# Compensated pathogenic deviations

Anja Barešić and Andrew C.R. Martin
Institute of Structural and Molecular Biology,
Division of Biosciences, University College London,
Darwin Building, Gower Street, London WC1E 6BT. UK

Deleterious or 'Disease-Associated Mutations' (DAMs) are mutations that lead to disease with high phenotype penetrance: they are inherited in a simple Mendelian manner, or, in the case of cancer, accumulate in somatic cells leading directly to disease. However, in some cases, the amino acid that is substituted resulting in disease, is the wild-type native residue in the functionally equivalent protein in another species. Such examples are known as 'Compensated Pathogenic Deviations' (CPDs) since, somewhere in the second species, there must be compensatory mutations that allow the protein to function normally despite having a residue which would cause disease in the first species. Depending on the nature of the mutations, compensation may occur in the same protein, or in a different protein with which it interacts. In principle, compensation may be achieved by a single mutation (most likely structurally close to the CPD), or by the cumulative effect of a number of mutations.

While it is clear that these effects occur in proteins, compensatory mutations are also important in RNA potentially having an impact in disease. As a much simpler molecule, RNA provides an interesting model for understanding mechanisms of compensatory effects, both by looking at naturally occurring RNA molecules and as a means of computational simulation.

This review surveys the quite limited literature that has explored these effects. Understanding the nature of CPDs is important in understanding traversal along fitness landscape valleys in evolution. It may also have applications in treating diseases that result from such mutations.

**Keywords:** Deleterious mutations; disease-associated mutations; epistasis; SNPs; co-evolution; co-adaptation
**Abbreviations:** PD - Pathogenic Deviation; CPD - Compensated Pathogenic Deviation; DAM - Disease-Associated Mutation; DM - Deleterious Mutation; OMIM - Online Mendelian Inheritance in Man; SNP - Single Nucleotide Polymorphism; SAAP - Single Amino Acid Polymorphism; FEP - Functionally Equivalent Protein

## Introduction

It has frequently been observed that, when deleterious single amino acid mutations are surveyed, mutated amino acid types with detrimental effects in one species are found as the native wild-type residue in homologous proteins of other species, with neutral effect on the fitness of the latter species. The most likely scenario explaining such observations is that the two homologous proteins provide slightly different structural environments for the same residue, thus compensating for the deleterious effect of the residue in the first protein. Generally people have looked at cases of human disease-causing 'deleterious' or 'disease-associated' mutations (DAMs) and observed that the mutant (disease-causing) amino acid is the native (wild-type) amino acid in another species. Such cases are known as 'Compensated Pathogenic Deviations' (CPDs).

Figure 1 shows an example of two DAMs in human antithrombin-III (ANT3), one of which is compensated and the other un-compensated. In the human protein, the mutations Ala416→Pro and Ala416→Ser both cause susceptibility to thrombophilia as a result of antithrombin III deficiency. Details of these mutations can be seen in Online Mendelian Inheritance in Man (OMIM) Entries 107300.0007 and 107300.0027 (`http://www.ncbi.nlm.nih.gov/omim/107300`).
While OMIM states that the mutation occurs at residue 384, this equates to residue 416 in the UniProtKB/SwissProt sequence (UniProtKB/SwissProt accession P01008). Our online resource at `http://www.bioinf.org.uk/omim/` provides a validated mapping of residue numbers in OMIM to UniProtKB/SwissProt residue numbers. As the alignment shows, this residue is a conserved alanine in all the sequences examined — neither proline nor serine is seen in any other species and the two disease-causing mutations seen in humans are therefore classified as 'pathogenic deviations' (PDs, see below). However, a mutation of Ala419→Val (as described in OMIM Entry 107300.0042, OMIM residue number 387), which also leads to antithrombin III deficiency, occurs

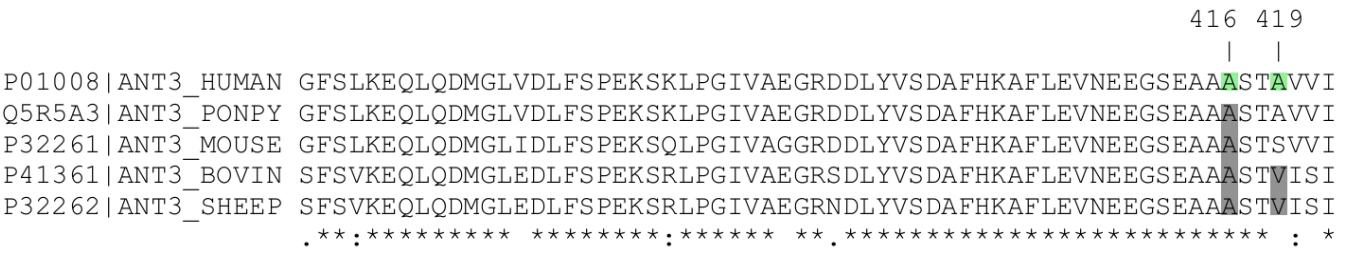


Figure 1: Examples of two disease-associated mutations (DAMs) reproduced from our structural analysis (10). The figure shows the alignment of the human antithrombin-III (ANT3) protein sequence with non-human functionally equivalent homologous proteins. Highlighted are columns 416 and 419 which represent an un-compensated pathogenic deviation (PD) and a compensated pathogenic deviation (CPD) respectively.

at a residue which is not conserved in the alignment. In fact, sheep and cows have a valine at this position in the native sequence and thus the Ala419→Val mutation in humans is classified as a CPD.

The question, therefore, is how do the sheep and bovine proteins function properly with a valine at position 419? Presumably, during the evolution of human, sheep and bovine ANT3 proteins from a common ancestor, some other amino acid difference(s) have occurred in the sheep and bovine proteins compared with the human protein that somehow compensate for what, in the human protein, is the negative effect of having a valine at position 419. How the compensation is achieved in this example is not clear.

