Published in final edited form as:

J Mol Biol. 2010 March 12; 396(5): 1410–1421. doi:10.1016/j.jmb.2009.11.075.

Lineage-Specific Differences in the Amino Acid Substitution Process

Snehalata Huzurbazar¹, Grigory Kolesov², Steven E. Massey², Katherine C. Harris², Alexander Churbanov², and David A. Liberles^{2,*}

¹Department of Statistics, University of Wyoming, Laramie, WY 82071, USA

²Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA

Abstract

In Darwinian evolution, mutations occur approximately at random in a gene, turned into amino acid mutations by the genetic code. Some mutations are fixed to become substitutions and some are eliminated from the population. Partitioning pairs of closely related species with complete genome sequences by average population size of each pair, we looked at the substitution matrices generated for these partitions and compared the substitution patterns between species. We estimated a population genetic model that relates the relative fixation probabilities of different types of mutations to the selective pressure and population size. Parameterizations of the average and distribution of selective pressures for different amino acid substitution types in different population size comparisons were generated with a Bayesian framework. We found that partitions in population size as well as in substitution type are required to explain the substitution data. Selection coefficients were found to decrease with increasingly radical amino acid substitution and with increasing effective population size.

To further explore the role of underlying processes in amino acid substitution, we analyzed embryophyte (plant) gene families from TAED (The Adaptive Evolution Database), where solved structures for at least one member exist in the Protein Data Bank. Using PAML, we assigned branches to three categories: strong negative selection, moderate negative selection/ neutrality, and positive diversifying selection. Focusing on the first and third categories, we identified sites changing along gene family lineages and observed the spatial patterns of substitution. Selective sweeps were expected to create primary sequence clustering under positive diversifying selection. Co-evolution through direct physical interaction was expected to cause tertiary structural clustering. Under both positive and negative selection, the substitution patterns were found to be nonrandom. Under positive diversifying selection, significant independent signals were found for primary and tertiary sequence clustering, suggesting roles for both selective sweeps and direct physical interaction. Under strong negative selection, the signals were not found to be independent. All together, a complex interplay of population genetic and protein thermodynamics forces is suggested.

© 2009 Elsevier Ltd. All rights reserved.

*Corresponding author. liberles@uwyo.edu.

Edited by M. Sternberg

Present addresses: S. E. Massey, Department of Biology, University of Puerto Rico, San Juan, PR 00931, USA; K. C. Harris, School of Natural Sciences, University of California, Merced, CA 95344, USA; A. Churbanov, Department of Biology, New Mexico State University, Las Cruces, NM 88003, USA.

Supplementary data

Supplementary data associated with this article can be found in the online version, at doi:10.1016/j.jmb.2009.11.075

Keywords

molecular evolution; protein structure; sequence–structure relationships; population genetics; selection

Introduction

Amino acid substitution is a complex process generating the observed sequences of homologous genes in different genomes. A number of factors influence amino acid substitution, but these factors are rarely considered together. Protein function is one potential driver of amino acid substitution, but residues directly responsible for binding specificity or catalysis represent a small fraction of total amino acids. Proteins fold into stable three-dimensional (3D) structures, typically with a moderate energy gap to other conformations, and this is potentially important in protein evolution. DNA-level processes from mutation and recombination to population-level actions on the protein-encoding gene also affect sequence evolution. Finally, rates of amino acid substitution are correlated with protein expression levels and there is an ongoing discussion of the underlying mechanism driving this correlation. Explanations for the correlation include those based on structure (folding) and those based on binding functions and this interplay will ultimately be discussed. The main thrust of this article will be to examine the interplay between mutation, linkage, and the fixation process on the one hand with protein thermodynamics on the other hand.

Recently, Sasidharan and Chothia¹ have characterized the amino acid substitution process across genomes, finding qualitatively similar patterns of amino acid substitution. They extrapolate a major selective role for protein thermodynamics in amino acid substitution with no discussion of mutational or population genetic forces. In this view, substitution is driving proteins to “structural cul-de-sacs”² where a level of stability (linked to folding rate) is reached, generating sites that then remain highly conserved through evolution. In this thermodynamic hypothesis, there is no expectation of lineage-specific differences in the substitution processes, as evolution is ultimately driving proteins to thermodynamic optima. Sasidharan and Chothia¹ do observe an evolutionary distance dependence to the amino acid substitution process.

The thermodynamic view presented above is one expectation of the unknown relationship between selection and thermodynamic stability. Another view of protein sequence evolution links substitution with maintenance of an optimal stability range.³ In this view, compensatory substitution drives protein sequence evolution in maintaining this optimal range. This selectionist model is linked to a population genetic model, where larger effective population size organisms have stronger selection and smaller effective population size species might be expected to have a greater variance in substitution with less deterministic evolution, but differences are balanced by the strong degrees of selection.

Another view is of lagom (metastable) proteins where the interplay between structure and protein sequence evolution involves a minimum stability threshold, but no selection against proteins being too stable has been generated.⁴ Neutral processes generate a metastable stability function about the threshold where a folding nucleus might be under stronger negative (conservative) selection, but represent a smaller number of residues (see Ref. 5 for a discussion). In this case, effective population size has negligible effects on the shape of the energy distribution, but does not rule out a role for positive diversifying selection dependent on effective population size in protein substitution. Consistent with this view, proteins may walk through neutral networks that are punctuated by adaptive shifts that change the accessibility of sites for substitution. This behavior may relate to heterotachy.

