).

STORI now considers cow, dog, horse, and opossum (Fig. 2C). Opossum does not yet have any assigned sequences. The window is at the M subfamily (Fig. 2C, left column). STORI BLASTs cow mu, dog mu, and horse alpha against each taxon in the window. As the left column of Fig. 2D shows, the score for cow mu gains 2, but an additional cow sequence, alpha, joins the cow quasi-subfamily M, due to the horse alpha query against cow. Dog results are analogous to cow’s. The horse alpha increases by 2, and the horse quasi-subfamily M picks up horse zeta (6). The opossum adds opossum zeta +2 and opossum alpha +1; like the horse database, the opossum database also lacks a mu sequence. When the window slides over to subfamily Z (Fig. 2C, right column), each of the three zeta sequence scores increase by 3, and opossum zeta gets assigned to opossum with a score of 3 (Fig. 2D, right column).

After the window slides over to any other subfamilies (ignored for simplicity) and repeats the sequence retrieval, STORI enters the pruning step. Opossum zeta is assigned to both the M subfamily and the Z subfamily, but its score is highest in the Z subfamily (Fig. 2D; $3 > 2$). As a result, the STORI iterator prunes opossum zeta from M (Fig. 2E, left column).

The window slides down by one taxon, and now considers dog, horse, opossum and mouse. Sequence retrieval and pruning repeat. The window advances one more taxon, and so on, until it reaches the bottom of the full taxon list. At this point, STORI randomizes the order of the taxa in the list, and takes a few additional steps explained below. The window returns to the top, and the process repeats until a user-defined time limit expires, or the iterator stops finding new sequences.



2A	M	Z
panda	**301769567 (+0)**	281341542 (+0)
cow	-1	-1
dog	-1	-1
horse	-1	-1

2B	M	Z
panda	**301769567 (+1)**	281341542 (+1)
cow	139947644 (+1)	297470342 (+1)
dog	359319827 (+1)	359319829 (+1)
horse	146149083 (+1)	167621441 (+1)

2C	M	Z
cow	139947644 (+1)	297470342 (+1)
dog	359319827 (+1)	359319829 (+1)
horse	146149083 (+1)	167621441 (+1)
opossum	-1	-1

2D	M	Z
cow	139947644 (+1)(+2) [a +1]	297470342 (+1)(+3)
dog	359319827 (+1)(+2) [a +1]	359319829 (+1)(+3)
horse	146149083 (+1)(+2) [z +1]	167621441 (+1)(+3)
opossum	334333440 (+2) [a +1]	334333440 (+3)

2E	M	Z
cow	139947644 (+1)(+2)	297470342 (+1)(+3)
dog	359319827 (+1)(+2)	359319829 (+1)(+3)
horse	146149083 (+1)(+2)	167621441 (+1)(+3)
opossum	-1	334333440 (+3)

Figure 2. Illustration of how STORI uses seed sequences, a sliding window, and the concept of ”most-popular best hits” to dynamically assign sequences to subfamilies.

SEARCHING FOR THE MIDDLE PATH

Described above is what each parallel, independent STORI iterator spends most of its time doing. STORI takes a few additional steps to produce reasonable orthology predictions. These steps occur within each parallel iterator as well as outside of the iterators. On the one hand, these steps usually prevent capture by local optima, and on the other, they prevent orthology predictions from careening off to an irrelevant part of subfamily space.

Once both iterators have finished, the STORI run controller finds the intersection of the results from each independent output, and provides the intersection set to each of a new pair of iterators. Rather than assign each seed sequence to its own quasi-subfamily, as STORI did for the first iterator pair, this initialization includes the previous iterator pair's predictions of which sequences are orthologous (members of the same subfamily). This new pair of parallel, independent iterators may provide similar output at first, since their inputs were identical. However, each iterator's full taxa list begins in a randomized order, which the iterators reshuffle every time their sliding windows reach the ends of their lists. This randomization causes the iterators to sample different areas of subfamily space.

Periodically, the STORI controller compares the results of the iterators, and assigns the pair a score reflecting how similar their orthology predictions are to one another. As these independent runs are executed, compared, and restarted with updated seed subfamilies, the similarity score stabilizes in the range of 90-100%. When the controller detects this stabilization, it labels the run as converged and stops further iteration.