Compensation of mutations is also important at the RNA level. Stable Watson-Crick base pairing in RNA can bring together remote parts of the molecule to form stable three-dimensional structures of functional importance. Thus mutations in the RNA must undergo compensatory events to maintain the necessary base pairing requiring the crossing of valleys on the fitness landscape. Not only has this been studied using real RNA sequences (1), but RNA has also been used in computational models designed to understand compensatory mutation (2, 3).

# Body of Review

The term 'compensated mutations' was introduced by Kimura (4), who demonstrated that two mutually compensatory mutations could become fixed in a population as a result of random genetic drift. Kimura defined 'compensatory neutral mutations' as linked deleterious mutations; in other words two mutations each of which, by itself, has a deleterious effect, but together have a neutral (or potentially even a beneficial) effect on overall fitness. The ability of one mutation to compensate for the pathogenic effects of another newly-introduced mutation is an important mechanism in evolution. Using the same analogy used by Wright (5) and used extensively by Dawkins (6), the fitness landscape can be viewed as mountains of high fitness separated by valleys of low fitness. Thus compensation of mutations allows bridging the valleys of low fitness.

## Terminology

Since the analysis and understanding of CPDs crosses the boundaries of structural and evolutionary biology, it is useful to define a number of terms that are used in the field before we go into any more discussion.

**Single nucleotide polymorphisms (SNPs)** are single DNA base changes. Strictly the term is applied only to instances where the mutation is observed in at least 1% of a 'normal' population. In other words they will either have a completely neutral phenotype, or a low penetrance phenotype where there is no clear Mendelian inheritance. Such SNPs may be involved in more complex conditions such as heart disease, or simply give a propensity towards disease through interaction with external factors. However, it should be noted that many people use the term SNP to refer to *any* single base change, even when no frequency data are available. In our work looking at the effects of mutation on protein structure (7), we tried to use the term SNP in the correct way (with the assumption that they do not lead to high-penetrance Mendelianly inherited disease) and contrasted these with mutations that do lead to disease. However, even dbSNP (8), the primary repository for SNP data at the National Center for Biotechnology Information (NCBI), includes data on lower frequency mutations.

SNPs may occur in coding or non-coding regions of DNA. Both coding (cSNPs) and non-coding SNPs (ncSNPs) may have effects on gene expression or mRNA splicing; cSNPs may (i) be synonymous in terms of the resultant amino acid (sSNP), (ii) lead to a premature stop or 'nonsense' codon (nSNP), or (iii) be non-synonymous (an nsSNP) resulting in a single amino acid change. See Figure 2.

**Single amino acid polymorphisms (SAAPs)**

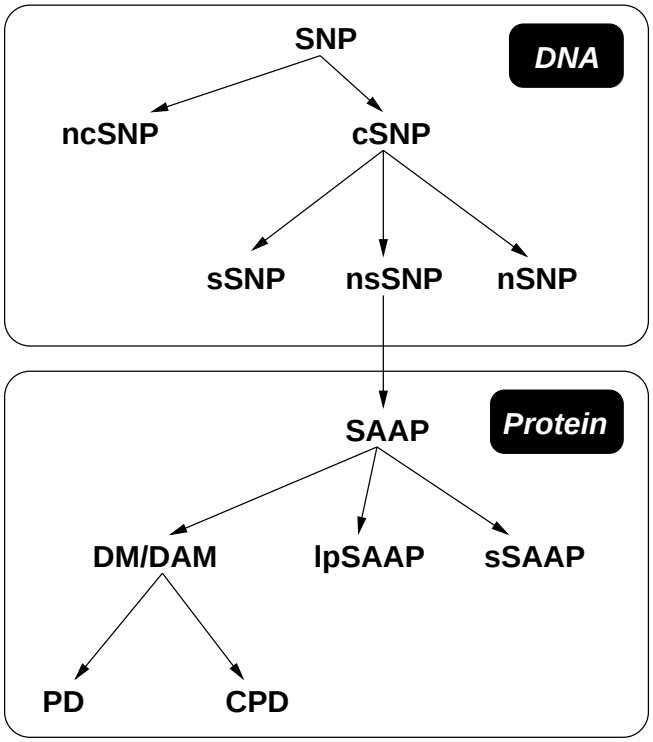


Figure 2: Hierarchy of SNPs, mutations and their effects. SNPs (defined in the general sense to mean any single base DNA mutation) can be non-coding (ncSNPs) or coding (cSNPs). cSNPs can be synonymous (sSNPs), nonsense (nSNPs), or non-synonymous (nsSNPs). nsSNPs result in a single amino acid polymorphism (SAAP) at the protein level. These can be phenotypically silent (sSAAP), low penetrance (lpSAAP), or high penetrance deleterious mutations (DMs) also known as disease-associated mutations (DAMs). A DAM can be compensated in another species (a compensated pathogenic deviation, CPD) or un-compensated (a pathogenic deviation, PD). Note that all forms of SNPs can have effects on expression as they may affect regulatory regions or splice sites. Note also that lpSAAPs form a continuum between phenotypically silent and high penetrance disease-associated mutations.

are single amino acid mutations resulting from nsSNPs. We use the term as defined by Hurst *et al.* (7) to apply both to mutations resulting from strictly-defined nsSNPs (i.e. those that occur in at least 1% of a normal population) and to deleterious muations (DAMs) as defined below. See Figure 2.

**Deleterious mutations** also referred to as **disease-associated mutations** (DAMs) (9) are SAAPs that result in high-penetrance disease phenotypes. In this review, we use the term to encompass both PDs and CPDs as defined below. See Figure 2.

**Pathogenic deviations (PDs)** is often used as a synonym for DAMs, but in the discussion of CPDs (see below), we generally refer to PDs as disease-causing mutations that are *not* observed to be compensated in any other species and that is the definition we use throughout this review. As discussed by Barešić *et al.* (10), this definition of PDs is not completely reliable since it is based on a negative observation. Mutations are classified as PDs rather than CPDs simply because the residue is not observed as the native residue in any other species, but until we have the sequence of every species, we cannot know conclusively that it is not compensated in at least one other species. See Figure 2 and column 416 in Figure 1 for an example of a PD.

