OPEN ACCESS Freely available online

PLoS one

Detecting Clusters of Mutations

Tong Zhou[1], Peter J. Enyeart[2], Claus O. Wilke[1,2]*

1 Center for Computational Biology and Bioinformatics, Section of Integrative Biology, University of Texas at Austin, Austin, Texas, United States of America, **2** Institute for Cell and Molecular Biology, University of Texas at Austin, Austin, Texas, United States of America

Abstract

Positive selection for protein function can lead to multiple mutations within a small stretch of DNA, i.e., to a cluster of mutations. Recently, Wagner proposed a method to detect such mutation clusters. His method, however, did not take into account that residues with high solvent accessibility are inherently more variable than residues with low solvent accessibility. Here, we propose a new algorithm to detect clustered evolution. Our algorithm controls for different substitution probabilities at buried and exposed sites in the tertiary protein structure, and uses random permutations to calculate accurate *P* values for inferred clusters. We apply the algorithm to genomes of bacteria, fly, and mammals, and find several clusters of mutations in functionally important regions of proteins. Surprisingly, clustered evolution is a relatively rare phenomenon. Only between 2% and 10% of the genes we analyze contain a statistically significant mutation cluster. We also find that not controlling for solvent accessibility leads to an excess of clusters in terminal and solvent-exposed regions of proteins. Our algorithm provides a novel method to identify functionally relevant divergence between groups of species. Moreover, it could also be useful to detect artifacts in automatically assembled genomes.

Citation: Zhou T, Enyeart PJ, Wilke CO (2008) Detecting Clusters of Mutations. PLoS ONE 3(11): e3765. doi:10.1371/journal.pone.0003765

Editor: Jason E. Stajich, University of California, Berkeley, United States of America

Received July 21, 2008; **Accepted** October 26, 2008; **Published** November 19, 2008

Funding: This work was supported by NIH grant R01 AI065960 and by a Reeder Centennial Fellowship in Systematic and Evolutionary Biology to C.O.W.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: cwilke@mail.utexas.edu

Introduction

Numerous methods have been proposed to identify positive selection in coding sequences [1–10]. These methods differ in their underlying assumptions, the data required to complete the analysis, and the type of conclusion that can be drawn. The most popular of these methods have in common that they are based in some form on the comparison of nonsynonymous to synonymous substitution frequencies, usually in the form of the ratio dN/dS. Their central premise is that synonymous substitutions are neutral and thus provide a baseline substitution rate to compare nonsynonymous substitutions against. Yet evidence is accumulating that synonymous substitutions often are not neutral. In particular, selection for translationally optimal codons operates from bacteria to mammals [11–17]. Other selective pressures on synonymous sites can arise from selection acting on mRNA secondary structure [18] or on exonic splicing enhancers [19]. For these and other reasons, dN/dS-based methods have been increasingly criticized [17,20–22]; in particular, a recent study showed that sites known to be under positive selection for function are often not identified by dN/dS methods and vice versa [23]. Methods to detect positive selection that do not rely on synonymous substitution rates, such as Fu's W [24] or Tajima's D [25], are generally based on allele frequencies and thus are sensitive to demographic events, e.g., recent population bottlenecks [26–28].

Wagner has recently proposed a new method to detect positive selection that uses neither synonymous substitution rates nor allele frequencies [10]. Wagner suggested that strong positive selection will lead to multiple substitutions in close proximity, that is, it will lead to a clustering of nonsynonymous mutations in sequence space. He developed a statistical method to identify such clusters of mutations, and identified several cases of strong clustering of mutations in a comparison of human and chimpanzee genes.

Wagner's method constitutes an innovative and novel approach to a longstanding problem. Unfortunately, it suffers from three limitations. First, the P value Wagner assigns to a mutation cluster generally underestimates the probability that the cluster would arise by chance if the null hypothesis were true. Second, by design, Wagner's method can detect at most one variation cluster per gene. Third, and most importantly, Wagner's method does not control for inhomogeneous substitution rates caused by protein structure. The solvent accessibility of a site influences its evolutionary conservation, with more exposed residues generally being less conserved [29–34], and a method that does not consider this difference in baseline selective constraints in its null distribution will tend to find spurious mutation clusters in solvent-exposed regions of proteins.

In this study, we propose a novel method to detect mutation clusters that alleviates all three drawbacks. We use this method to locate mutation clusters in three groups of species: bacteria (*Escherichia coli* vs. *Salmonella enterica*), fly (*Drosophila melanogaster* vs. *Drosophila obscura*), and mammals (primates vs. rodents). We analyze the properties of the clusters we find and discuss how some of these clusters may affect protein function.

Results

A novel algorithm to detect mutation clusters

We begin by briefly reviewing Wagner's approach [10]. Wagner based his method on the probability $p(d_{i,k})$ that, under the null hypothesis that all sites are equally likely to be mutated, k mutations arise within $d_{i,k}$ or fewer residues, starting at the position of mutation i. The probability $p(d_{i,k})$ can be calculated from the

gamma distribution. For a given gene, Wagner calculated $p(d_{i,k})$ for all possible contiguous sets of mutations in that gene, found the set with the minimum $p(d_{i,k})$, and referred to this set as the gene's mutation cluster. He used the $p(d_{i,k})$ value of this cluster, that is, $P_P = \min_{i,k} p(d_{i,k})$, as the cluster's P value.

Wagner's approach has two (arguably minor) statistical problems. First, because of the minimization procedure, P_P is not the probability that the associated cluster would arise if the null hypothesis were true. The probability $p(d_{i,k})$ measures the likelihood that, under the null hypothesis, a randomly chosen contiguous set of k mutations falls within at most $d_{i,k}$ residues. Consequently, P_P underestimates the probability that the most-clustered set of mutations (i.e., the cluster corresponding to $\min_{i,k} p(d_{i,k})$) falls within at most $d_{i,k}$ residues by chance alone. We can make this reasoning more intuitive by considering the general situation of a set of multiple events that occur according to some probability distribution. The most extreme of these events has a higher probability of being extreme than each event has individually. Second, by focusing on the cluster with the minimum $p(d_{i,k})$, Wagner can never detect more than one cluster per gene, even if a second, highly significant cluster is present.

It would be straightforward to fix these two statistical problems with a minor modification to Wagner's approach. But we are here primarily interested in a third, more fundamental limitation. We believe that the null hypothesis of a single, homogeneous mutation probability throughout the protein does not reflect biological reality and will lead to spurious mutation clusters. It is well known that amino-acid substitution rates correlate with solvent accessibility [29–33,35]. Substitutions at buried sites are more likely to be disruptive than substitutions at exposed sites, and are therefore more strongly selected against. If we don't control for this effect when searching for mutation clusters, we are likely to identify clusters in highly variable and relatively unimportant loop regions. It is unlikely that such clusters represent positive selection; they simply represent regions of weak selective constraint.

We now describe a method to detect mutation clusters that controls for solvent accessibility and that does not suffer from the two statistical problems outlined above. Instead of building our algorithm on the probability that, under the null hypothesis, k mutations arise within n or fewer residues, we consider instead the probability that k or more mutations fall within exactly n residues. This probability is binomial. (See Methods for details. Note that we use n instead of Wagner's d throughout the remainder of this paper.) By keeping the number (and location) of the residues fixed, we can easily generalize our algorithm to situations where different sites have different mutation probabilities. In the present work, we distinguish only between buried and exposed sites, but more complicated models would be feasible.