Heterotachy⁶ is a class of statistical models that include parameterization for site-specific shifts in evolutionary rates over time. Guindon *et al.*⁷ and Blanquart and Lartillot⁸ have also recently published related models that enable site-specific evolutionary rates to shift in a lineage-specific manner. Ultimately, it is unclear what either the energetic rules or the structural and functional constraints underlying these heterogeneous models are. Heterotachy is described by Lopez *et al.*⁶ as a neutral process, but is used by Gu⁹ to predict functional shifts. It may be that the accessibility of sites for substitution changes neutrally, but changes faster under positive diversifying selection leading to a time-dependent transition from a rates across sites model that is sped up by positive diversifying selection. It may also be that the structural cul-de-sacs² presented in the thermodynamic view can be evolutionarily transient on long time scales, where structural transitions between alternative stabilizing elements separated by a large energy threshold represent modulators of evolutionary rate that eventually drive the rate shifts under negative selection observed by Lopez *et al.*⁶

From population genetic theory, effective population size is crucial in selecting for advantageous mutations and against deleterious ones. Mutations appear in populations in proportion to the effective population size and are fixed through the interplay of the selective effects of the mutation and the inverse of the effective population size. Therefore, large effective population size species will have stronger selective effects than small effective population size species. Ignoring any structural effects would lead to a strong dependence between protein substitution and effective population size. Under single-mutation landscape kinetics, a greater fraction of changes will be explored in larger effective population size species and the role of stochastic processes will be reduced. Furthermore, with higher mutation rates and larger effective population sizes, one enters the regime of multiple mutation kinetics, where multiple changes are simultaneously accessible segregating in a population, allowing exploration of otherwise disfavored parts of sequence space (such as those separated by single mutants with high-energy barriers).

To the extent that most substitutions are neutral (and independent of structure; see Ref. ¹⁰ for a discussion), under the neutral theory, neutral substitutions are fixed with the inverse of the effective population size, canceling out population-level effects governing their introduction to the population. Where there is selection, in addition to structural effects on the substitution properties, advantageous substitutions can pull neutral and slightly deleterious substitutions to fixation (selective sweeps), leading to regions of proteins where there has been more substitution for those that are close in space and connected through the backbone, but not those that are close in space but distant in primary sequence. The reverse can happen under negative selection, where linkage to a deleterious change leads to a higher probability of elimination for changes that are neutral or slightly advantageous. Therefore, linkage is expected to leave a signature in protein regions that are proximal in primary sequence independent of tertiary structure. Alternatively, thermodynamic direct physical interaction (ignoring effects that are mediated allosterically) is expected to leave a signal of proximity in tertiary structure independent of distance in primary sequence.

Furthermore, as suggested above, other processes and subtleties may influence this process, including lineage-specific recombination and mutation rates. Gene expression levels of individual proteins have been suggested as a crucial determinant of selection intensity and evolutionary rate.^{11–13} One underlying hypothesis driving this process is the generation of overly stable proteins that are resistant to mistranslation-induced mis-folding, potentially for energetic reasons.^{14,15} Another potential explanation is one driven by mass action of increased concentration that is expected to lead to spurious protein–protein interaction. Given that specificity of protein–protein interaction appears to be under strong selection but difficult to achieve, typical specificities of proteins (inversely corresponding to rates of misbinding) are lower than rates of mistranslation, potentially making this effect more powerful (see Ref. ¹⁶).

Further testing of these hypotheses will shed more light on the role of protein expression in influencing protein substitution rates and patterns.

Other structure-dependent hypotheses would include the protein size (surface area-to-volume ratios) and the functional density of binding sites on the surface (contact density).^{17,18} These effects would make the lineage-specific evolutionary process dependent on the fold distribution in the lineage and the complexity of the protein–protein interaction network. Underlying these views mechanistically, however, are the roles of folding and binding thermodynamics and will generate signals of a thermodynamic nature.

Given the different views of the interplay between protein structure, effective population size, and the protein substitution process, this study will address the signals for direct physical interaction and for selective sweeps and eliminations (close spatial proximity of substitutions) in a gene family data set derived from plant genomes. These will be analyzed in two data sets, one where positive diversifying selection is suggested and another where strong negative selection is indicated.

A second part of the study will examine the lineage-specific substitution patterns in genome pairs with varying average effective population sizes. Comparisons include rice, primate, rodent, *Drosophila*, and bacterial genomes of close sister taxa. Amino acid substitutions are subdivided into categories on the basis of how different the amino acid side-chain properties are and fit to a population genetic model to examine the influence of amino acid property change and effective population size on the amino acid substitution process.

Results and Discussion

Several models accounting for the roles of various biological forces, including population-level effects mediated by selection and the role of structural constraint on protein sequence evolution, have been presented. These models make different predictions about variations in the nature of substitution dependent on lineage-specific properties. Previously, an analysis of evolution through structure has found that substitutions were clustered in 3D space¹⁹ and that the substitution pattern was more relaxed on the surface and dependent on secondary structure.^{20–22} Further, the rodent lineage was previously shown to contain more radical substitution than either the primate or the artiodactyl lineage.²³ A combination of several of these scientific trajectories was carried out in the estimation of lineage-specific substitution matrices and extended to bacteria, where the effective population size is much larger than in rodents or primates.

As it is well known that the distance of the sequences used in the calculation of the substitution matrix can affect the observed substitution patterns, genomic comparisons with closely aligned distributions were selected. Evolutionary rate itself may affect the apparent biophysical nature of substitution in that it is well known that extrapolation of changes at high sequence identity does not reproduce the changes observed at lower sequence identity and a faster evolutionary rate may mimic this process at equivalent distance by introducing more multiple substitutions simultaneously.^{1,24} This phenomenon is reflected in population genetics as the transition from small $N_e\mu$ to large $N_e\mu$. It corresponds with jumps to alternative parts of sequence and structure space, rate shifts as structural units change, and ultimately jumps in fold space according to the energy gap model for the divergence of protein folds enabling mutations that would not otherwise be tolerated.²⁵ The logic behind this is simple in that increases in effective population size and in mutation rate increase the number of alleles that segregate with multiple mutations. To the extent that multiple mutations interact nonadditively and that there is pleiotropic constraint on residues for folding and/or functional reasons, this enables transitions through barriers in fitness landscapes that may not be surmountable by single mutations in isolation.

Increased $N_e\mu$ also alters the dynamics of exploration of sequence space, enabling faster and wider exploration of sequences within a population, although this second effect is expected to be less discrete in its effects.