**Compensated pathogenic deviations (CPDs)** have also been referred to as 'potential compensated mutations' (9). Their existence was first discussed by Kimura (4), who termed them 'compensatory neutral mutations' while the term CPD was first defined by Kondrashov *et al.* (11). A CPD is a SAAP (as defined above) associated with a disease phenotype (i.e. a DAM), usually in a human protein, where the mutated amino acid type is found as the native (phenotypically neutral) residue at the same position in an orthologue of another species. See Figure 2 and column 419 in Figure 1 for an example of a CPD.

**Functionally equivalent proteins (FEPs)** are orthologues which have maintained the same function during evolution, as discussed by McMillan & Martin (12). Homologous genes (or proteins) have descended from a common ancestor while orthologues are the subset of homologues that arise from speciation events (13). However, if two species have diverged sufficiently, the function of one of the pair of orthologous proteins may diverge. For example, Shibata *et al.* (14) showed that although the general function of exportin-5 proteins (nuclear export of miRNAs and tRNAs) is conserved across different species, substrate specificity varies.

**Co-evolution.** At the molecular level, evolution of each protein molecule, is affected by (potentially numerous) interaction partners and environmental factors. When a similar evolutionary pattern is detected for the two molecules, they are said to be **co-evolving**. This shared evolutionary history can be a consequence of their co-adaptation, shared cellular pathway or localization, or a shared expression pattern (15). In examining CPDs, we are only interested in the first of these — the co-adaptation of two amino acids which affect each other's evolutionary paths.

**Epistasis** is defined as the effects of one gene being modified by one (or several) other genes (sometimes termed 'modifier genes'). Typically

the phenotype of one gene (the 'epistatic' gene) is expressed while the other (the 'hypostatic' gene) is altered or suppressed. This interaction of *different* genetic loci contrasts with normal Mendelian effects, where one allele is 'dominant' over another 'recessive' allele at the *same* locus. In a more general way, epistasis is defined as an inter-dependence between two gene loci as discussed by Cordell (16). In the context of population genetics, 'epistasis' refers to the interaction between alleles at different loci in such a way that the effect on the individual cannot be predicted from merely adding up effects of interacting loci. In the case of CPDs we are interested in the change of fitness of a protein caused by a change of a single amino acid. Fitness may be modified by differences (i.e. amino acid changes) at other locations. While strictly the term 'epistasis' should be applied only to changes in other proteins, when discussing CPDs it is further generalized to refer to changes at other locations within the *same* protein.

**Sign epistasis** also known as **fitness reversal** refers to the situation in which there is a deleterious mutation which co-evolves with a mutation having an epistatic effect that more than compensates for the deleterious effects of the other mutation. Thus the overall fitness change becomes positive (or at least neutral) rather than negative. Sign epistasis facilitates sampling protein space for novel amino acid combinations and provides a mechanism of escape from local fitness minima (2). In some cases, 'fitness reversal' may be used as a more general term (perhaps influenced by epigenetic effects) while sign epistasis refers specifically to the effect of compensatory mutations. In this review, we use the terms interchangeably.

## RNA as a model of compensation

While it is clear that protein and RNA are very different molecules, the simple nature of RNA models has, in general, been widely applied to study evolution. Understanding the importance of compensatory events during evolution is no exception. RNA consists of just four nucleotides: adenine (A), guanine (G), cytosine (C), and uracil (U) and, just as in DNA, stable Watson-Crick pairing can occur between A and U, and between G and C. This can bring relatively remote parts of the molecule together to form stable three-dimensional structures composed of features such as 'stems' (helical base-paired regions) and unpaired regions which form 'loops' (at the end of a stem) or 'bulges' (in the middle of a stem). One can view the RNA sequence as being a 'genotype' while the manifestation of a stable folded structure is the 'phenotype'. The simple nature of RNA folding means that it can be simulated in a computer with a high degree of accuracy using freely available software (see for example Zuker's MFold software (17, 18) and Schuster's ViennaRNA (19)). More recent software can even predict the shapes of RNA molecules during interactions with other molecules (for example, the work of Hofacker (20) and of Matthews (21, 22)).

Computational models of RNA evolution typically simulate a large population of RNA molecules and apply the standard strategy of random mutation followed by natural selection. On the basis that most functional RNA molecules have shapes that are extremely conserved throughout evolution, since shape has a dominant rôle in determining function (23), the fitness of an RNA molecule is determined by predicting its shape and then applying a fitness function based on similarity to some predetermined ideal target shape. Having evaluated the fitness, molecules are allowed to replicate in proportion to their fitness and, during the replication, random mutations are allowed to occur.

The application of RNA models to understanding evolution is reviewed by Cowperthwaite and Meyers (3) and, in an earlier paper, Cowperthwaite *et al.* (2) used these models to examine fitness reversal. They observed that RNA mutations that can be regarded as 'pathogenic' in the model system accumulate more rapidly than expected based on their effect on overall population fitness. Furthermore, they observed that the drop in fitness was not as severe as would be expected based on the accumulation of deleterious variations. Since deleterious effects were not additive, compensatory events were clearly occurring. Indeed, mutations that initially were deleterious accumulated at nearly the same rate as mutations that were immediately beneficial and fixations of more than half of the initially deleterious mutations led to fitness reversals. The fixation of initially deleterious mutations led to a substantial positive effect on the total fitness of the genome. When other mutational events such as 'hitchhiking' and random drift were considered, their model showed that some 80% of PDs were fixed through fitness reversal, or co-adaptation with a compensatory mutation.

In a related study, but using real sequences rather than computer simulations, Meer *et al.* (1) attempted to address the question of whether valleys on the fitness landscape (corresponding to low-fitness genotypes) can be crossed in order to reach isolated fitness peaks. In particular, they examined the switch between AU and GC Watson-Crick nucleotide pairs at equivalent sites in the mitochondrial tRNA stem regions in 83 mammalian species. Clearly, to switch from an AU pair to a GC pair either needs A→G and U→C mutations to occur simultaneously (thus jumping from one fitness peak to another — an unlikely event), or requires one mutation to occur before the other thus passing through a valley of

low fitness where there will be a Watson-Crick mismatch. Because of the need to traverse low-fitness valleys, they found that these 'Watson-Crick switches' occurred 30–40 times more slowly than did pairs of neutral substitutions (where base pairing was not a factor). However, they found that substitutions leading to a Watson-Crick switch were strongly correlated. They were able to estimate the depths of the fitness valleys and showed that AC intermediates are slightly more deleterious than GU intermediates. Nevertheless, the compensatory evolutionary events that do occur must proceed via rare disfavoured intermediate variants that never become fixed in the population.