Our algorithm assumes that we are given two pieces of information for each gene to be analyzed: the location of all amino-acid mutations in the gene, and the solvent accessibility (measured as either buried or exposed) of each residue in the translated and folded protein. We then calculate the fraction of mutations for buried and exposed sites, f_b and f_e, and use these values to parameterize our binomial model. Thus we calculate the probability $q(k; n_e, n_b, f_e, f_b)$ to observe at least k mutations within a given stretch of n residues composed of n_b buried and n_e exposed residues (Eq. 3 in Methods). We calculate $q(k; n_e, n_b, f_e, f_b)$ for all possible contiguous sets of mutations (i.e., possible clusters) in the gene, and for each mutation, record the minimum-$q(k; n_e, n_b, f_e, f_b)$ value of all possible clusters starting with this mutation. We refer to the minimum-$q(k; n_e, n_b, f_e, f_b)$ value as Q, and to the total set of Q values as Q-landscape (Fig. 1). Local minima in the Q-landscape correspond to potential mutation clusters. We then discard all potential clusters that overlap with any other potential cluster with lower Q (Fig. 1).

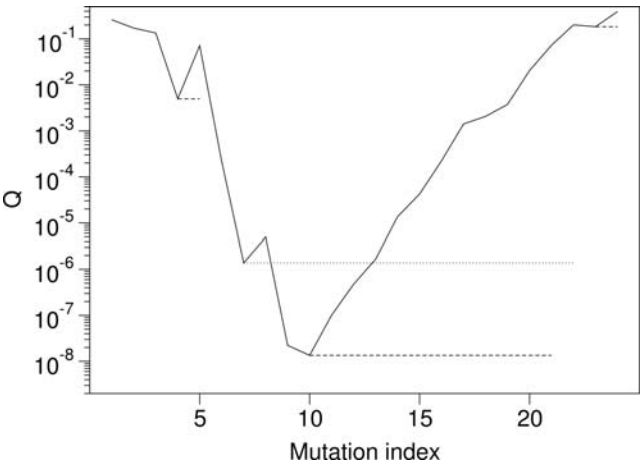

Figure 1. Q-landscape of the *E. coli* gene *frdA* (fumarate reductase flavoprotein subunit). Q has four local minima. Thus we have four potential clusters, starting at mutation numbers 4, 7, 10, and 23. The horizontal lines show the range of each potential cluster. The second cluster (dotted line) overlaps with the third cluster, which has lower Q. Therefore, we exclude the second cluster and obtain three potential mutation clusters (dashed lines). After correction for multiple testing, only one significant cluster remains, the one starting at position 10.
doi:10.1371/journal.pone.0003765.g001

We now have a set of potential clusters for the gene, and the next step is to calculate a P value for each cluster. For a given cluster for which we want to calculate P, we use the Q value defined above as test statistic, and denote it as Q_c. We then randomly reshuffle the mutations in the gene, repeat our analysis of finding non-overlapping clusters at minima in the Q-landscape, and record the frequency with which $Q_c < Q_c$. This frequency is the cluster's P value.

Because we are finding many potential clusters (we may find multiple clusters per gene, and we are analyzing hundreds of genes), we use the false-discovery-rate correction [36] to correct for multiple testing. We refer to the corrected P value as P_M, and to the uncorrected P value as P_U. Throughout this work, we consider potential clusters with $P_M < 0.05$ as significant.

Mutation clusters in bacteria, fly, and mammals

In principle, we can apply our method to any pair of orthologous sequences, such as a human sequence and the corresponding ortholog in macaque. But when we carried out genome-wide scans for clusters of mutations between pairs of species, we found numerous clusters that, upon closer inspection, appeared to be artifacts in one of the species. In particular, we found numerous clusters in the macaque genome that seemed to stem from errors in the assembly of the draft genome rather than representing true sequence differences (data not shown). Therefore, we decided to compare pairs of species and considered only those mutations that were conserved within each pair but differed among pairs.

We carried out scans for mutation clusters in bacteria (two species of *E. coli* compared to two species of *S. enterica*), fly (two species of the group *D. obscura* compared to two species of the group *D. melanogaster*), and mammals (two species of primates compared to two species of rodents). See Fig. 2 for details. To obtain the solvent accessibility data required for our analysis, we mapped all sequences to homologous sequences with known

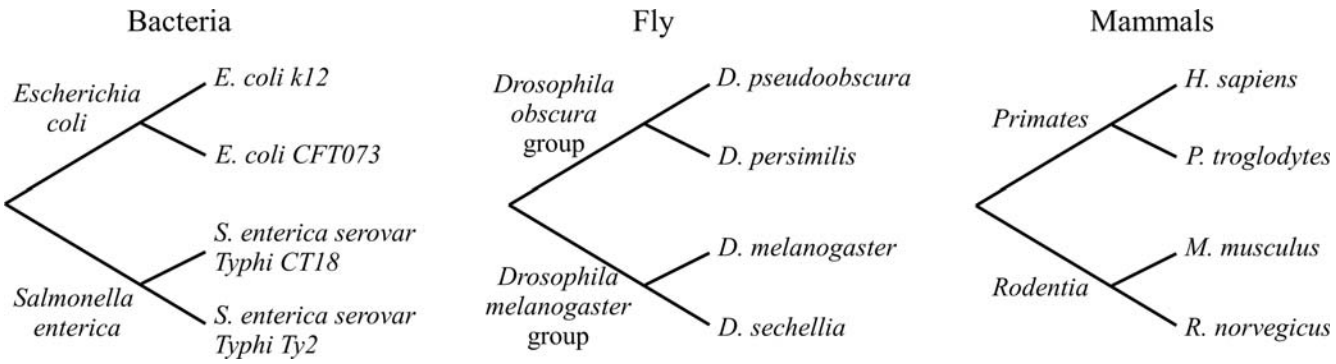

Figure 2. Species considered in this work. For each set of four species, we only considered mutations that were conserved within the upper and lower branch but differed between these two branches, and searched for clustered occurrences of these mutations. Branch lengths are not to scale.
doi:10.1371/journal.pone.0003765.g002

structure in the PDB, and discarded genes for which we could not find any reliable structure information. The final data sets contained 356, 99, and 246 genes for bacteria, fly, and mammals, respectively.

Controlling for differences in evolutionary rate at buried and exposed sites as described above, we found a total of 1255, 246, and 868 potential mutation clusters in bacteria, fly, and mammals. Of these, in bacteria there were 31 significant clusters after correction for multiple testing. There were 181 clusters for which the uncorrected P_U was less than 5%. In fly, there were 6 significant clusters (48 with $P_U<0.05$). In mammals, there were 5 significant clusters (87 clusters with $P_U<0.05$).

Statistical properties of mutation clusters

We next analyzed whether the mutation clusters differed in some aspect from the protein regions that did not display clustered mutations. First, we considered solvent accessibility. We calculated the fraction of mutations at buried sites within and outside of mutation clusters, and carried out a paired *t*-test to determine whether clustered mutations were more or less buried than non-clustered mutations. We jointly considered all mutation clusters for all species groups, as long as there was at least one mutation in the gene outside the cluster, and found a mean difference in the fraction of buried sites in clustered and non-clustered mutations of 0.061 ($P=0.077$, $n=41$). Thus, after controlling for solvent accessibility, mutation clusters are roughly equally likely to appear in buried or in exposed regions of the protein.

Second, we tested whether mutation clusters were associated predominantly with specific secondary structure motifs. We computed the fraction of sites with secondary structure of the types helix, sheet, turn, and coil, both inside and outside of mutation clusters, and found no significant differences (paired *t*-test $P=0.855$ for helix, $P=0.392$ for sheet, $P=0.882$ for turn, and $P=0.454$ for coil, $n=42$). Therefore, secondary structure composition does not seem to affect the location of mutation clusters.

Finally, we considered physicochemical distance for clustered and non-clustered mutations. Some authors have proposed that positive selection leads to physicochemically radical amino acid replacements [37,38] (but see [39–41]). Here, we considered five amino acid properties that have been found to correlate with rates of amino acid replacement [42,43]: composition of the side chain, polarity, and molecular volume [44], as well as hydropathy [45] and isoelectric point [46]. For each of these properties, we calculated for each gene the mean distance for mutations within a cluster and mutations outside of the cluster, and then tested for a non-zero mean distance using a paired *t*-test. We found one marginally significant result: mutations within clusters tend to have a more radical molecular-volume change than mutations outside of clusters ($P=0.024$, $n=41$), but the magnitude of the effect was small. The mean difference in the absolute volume change for mutations inside and outside of clusters was 3.78. The molecular-volume scale ranges from 3 (glycine) to 170 (tryptophan), with the majority of amino acids falling between 30 and 130 [44]. For the other four properties, differences were not significant (side chain composition, $P=0.816$; polarity, $P=0.157$; hydropathy, $P=0.134$; isoelectric point, $P=0.167$).