Pairwise amino acid substitution matrices

With these considerations, genome pairs were systematically searched to identify pairs with a maximum in the frequency of pairwise amino acid identity >98% and these can be seen in Supplementary Fig. 1. These pairs included *Escherichia coli*, *Drosophila*, rodents, primates, and rice. The sequences from rice, human–chimp, and *E. coli* were slightly more closely related than those of the other comparisons. However, as these lie at opposite ends of the effective population size analysis, this will not affect the modeling that has been done. From these pairs of very similar proteins, genome-specific substitution matrices were calculated and substitutions binned by physical chemical similarity according to the Grantham matrix.²⁶

Development and application of a population genetic model

With the development of a population genetic model described below with effective population size as an important factor in strength of selection, genome pairs with divergent effective population sizes were needed. While it has been reported that *E. coli* has effective population size multiplied by mutation rate that is several orders of magnitude larger than for mammals, the same study reported little difference between this value among human, chimpanzee, and mouse.²⁷ This last point contrasts with other studies that have linked higher mutation rates, reduced generation times, and higher effective population sizes in rodents compared with primates, although the difference is much less dramatic than that comparing either mammalian pair with *E. coli* genomes.^{27,28} In parameterizing our model, literature values of N_e were adopted from references listed in the Methods section and are shown in Table 1.

Previous comparisons have taken the view that nonsynonymous substitutions are more likely to be selected than synonymous substitutions and used comparisons of the nonsynonymous to synonymous nucleotide substitution rates to characterize differences in selection between primates and rodents using a population genetic model.^{28–32} A recent study using the same two *Drosophila* genomes compared here found evidence for weak positive diversifying selection dominating among fixed changes.³³ This is related to our analysis, but does not directly address the types of substitutions that occur in such species with the added variables of amino acid substitution type and ultimately the interplay between these with selection and protein structure.

Dating back to the seminal work of Kimura,³⁴ a simple population genetic model relates the probability of fixation of a mutation to effective population size and strength of selection. Previously, Halpern and Bruno³⁵ have used Kimura's fixation probability to model position-specific amino acid frequencies to improve the measurement of protein distances. Nielsen and Yang³⁶ have also used Kimura's formulation to examine the relationship between dN/dS ratios and selection coefficients in primate mitochondrial and viral data sets. Thorne *et al.*³⁷ have examined the substitution process in RNA and protein structures in terms of a population genetic model to examine the relationship between evolutionary rate and relative fitness. Also extending on Kimura's framework to examine interspecific rather than population genetic data, we have built a model to look at the relative probabilities of fixation of different types of mutations based on the relative mutation rates, which can be obtained from the genetic code at the 1% total divergence level (with derived probabilities for double and triple mutations). The model then (as described in the Methods section in more detail) relates the relative probability of fixation to the observation of occurrence in different genomes, the effective population size of each genome pair, and the parameterized selection coefficient for the class of substitutions.

As seen in Table 2, the simplest models with a single bin of N_e , but multiple bins of amino acid substitution, fit the data poorly. Models with three or seven bins of effective population size explained the underlying data much better, even when accounting for the increased number of parameters using deviance information criterion (DIC)³⁸ as implemented in WinBUGS.³⁹ From the parameterization of selective coefficients that was observed (Table 3), several trends were clear from the best supported models (Fig. 1). The strength of selection decreases with increasing amino acid substitution severity within a given effective population size bin. The strength of selection also decreases with increasing effective population size on a given Grantham matrix bin. Both of these results can be explained, but do not fit the conventional wisdom.

Roles of radical and conservative substitutions

Decreasing selection on substitutions of amino acids of increasing difference in physical properties may be explained by the differing roles for different types of amino acid substitution. It may be that positive diversifying selection acts mostly on subtle changes in functionally important protein regions, such as the hydrophobic core and binding pockets, where more radical changes cannot be tolerated. The radical changes that are tolerated rather than eliminated may be more likely to be neutral. This interpretation has implications for dR/dC (the ratio of the rates of radical and conservative amino acid substitution, where each is assessed by Grantham matrix value), a common test for positive diversifying selection.^{23,40} If radical substitutions (as opposed to mutations) are neutral and positive and negative selection both tolerate more conservative changes, then dR/dC will lack the power to detect positive diversifying selection. A first step in evaluating this will be to trace how dR/dS (the rate of accumulation of radical changes normalized by mutational opportunity as reflected in the rate of accumulation of synonymous changes at the DNA level) compares with dC/dS (the ratio of the rates of accumulation of conservative amino acid substitutions and of synonymous nucleotide substitutions) for closely related proteins under different dN/dS and how this differs between different protein folds (work in progress). The relationship between dN/dS and amino acid level rate shifts is still unclear, and it may be that radical changes, especially in the core, promote changes in coevolutionary patterns, accessibility of new parts of sequence space, and potentially new folds (as in the energy gap model of Dokholayan and Shakhnovich).²⁵ An alternative explanation of the trend observed with radical substitution showing more neutral parameterization of selection coefficients is that these changes are indeed more radical in opposite directions and an averaging effect is observed.

Population size effects

The need for multiple bins of N_e shows that the original population genetic model was overly simple. While there may be demographic explanations for this (differences in variance in effective population size between lineages in the comparison), other functional explanations should be considered too. At smaller effective population sizes, there may be compensation by stronger selective effects of individual changes. This can be considered in the context of both systems biology and protein fold distributions. Many proteins arise via gene duplication. Evidence suggests that even in mammals, gene duplicates are fixed via neofunctionalization⁴¹ and that subfunctionalized genes are eventually neofunctionalized.^{42,43} The distributions of protein folds are known to be different in small and large effective population size organisms.⁴⁴ While a random fold may easily neofunctionalize in a large effective population size genome, evolution may select only the folds that neofunctionalize most easily in small effective population size genomes, resulting in a fold distribution that is more sensitive to selection (positive and negative) on amino acid substitution in small effective population size organisms. Another potential explanation is that of linkage. Nevo *et al.*⁴⁵ have suggested that linkage can amplify the effects of selection in small effective population size species. With these considerations, we will now examine the roles of direct physical interaction

and of linkage on patterns of co-evolution under both negative and positive diversifying selection in plant gene families.

Spatial distributions of substitution

The first question to ask was whether distributions of co-substitutions under negative selection and under positive diversifying selection were randomly distributed at the primary sequence and tertiary structural levels (see Fig. 2). As seen in Fig. 3, they are not randomly distributed. Using the Kolmogorov–Smirnov (K-S) test, the random distributions for positive diversifying selection can be rejected at the $p < 10^{-169}$ level in primary sequence and $p = 4 \times 10^{-169}$ level in 3D structure. Under negative selection, random distributions can be rejected at the $p = 3 \times 10^{-4}$ level in primary sequence and at the $p = 7 \times 10^{-6}$ level in 3D structure.