## Analysis of compensatory events in proteins

As we have seen, computer simulations in RNA, and studies of RNA molecules, have shown that compensatory events do indeed allow traversal of valleys in the fitness landscape. RNA, having only 4 nucleotides is clearly a much simpler system than proteins composed of 20 amino acids, but we know that compensatory events must also occur in proteins. It is hard to say whether the fact that there are 20 amino acids with a wide variety of chemical and physical properties makes it harder or easier to compensate in proteins than in RNA. On the one hand, the subtlety and complexity of interactions made by amino acids may mean that compensatory events are difficult; on the other hand, a change that is damaging might be quite small in nature and therefore only need a small compensatory event, perhaps by a conservative substitution in a nearby amino acid. The compensating event may (if it happens first) not have a particularly negative effect.

Over the past decade, a number of groups have started to look at CPDs in proteins, but while the definition of a CPD is the same, different approaches have been taken to gathering CPD data.

CPDs are identified by (i) identifying missense mutations that lead to disease (generally in humans), (ii) identifying a set of homologous proteins, (iii) performing a multiple alignment of the human sequences with the homologous sequences, (iv) identifying cases where the pathogenic mutation is observed as the native residue in at least one other species. Thus, not surprisingly, datasets of CPDs are highly dependent on (i) the alignment building method, (ii) the thresholds used to detect homologous proteins, and (iii) the choice of species to be tested for homologues. Several methods are summarized in Table I showing a variety of species, cutoffs for identifying homologies to be included in the dataset, and multiple sequence alignment methods. In particular, Poon's approach (24) was rather different from the others. They analyzed deleterious missense mutations from a range of proteins in different species. Rather than use a sequence-comparison approach as used in the other datasets, they analyzed data from publications identified using relevant keyword searches. Thus their data show a very high fraction of deleterious mutations that are compensated because their analysis focussed only on these mutations. In addition they considered mutations introduced with mutagenesis-inducing agents as well as evolutionary events.

Once the data have been collected, some authors performed various analyses to compare and contrast compensated mutations with the rest of the dataset to try to understand whether the nature of mutations that are seen to be compensated (CPDs) is different from those that are not seen to be compensated (PDs). As described in the definitions above, while we use the term PDs strictly to refer to uncompensated mutations, the identification of a clear uncompensated set is not completely rigorous as it is based on a negative observation. Thus the fact that no compensatory event has been observed may simply be because a species that has a compensatory event has not yet been sequenced. Similarly, sequence quality is always a concern (25) and it is possible that apparent CPDs are actually a result of sequencing errors.

Excluding the Poon dataset which is deliberately biased towards compensated mutations, Table I shows that the fraction of disease-causing mutations that are seen to be compensated varies from 0.14% in the Zhang dataset (62/44348) to 19.5% in the Barešić dataset (453/2328), our contribution to this field. This clearly shows that the number of compensatory events is correlated with the evolutionary distance between the species considered. In the Zhang dataset, only humans, chimpanzees and neanderthals were examined, while the high fraction in the Barešić dataset results from the fact that no limit was applied to the divergence of the homologous sequences. As sequences diverge more, the environment around any given residue is likely to be more different and therefore a residue change is more likely to be tolerated, or indeed, required. Kondrashov found that, when using a dataset containing only homologues with at least 50% identity to the reference sequence, on average around 1 in 10 disease-causing mutations is seen to be compensated in other species (11, 26). In contrast, alignments of recently diverged sequences (e.g. three Dipteran genomes (26), or chimpanzee, neanderthal and modern human (9)) show far fewer CPDs (0.4% in the Kulanthinal dataset and 0.14% in the Zhang dataset).

The motivation for not using any sequence identity threshold in our work (10) was that we wished to compare the local structural effects of mutations that could be compensated with those that could not. Therefore having a set of CPDs that was as broad as possible meant that our uncompensatable PD dataset was likely to be more accurate. The dataset was built using only 'functionally equivalent proteins' (FEPs as defined by McMillan *et al.* (12)). Thus while other groups identify homologues using a BLAST (27) search with default parameters (11), or manually curated alignments

Table I: Datasets of compensated pathogenic deviations described in the literature.

| Dataset | Species | Identity cut-off | Alignment method | Human proteins | # DAMs | # CPDs |
|---|---|---|---|---|---|---|
| Kondrashov (11) | Any | | CLUSTALW | 32 | 4880$^{+}$ | 608 |
| | mammals$^{\dagger}$ | >50% | | 3 | | 20 |
| Kulathinal (26) | Diptera | | | 475$^{\circ}$ | 1527 | 6 |
| Ferrer-Costa (33)$^{*}$ | Any | ≥10% (>60%) | Pfam | 287 (24) | 9334 | 1658 (52) |
| | mammals | | | 184 | | 847 |
| Barešić (10) | Any | None$^{\star}$ | MUSCLE | 245 | 2328 | 453 |
| Zhang (missense) (9) | Human, neanderthal, chimpanzee | | ANFO | 2628 | 44348 | 62 |
| Poon (24) Set A | Any | | | 43$^{\ddagger}$ | 115 | 88 |
| Poon (24) Set B | Any | | | 17$^{\ddagger}$ | 59 | 49 |

The Poon Set-A includes mutations brought about by mutagenic agents while Set-B does not.
$^{+}$ Precise numbers are somewhat unclear. They report 608 CPDs and that this is approximately 10% of DAMs. In Table 1 of their paper, there are 4272 'known missense' mutations which we believe to be PDs since the last row of the table has more CPDs that 'known missense' mutations. This makes a total of 4880 (4272+608) DAMs
$^{\dagger}$ Kondrashov tested all found orthologues (with no sequence identity threshold) for CPDs and then switched to mammalian-only orthologues to identify compensatory mutations
$^{\circ}$ In the Kulathinal dataset, the reference species is *D. melanogaster*
$^{*}$ Numbers in parentheses refer to the CPDs with structural data available, used for structural analysis
$^{\star}$ Functional-equivalence among homologues used instead of a sequence identity threshold
$^{\ddagger}$ There is no reference species in the work of Poon *et al.*

from Pfam (28), we selected all orthologues where function has been conserved as defined by annotations in UniprotKB/SwissProt. These data are available in our FOSTA database (29).