The effect of solvent accessibility on cluster location

Buried residues experience more purifying selection than exposed residues [29–33]. Therefore, an algorithm that doesn't control for solvent accessibility should find mutation clusters predominantly in exposed areas of the protein.

To determine the effect of solvent accessibility on the mutation clusters, we repeated our analysis but ignored protein structure, by artificially assigning to all residues the "buried" status in our cluster detection program. (We could have chosen the "exposed" status with identical results. What matters is that all sites have the same status.) In this case, we found 47 significant clusters in bacteria, 12 in fly, and 6 in mammals. We then calculated the fraction of buried sites within each significant cluster, and compared this fraction for clusters determined with and without controlling for solvent accessibility. We found that clusters determined without controlling for solvent accessibility tend to have fewer buried sites, i.e., are more exposed (two-sample *t*-test on pooled data from all three species groups, $P<0.001$, see also Fig. 3).

For bacteria, we also used our algorithm to calculate mutation clusters for all ORFs, regardless of whether we had protein structures for them or not, again artificially treating all residues as buried. We found 1070 significant clusters out of 13642 potential clusters. We then determined where within each coding sequence the mutation clusters were located, and found that the distribution of cluster locations along the coding sequence was not uniform (χ^2-test, $P<10^{-10}$). Clusters appeared more frequently on the termini (within 10% of total sequence length), as shown in Fig. 4. This result agrees with our hypothesis that solvent exposure can lead to spurious mutation clusters. Terminal regions of proteins (i.e., the N- and C-termini) are predominantly located on the protein surface and are exposed to the solvent [47–49]. By contrast, when controlling for solvent accessibility, we found only 2 out of 20 significant clusters within 10% (in terms of total sequence length)

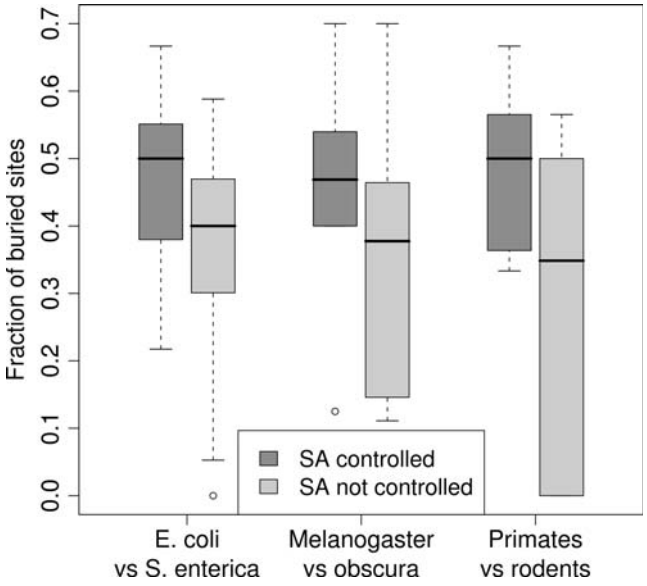

Figure 3. Fraction of buried sites in significant clusters obtained by either controlling or not controlling for solvent accessibility (SA). If solvent accessibility is not controlled for, many of the resulting clusters are located in exposed regions of the protein.
doi:10.1371/journal.pone.0003765.g003

of the C-terminus, and none out of 12 significant clusters within 10% of the N-terminus. For this analysis, we considered only those proteins for which the terminal region of interest was not truncated in the PDB structure. Therefore, we had only 20 clusters whose location we could determine relative to the C-terminus, and 12 relative to the N-terminus.

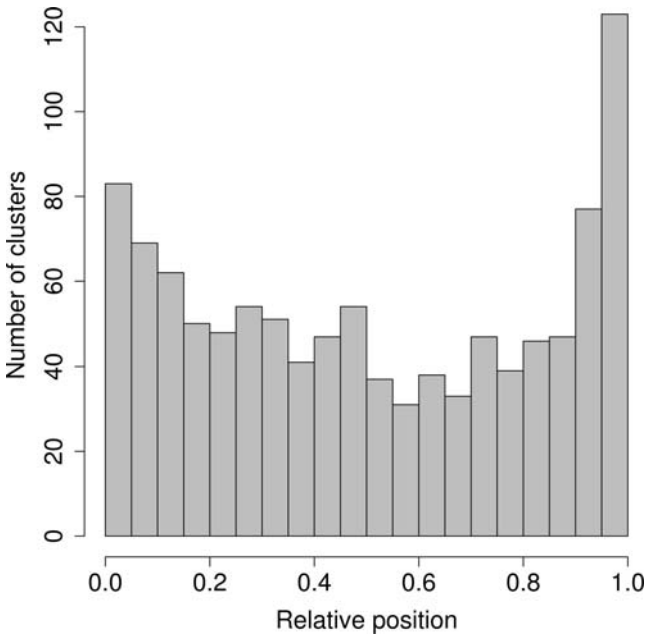

Figure 4. Distribution of cluster positions, for *E. coli* clusters found without controlling for solvent accessibility. The relative cluster position was calculated by dividing the cluster's central coordinate by the total sequence length. The cluster positions are not uniformly distributed, and clusters are most frequent in terminal regions of proteins.
doi:10.1371/journal.pone.0003765.g004

Example mutation clusters

Supplementary Tables S1, S2, and S3 list the significant mutation clusters in the three species groups. The mutation clusters span on average 9.2%, 7.3%, and 3.8% (numbers are for bacteria, fly, and mammals) of the coding regions in which they appear. Supplementary Figures S1.1–S1.39 show how each cluster maps onto the corresponding protein structure. The figures also show each cluster in the context of a multi-species sequence alignment. Supplementary Text S1 provides an overview over all supplementary materials.

We now discuss two examples of mutation clusters. The first example is the human enzyme carbonyl reductase 3 (CBR3, ENSG00000159231), which catalyzes the NADPH-dependent reduction of a variety of xenobiotic ketones and quinones [50,51]. The mutation cluster in CBR3 runs from position 239 to position 244. It is fully conserved within primates and within rodents, but differs at all amino acid positions between these two groups (Fig. 5). The same region is also highly variable in other species; the sequences of other vertebrates share little similarity with either the primate or the rodent sequence in the cluster region (Fig. 5).

The tertiary structure of CBR3 is shown in Fig. 6. In this protein, it is known that the peptide region in the C-terminal half constitutes the outer walls of the substrate-binding cleft of the active site, which provides specific interactions that are critical to the selectivity of substrates and to the mechanism of molecular recognition by the enzyme [52]. The mutation cluster detected by our method is located in the substrate entry-loop between β-sheet F (βF) and α-helix G (αG), which tightens on the substrate upon its entry into the active site providing additional substrate-specific interactions [52–55]. It is also reported that this entry-loop is tighter in human CBR3 in comparison with porcine testicular carbonyl reductase (PTCR), which shares about 70% sequence identity with CBR3 [56]. Conceivably, the mutation cluster influences the docking and/or release of the cofactor during enzymatic catalysis [56].

The second example is the β subunit of nitrate reductase A (NarH, b1225) in *E. coli*. The mutation cluster runs from position 133 to position 164, and is completely conserved within both *E. coli* and *S. enterica* (Fig. 7). Among the two groups, the cluster region has 66% sequence similarity, while the entire gene has 93% sequence similarity. *Shigella* sequences in the cluster region are identical to the *E. coli* sequences (Fig. 7), in agreement with the notion that *Shigella* strains are clones of *E. coli* [57].