Within each iterator, STORI checks for "orphan" (aka pseudo-orthologous) sequences with a score of 1. This check occurs each time the sliding window reaches the bottom of the full taxa list. When it identifies an orphan, STORI moves the sequence to a new subfamily. A score of 1 means that only one of that sequence's presumed orthologs chose it as a best hit; in this scenario paralogy may be more probable than orthology.

A particular subfamily's orthologs may present in only a small fraction of the taxa. In this case, the "sparsely populated" subfamily will compete for representation with a paralogous "abundant" subfamily whose orthologs present in a large fraction of the taxa. For example, if mu-hemoglobin only presents in 3/10 taxa, then mu queries may return alpha best hits, which in a subsequent iteration push out the mu sequences. The pruning step often reassigns alpha to the alpha subfamily, however, this reassignment only occurs when a taxon's alpha subfamily is assigned the alpha sequence and its score is higher than that of the misplaced alpha. Before convergence, we do not expect subfamilies to be fully populated, so the pruning step does not always make correct reassignments. To prevent abundant subfamilies from outcompeting sparse subfamilies, STORI can boost the seed sequence scores when initializing the first pair of iterators. Using a single parameter, the user defines a range for all initial seed scores, specifying the extent to which the seeds persist for multiple taxa list traversals within the first iteration pair.

After the STORI iterator reassigns orphans to new subfamilies, it decreases the seed scores by the value of the highest non-seed score, and resets the non-seed scores to 2. This reset enables non-seed sequences to move between subfamilies in the next taxa list traversal.

Not all sequences with a score of 1 are truly orphans; sometimes BLAST searches yield several different best hits each scoring 1. In this case, STORI incorrectly decides that a sequence is an orphan and moves it to a new subfamily. As a result, two or more subfamilies may develop although only one should exist. To prevent subfamily scattering, a merge step may execute once the sliding window hits the bottom of the taxa list. This step compares every subfamily with every other subfamily, and merges all subfamilies above a similarity threshold. To avoid undermining the boosted seed scores, merges only occur when the seed scores and non-seed scores have similar magnitudes.

After the merge step, each iterator sorts its subfamilies by number of member sequences. If the number of subfamilies is larger than the maximum allowable subfamilies, then the iterator deletes the smallest subfamilies until the number of subfamilies does not exceed the maximum allowable.

Using a viewer program, the user can view run results, as well as assign natural language annotation to each subfamily.

Typically, repeated STORI runs with the same starting conditions produce similar but not identical ("quasi-deterministic") results. In contrast, repeated EdgeSearch runs with the same starting conditions produce identical (deterministic) results [31]. This attribute of STORI seems unsurprising given how often this method uses randomized conditions.

FUNCTION WITHOUT PURPOSE

Can a creator know with certainty how others will use her creation? We could use a mousetrap as a paperweight even though it isn't a paperweight [37]. Bacteria once used the same proteins for two different functions: locomotion and host-cell manipulation (7). We have hypotheses about how others will use STORI. And, assuming that no startup plans to fail, 90% of web startup founders have business hypotheses that are wrong [36]. Planning is essential, resources are finite, and it is hard to make predictions, especially about the future. Could obsessing over usefulness, significance, and attainability undermine these qualities rather than develop them [33]?

In certain situations, STORI is faster and more flexible than methods reliant on all-all sequence comparison [58]. STORI enables users to swap out any number of taxa without needing to recalculate symmetric best hits. The accuracy of STORI compares favorably to the accuracy of manual ortholog retrieval for a specific set of protein subfamilies (the ribosomal proteins) [58].

STORI attempts to provide a general solution to the practical problem of timely ortholog retrieval. We do not know how general this solution actually is. STORI's accuracy when retrieving non-ribosomal subfamilies is unclear. Anecdotally, STORI does a reasonable job with kinases, globins, and tRNA synthetases. Systematic testing of STORI on a variety of families, taxa sets, and run parameters would improve the algorithm's usefulness.

Another shortcoming of STORI is that it does not explicitly resolve orthologous groups – i.e., situations where a sequence in one species is orthologous to more than one sequence in another species (12). STORI output is a simple table, with rows for species and columns for subfamilies. As a result, STORI does not presently resolve co-orthologies. We designed STORI to retrieve co-orthologs (aka lineage-specific expansions), but not to cluster them as the COG algorithm does. In situations of co-orthology, STORI needs more testing.