## Properties of compensated mutations and mechanisms of compensation

Maintaining a functional protein requires a delicate balance between the residues present in order to obtain proteins having a narrow range of thermodynamic stability, a range of $\Delta G$ from $-3$ to $-10$ kcal/mol. If the stability is any lower, then the protein will start to unfold, becoming a target for degradation; higher stability means that the protein cannot be turned over effectively and therefore often becomes unresponsive to cell-regulation or may lose its activity (30). In addition, mutated proteins of both lower and higher stability than optimal often show increased propensity for aggregation, although aggregation potential is not solely dependent on protein stability. Amino acid substitutions result in an average $\Delta\Delta G$ of 0.5–5 kcal/mol (30), so it is clear that most SAAPs will have a significant effect on protein stability and consequently protein function and the individual's fitness.

From a structural perspective, compensated mutations have been shown to have less damaging effects than uncompensated mutations. Henikoff and Henikoff (31) created the BLOSUM amino acid substitution matrices from around 2000 blocks of aligned sequence segments from more than 500 groups of related proteins to show how frequently one amino acid can substitute for another in homologous proteins. These matrices were designed for use in protein sequence alignment and are familiar to most biologists as the default similarity matrix for use with the BLAST sequence searching tool (32). Ferrer-Costa *et al.* (33) showed that CPDs show significantly larger BLOSUM62 scores than PDs — in other words the amino acid replacements observed in CPDs are more frequently observed to occur in general in homologous proteins, while the replacements seen in PDs are less commonly observed between homologous proteins. They also found that CPDs are characterized by less extreme changes in amino acid volume and hydrophobicity when compared with uncompensated PDs.

In our own work (10), we examined 14 different local structural effects covering stability and folding of the protein, as well as binding effects and functional annotations. We found that CPDs are less likely to display any of these effects, especially if the structural effect is likely to require several consecutive compensatory mutations for full fitness reversal rather than it being possible to compensate using a single substitution. For example, a buried mutation, where a small residue is replaced by a larger residue, could cause a clash. However, while it is theoretically possible that a single mutation could do so, compensation of a clash is most likely to be achieved by making a number of smaller changes. Both Ferrer-Costa *et al.* (33) and Barešić *et al.* (10) found that CPDs have a higher average solvent accessibility. In other words,

they are much more likely to be found on, or near, the protein surface.

## Compensatory mutations in evolution

In the context of evolution, compensated mutations become fixed in a population through 'co-adaptation', or more precisely, through 'sign epistasis' as defined above. At the protein level, depending on the context and rôle of the deleterious mutation ($D$), the compensatory mutation ($C$) can be on the same protein, or on an interacting partner protein. The compensatory mutation, $C$, may have no effect on fitness, or may itself be somewhat deleterious, but at such a level that it can exist in the population. However, the main feature of $C$ is that, when it co-occurs together with the deleterious mutation, $D$, it reverses the negative fitness effect of $D$ to a neutral or positive one and, if $C$ by itself has any deleterious effect, the combination of $C$ and $D$ will also have a neutral or positive effect. Thus during evolution, when fitness landscapes are explored, compensation provides a path through the valleys of lower fitness, allowing individuals to travel from one peak to another (5).

As discussed above, numerous cases of compensation have been identified and documented in proteins (10, 9, 11). While a classic compensatory event to achieve fitness reversal would result from $C$ being a single amino acid change, in proteins it is also perfectly possible — and indeed more likely — for $C$ to consist of a complete change in environment from multiple amino acid changes.

Poon *et al.* (24) set out to study how many different compensatory mutations act on a given deleterious mutation. They performed a maximum-likelihood analysis of experimental data collected from literature on suppressor mutations (which are equivalent to compensatory mutations) to determine the shape of the statistical distribution for the number of compensatory mutations per deleterious mutation. They found that the data were best explained by an L-shaped gamma distribution which predicted an average of 11.8 compensatory mutations per deleterious mutation in order to achieve full sign epistasis and compensate for the deleterious effect of a DAM (24). Interestingly, they also found that, when they partitioned the data into viruses, prokaryotes and eukaryotes, there was a significant improvement in the fit to the model: on average, there were fewer compensatory mutations in viruses than in prokaryotes or eukaryotes. They suggested that the differences in genome size and gene length in viruses compared with prokaryotes and eukaryotes means that the number of possible interactions within and between gene products is constrained.

In our more recent structural study (10), we showed that CPDs are surrounded by significantly more diverged residues than PDs. As described above, we created sequence alignments of functionally equivalent homologous proteins for each instance in which a human deleterious mutation (DAM) is known (typically identified from OMIM (34, 35), but also from a number of locus-specific mutation databases (7)). The DAMs were then assigned as CPDs or PDs depending on whether the damaging mutant residue was observed as the native in another species. Where a structure was available for the human protein, we identified amino acids within an 8Å sphere around the DAM. Having identified these structural neighbours in the human protein which form the environment surrounding the DAM, we mapped their positions back onto the sequence alignment. We were then able to calculate the fraction of these structurally neighbouring residues that were mutated in each of the sequences when compared with the human sequence. For CPDs this was done just with the sequences in which compensation was observed, while for PDs it was done for each sequence in the alignment. We then plotted this local fraction of mutated residues against the overall (whole protein) sequence identity between the human and non-human sequence.