With random causes excluded, the underlying patterns of the distribution were explored in further detail. It is expected that direct physical interaction of residues in structure due to thermodynamic effects will lead to clustering in 3D space. Similarly, it is expected that selective sweeps will lead to fixation of linked residues that are proximal in primary sequence. Therefore, a distribution of substitutions with the same distances in primary sequence as the underlying data was mapped onto 3D structures and a distribution of substitutions with conserved distances in 3D structures was mapped onto primary sequence. For the positively selected data, the 1D random distribution perfectly explained the 1D data and the 3D random distribution perfectly explained the 3D data. When the randomized 1D data were mapped onto 3D structure and compared with the actual distribution, the two data sets originating from the same distribution could be rejected at the $p = 10^{-2}$ level. When the randomized 3D data were mapped onto primary sequence and compared with the actual distribution, the two data sets originating from the same distribution could be rejected at the $p = 10^{-9}$ level. From this analysis (Fig. 4a and b), we can detect independent signals for selective sweeps and for thermodynamic co-evolution, indicating again that both processes are operating and that there is a signal of both in co-substitution data. Interestingly, there is a more significant signal for selective sweeps than for protein thermodynamics when positive diversifying selection is operating. This corresponds with the observation in Table 3 of selective coefficients that are not strong enough to operate independent of effective population size ($s > 50/N_e$).

Under negative selection, where random substitution patterns could also be rejected, the same tests were run. With strong negative selection, purifying sweeps for residues under the strongest purifying selection eliminating linked changes are expected. In this case, neutral or slightly beneficial changes will be eliminated due to linkage effects. This contrasts with selective sweeps, where strongly beneficial changes pull neutral or slightly deleterious changes to fixation. The linkage effect under strong negative selection is expected to be a weaker effect than selective sweeps are.

In fact, the randomized 3D substitutions mapped onto primary sequence explained the primary sequence substitution patterns at the $p = 0.7$ level (Fig. 4c). The thermodynamic hypothesis is expected to remain the same under negative and positive diversifying selection. However, the randomized primary sequences mapped onto 3D structure explained the 3D data at the $p = 0.43$ level (Fig. 4d). Even when controlling for secondary structure and primary sequence effects due to co-occurrence in a secondary structural unit (Supplementary Materials 4), the distributions of co-substitution at the primary and tertiary sequence levels were not independent. It is probably due to a lack of statistical power that the underlying cause of the correlated distributions cannot be ascertained in this analysis, which does clearly reject random substitution patterns. Future work will involve the use of more powerful parametric methods to examine the interplay between 1D and 3D effects.

Conclusions and further discussion

In this study, we have dissected the role of effective population size on selection for different types of amino acid substitutions and further examined the signals for clustering in primary sequence (linkage) and in 3D structures (direct physical interaction of amino acids). From this, a relatively complex picture of the interplay between population genetic forces and thermodynamic forces has emerged. Selective coefficients on amino acid substitution are in the range where stochastic effects are possible and selection does not completely dominate more neutral processes. Selective coefficients decline with increasing effective population size and with increasingly radical amino acid substitution. It may be that positive diversifying selection acts through less radical change that is less disruptive to functionally important regions of the protein.

The declining selection with increasing effective population size may reflect an interplay at several levels between strength of selection, linkage, other population level effects, protein folds, and substitution patterns. Compensation has been suggested as a mechanism to explain this effect. In terms of underlying process, compensation can mean several things, from selective effects at the systems level to population-level effects. The time spent segregating for any given mutation is dependent on the effective population size and on the selective effect. As a mutation segregates for longer, it is more likely to become linked to other mutations, especially as mutation rate increases. This reflects the transition from single-hit kinetics to multiple-hit kinetics and the onset of Hill–Robertson effects. This transition, however, is not necessarily equally dependent on the global effective population size, as mutations that segregate for longer in small effective population size organisms behave more like mutations in larger effective population size organisms. As Nevo *et al.*⁴⁵ have suggested, linkage itself can compound the effects, leading to stronger selection in small effective population size species.

On top of this is the role of protein structure. It has been suggested above on the basis of the gene duplication process and neofunctionalization as a mechanism that the fold distribution may be systematically different between large and small effective population size species, resulting in more sensitive, more designable folds in such lineages. It is known that not only are fold identities different in different lineages, but so are other factors, including protein size and compactness that effect designability.⁴⁴

When positive diversifying selection is occurring, it is clear that there are strong independent signals for linkage and for direct physical interaction. When negative selection is occurring, co-substitution is clearly nonrandom, but the distributions for linkage and direct physical interaction could not be significantly separated. When this is all put together, it is clear that the interplay between population genetic forces and protein thermodynamics is complex, but that it needs to be considered in understanding amino acid substitution.

Lastly, the interspecific population genetic model that has been generated can be applied generally to problems in evolutionary structural biology. The analysis here simplifies protein structure to a substitution matrix, but the underlying structural dynamics are more complex. Over evolutionary time, different types of structural units are interconverted (e.g., between Cys–Cys bridges, hydrophobic interactions, charge interactions, cation– π interactions, etc.) and the rates and selective pressures governing these interconversions have not been studied. The model is therefore amenable to more complex and realistic analysis of structural evolution, as well as to other problems in the linkage of genotype to phenotype.