STORI helps clear a bottleneck encountered when studying the evolution of proteins: retrieving orthologous protein sequences from a custom set of taxa. STORI may be used for exploratory data analysis, at the beginning of a workflow that culminates with biochemical tests of synthesized proteins ([11] details such a workflow). Intermediate steps could include building phylogenetic trees of STORI output to refine its orthology and paralogy predictions.

Using proteins, and other subcellular components, biologists are beginning to create purposeful, useful, and economically significant functions. These creations have no anthropogenic precedent. For example, researchers created an experimental gout treatment [30], enzymes functional at extreme temperature and pH [19, 48, 68], and ancestral RNA molecules relevant to the origin of life [34]. Other *in vitro* work shows how paralogous ATPase proteins, whose dysfunction may cause osteoporosis, evolved specific and complementary roles after their parent ancestral gene duplicated [16].

One of the first studies to apply the descent-with-modification paradigm to molecular sequences used hemoglobin as its case-in-point [47]. In the 50 years since this seminal work, molecular evolution established that all known life on Earth descended from a last universal common ancestor [5]. This astounding realization is only part of an incomplete model for the emergence of terrestrial life. Pauling and Zuckerkandl speculated that all *genes* descended from a common ancestral *gene* [47]. Other researchers have since developed related hypotheses and begun testing them with sequence analysis [27, 17, 55, 3, 61, 62]. Such hypotheses about events three to four billion years ago are difficult if not impossible to test. In and of itself, the endeavor to test is part of what makes research in molecular evolution fun. Furthermore, basic research becomes practically useful in the long term [8].

As far as we know, non-anthropogenic proteins cannot have *purpose*. However, all proteins have *function*. On Earth, evolution is four billion years old. Perhaps, without intending to, this enduring process found solutions that present-day humans will use purposefully. The evolutionary history of functional diversification is a story worth telling.

STORI is available here: `https://github.com/jgstern/STORI_singlenode`

ACKNOWLEDGMENTS

Funding from the National Aeronautics and Space Administration's Astrobiology Institute (E.A.G), the Georgia Institute of Technology Presidential Fellowship, and NASA's Earth and Space Science Fellowship (grant NNX10AT21H; J.G.S) supported STORI development. This work used the Georgia Institute of Technology's Partnership for an Advanced Computing Environment (PACE), and the National Science Foundation's Extreme Science and Engineering Discovery Environment (XSEDE TG-MCB120032). Molecular graphics and analyses were performed with the UCSF Chimera package. Chimera is developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIGMS P41-GM103311). Thank you to Stanley H. Stern, Judith Gallant, and Kenneth S. Gallant for providing excellent feedback on earlier drafts of this article.

AUTHOR CONTRIBUTIONS

Joshua G. Stern conceived and designed the experiments, performed the experiments, analyzed the data, contributed materials/analysis tools, wrote the paper, prepared figures and/or tables, performed the computation work, and reviewed drafts of the paper.

Eric A. Gaucher contributed materials/analysis tools, reviewed drafts of the paper, and provided advice and supervision.

REFERENCES

[1] Aljurf, M., Ma, L., Angelucci, E., Lucarelli, G., Snyder, L., Kiefer, C., Yuan, J., and Schrier, S. (1996). Abnormal assembly of membrane proteins in erythroid progenitors of patients with beta-thalassemia major. *Blood*, 87(5):2049–2056.

[2] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402.

[3] Anantharaman, V., Aravind, L., and Koonin, E. V. (2003). Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins. *Current opinion in chemical biology*, 7(1):12–20.

[4] Ärnlöv, J., Larsson, A., et al. (2014). Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the global burden of disease study 2013. *The Lancet*.

[5] Battistuzzi, F. U. and Hedges, S. B. (2009). A major clade of prokaryotes with ancient adaptations to life on land. *Molecular biology and evolution*, 26(2):335–343.

[6] Bauer, D. E. and Orkin, S. H. (2011). Update on fetal hemoglobin gene regulation in hemoglobinopathies. *Current opinion in pediatrics*, 23(1):1.

[7] Blount, Z. D., Barrick, J. E., Davidson, C. J., and Lenski, R. E. (2012). Genomic analysis of a key innovation in an experimental escherichia coli population. *Nature*, 489(7417):513–518.