We found that this environmental 'sphere' compensation appeared on average to occur as a result of random drift in the sequence. We fitted a straight line to the data imposing the biologically obvious constraint that the line had to pass through the 100% identity, zero mutations point — if the sequences are 100% identical then there can be no mutations within the local environment. Allowing for the fact that sequence identity ranges from 0–100% while our fraction of mutations scale runs from 0–1 (and that one scale is scoring conservation while the other is scoring mutations), this fitting revealed a slope of 1.007 for CPDs and 0.9 for PDs. The slope of $\sim 1$ for CPDs implies that the environment around a CPD is mutated at the same rate as the sequence overall such that compensation occurs as a result of random drift in the sequence. On the other hand, the environment around PDs is more conserved than the sequence as a whole.

While this 'sphere compensation' is probably the most common compensatory mechanism in proteins, the alternative classical 'one-on-one compensation' can also occur where one deleterious SAAP is compensated by another single mutation in the structural vicinity. This type of compensation is easier to detect, especially in analyses where only recently diverged homologues are considered (9, 11, 26). Two examples of one-on-one compensations are shown in the Case Study presented below.

Poon *et al.*'s study (24) also investigated whether compensatory mutations are intragenic (i.e. occur within the same gene, and hence the same protein chain, as the deleterious mutation) or intergenic (i.e. occur within a different gene and protein chain from the DAM). Overall, from their dataset of 129 CPDs, they found that the majority (78%) of compensatory mutations were intragenic suggesting that the complexity of interactions between proteins is likely to be less important than the complex-

ity of the protein itself. However, when they studied different taxa separately, they found that compensation is much less likely to be intragenic in viruses (69%) than in prokaryotes (92%) and eukaryotes (90%). They proposed that this is probably a result of the fact that viral genes tend to be shorter, thus limiting the number of internal interactions.

Work performed by Povolotskaya and Kondrashov (36) suggests that compensated pathogenic deviations are unidirectional drivers of evolution; once compensation occurs, it is unlikely that sequences will revert to the original wild-type state. Their investigation of divergence of proteins in sequence space showed that, at any given point in time, only 2% of all possible missense mutations are allowed in order to avoid non-functional protein products. If we assume that only one missense mutation at a time can be introduced into the sequence, then we can consider how this observation affects a protein chain consisting of 100 residues. For every residue there are 19 possible substitutions, so at any one time 1900 ($100 \times 19$) mutational events could occur. Given that 2% of these are 'allowed', 38 missense mutations will result in a functional protein. Let us assume that an allowed mutation of residue $X$ to residue $Y$ occurs at position $n$ (i.e. $X_n \rightarrow Y$). At the next step, there will again be 38 allowed missense mutations, one of which would be the reversal of the mutation that occurred in the previous step (i.e. $Y_n \rightarrow X$). Thus there is a 1 in 38 (2.6%) chance that this will occur, but a 97.4% chance that another mutation will occur. From this statistic, we do not know how the 38 allowed mutations will be distributed across the 100 amino acid positions of the protein. Thus the second mutation could result in $Y_n \rightarrow Z$, but in general it is much more likely that the mutation will occur at a location $m$ that is different from $n$. Thus we are much more likely to obtain a double mutant after the second step than we are to obtain a reversion to the original sequence or to introduce a different amino acid at position $n$. Consequently, subsequent mutational events will lead to a drift away from the original sequence and it is intuitive that compensation will be observed significantly more often than reversal to the original sequence.

The question remains as to the timeline of compensatory events. As we discussed in our previous work (10), DePristo *et al.* (30) proposed two hypotheses of CPD evolution based on models of biophysical properties. In the first scenario, a compensatory mutation $C$ is phenotypically neutral and stable, thus fixing itself quickly in the population. The deleterious mutation $D$ is unstable, and can only become fixed in the population if it occurs *after* the compensatory mutation $C$. Thus $D$ will exist as a CPD because of the compensatory effect of $C$. In the second model, both $D$ and $C$ are individually deleterious, but either can exist in the population at low levels; it is known that small frequencies of low-fitness mutations exist in large populations. Consequently if $D$ is present in the population at low levels, then $C$ can occur later and fix the $D$–$C$ genotype in the population because of the epistatic effect of the mutant-pair. Cowperthwaite *et al.* (2), in their *in silico* RNA models discussed earlier, confirmed that the deleterious mutation, $D$, can occur first. Equally the compensatory mutation $C$ can be present in the population at low levels and $D$ can occur later leading to fixation of the $D$–$C$ genotype in the same way. A less likely, but possible, scenario is that both $C$ and $D$ occur simultaneously. Provided the mutation rate is sufficiently high, epistatic selection with compensatory mutations is the most prevalent mechanism for fixation of otherwise deleterious mutations.

## Artificial compensatory events

With recent advances in sequencing technology (37), sequencing large amounts of genomic data is becoming cheaper, faster and more accessible, providing new opportunities in biomedical research. Genome-wide association studies (GWAS) are becoming more and more widespread, associating mutations with both high and low penetrance disease phenotypes. An important area of interest is the ability to predict whether a given mutation — particularly a SAAP — will lead to disease. Numerous tools have been developed both to analyze the local structural effects of mutations and to predict whether mutations will be damaging, many of these working mostly at the sequence level. Among these are SAAPdb (7), SNPs3D (38), stSNP (39), ModSNP (40), MutDB (41), LS-SNP (42), TopoSNP (43), SIFT (44), SNPeffect (45), PolyPhen (46, 47), subPSEC (48) and nsSNPAnayzer (49).

Recently, Critical Assessment of Genome Interpretation (CAGI) (50), a community experiment to assess computational methods for predicting the phenotypic impacts of genomic variation objectively, organized by Steve Brenner, John Moult and Susanna Repo, was run for the first time. Participants were provided with genetic variant data and asked to make predictions of the resulting molecular, cellular, or organismal phenotype. Results from over 100 prediction submissions from 8 countries were evaluated against experimental data by independent assessors and discussed at a workshop in December 2010 (see `http://genomeinterpretation.org/`).