E. coli can use nitrate as an electron acceptor for anaerobic growth [58,59]. This oxidoreduction is catalyzed by nitrate reductase A (NarGHI), which is a membrane-bound complex of three subunits coded by three genes, NarG, NarH, and NarJ. NarH is an [Fe-S]-cluster-containing electron transfer subunit [59,60]. The mutation cluster found in NarH is located in the motif which is thought to have an important function in defining subunit-subunit interactions within the overall structure of NarGHI and to provide additional shielding of [Fe-S] clusters from the aqueous milieu [60] (Fig. 8).

Other clusters that were readily determined to be in locations important to the structure and/or function of the protein are as follows. The cluster in the *E. coli* gene *pepN* lies in the substrate-recognition domain of the protein [61]. The cluster in the fly gene *CkIIβ2* partly overlaps with an acidic loop that is important for modulating autophosphorylation and the overall activity of the protein [62]. The cluster in the mammalian gene *NEDD8* is completely conserved in all vertebrates but rodents. Non-rodent vertebrates have a lysine at position four which forms a salt bridge with the glutamic acid at position 12 [63]. In rodents, the lysine is

Figure 5. Multiple sequence alignment of the human protein CBR3 and its orthologs in chimpanzee, macaque, mouse, rat, cow, cat, chicken, zebrafish and *Xenopus tropicalis*. The mutation cluster spans from position 239 to position 244 and is marked by the symbol X.
doi:10.1371/journal.pone.0003765.g005

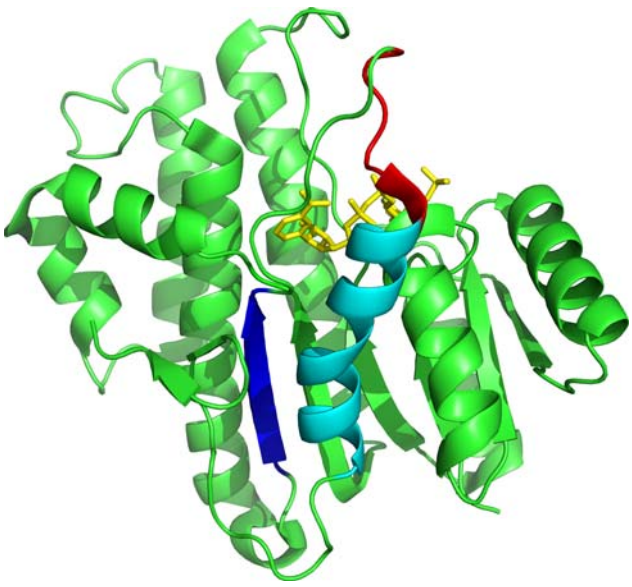

Figure 6. The tertiary structure of CBR3. The mutation cluster is shown in red, β-sheet F is shown in blue, and α-helix G is shown in cyan, while the remainder of the protein is shown in green. Coenzyme NADPH is shown in yellow.
doi:10.1371/journal.pone.0003765.g006

replaced by another glutamic acid, which disrupts the salt bridge and likely alters the protein structure.

Comparison with dN/dS-based methods

As discussed in the introduction, the most commonly used methods to detect positive selection rely on high dN/dS values. Therefore, for genes for which we found clusters, we also carried out dN/dS-based analyses.

First, for the mammalian clusters, we determined whether our mutation clusters coincided with sites predicted to have elevated dN/dS according to Bayes Empirical Bayes Inference [9], as published in the Human PAML Browser [64] (http://mendel.gene.cwru.edu/adamslab/pbrowser.py). Of the five genes for which we identified mutation clusters, the PAML Browser contained results for only three (Ensembl IDs ENSG00000105220, ENSG00000129559, ENSG00000198951). For neither of these did we find that mutation clusters overlapped with sites predicted to have dN/dS.

Second, we compared our results to results obtained by a 3D sliding window method [7,65,66]. We carried out this analysis using the SWAKK web server [66] (http://oxytricha.princeton.edu/SWAKK/), using a 3D window size of 10Å. Because this method can only work on pairs of sequences, we compared *E. coli* K12 with *S. enterica* CT18 for bacteria, *D. melanogaster* with *D. persimilis* for fly, and human with mouse for mammals. As with Bayes Empirical Bayes Inference, the mammalian clusters did not overlap with regions predicted to have $dN/dS>1$. In contrast, eight of the bacterial clusters and one fly cluster coincided with regions with $dN/dS>1$ (Fig. 9).

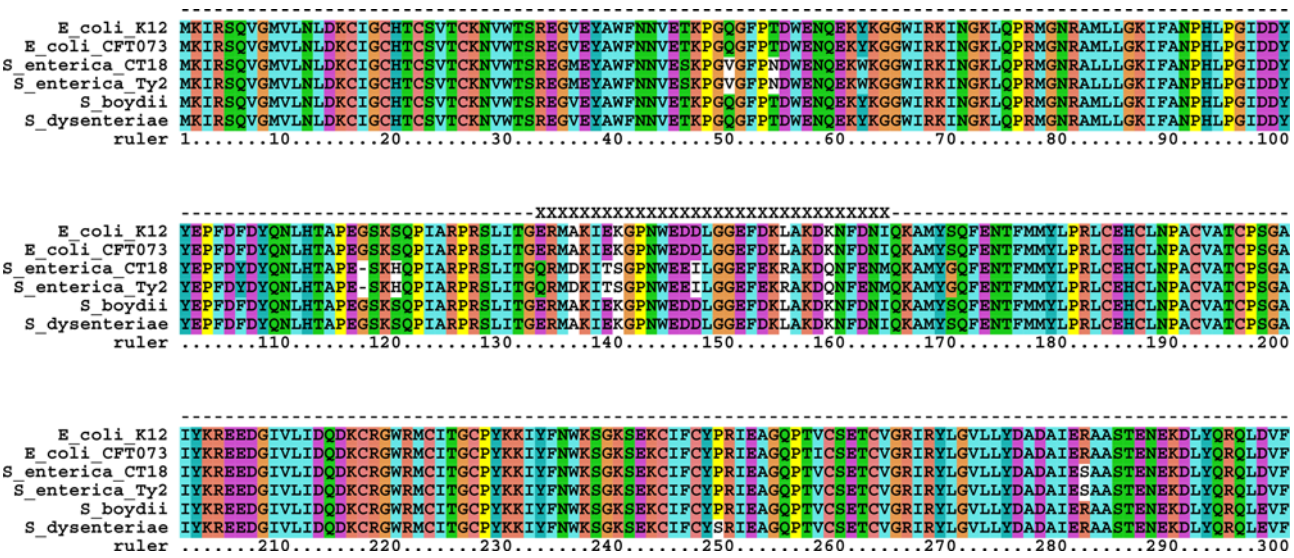

Figure 7. Multiple sequence alignment of the *E. coli* protein NarH and its orthologs in *S. enterica*, *Shigella boydii* and *Shigella dysenteriae*. The mutation cluster spans positions 133 to 164 and is marked by the symbol X.
doi:10.1371/journal.pone.0003765.g007

Discussion

We have presented a new method to discover clustered evolution in protein-coding sequences. Our method takes into account the increased variability of exposed residues relative to buried residues and finds clusters of mutations that are unlikely to have arisen by chance given their composition of buried and exposed residues. The method can find multiple clusters in a single gene, and uses permutation tests to assign accurate *P*-values to each cluster.

We have used this method to search for mutation clusters in bacteria, fly, and mammals. By and large, we have found that mutation clusters are not particularly common. We found a total of 31 clusters in 356 bacterial genes, 6 clusters in 99 fly genes, and 5 clusters in 246 mammalian genes. However, some of the clusters we found were striking. For example, in the *E. coli* fumarate reductase flavoprotein subunit (FrdA, b4154), nearly half of the sequence differences relative to the *S. enterica* ortholog fall into the mutation cluster, which nonetheless spans only 5% of the entire protein. Several of the clusters we have identified seem to be located in or adjacent to active sites or otherwise functionally relevant regions of the protein. We therefore expect that a good fraction of the clusters we found reflect functional divergence between the species groups we compared. Unfortunately, we did not find a single example where the corresponding protein had been experimentally characterized in both species. Therefore, at this point we can only speculate about the meaning of the clusters.