Methods

Identification of genome pairs for analysis

Protein sequences were extracted from a collection of species that were candidates to be sister taxa. Using BLASTp, we collected all domain matches with bit scores >60 , E value $s < 1.0$ and $\%ID > 90\%$. Histograms of $\%ID$ were generated for all species pairs and those with a peak $>98\%$ were selected (see Supplementary Fig. 1). These included the following pairs: *E. coli* K12–CFT30, *Oryza sativa* ssp. *indica*–*japonica*, *Drosophila melanogaster*–*Drosophila simulans*, *Mus musculus*–*Rattus norvegicus*, *Homo sapiens*–*Pan troglodytes*, *Pan troglodytes*–*Macaca mulatta*, and *Homo sapiens*–*Macaca mulatta*. From these BLAST hits, pairwise sequence alignments were generated using MUSCLE.⁴⁶

Effective population sizes

The average pairwise effective population sizes (assuming rough constancy through evolution) are listed in Table 1. These values are taken from the following references: human,⁴⁷ chimpanzee⁴⁸, Rhesus macaque,⁴⁹ mouse,⁵⁰ *Drosophila melanogaster*,⁵¹ *Drosophila simulans*,⁵⁰ rice,⁵² and *E. coli*.⁵³ The values obtained broadly agree with values reported by Lynch and Conery.²⁷

Generation of substitution matrices

Each individual global pairwise alignment that had $>90\%$ sequence identity was used in the generation of a substitution matrix according to the procedure of Jones *et al.*⁵⁴ based on pairwise sequence comparisons without inference of ancestral sequences. These matrices were normalized to PAM₁ matrices. Individual transitions in the matrix were partitioned by Grantham matrix score,²⁶ a matrix characterizing the chemical properties of the amino acids. These partitioned transitions gave vectors of counts, $C_i = (C_{i1}, \dots, C_{ik})$, for the i th population and k partitions. The normalized matrices can be found as Supplementary Materials 2.

Normalizing by mutational opportunity

In addition to characterizing the observed amino acid changes, we evaluated the mutational opportunity for each type of change. At 1% total amino acid substitution and an average dN/dS ratio of 0.20 (see Ref. 20), a weighting of 95% single mutations, 4% double mutations, and 1% triple mutations was evaluated through the genetic code to determine relative mutational opportunities. This calculation assumes equal nucleotide frequencies generating amino acid frequencies through the genetic code as a simplifying assumption.

Modeling the fixation of amino acid transitions

A model originally derived by Kimura³⁴ relates the probabilities of mutations occurring to be fixed in a population dependent on the effective population size (N_e) and strength of selection (s), given that a mutation has occurred.

$$F = (1 - e^{-2s}) / (1 - e^{-4N_e s})$$

When comparing substitution patterns in pairs of genomes, one observes fixed substitutions rather than mutations. When normalized for mutational opportunity, a model can be generated that examines the relative occurrence of fixed substitutions by their relative probabilities of fixation, given the model above.

The following model then relates the relative probability of fixation (F_{ij}) to the selection parameters (S_{ij}) via the effective population sizes ($N_{e,i}$) and relative mutational opportunities

(μ_j) , where j indexes the Grantham-matrix-based partitions or categories and i indexes the populations,

$$F_{ij} = \left\{ (\mu_j S_{ij}) / (1 - e^{-4N_{e,i}S_{ij}}) \right\} / \left(\sum_j \left\{ (\mu_j S_{ij}) / (1 - e^{-4N_{e,i}S_{ij}}) \right\} \right)$$

The equation assumes that each site is behaving independently of other sites. In general, we model the above counts of amino acid substitutions, $C_i = (C_{i1}, \dots, C_{ik})$ as having a k -dimensional multinomial distribution with cell probabilities F_{ij} . Thus, $C_i = (C_{i1}, \dots, C_{ik}) \sim \text{Mult}(F_{i1}, \dots, F_{ik})$, and we obtain the binomial for $k=2$. The above nonlinear model is then incorporated for the parameters of the multinomial, F_{ij} , with the goal of estimating the unknown selection parameters, S_{ij} . This generalized nonlinear model is estimated in a Bayesian framework with constraints on the S_{ij} used to formulate the prior distributions, $S_{ij} \sim \text{Uniform}[1/(20N_{e,i}), 50/N_{e,i}]$. These constraints reflect the parameterization where there is power to select based on the effective population size. Data on the effective population sizes and relative mutational opportunities are taken to be fixed. WinBUGS³⁹ was used to fit the above model for different scenarios as the number of categories was changed from $k=2$ to 3 and the S_{ij} were constrained to be the same for all the seven populations (one group), or allowed to vary within population groups (three groups and seven groups). The two categories corresponded to Grantham matrix values within the intervals $\{(5,90), (91,216)\}$ and three categories corresponded to $\{(5,70), (71,91), (91,216)\}$. The three population groups consisted of *O. sativa* ssp. *indica-japonica* in the first, human-chimp, human-macaque, chimp-macaque, and mouse-rat in the second with *Drosophila* species and *E. coli* species forming the third.

The models were fit within WinBUGS using Markov chain Monte Carlo, and model comparison was carried out using DIC.³⁸ Various components of the DIC measure are also presented with our results, the main idea being that within a set of models, those with smaller values of DIC will be better at predicting a replicate data set. Note that since DIC is a Bayesian measure, it is based on a log-likelihood that is evaluated at the posterior means of the parameters. Plots of observed relative probabilities *versus* estimated relative probabilities were also examined to assess model fit. From among the models examined, the saturated models, where the S_{ij} are allowed to vary across the seven populations, fit the best. Grouping the populations leads to a smaller number of S_{ij} parameters, and model selection consists of picking a model with a smaller number of parameters that fits as adequately as a saturated model.

Detecting positive and negative selection in plant genomes

Phylogenetic trees, nucleotide, and protein multiple sequence alignments for embryophyte gene families were extracted from The Adaptive Evolution Database (TAED).²⁰ Each multiple alignment was mapped to Protein Data Bank (PDB) databases using a pairwise alignment with the best matching sequence in the multiple sequence alignment.

For each gene family that had reliable PDB mapping, PAML⁵⁶ was used to detect selection levels on each individual branch (free ratios model). For branches where $dN/dS > 1$ was detected, the PAML branches-and-sites model was run to detect sites under positive diversifying selection with posterior probability > 0.9 . The sites on branches that have scored probability 0.9 to be under positive diversifying selection along branches where K_a/K_s was > 1 when averaged across all sites were chosen for subsequent analysis in the positive diversifying selection data set. Branches with $dN/dS < 0.2$ were selected as the negative selection data set. Taking the ancestral sequences generated in TAED and using subtractive parsimony, branches

with at least two substitutions were selected and these sites were taken for subsequent analysis in the negative selection data set.

The free ratios model was taken as a model to detect lineage-specific selection and was not contrasted with a null model. The free ratio model dN/dS values are strongly correlated with those obtained by parsimonious ancestral sequence reconstruction and counting along branches. For the purposes of this analysis, false positives were not more problematic than false negatives and the analysis was intended to put each branch in the most appropriate selection bin.