One of the prediction datasets was particularly interesting in the context of compensated mutations. A dataset of p53 mutations (see `http://genomeinterpretation.org/content/p53/`) was designed to test prediction of 'cancer rescue mutations'. p53 is a tumour suppressor protein which plays a central rôle in detecting DNA damage, slowing the cell cycle to allow DNA repair enzymes to do their work (51), or if DNA damage is too severe, triggering

programmed cell death (apoptosis) (52, 53). If p53 is rendered non-functional as a result of mutation, this central checkpoint is lost, allowing other mutations to accumulate in the DNA eventually leading to cancer. Unusually for tumour suppressor genes, in which most mutations tend to be frameshifts or nonsense codons, the majority of mutations in p53 are single DNA base changes resulting in a SAAP. In some cases, mutations at second intragenic sites are known to rescue the function reactivating otherwise inactive p53 (54, 55, 56) and are therefore acting as compensatory mutations. Rick Lathrop's group at the University of California, Irvine, has been performing a complete functional census of these cancer rescue mutations (57). In this case, the aim of the CAGI prediction experiment was to predict whether a given mutation is able to rescue the function of p53 and thus act as a compensatory mutation (58, 59). While the results of the CAGI prediction experiment have not been published at the time of writing this review, we suspect the field of compensation prediction will progress significantly in the near future. If a disease-associated deleterious mutation is amenable to compensation (i.e. it is seen to be a CPD), it is likely that other (non-mutational) mechanisms of compensation may also be applied. For example, Alan Fersht's group in Cambridge has shown that some p53 mutations can be compensated by binding small peptides that stabilize the p53 core domain (60, 61). More recently, small molecules which are more likely to be usable drug leads have been used successfully in the same way (62, 63, 64, 65).

## Case study: two compensated mutations and their environment

There are many examples of compensation which include the p53 rescue mutations described above where, in some cases, crystal structures have been solved to study the mechanism of compensation (55). Here we will discuss two examples of compensated mutations. First, a CPD in human GTP cyclohydrolase (GTPCH) is presented, with an obvious destabilizing effect on the protein structure, while a compensating mutation has a stabilizing effect restoring enzyme activity. The second example, in ornithine transcarbamylase (OTC), is less obvious at the structural level, despite being confirmed by *in vitro* enzyme activity experiments.

GTPCH, encoded by the gene GCH1, plays a rôle in the folate and biopterin biosynthesis pathways and hydrolyses guanosine triphosphate (GTP) to form 7,8-dihydroneopterin-3′-triphosphate. This is the first step in the biosynthesis of tetrahydrobiopterin, an essential cofactor required by aromatic amino acid hydroxylase (AAAH) and nitric oxide synthase (NOS). These, in turn, are involved in the biosynthesis of monoamine neurotransmitters such as serotonin, melatonin, dopamine, noradrenaline, adrenaline and nitric oxide. Mutations are associated with phenylketonuria (PKU) and hyperphenylalaninemia (HPA), as well as levodopa-responsive dystonia.

Figure 3a shows the whole wild-type GTP cyclohydrolase I which consists of five identical chains, with mutually parallel C′ helices stabilizing the pentameric structure (66). The images, rendered with PyMol (`http://www.pymol.org/`), are based on coordinates obtained from Protein Databank entry 1FB1 accessible online at `http://www.pdb.org/` (67). Figure 3b shows detail of the wild-type residues that are mutated (residues 249 and 250). The wild-type Arg249 in one chain and Ser250 in the next chain form a tight ring-like structure.

An Arg249→Ser mutation is associated with disease, causing a severe decrease in enzyme activity and resulting in recessive levodopa-responsive dystonia (OMIM:600225.0016). Figure 3c shows the effect of introducing an Arg249→Ser mutation in all five chains modelled using the minimum perturbation protocol (68) implemented in the program Mutmodel (69). The non-covalent interactions between residues 249 and 250 are reduced, presumably destabilizing the complex and leading to disease. However, the functionally-equivalent protein in *Rickettsia bellii* has a serine at 249 in the wild-type enzyme, but also has a compensatory lysine at 250, which is also modelled into the structure in Figure 3d restoring, and indeed enhancing, the ring-like set of interactions.

A less clear example of a compensatory mutation is seen in Ornithine transcarbamylase (OTC) which catalyzes the reaction between carbamoyl phosphate and ornithine to form citrulline and phosphate. In prokaryotes and plants, it is involved in arginine biosynthesis, while in mammals it is a key enzyme of the urea cycle. Figure 4a shows one monomer of the enzyme which exists as a trimer. The structure for OTC in the Protein Databank shows only a monomer (PDB ID: 1OTH), but the assembly of the biologically relevant trimer can be obtained from PISA (70) available online at `http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html`. OTC deficiency, although a rare condition occurring in around 1 in 80,000 births, is the most common disorder of the urea cycle which removes ammonia from the body. Mutations in OTC lead to an accumulation of toxic ammonia which can lead to developmental delay and mental retardation, progressive liver damage, skin lesions, poorly-controlled breathing, seizures, coma and death.

Figure 4b shows helix 3 from PDB entry 1OTH and highlights residues 125 and 135 in red and green respectively (71). Thr125→Met is a known disease causing mutation in humans resulting in lethal neonatal congenital hyperammonemia (OMIM:311250). Suriano *et al.* (72) showed that the human enzyme with the Thr125→Met mutation has a negligible rate of enzyme activity in *in vitro* constructs. However this mutation is a CPD as