[8] Brock, T. D. (1997). The value of basic research: discovery of thermus aquaticus and other extreme thermophiles. *Genetics*, 146(4):1207.

[9] Buchanan, B. G., Mitchell, T. M., Smith, R. G., and Johnson Jr, C. R. (1979). Models of learning systems. Technical report, DTIC Document.

[10] Chen, F., Mackey, A. J., Vermunt, J. K., and Roos, D. S. (2007). Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PloS one*, 2(4):e383.

[11] Cole, M. F. and Gaucher, E. A. (2011). Exploiting models of molecular evolution to efficiently direct protein engineering. *Journal of molecular evolution*, 72(2):193–203.

[12] Döbler, J. and Bertles, J. F. (1968). The physical state of hemoglobin in sickle-cell anemia erythrocytes in vivo. *The Journal of experimental medicine*, 127(4):711–716.

[13] Felsenstein, J. and Felenstein, J. (2004). Inferring phylogenies.

[14] Fermi, G., Perutz, M., Shaanan, B., and Fourme, R. (1984). The crystal structure of human deoxyhaemoglobin at 1.74 å resolution. *Journal of molecular biology*, 175(2):159–174.

[15] Finch, J., Perutz, M., Bertles, J., and Döbler, J. (1973). Structure of sickled erythrocytes and of sickle-cell hemoglobin fibers. *Proceedings of the National Academy of Sciences*, 70(3):718–722.

[16] Finnigan, G. C., Hanson-Smith, V., Stevens, T. H., and Thornton, J. W. (2012). Evolution of increased complexity in a molecular machine. *Nature*, 481(7381):360–364.

[17] Fox, G. E. (2010). Origin and evolution of the ribosome. *Cold Spring Harbor perspectives in biology*, 2(9):a003483.

[18] Gaucher, E. A. (2007). Ancestral sequence reconstruction as a tool to understand natural history and guide synthetic biology: realizing and extending the vision of Zuckerkandl and Pauling. *Liberles [83]*, pages 20–33.

[19] Gaucher, E. A., Govindarajan, S., and Ganesh, O. K. (2008). Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature*, 451(7179):704–707.

[20] Gong, L., Parikh, S., Rosenthal, P. J., and Greenhouse, B. (2013). Biochemical and immunological mechanisms by which sickle cell trait protects against malaria. *Malaria journal*, 12(1):317.

[21] Gophna, U., Ron, E. Z., and Graur, D. (2003). Bacterial type iii secretion systems are ancient and evolved by multiple horizontal-transfer events. *Gene*, 312:151–163.

[22] Harms, M. J. and Thornton, J. W. (2013). Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nature Reviews Genetics*, 14(8):559–571.

[23] Harrington, D. J., Adachi, K., and Royer, W. E. (1997). The high resolution crystal structure of deoxyhemoglobin s. *Journal of molecular biology*, 272(3):398–407.

[24] Hoffmann, F. G., Opazo, J. C., and Storz, J. F. (2010). Gene cooption and convergent evolution of oxygen transport hemoglobins in jawed and jawless vertebrates. *Proceedings of the National Academy of Sciences*, 107(32):14274–14279.

[25] Hoffmann, F. G. and Storz, J. F. (2007). The αd-globin gene originated via duplication of an embryonic α-like globin gene in the ancestor of tetrapod vertebrates. *Molecular biology and evolution*, 24(9):1982–1990.

[26] Hundahl, C., Fago, A., and Weber, R. E. (2003). Effects of water activity on oxygen-binding in high-molecular weight, extracellular invertebrate hemoglobin and hemocyanin. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, 136(1):83–90.

[27] Jue, R., Woodbury, N., and Doolittle, R. (1980). Sequence homologies among e. coli ribosomal proteins: evidence for evolutionarily related groupings and internal duplications. *Journal of molecular evolution*, 15(2):129–148.

[28] Kacar, B. and Gaucher, E. A. (2013). Experimental evolution of protein-protein interaction networks. *Biochemical Journal*, 453(3):311–319.

[29] Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics 1. *Annu. Rev. Genet.*, 39:309–338.

[30] Kratzer, J. T., Lanaspa, M. A., Murphy, M. N., Cicerchi, C., Graves, C. L., Tipton, P. A., Ortlund, E. A., Johnson, R. J., and Gaucher, E. A. (2014). Evolutionary history and metabolic insights of ancient mammalian uricases. *Proceedings of the National Academy of Sciences*, 111(10):3763–3768.