Evaluating positive diversifying selection in primary and tertiary sequence contexts

The following formula was used to obtain the probability of a given number of residues or more to coincide within a given volume (or within a given sequence window):

$$p = \frac{\sum_{l=t}^{m=\min(s,k)} \binom{s}{l} \binom{N-s}{k-l}}{\binom{n}{k}}$$

where N is total number of residues available, k is a number of given residues, s is the total number of residues within a given volume (sequence window), and t is the number of given residues co-occurring within a volume. Probability is conditional on a given residue, that is, when a volume or window centered on a given residue was calculated in a similar way. This is shown in Fig. 2. Calculations at the primary sequence level were performed with the simplifying assumption that the genes were intron free. Exon boundaries represent a small percentage of the total gene length and the simplifying assumption is expected to have only a small effect on the underlying probability distributions.

The volume enclosing clusters was chosen as a sphere with two residues being end points of a diameter (all residue pairs considered). Other enclosures were considered with consistent results (see Ref. 57 for a discussion of evolutionary properties in the context of sphere size and Ref. 58 for further discussion on detecting 3D clusters of residues).

Uncontrolled random distributions of residues in primary sequence and in tertiary structure were generated to test for randomness of observed distributions. Then, for each per-branch set of residues under positive diversifying selection, we generated a random set of residues that had clustering properties (distances) that were the same to clusters in either primary sequence or 3D-structure. This is trivial to do on a primary sequence (e.g., shifting a cluster in the sequence by a random number). To generate such clusters on 3D structures, a Monte Carlo approach was employed. We used the standard Metropolis algorithm with the objective function being the mean square deviation of the distance matrix of generated cluster from distance matrix of the original cluster (see Fig. 5 for an illustration of this).

For clusters generated in 1D, probabilities were calculated as above. The control distributions obtained were tested against the distribution of probabilities generated from the positively and negatively selected sequences. This was a control to ensure that properties of the clusters are reproduced properly and that the distributions are nonrandom. We used the standard two-sample K-S test to compare the distributions. Sequence clusters were then mapped to the corresponding PDB structures, and the probabilities were calculated and then compared to the set of probabilities from positively and negatively selected sites using the K-S test. The same procedure was repeated in reverse, where clusters were generated maintaining their properties in 3D and then mapped back to the sequence. All tests for randomness and independence of

distribution in 1D and in 3D for positive and negative selection were compared with the K–S test.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Abbreviations used

DIC	deviance information criterion
K-S	Kolmogorov–Smirnov

Acknowledgments

We are grateful to two anonymous reviewers, Rasmus Nielsen, Jay Storz and Jan Kubelka for helpful comments, and Johan Grahnén and Sanaa Ahmed for reading an early draft of this paper. The work was supported by an NIH-INBRE institutional award to University of Wyoming.

References

1. Sasidharan R, Chothia C. The selection of acceptable protein mutations. *Proc. Natl Acad. Sci. USA* 2007;104:10080–10085. [PubMed: 17540730]
2. Hamill SJ, Cota E, Chothia C, Clarke J. Conservation of folding and stability within a protein family: the tyrosine corner as an evolutionary cul-de-sac. *J. Mol. Biol* 2000;295:641–649. [PubMed: 10623553]
3. DePristo MA, Weinreich DM, Hartl DL. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat. Rev. Genet* 2005;6:678–687. [PubMed: 16074985]
4. Taverna DM, Goldstein RA. Why are proteins marginally stable? *Proteins* 2002;46:105–109. [PubMed: 11746707]
5. Shakhnovich E. Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. *Chem. Rev* 2006;106:1559–1588. [PubMed: 16683745]
6. Lopez P, Casane D, Philippe H. Heterotachy, an important process of protein evolution. *Mol. Biol. Evol* 2002;19:1–7. [PubMed: 11752184]
7. Guindon S, Rodrigo AG, Dyer KA, Huelsenbeck JP. Modeling the site-specific variation of selection patterns along lineages. *Proc. Natl Acad. Sci. USA* 2004;101:12957–12962. [PubMed: 15326304]
8. Blanquart S, Lartillot N. A site-and time-heterogeneous model of amino acid replacement. *Mol. Biol. Evol* 2008;25:842–858. [PubMed: 18234708]
9. Gu X. Maximum-likelihood approach for gene family evolution under functional divergence. *Mol. Biol. Evol* 2001;18:453–464. [PubMed: 11264396]
10. Nei M. Selectionism and neutralism in molecular evolution. *Mol. Biol. Evol* 2005;22:2318–2342. [PubMed: 16120807]
11. Duret L, Mouchiroud D. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol. Biol. Evol* 2000;17:68–74. [PubMed: 10666707]
12. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. Why highly expressed proteins evolve slowly. *Proc. Natl Acad. Sci. USA* 2005;102:14338–14343. [PubMed: 16176987]
13. Drummond DA, Raval A, Wilke CO. A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol* 2006;23:327–337. [PubMed: 16237209]
14. Drummond DA, Wilke CO. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 2008;134:341–352. [PubMed: 18662548]
15. Wagner A. Energy constraints on the evolution of gene expression. *Mol. Biol. Evol* 2005;22:1365–1374. [PubMed: 15758206]