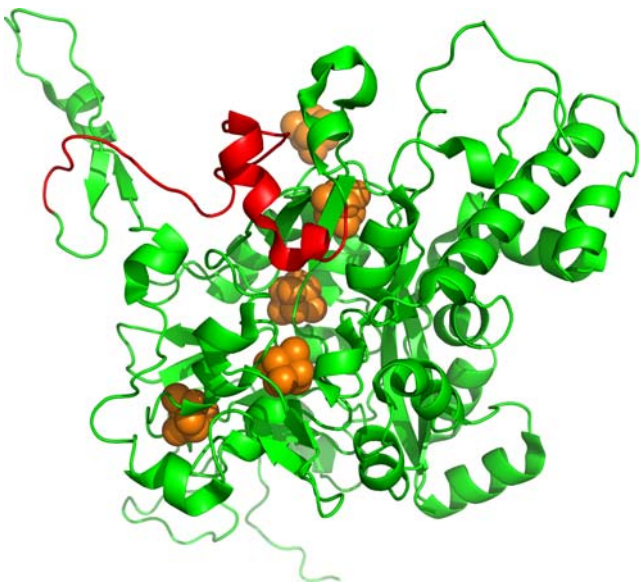

Figure 8. The tertiary structure of NarH. The mutation cluster is shown in red, while the remainder of the protein is shown in green. [Fe-S] clusters are shown in orange.
doi:10.1371/journal.pone.0003765.g008

By controlling for solvent accessibility, we avoid detecting spurious clusters that only reflect inherent variability differences along the protein sequence. Yet controlling for solvent accessibility does not preclude the possibility that clusters arise more frequently in either buried or exposed regions. For example, if mutations in exposed regions tended to be distributed uniformly along the protein sequence whereas mutations in buried regions tended to be clustered together, we would find an excess of mutation clusters in buried regions even after controlling for solvent accessibility. We therefore tested whether mutation clusters were particularly likely to appear in either buried or exposed regions, and found no such signal. Neither did we find a propensity of clusters to appear in specific secondary structure elements.

We also tested whether mutations in clusters were more physicochemically radical. We found a weak signal for molecular volume, but no signal for side-chain composition, polarity, hydropathy, or isoelectric point. This result seems to support the notion that adaptive evolution does not coincide with more radical amino-acid replacements [39–41]. Because we had a relatively small data set of only 42 significant mutation clusters, we had limited statistical power to detect differences between mutations inside and outside of clusters. Therefore, our results do not imply that clustered mutations are completely indistinguishable from non-clustered mutations. However, they do imply that any difference between these two types of mutations must be minor.

We have found that controlling for solvent accessibility is crucial to avoid detecting clusters that simply reflect highly variable loop

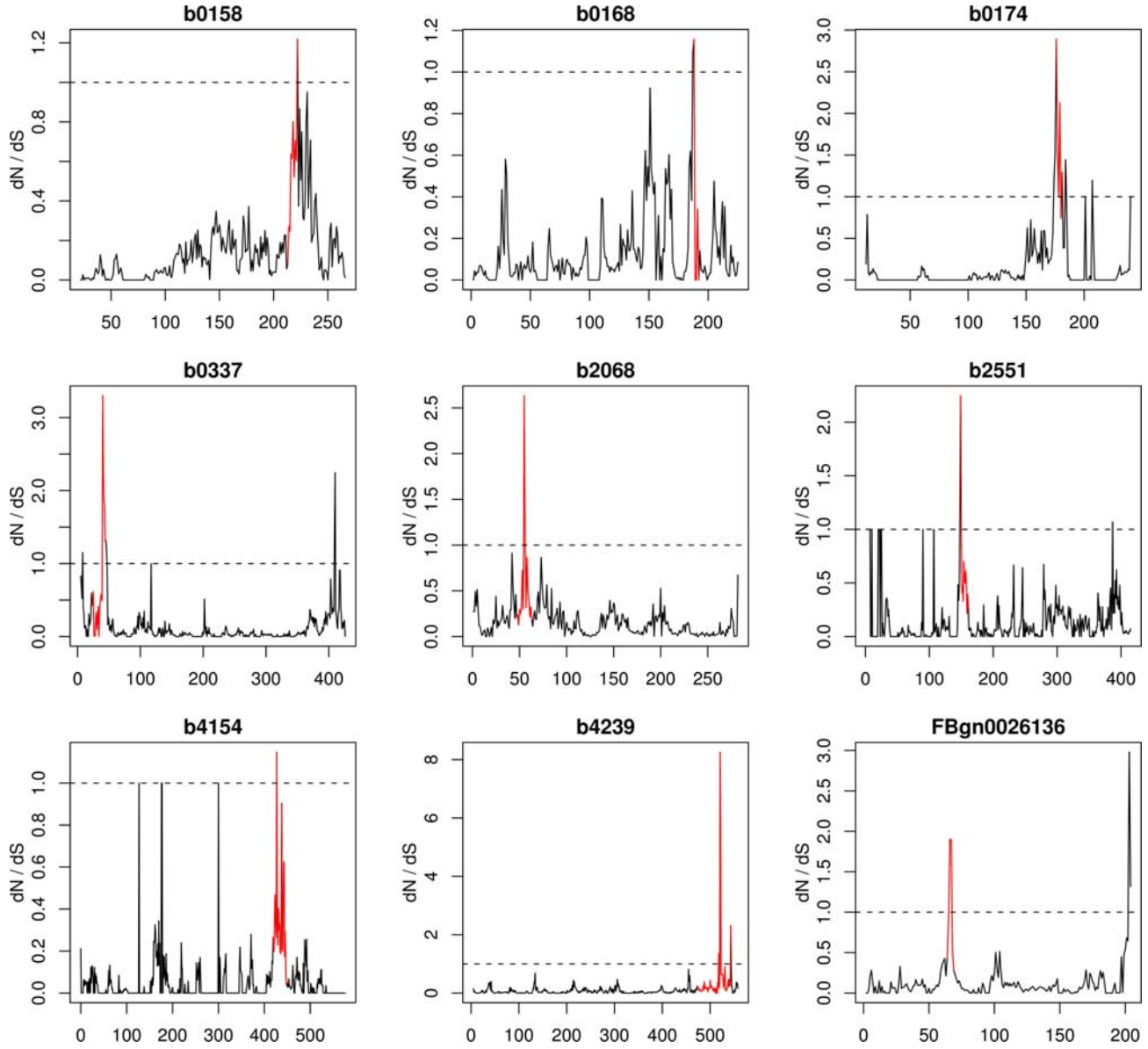

Figure 9. *dN*/*dS* in 3D window versus the residue number at the center of the 3D window. The red coloration indicates residues which we identified as being part of a mutation cluster. The dashed line indicates *dN*/*dS* = 1. We show 3D-window analyses for all cases in which a mutation cluster we identified coincided with a *dN*/*dS*-value above 1. There are eight such cases for bacteria (b0158–b4239), one in fly (FBgn0026136), and none in mammals.
doi:10.1371/journal.pone.0003765.g009

regions or other highly variable exposed regions of the protein. If we do not control for solvent accessibility, we find an excess of mutation clusters in highly variable terminal regions of the proteins, and we find clusters that are predominantly comprised of solvent-exposed residues.

One issue we encountered repeatedly when we initiated this work was the emergence of spurious clusters in two-species comparisons. The most extreme case arose in a comparison of human to macaque genes, where we found multiple putative clusters that could be traced to problems with the macaque draft genome sequence. To avoid such problems, we decided to base our analysis on a comparison of pairs of species, and considered as mutated only those sites that were conserved within each pair but differed among pairs. Conversely, our method might be useful for quality control in the automated assembly of newly sequenced species. Any cluster that shows up in a pairwise comparison of a newly sequenced species and a closely related species should be considered suspicious. These clusters could then be double-checked manually for accuracy.