[31] Kristensen, D. M., Kannan, L., Coleman, M. K., Wolf, Y. I., Sorokin, A., Koonin, E. V., and Mushegian, A. (2010). A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics*, 26(12):1481–1487.

[32] Kuzniar, A., van Ham, R. C., Pongor, S., and Leunissen, J. A. (2008). The quest for orthologs: finding the corresponding gene across genomes. *Trends in Genetics*, 24(11):539–551.

[33] Loeb, A. (2012). How to nurture scientific discoveries despite their unpredictable nature. *arXiv preprint arXiv:1207.3812*.

[34] Lu, Q. and Fox, G. E. (2011). Resurrection of an ancestral 5s rrna. *BMC evolutionary biology*, 11(1):218.

[35] Lynch, M. (2013). Evolutionary diversification of the multimeric states of proteins. *Proceedings of the National Academy of Sciences*, 110(30):E2821–E2828.

[36] Marmer, M., Bjoern, L., Dogrultan, E., and Berman, R. (2011). Startup genome report. *Berkley University and Stanford University, Tech. Rep*.

[37] Miller, K. R. (2008). *Only a theory: Evolution and the battle for America's soul*. Penguin.

[38] Milo, R. (2013). What is the total number of protein molecules per cell volume? a call to rethink some published values. *Bioessays*, 35(12):1050–1055.

[39] MUSHEGIAN, A. (2005). Protein content of minimal and ancestral ribosome. *RNA*, 11(9):1400–1406.

[40] Mushegian, A. (2014). Reconstructing gene content in the last common ancestor of cellular life: is it possible, should it be done, and are we making any progress? *bioRxiv*.

[41] Nelson, D. L. and Cox, M. M. (2005). *Lehninger principles of biochemistry*. W. H. Freeman and Company.

[42] Noguchi, C. T., Schechter, A. N., Haley, J. D., and Abraham, D. J. (2003). Inhibition of sickle hemoglobin polymerization as a basis for therapeutic approaches to sickle-cell anemia. *Burger's medicinal chemistry and drug discovery*.

[43] Oneal, P. A., Gantt, N. M., Schwartz, J. D., Bhanu, N. V., Lee, Y. T., Moroney, J. W., Reed, C. H., Schechter, A. N., Luban, N. L., and Miller, J. L. (2006). Fetal hemoglobin silencing in humans. *Blood*, 108(6):2081–2086.

[44] Opazo, J. C., Hoffmann, F. G., and Storz, J. F. (2008a). Differential loss of embryonic globin genes during the radiation of placental mammals. *Proceedings of the National Academy of Sciences*, 105(35):12950–12955.

[45] Opazo, J. C., Hoffmann, F. G., and Storz, J. F. (2008b). Genomic evidence for independent origins of

β-like globin genes in monotremes and therian mammals. *Proceedings of the National Academy of Sciences*, 105(5):1590–1595.

[46] Pallen, M. J., Beatson, S. A., and Bailey, C. M. (2005). Bioinformatics, genomics and evolution of non-flagellar type-iii secretion systems: a darwinian perpective. *FEMS microbiology reviews*, 29(2):201–229.

[47] Pauling, L. and Zuckerkandl, E. (1963). Chemical paleogenetics. *Acta chem. scand*, 17:9–16.

[48] Perez-Jimenez, R., Inglés-Prieto, A., Zhao, Z.-M., Sanchez-Romero, I., Alegre-Cebollada, J., Kosuri, P., Garcia-Manyes, S., Kappock, T. J., Tanokura, M., Holmgren, A., Sanchez-Ruiz, J. M., Gaucher, E. A., and Fernandez, J. M. (2011). Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nat Struct Mol Biol*, 18(5):592–596.

[49] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004). Ucsf chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13):1605–1612.

[50] Piel, F. B., Patil, A. P., Howes, R. E., Nyangiri, O. A., Gething, P. W., Williams, T. N., Weatherall, D. J., and Hay, S. I. (2010). Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. *Nature Communications*, 1:104.

[51] Rees, D. C., Williams, T. N., and Gladwin, M. T. (2010). Sickle-cell disease. *Lancet*, 376:2018–31.

[52] Rochette, J., Craig, J., and Thein, S. (1994). Fetal hemoglobin levels in adults. *Blood reviews*, 8(4):213–224.