16. Zhang J, Maslov S, Shakhnovich EI. Constraints imposed by non-functional protein–protein interactions on gene expression and proteome size. *Mol. Syst. Biol* 2008;4:210. [PubMed: 18682700]
17. Zuckerkandl E. Evolutionary processes and evolutionary noise at the molecular level. I. Functional density in proteins. *J. Mol. Evol* 1976;7:167–183. [PubMed: 933174]
18. Lin YS, Hsu WL, Hwang JK, Li WH. Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. *Mol. Biol. Evol* 2007;24:1005–1011. [PubMed: 17264066]
19. Yu J, Thorne JL. Testing for spatial clustering of amino acid replacements within protein tertiary structure. *J. Mol. Evol* 2006;62:682–692. [PubMed: 16752209]
20. Roth C, Liberles DA. A systematic search for positive selection in higher plants (Embryophytes). *BMC Plant Biol* 2006;6:12. [PubMed: 16784532]
21. Koshi JM, Goldstein RA. Context-dependent optimal substitution matrices. *Protein Eng* 1995;8:641–645. [PubMed: 8577693]
22. Illergård K, Ardell DH, Elofsson A. Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins* 2009;77:499–508. [PubMed: 19507241]
23. Zhang J. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. *J. Mol. Evol* 2000;50:56–68. [PubMed: 10654260]
24. Benner SA, Cohen MA, Gonnet GH. Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Eng* 1994;7:1323–1332. [PubMed: 7700864]
25. Dokholyan NV, Shakhnovich EI. Understanding hierarchical protein evolution from first principles. *J. Mol. Biol* 2001;312:289–307. [PubMed: 11545603]
26. Grantham R. Amino acid difference formula to help explain protein evolution. *Science* 1974;185:862–864. [PubMed: 4843792]
27. Lynch M, Conery JS. The origins of genome complexity. *Science* 2003;302:1401–1404. [PubMed: 14631042]
28. Ohta T. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J. Mol. Evol* 1995;40:56–63. [PubMed: 7714912]
29. Nachman MW. Y chromosome variation of mice and men. *Mol. Biol. Evol* 1998;15:1744–1750. [PubMed: 9866208]
30. Weinreich DM. The rates of molecular evolution in rodent and primate mitochondrial DNA. *J. Mol. Evol* 2001;52:40–50. [PubMed: 11139293]
31. Seo TK, Kishino H, Thorne JL. Estimating absolute rates of synonymous and non-synonymous nucleotide substitution in order to characterize natural selection and date species divergences. *Mol. Biol. Evol* 2004;21:1201–1213. [PubMed: 15014159]
32. Kryukov GV, Schmidt S, Sunyaev S. Small fitness effect of mutations in highly conserved non-coding regions. *Hum. Mol. Genet* 2005;14:2221–2229. [PubMed: 15994173]
33. Sawyer SA, Parsch J, Zhang Z, Hartl DL. Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. *Proc. Natl Acad. Sci. USA* 2007;104:6504–6510. [PubMed: 17409186]
34. Kimura M. On the probability of fixation of mutant genes in a population. *Genetics* 1962;47:713–719. [PubMed: 14456043]
35. Halpern AL, Bruno WJ. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol* 1998;15:910–917. [PubMed: 9656490]
36. Nielsen R, Yang Z. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol. Biol. Evol* 2003;20:1231–1239. [PubMed: 12777508]
37. Thorne JL, Choi SG, Yu J, Higgs PG, Kishino H. Population genetics without intraspecific data. *Mol. Biol. Evol* 2007;24:1667–1677. [PubMed: 17470435]
38. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit (with discussion). *J. R. Stat. Soc., Ser. B* 2002;64:583–640.
39. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. *Stat. Comput* 2000;10:325–337.

40. Hughes AL, Ota T, Nei M. Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. *Mol. Biol. Evol* 1990;7:515–524. [PubMed: 2283951]
41. Hughes T, Liberles DA. The pattern of evolution of smaller-scale gene duplicates in mammalian genomes is more consistent with neo-than subfunctionalisation. *J. Mol. Evol* 2007;65:574–588. [PubMed: 17957399]
42. He X, Zhang J. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 2005;169:1157–1164. [PubMed: 15654095]
43. Rastogi S, Liberles DA. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol. Biol* 2005;5:28. [PubMed: 15831095]
44. Gerstein M, Levitt M. A structural census of the current population of protein sequences. *Proc. Natl Acad. Sci. USA* 1997;94:11911–11916. [PubMed: 9342336]
45. Nevo E, Kirzhner V, Beiles A, Korol A. Selection versus random drift: long term polymorphism persistence in small populations (evidence and modelling). *Philos. Trans. R. Soc. London, Ser. B* 1997;352:381–389.
46. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 2004;5:113. [PubMed: 15318951]
47. Takahata N, Satta Y, Klein J. Divergence time and population size in the lineage leading to modern humans. *Theor. Popul. Biol* 1995;48:198–221. [PubMed: 7482371]
48. Stone AC, Griffiths RC, Zegura SL, Hammer MF. High levels of Y-chromosome nucleotide diversity in the genus. *Pan. Proc. Natl Acad. Sci. USA* 2002;99:43–48.
49. Hernandez RD, Hubisz MJ, Wheeler DA, Smith DG, Ferguson B, Rogers J, et al. Demographic histories and patterns of linkage disequilibrium in Chinese and Indian rhesus macaques. *Science* 2007;316:240–243. [PubMed: 17431170]
50. Eyre-Walker A, Keightley PD, Smith NGC, Gaffney D. Quantifying the slightly deleterious mutation model of molecular evolution. *Mol. Biol. Evol* 2002;19:2142–2149. [PubMed: 12446806]
51. Loewe L, Charlesworth B. Background selection in single genes may explain patterns of codon bias. *Genetics* 2007;175:1381–1393. [PubMed: 17194784]
52. Zhu Q, Zheng X, Luo J, Gaut BS, Ge S. Multilocus analysis of nucleotide variation of *Oryza sativa* and its wild relatives: severe bottleneck during domestication of rice. *Mol. Biol. Evol* 2007;24:875–888. [PubMed: 17218640]
53. Berg OG. Selection intensity for codon bias and the effective population size of *Escherichia coli*. *Genetics* 1996;142:1379–1382. [PubMed: 8846914]
54. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci* 1992;8:275–282. [PubMed: 1633570]
55. Roth C, Betts MJ, Steffansson P, Saelensminde G, Liberles DA. The Adaptive Evolution Database (TAED): a phylogeny based tool for comparative genomics. *Nucleic. Acids Res* 2005;33:D495–D497. Database issue. [PubMed: 15608245]
56. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol* 2007;24:1586–1591. [PubMed: 17483113]
57. Berglund AC, Wallner B, Elofsson A, Liberles DA. Tertiary windowing to detect positive diversifying selection. *J. Mol. Evol* 2005;60:499–504. [PubMed: 15883884]
58. Zhou T, Enyeart PJ, Wilke CO. Detecting clusters of mutations. *PLoS One* 2008;3:e3765. [PubMed: 19018282]

Fig. 1.