Our method has several limitations. First, we require solved protein structures for every gene we analyze. This requirement severely limits the size of the data sets we can analyze. One possibility to alleviate this limitation would be to use computationally predicted solvent accessibilities for those genes or parts of genes for which no solved protein structure is available. The drawback of this approach is that these computational predictions are typically only 70%–80% accurate [67–71], and it is not clear how incorrectly predicted solvent accessibility would affect the clusters found by our algorithm.

Second, and more importantly, our method finds clusters in the protein's primary structure (i.e., its sequence). Mutation clusters in

the primary structure will generally map to clusters in the tertiary structure [10], but the converse is not necessarily true. A cluster in tertiary structure can conceivably consist of mutations that are distant in primary structure. Such clusters would be missed by our method. It would, however, be straightforward to modify our method so that it does apply to 3D space. Instead of searching for clusters in consecutive stretches of the amino-acid sequence, we would have to consider spheres with varying radii centered around the mutations in the protein. Eq. 3 would still apply if we interpreted n_b and n_e as the total number of buried and exposed amino acids in the sphere, and f_b and f_e as the fraction of mutations at buried and exposed sites in the protein. All other aspects of our method would transfer to the 3D case without change.

Third, separating all residues into groups of either buried or exposed residues may be too coarse. Sequence conservation varies continuously with solvent accessibility [31], and hence there may still be significant variation in the mutation probabilities within all residues we considered exposed or buried. Moreover, we considered only the solvent accessibility in a protein's tertiary structure. However, residues that are solvent-exposed in tertiary structure but buried in quaternary structure tend to be more conserved than residues that remain always exposed [35,72,73]. In principle, all these drawbacks can be alleviated by introducing additional classes of residues, say partially exposed residues, or exposed residues in contact with other proteins. The problem with such an approach is that with any additional residue class that we introduce, it becomes harder to reliably estimate the mutation frequency within that class. One possibility would be to combine our approach with evolutionary trace methods [74]. Evolutionary trace methods aim to identify functional sites in proteins by locating regions of high sequence conservation in large multiple sequence alignments, whereas our approach does the opposite. It finds regions with unusually high sequence variability. It would be possible to use a method similar to the evolutionary trace to calculate a background variability of each site, and then use a method similar to ours to search for clusters of mutations that are particularly unlikely to arise under this background level of variation.

When comparing mutation clusters to results from *dN/dS*-based methods, we found that approximately 20% of the clusters we identified coincided with regions with $dN/dS>1$, while the remainder did not. What should we have expected for this comparison? One significant difference between mutation clusters and the *dN/dS*-based methods is that the latter use an absolute standard, i.e., they search for sites or regions with $dN/dS>1$, whereas our method finds regions in which *dN* is elevated compared to the rest of the gene. For example, if a gene has $dN/dS=0.01$ throughout, apart from a small region with $dN/dS=0.8$, and assuming that the difference in *dN/dS* is caused by a change in *dN* and not *dS*, the region with $dN/dS=0.8$ would likely be identified as a mutation cluster by our method but would not register in screens for $dN/dS>1$. On the other hand, because we identify mutation clusters purely based on nonsynonymous mutations, post-hoc testing for elevated *dN/dS* in cluster regions suffers from ascertainment bias. In other words, we expect to find cases with $dN/dS>1$ simply because of the way in which we carried out our analysis, and we would obtain this result even in simulated data sets generated with completely homogeneous substitution rates and without any positive selection.

Do the clusters we identify actually represent positive selection, or might they just reflect relaxed selective pressures? We concede that the latter is a realistic possibility. Even though we certainly removed some regions of relaxed selection by considering separately the more and less variable regions in each protein, we have no guarantee that the remaining clusters are not caused by relaxed selection. In fact, positive and relaxed selection can lead to very similar patterns of evolution. For instance, significant divergence in the active site of a protein could indicate adaptation to a new enzymatic function, but it could also indicate loss of function. An example of the latter case would be a protein whose main importance has become structural, as has happened with crystallins [75]. As recent work on the *dN/dS* method has shown [23], reliable identification of positive selection by purely statistical methods is extremely difficult. For these reasons, we believe that the main purpose of our method is to identify unusual patterns of sequence divergence. The mechanisms by which these patterns arose will have to be determined separately, most likely by direct biochemical experimentation.

Materials and Methods

Genomic and structural data

For bacteria, we obtained orthologs between *E. coli K12*, *E. coli CFT073*, *S. enterica CT18*, and *S. enterica Ty2* from TIGR's Comprehensive Microbial Resource's multi-genome homology comparison tool (http://cmr.tigr.org/). For fly, we obtained orthologs between *D. melanogaster*, *D. sechellia*, *D. persimilis*, and *D. pseudoobscura* from the *Drosophila* 12-genome project AAAWiki at http://rana.lbl.gov/drosophila/. For mammals, we obtained orthologs between *H. sapiens*, *P. troglodytes*, *M. musculus*, and *R. norvegicus* from Biomart through the Ensembl Homology track (http://www.ensembl.org/). For each group of orthologs, we obtained aligned nucleotide sequences based on the alignment of the peptide sequences, which we generated with MUSCLE [76]. We excluded from our data set those ortholog pairs for which less than 80% of either sequence could be aligned to the other sequence. Then we determined from the alignments the number and coordinates of all amino acid changes that had occurred between the species pairs of each group (bacteria, fly, and mammals). In other words, we considered only mutations shared by the species pairs. We excluded from this count all sites at which at least one sequence had an indel (alignment gap). For genes with multiple transcripts, we based our analysis on the longest transcript that could be aligned to a PDB structure (see next paragraph). Moreover, to be conservative, we considered only those sites as mutated for which no transcript in one species pair agreed with any transcript in the other species pair.

We matched sequences to protein structures using the GTOP (Genomes TO Protein structures and functions) database [77]. For a given match in the GTOP database, if the region of similarity was longer than 80% of the protein length and the sequence identity was larger than 40% of the sequence in the Protein Data Bank (PDB), the match was saved for further calculation. This process yielded 777, 795, and 860 matches in *E. coli*, *D. melanogaster*, and *H. sapiens*, respectively.

For each protein with a match, the corresponding 3D structural information was obtained from the PDB. We aligned the orthologs plus the sequence of the corresponding PDB structure with MUSCLE, and then calculated the percent solvent-accessible surface area for each orthologous residue position using the DSSP (Dictionary of Protein Secondary Structure) program [78]. We normalized these results by the reference surface areas of an extended Gly-X-Gly peptide [79]. We considered residues with less than 25% relative solvent accessibility as buried. We also calculated the secondary structure for each aligned residue position using the DSSP program [78]. We simplified our data set by keeping track of only four types of secondary structure elements: helix (DSSP class H), sheet (DSSP class E), turn (DSSP classes S and T), and coil (DSSP classes B, G, I, and '.').

We excluded from our analysis those alignments in which there was at least one site without known solvent accessibility. Our final datasets contained 356, 99, and 246 orthologs for bacteria, fly, and mammals, respectively.

Computational method for cluster detection

Under neutrality, all sites in a protein of length l with m amino acid mutations are equally likely to have been mutated. Therefore, the m mutations should be evenly distributed over the entire protein. In this case, the probability of getting exactly k mutations in n successive residues is given by the binomial distribution,

$$p(k;n,f)=\binom{n}{k}f^k(1-f)^{n-k}, \quad (1)$$

where $\binom{n}{k}$ is the binomial coefficient, and we define $f=m/l$. The probability $q(k; n, f)$ that the number of mutations in n successive residues is no less than k is equal to

$$q(k;n,f)=1-\sum_{i=0}^{k-1}p(i;n,f)=1-I_{1-f}(n-k+1,k), \quad (2)$$

where $I_{1-f}(n-k+1, k)$ is the regularized incomplete beta function [80].