[53] Rodgers, D. W., Crepeau, R. H., and Edelstein, S. J. (1987). Pairings and polarities of the 14 strands in sickle cell hemoglobin fibers. *Proceedings of the National Academy of Sciences*, 84(17):6157–6161.

[54] Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. (2012). Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, 61(3):539–542.

[55] Root-Bernstein, M. and Root-Bernstein, R. (2015). The ribosome as a missing link in the evolution of life. *Journal of theoretical biology*, 367:130–158.

[56] Scott, M., Van den Berg, J., Repka, T., Rouyer-Fessard, P., Hebbel, R., Beuzard, Y., and Lubin, B. (1993). Effect of excess alpha-hemoglobin chains on cellular and membrane oxidation in model beta-thalassemic erythrocytes. *Journal of Clinical Investigation*, 91(4):1706.

[57] Shaked, N. T., Satterwhite, L. L., Telen, M. J., Truskey, G. A., and Wax, A. (2011). Quantitative microscopy and nanoscopy of sickle red blood cells performed by wide field digital interferometry. *Journal of biomedical optics*, 16(3):030506–030506.

[58] Stern, J. G. (2013). STORI: Selectable Taxon Ortholog Retrieval Iteratively. Master's thesis, Georgia Institute of Technology.

[59] Storz, J. F., Opazo, J. C., and Hoffmann, F. G. (2013). Gene duplication, genome duplication, and the functional diversification of vertebrate globins. *Molecular phylogenetics and evolution*, 66(2):469–478.

[60] Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997). A genomic perspective on protein families. *Science*, 278(5338):631–637.

[61] Tawfik, D. S. (2006). Loop grafting and the origins of enzyme species. *SCIENCE-NEW YORK THEN WASHINGTON-*, 311(5760):475.

[62] Tokuriki, N. and Tawfik, D. S. (2009). Protein dynamism and evolvability. *Science*, 324(5924):203–207.

[63] Vankayala, S. L., Hargis, J. C., and Woodcock, H. L. (2012). Unlocking the binding and reaction mechanism of hydroxyurea substrates as biological nitric oxide donors. *Journal of chemical information and modeling*, 52(5):1288–1297.

[64] Walkiewicz, K., Cardenas, A. S. B., Sun, C., Bacorn, C., Saxer, G., and Shamoo, Y. (2012). Small changes in enzyme function can lead to surprisingly large fitness effects during adaptive evolution of antibiotic resistance. *Proceedings of the National Academy of Sciences*, 109(52):21408–21413.

[65] Wall, D., Fraser, H., and Hirsh, A. (2003). Detecting putative orthologs. *Bioinformatics*, 19(13):1710–1711.

[66] Yang, Z. and Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*, 13(5):303–314.

[67] Zhao, Z.-M., Reynolds, A., and Gaucher, E. (2011). The evolutionary history of the catenin gene family during metazoan evolution. *BMC Evolutionary Biology*, 11(1):198.

[68] Zhou, Y., Asahara, H., Gaucher, E. A., and Chong, S. (2012). Reconstitution of translation from

thermus thermophilus reveals a minimal set of components sufficient for protein synthesis at high temperatures and functional conservation of modern and ancient translation components. *Nucleic acids research*, page gks568.

[69] Zhou, Y. and Landweber, L. F. (2007). Blasto: a tool for searching orthologous groups. *Nucleic acids research*, 35(suppl 2):W678–W682.

ENDNOTES

1. In practice we do the BLAST searches slightly differently from this description, but the principle is the same. We use PSI-BLAST to query a single FASTA file containing every sequence from every taxon against a single BLAST database containing every sequence from every taxon. As per the COGsoft README, this query should be performed in duplicate: once with low-complexity filtering turned on and once with it turned off. To the best of our knowledge, -num_descriptions and -num_alignments flags must be at least the number of taxa.

2. Not every sequence has a best hit; below a certain alignment score threshold, similarity is coincidental rather than evolutionary. Moreover, not every best hit is symmetric. Suppose that alpha-hemoglobin in a jawed vertebrate is related to a globin-like protein in a jawless vertebrate (who doesn't have alpha/beta paralogs). However, jawed beta-hemoglobin is also related to jawless globin, and the beta-hemoglobin/globin sequence alignment has a higher similarity score than the alpha-hemoglobin/globin alignment. In this scenario, the jawless globin will be the best hit for jawed alpha, but the jawed beta will be the best hit for the jawless globin.