(a) A population genetic model with two bins of Grantham matrix values and seven bins of N_e is used to generate selective coefficients for each bin. The trend in the selective coefficient relative to Grantham bin and N_e is observed. (b) A population genetic model with three bins of Grantham matrix values and three bins of N_e is used to generate selective coefficients for each bin. The trend in the selective coefficient relative to Grantham bin and N_e is observed.

$$p = \frac{\sum_{l=t}^{m=\min(s,k)} \binom{s}{l} \binom{N-s}{k-l}}{\binom{N}{k}}$$

Fig. 2. The probability of observing clusters of substituted or positively selected codons/amino acids in primary sequence and in tertiary structure is shown. This was used to generate probability distributions for clusters under different sets of assumptions.

Fig. 3.

(a) The probability histogram of clusters of residues found in primary sequence under strong negative selection is compared to the distribution that would be observed if substitution were random. The calculation of these probabilities is described in the Methods section and is shown in Fig. 2. The distributions are then compared using a K-S test. (b) The probability histogram of clusters of residues found in tertiary structure under strong negative selection is compared to the distribution that would be observed if substitution were random. (c) The probability histogram of clusters of residues found in primary sequence under positive diversifying selection is compared to the distribution that would be observed if substitution were random. (d) The probability histogram of clusters of residues found in tertiary structure under positive diversifying selection is compared to the distribution that would be observed if substitution were random.

Fig. 4.

(a) The probability histogram of clusters of residues found in primary sequence under strong negative selection is compared to the distribution that would be observed if substitution were random in 3D structure. The calculation of these probabilities is described in the Methods section and is shown in Fig. 2. The distributions are then compared using a K-S test. (b) The probability histogram of clusters of residues found in tertiary structure under strong negative selection is compared to the distribution that would be observed if substitution were random in primary sequence. (c) The probability histogram of clusters of residues found in primary sequence under positive diversifying selection is compared to the distribution that would be observed if substitution were random in 3D structure. (d) The probability histogram of clusters of residues found in tertiary structure under positive diversifying selection is compared to the distribution that would be observed if substitution were random in primary sequence.

Fig. 5.

To examine the effects of 3D clustering on primary sequence substitution patterns, random distributions of substitutions in 3D that maintained clustered distances were needed. An observed 3D cluster that is also a cluster in primary sequence is compared with a random 3D cluster that is less tightly clustered in primary sequence.

Table 1

Effective population sizes of species pairs derived from the literature (these values were used in the modeling study)

Index <i>i</i>	N_e	Species
1	1500	Rice
2	20,500	Human–chimp
3	41,500	Human–macaque
4	52,000	Chimp–macaque
5	161,000	Mouse–rat
6	2,400,000	<i>Drosophila</i>
7	5,000,000	<i>Ecoli</i>

Using WinBugs,²⁶ population genetic models with varying numbers of categories of Grantham matrix value and groups of N_e were fit to the substitution matrix data for a series of genome pairs

Table 2

Categories	Groups	Dbar	Dhat	pD	DIC
3	1	3303.450	3300.440	3.009	3306.460
3	3	305.192	298.231	6.961	312.153
3	7	162.216	148.194	14.022	176.237
2	1	4727.250	4726.250	0.999	4728.250
2	3	172.724	168.828	3.897	176.621
2	7	84.015	77.021	6.994	91.009

DIC²⁰ was used to differentiate parameter-rich models from simpler models and showed support for models with many groups of N_e .

Table 3

Parameterizations of s values for the population genetic models, where fit to the data is shown in Table 2

Parameter	Mean	SD	2.5%	Median	97.5%
<i>A. Model with seven groups of N_e and two categories of Grantham matrix value</i>					
$s_{1,1}$	0.02034	0.006981	0.005127	0.02171	0.02964
$s_{1,2}$	0.01229	0.004219	0.003101	0.01313	0.0179
$s_{2,1}$	0.001659	5.575E-4	4.643E-4	0.001764	0.002409
$s_{2,2}$	9.354E-4	3.146E-4	2.609E-4	9.948E-4	0.00136
$s_{3,1}$	8.401E-4	2.679E-4	2.532E-4	8.936E-4	0.001192
$s_{3,2}$	3.794E-4	1.21E-4	1.146E-4	4.037E-4	5.378E-4
$s_{4,1}$	6.407E-4	2.244E-4	1.832E-4	6.797E-4	9.498E-4
$s_{4,2}$	2.817E-4	9.875E-5	8.047E-5	2.987E-4	4.182E-4
$s_{5,1}$	2.234E-4	6.318E-5	8.259E-5	2.347E-4	3.071E-4
$s_{5,2}$	7.303E-5	2.065E-5	2.7E-5	7.675E-5	1.003E-4
$s_{6,1}$	1.296E-5	5.617E-6	1.686E-6	1.398E-5	2.058E-5
$s_{6,2}$	4.773E-6	2.07E-6	6.169E-7	5.146E-6	7.584E-6
$s_{7,1}$	6.643E-6	2.361E-6	1.58E-6	6.989E-6	9.87E-6
$s_{7,2}$	2.076E-6	7.415E-7	4.922E-7	2.184E-6	3.123E-6
<i>B. Model with three groups of N_e and three categories of Grantham matrix value</i>					
$s_{1,1}$	0.02282	0.005672	0.009906	0.02433	0.02976
$s_{1,2}$	0.01509	0.003759	0.006542	0.01608	0.01977
$s_{1,3}$	0.01286	0.003198	0.005578	0.0137	0.01678
$s_{2,1}$	2.33E-5	3.407E-7	2.265E-5	2.33E-5	2.398E-5
$s_{2,2}$	1.569E-5	2.551E-7	1.52E-5	1.569E-5	1.62E-5
$s_{2,3}$	7.065E-6	1.313E-7	6.813E-6	7.064E-6	7.327E-6
$s_{3,1}$	7.633E-6	1.702E-6	3.795E-6	7.937E-6	9.913E-6
$s_{3,2}$	4.164E-6	9.3E-7	2.067E-6	4.328E-6	5.419E-6
$s_{3,3}$	2.533E-6	5.654E-7	1.26E-6	2.634E-6	3.293E-6