Now assume that the protein is subdivided into buried and exposed residues, and that the mutation probability differs among these two classes of residues. We denote the number of exposed residues by l_e and the number of buried residues by l_b, with $l_e+l_b=l$. Assume that there are m_e mutations at exposed sites and m_b mutations at buried sites, with $m_e+m_b=m$. Then, for a stretch of n residues with n_e exposed residues and n_b buried residues ($n=n_e+n_b$), the probability that these n residues contain at least k mutations, given that mutations are equally likely at all exposed and all buried sites, becomes

$$\begin{aligned} &q(k;n_e,n_b,f_e,f_b)\\ &=1-\sum_{i=0}^{k-1}\left[p(i;n_e,f_e)\sum_{j=0}^{k-1-i}p(j;n_b,f_b)\right]\\ &=1-\sum_{i=0}^{k-1}\left[\binom{n_e}{i}f_e^i(1-f_e)^{n_e-i}I_{1-f_b}(n_b-k+1+i,k-i)\right], \end{aligned} \quad (3)$$

where $f_e=m_e/l_e$ and $f_b=m_b/l_b$.

Using Eq. 3, we calculate $q(k; n_e, n_b, f_e, f_b)$ for all possible contiguous sets of mutations. Assume that the m mutations are located at positions $x_1, x_2, \ldots, x_m$ in the gene. We first consider all possible clusters starting with the first mutation at position x_1. The corresponding sets of mutations are $\{x_1, x_2\}$, $\{x_1, x_2, x_3\}$, …, $\{x_1, \ldots, x_m\}$. The set with the minimum $q(k; n_e, n_b, f_e, f_b)$ is recorded as C_1, and the corresponding $q(k; n_e, n_b, f_e, f_b)$ as Q_1. We then repeat this procedure for sets starting at position x_2, x_3, and so on. This procedure yields mutation sets $C_2, C_3, \ldots, C_{m-1}$ with associated minimum-$q(k; n_e, n_b, f_e, f_b)$ values $Q_2, Q_3, \ldots, Q_{m-1}$. The Q-landscape plots the Q_i values ($i=1, 2, \ldots, m-1$) against the corresponding mutation index i (see Fig. 1). Local minima in this landscape represent possible mutation clusters, and we discard all sets of mutations that overlap with other sets having lower Q values.

Q values are probabilities, but they do not correspond to the probability that a given cluster arises by chance in the context of the other mutations present in the gene. In other words, we cannot equate a cluster's Q value with the cluster's P value. We calculate P values by interpreting Q as our test statistic. (In this context, we add the subscript s to Q.) For a given cluster with test statistic Q_s in a given gene, we carry out at least 10^4 independent, random reshufflings of the mutations, keeping the number of mutations at buried and exposed sites constant. For each reshuffled set of mutations, we repeat our procedure of identifying non-overlapping sets of mutations starting at local minima in the Q-landscape, and record whether $Q_s<Q$ for these sets. The total fraction of times that $Q_s<Q$ is the cluster's P value. We refer to this value as P_U, because it has not been corrected for multiple testing. We then carry out a false-discovery-rate correction [36] on the P_U values for all potential clusters in a given species, and record the corrected values as P_M. Clusters with $P_M<0.05$ are significant and are unlikely to have arisen by chance.

We implemented this algorithm as a C program called "ClusterExplorer". The program's source code is available as part of the online supplementary materials for this paper.

Supporting Information

Text S1 Summary of supplementary materials.
Found at: doi:10.1371/journal.pone.0003765.s001 (0.02 MB PDF)

Table S1 Significant mutation clusters in bacteria.
Found at: doi:10.1371/journal.pone.0003765.s002 (0.03 MB PDF)

Table S2 Significant mutation clusters in fly.
Found at: doi:10.1371/journal.pone.0003765.s003 (0.03 MB PDF)

Table S3 Significant mutation clusters in mammals.
Found at: doi:10.1371/journal.pone.0003765.s004 (0.03 MB PDF)

Figures S1 Supporting figures S1.1–S1.39.
Found at: doi:10.1371/journal.pone.0003765.s005 (5.48 MB PDF)

Acknowledgments

We thank Eugene Koonin for alerting us to the pitfalls of pairwise sequence comparisons. The work benefited from computing resources provided by the Texas Advanced Computing Center at UT.

Author Contributions

Conceived and designed the experiments: TZ COW. Performed the experiments: TZ. Analyzed the data: TZ PJE COW. Contributed reagents/materials/analysis tools: TZ PJE. Wrote the paper: TZ PJE COW.

References

1. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 351: 652–654.
2. Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148: 929–936.
3. Suzuki Y, Gojobori T (1999) A method for detecting positive selection at single amino acid sites. Mol Biol Evol 16: 1315–1328.
4. Kreitman M (2000) Methods to detect selection in populations with applications to the human. Annual Rev Genomics Human Gen 1: 539–559.