3. A separate class of methods relies on phylogenetic reconstruction, but we do not know of a way to automate such approaches.

5. For simplicity we assume that the misplaced alpha does not get reassigned to a third quasi-family of alpha-hemoglobins.

6. Horse alpha against horse returned horse alpha; dog mu against horse returned horse alpha; and cow mu against horse returned horse zeta. Depending on the taxa, mu can be most similar to zeta, or most similar to alpha.

7. More precisely, some flagellar proteins are homologous with some type-III secretion system proteins [21]. The story is complex and merits further study [46].

10. DNA evolves over time. Since DNA codes for proteins, proteins also evolve. An A to G substitution occurs more frequently than an A to C substitution (i.e., relative substitution rates depend on which character is replacing which). Moreover, different positions along the DNA chain have differing importance to the biological function of the gene product. As a result, the rate at which any mutation occurs depends on its position along the DNA chain. Important parts of a protein evolve slower than unimportant parts. Quantifying the similarity between a pair of DNA (or RNA, or protein) sequences from different species or different paralogs requires us to model how fast the positions are mutating relative to one another. Moreover, when a mutation does occur at a position, we need to model specifically which characters of the alphabet were involved; no matter the position, different types of mutation have different probabilities. In most cases we don't witness the evolution, so we can't know the positional rate heterogeneity, or the relative character substitution rates, or how the evolutionary distance between panda alpha-hemoglobin and cow alpha-hemoglobin compares to that between panda alpha and human alpha. All of these unknowns are parameters of our evolutionary model, and we are interested in finding the parameters under which the data (the aligned molecular sequences descended from a common ancestor) are most likely.

12. For example, alpha-hemoglobin and theta-hemoglobin in mice are co-orthologs with alpha-hemoglobin in flamingoes. This relationship exists because the ancestral mouse alpha duplicated to produce ancestral mouse theta, before the present day but after mice and flamingoes diverged [25]. Alpha and theta in present-day mice are descendants of alpha and theta from the mouse ancestor. Since the duplication occurred along the mouse branch, flamingoes inherited only alpha. In the last common ancestor of flamingo and mouse, only one gene – ancestral alpha – descended and modified into flamingo alpha, mouse alpha, and mouse theta.

14. We often refer to protein types, classes, or families in the singular. Depending on context, hemoglobin could mean a single molecule or all hemoglobin-type molecules. Moreover we've intentionally chosen the word subfamily rather than family to refer to a set of orthologs. A protein family

usually encompasses several paralogous subfamilies; e.g. alpha, beta, zeta, and epsilon hemoglobins are subfamilies of the hemoglobin family. Subfamily is still imprecise, since hemoglobin is a subfamily of globins and not all hemoglobin sequences are orthologs. The STORI source code uses ’family’ in many contexts where ’subfamily’ would have been less wrong.

15. Cellular machinery reads DNA three nucleotides at a time, and each triplet corresponds to one of the 20 amino acids. More precisely, 61 out of the 64 possible nucleotide triplets each codes for one of the 20 standard amino acids. The other three possible triplets (TAG, TGA, and TAA) tell the cellular machinery that it has reached the end of the protein. The triplet TAG indicates the beginning of a protein and also codes for the amino acid methionine. These codes are highly conserved throughout known life with a few exceptions.

16. However, evolution does not only solve problems; it also creates problems. As a case-in-point, many *healthy* people *without* sickle-cell anemia produce the sickling variant of beta-hemoglobin protein. These individuals possess the *sickle-cell trait*, meaning they have one copy of the normal beta-hemoglobin gene and one copy of the sickling gene. Because people with the sickle-cell trait produce both the sickling variant *and* the normal variant, their hemoglobin does not cause them health issues. The sickling beta-hemoglobin is actually an *adaptation* to protect people in places where malaria is common; this variant makes red blood cells inhospitable to the malaria parasite [20, 50]. Evolution solved one problem, and created another. People with the sickle-cell trait are less susceptible to malaria. However, children of parents who *both* carry the sickle-cell trait sometimes inherit the sickling hemoglobin gene from both parents; since these offspring are incapable of producing any normal beta-hemoglobin, they have sickle-cell anemia.