5. Smith NGC, Eyre-Walker A (2002) Adaptive protein evolution in *Drosophila*. Nature 415: 1022–1024.
6. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837.
7. Suzuki Y (2004) Three-dimensional window analysis for detecting positive selection at structural regions of proteins. Mol Biol Evol 21: 2352–2359.
8. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, et al. (2005) Natural selection on protein-coding genes in the human genome. Nature 437: 1153–1157.
9. Yang Z, Wong WSW, Nielsen R (2005) Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol 22: 1107–1118.
10. Wagner A (2007) Rapid detection of positive selection in genes and genomes through variation clusters. Genetics 176: 2451–2463.
11. Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J Mol Biol 151: 389–409.
12. Akashi H (1994) Synonymous codon usage in *Drosophila melanogaster*: Natural selection and translational accuracy. Genetics 136: 927–935.
13. Stenico M, Lloyd AT, Sharp PM (1994) Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. Nucl Acids Res 22: 2437–2446.
14. Stoletzki N, Eyre-Walker A (2007) Synonymous codon usage in *Escherichia coli*: Selection for translational accuracy. Mol Biol Evol 24: 374–381.
15. Hirsh AE, Fraser HB, Wall DP (2005) Adjusting for selection on synonymous sites in estimates of evolutionary distance. Mol Biol Evol 22: 174–177.
16. Drummond DA, Raval A, Wilke CO (2006) A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol 23: 327–337.
17. Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134: 341–352.
18. Chamary JV, Hurst LD (2005) Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biology 6: R75.
19. Parmley JL, Chamary JV, Hurst LD (2006) Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol Biol Evol 23: 301–309.
20. Hughes AL (2007) Looking for Darwin in all the wrong places: The misguided quest for positive selection at the nucleotide sequence level. Heredity 99: 364–373.
21. Hughes AL (2008) The origin of adaptive phenotypes. Natl Acad Sci USA 105: 13193–13194.
22. Hughes AL, Friedman R (2008) Codon-based tests of positive selection, branch lengths, and the evolution of mammalian immune system genes. Immunogenetics 60: 495–506.
23. Yokoyama S, Tada T, Zhang H, Britt L (2008) Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. Proc Natl Acad Sci USA 105: 13480–13485.
24. Fu Y (1996) New statistical tests of neutrality for DNA samples from a population. Genetics 143: 557–570.
25. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.
26. Wall JD, Przeworski M (2000) When did the human population size start increasing? Genetics 155: 1865–1874.
27. Bamshad M, Wooding SP (2003) Signatures of natural selection in the human genome. Nature Rev Genet 4: 99–111.
28. Ometto L, Glinka S, De Lorenzo D, Stephan W (2005) Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. Mol Biol Evol 22: 2119–2130.
29. Koshi JM, Goldstein RA (1995) Context-dependent optimal substitution matrices. Protein Engineering 8: 641–645.
30. Goldman N, Thorne JL, Jones DT (1998) Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics 149: 445–458.
31. Mirny LA, Shakhnovich EI (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol 291: 177–196.
32. Dean AM, Neuhauser C, Grenier E, Golding GB (2002) The pattern of amino acid replacements in α/β-barrels. Mol Biol Evol 19: 1846–1864.
33. Bloom JD, Drummond DA, Arnold FH, Wilke CO (2006) Structural determinants of the rate of protein evolution in yeast. Mol Biol Evol 23: 1751–1761.
34. Choi SC, Hobolth A, Robinson DM, Kishino H, Thorne JL (2007) Quantifying the impact of protein tertiary structure on molecular evolution. Mol Biol Evol 24: 1769–1782.
35. Kim PM, Lu LJ, Gerstein MB (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. Science 314: 1882–1883.
36. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Royal Stat Soc B 57: 289–300.
37. Hughes AL, Ota T, Nei M (1990) Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. Mol Biol Evol 7: 515–524.
38. Zhang J (2000) Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. J Mol Evol 50: 56–68.
39. Dagan T, Talmor Y, Graur D (2002) Ratios of radical to conservative amino acid replacement are affected by mutational and compositional factors and may not be indicative of positive Darwinian selection. Mol Biol Evol 19: 1022–1025.
40. Smith NGC (2003) Are radical and conservative substitution rates useful statistics in molecular evolution? J Mol Evol 57: 467–478.
41. Hanada K, Gojobori T, Li WH (2006) Radical amino acid change versus positive selection in the evolution of viral envelope proteins. Gene 385: 83–88.
42. Xia X, Li WH (1998) What amino acid properties affect protein evolution? J Mol Evol 47: 557–564.
43. McClellan DA, McCracken KG (2001) Estimating the influence of selection on the variable amino acid sites of the cytochrome *b* protein functional domains. Mol Biol Evol 18: 917–925.
44. Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185: 862–864.
45. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157: 105–132.
46. Zimmerman JM, Eliezer N, Simha R (1968) The characterization of amino acid sequences in proteins by statistical methods. J Theor Biol 21: 170–201.
47. Christopher JA, Baldwin TO (1996) Implications of N and C-terminal proximity for protein folding. Proc Natl Acad Sci USA 102: 1053–1158.
48. Chung JJ, Shikano S, Hanyu Y, Li M (2002) Functional diversity of protein C-termini: more than zipcoding. Trends Cell Biol 12: 146–150.
49. Jacob E, Unger R (2006) A tale of two tails: why are terminal residues of proteins exposed? Bioinformatics 23: e225–e230.
50. Oppermann U (2007) Carbonyl reductases: The complex relationships of mammalian carbonyl- and quinone-reducing enzymes and their role in physiology. Annu Rev Pharmacol Toxicol 47: 293–322.
51. Matsunaga T, Shintani S, Hara A (2006) Multiplicity of mammalian reductases for xenobiotic carbonyl compounds. Drug Metab Pharmacokinet 21: 1–18.
52. Ghosh D, Sawicki M, Pletnev V, Erman M, Ohno S, et al. (2001) Porcine carbonyl reductase. structural basis for a functional monomer in short chain dehydrogenases/reductases. J Biol Chem 276: 18457–18463.
53. Ghosh D, Wawrzak Z, Weeks C, Duax W, Erman M (1994) The refined three-dimensional structure of 3 alpha,20 beta-hydroxysteroid dehydrogenase and possible roles of the residues conserved in short-chain dehydrogenases. Structure 2: 629–640.
54. Tanaka N, Nonaka T, Tanabe T, Yoshimoto T, Tsuru D, et al. (1996) Crystal structures of the binary and ternary complexes of 7 alpha-hydroxysteroid dehydrogenase from *Escherichia coli*. Biochemistry 35: 7715–7730.
55. Sawicki MW, Erman M, Puranen T, Vihko P, Ghosh D (1999) Structure of the ternary complex of human 17β-hydroxysteroid dehydrogenase type 1 with 3-hydroxyestra-1,3,5,7-tetraen-17-one (equilin) and $NADP^+$. Proc Natl Acad Sci USA 96: 840–845.
56. Sukhwinder SL, Ghosh D, Blanco JG (2005) Functional significance of a natural allelic variant of human carbonyl reductase 3 (CBR3). Drug Metab Dispos 33: 254–257.
57. Fukushima M, Kakinuma K, Kawaguchi R (2002) Phylogenetic analysis of *Salmonella*, *Shigella*, and *Escherichia coli* strains on the basis of the *gyrB* gene sequence. J Clin Microbiol 40: 2779–2785.
58. Ruiz-Herrera J, DeMoss JA (1969) Nitrate reductase complex of *Escherichia coli* k-12: participation of specific formate dehydrogenase and cytochrome b_1 components in nitrate reduction. J Bacteriol 99: 720–729.
59. Giordani R, Buc J (2004) Evidence for two different electron transfer pathways in the same enzyme, nitrate reductase a from *Escherichia coli*. Eur J Biochem 271: 2400–2407.
60. Bertero MG, Rothery RA, Palak M, Hou C, Lim D, et al. (2003) Insights into the respiratory electron transfer pathway from the structure of nitrate reductase A. Nat Struct Biol 10: 681–687.
61. Ito K, Nakajima Y, Onohara Y, Takeo M, Nakashima K, et al. (2006) Crystal structure of aminopeptidase N (proteobacteria alanyl aminopeptidase) from *Escherichia coli* and conformational change of methionine 260 involved in substrate recognition. J Biol Chem 281: 33664–33676.
62. Niefind K, Guerra B, Ermakowa I, Issinger OG (2001) Crystal structure of human protein kinase CK2: insights into basic properties of the CK2 holoenzyme. EMBO J 20: 5320–5331.
63. Rao-Naik C, delaCruz W, Laplaza JM, Tan S, Callis J, et al. (1998) The Rub family of ubiquitin-like proteins. J Biol Chem 273: 34976–34982.
64. Nickel GC, Tefft D, Adams MD (2008) Human PAML browser: a database of positive selection on human genes using phylogenetic methods. Nucleic Acids Res 36: D800–D808.
65. Berglund AC, Wallner B, Elofsson A, Liberles DA (2005) Tertiary windowing to detect positive diversifying selection. J Mol Evol 60: 499–504.
66. Liang H, Zhou W, Landweber LF (2006) SWAKK: a web server for detecting positive selection in proteins using a sliding window substitution rate analysis. Nucleic Acids Res 34: W382–W384.
67. Rost B, Sander C (1994) Conservation and prediction of solvent accessibility in protein families. Proteins 20: 216–226.
68. Thompson MJ, Goldstein RA (1996) Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. Proteins 25: 38–47.
69. Pollastri G, Baldi P, Fariselli P, Casadio R (2002) Prediction of coordination number and relative solvent accessibility in proteins. Proteins 47: 142–153.
70. Adamczak R, Porollo A, Meller J (2004) Accurate prediction of solvent accessibility using neural networks based regression. Proteins 56: 753–767.

71. Nguyen MN, Rajapakse JC (2005) Prediction of protein relative solvent accessibility with a two-stage SVM approach. Proteins 59: 30–37.
72. Mintseris J, Weng Z (2005) Structure, function, and evolution of transient and obligate protein–protein interactions. Proc Natl Acad Sci USA 102: 10930–10935.
73. Eames M, Kortemme T (2007) Structural mapping of protein interactions reveals differences in evolutionary pressures correlated to mRNA level and protein abundance. Structure 15: 1442–1451.
74. Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 257: 342–358.
75. Jörnvall H, Persson B, Du Bois GC, Lavers GC, Chen JH, et al. (1993) ζ-crystallin versus other members of the alcohol dehydrogenase super-family. FEBS Lett 322: 240–244.
76. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.
77. Kawabata T, Fukuchi S, Homma K, Ota M, Araki J, et al. (2002) GTOP: a database of protein structures predicted from genome sequence. Nucleic Acids Res 30: 294–298.
78. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577–2637.
79. Creighton TE (1992) Proteins: Structures and Molecular Properties. New York: Freeman, 2 edition. 142 p.
80. Abramowitz M, Stegun IA (1965) Handbook of Mathematical Functions. New York: Dover, 2 edition. 263 p.