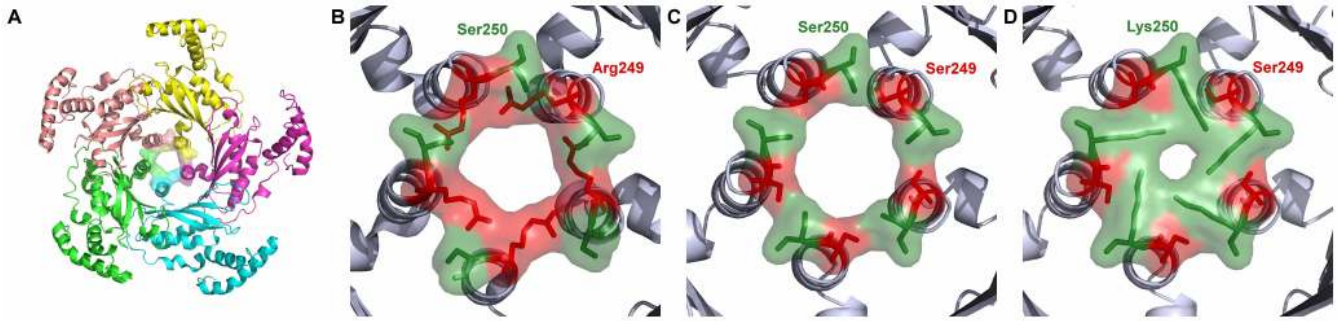


Figure 3: Compensated mutation in human GTP cyclohydrolase I. Residues 249 and 250 are shown with a surface in all 5 chains. **A)** Structure of the wild-type homo-pentamer with each chain shown in a different colour. **B)** Zoomed view of residues 249 and 250 from all five chains with the residues shown in green and red, respectively. **C)** The disease-causing Arg249→Ser mutation modelled into all five chains. **D)** The compensatory Ser250→Lys mutation modelled into all five chains as well as the Arg249→Ser mutation.

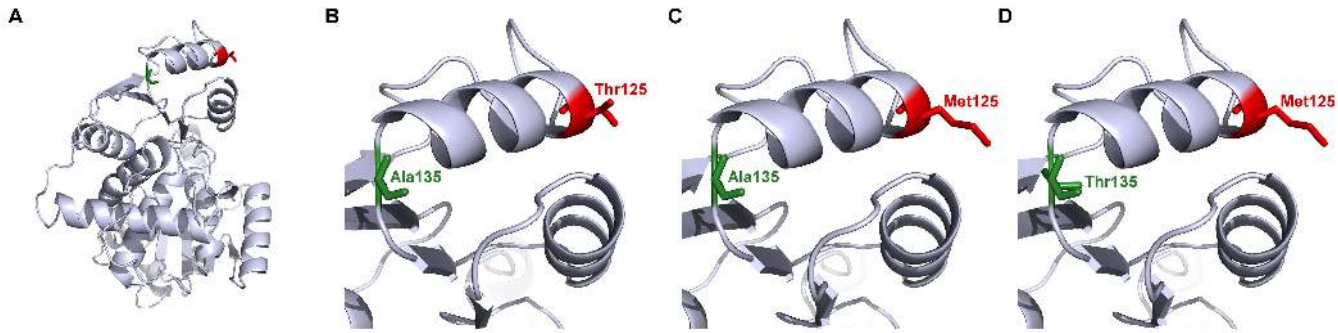


Figure 4: Compensated mutation in human ornithine transcarbamylase. **A)** Structure of wild-type human OTC. **B)** View on helix 3, with residues 125 and 135 shown in red and green respectively. **C)** The disease-causing Thr125→Met mutation, modelled structure. **D)** A compensatory Ala135→Thr in addition to the Thr125→Met mutation.

Met is the native residue in chimpanzees. The only other residue which differs between human and chimp OTC is residue 135 where there is an Ala→Thr mutation which must compensate for the deleterious effect of the Thr125→Met mutation. However the mechanism of compensation is unclear as Figure 4 shows that residues 125 and 135 are not in direct contact and this is also the case in the trimer. However, as previously suggested by Azevedo *et al.* (73), the presence of Thr125 might be crucial at the end of helix 3 since this helix is involved in trimerization of human OTC, and the chimpanzee compensates for its loss by having a threonine introduced at the other end of helix 3 (at position 135), restoring enzyme activity to rates similar to human wild-type. Interestingly, Suriano *et al.* (72) also suggested that the ancestral genotype may have had threonines at both positions 125 and 135 and had a higher enzyme activity than either the human or chimpanzee enzymes. If this is the case, then this mutation is an example of two species starting to explore fitness ridges, in search of another local optimum.

# Expert Opinion

In conclusion, while there is also the possibility that epigenetic effects can also be compensatory (i.e. some difference in the non-protein environment), compensation of deleterious mutations through epistatic protein mutations is a very common effect. The frequency of these compensatory mutations depends on the time elapsed from the common ancestor and the data in Table I show that there is a correlation between the frequency of CPDs and the diversity of the homologues used to detect CPDs. For example, our dataset (10) (where we apply no constraint on the sequence identity between functionally equivalent homologues) shows a higher ratio of CPDs compared with the dipteran-only (26) or mammalian-only (9, 11, 33) datasets.

Study of the evolution of RNA molecules and *in silico* models of RNA evolution show clear examples of one-on-one compensation (2). While compensation in proteins is often more complex, involving multiple compensatory events changing the environment in which a residue exists, there are also several examples of one-on-one compensation including 'cancer rescue' mutations in p53.

As shown by DePristo *et al.* (30), any mutation has an average effect on protein stability ($\Delta\Delta G$) of around

0.5–5 kcal/mol. Restoring protein stability and hence regulated activity, will often need compensatory mutations that restore stability to the acceptable range of free energies. From a structural analysis perspective, compensated mutations are preferentially on the protein surface (10, 33). As shown by Poon *et al.* (24), compensatory events are most commonly intragenic, so the surface location is likely to be a result of it being easier to accumulate compensatory events (probably before the CPD mutation occurs) rather than it being anything to do with interactions with other proteins. In addition CPDs have 'milder' effects on the protein structure than uncompensated mutations (10, 33) and tend to be more conservative in nature (33).

# Outlook

CPDs will continue to be an interesting area of research in understanding evolution and traversal of the fitness landscape. As more species are sequenced, the identification of true PDs will become more accurate. This will allow us to compare CPDs and PDs in a more accurate manner and therefore understand more completely which mutations can be easily compensated and which cannot. The CAGI experiment described above has led the way with the challenge of predicting which mutations will be 'cancer rescue' mutations in p53 and this will become a more significant area of research. The fact that certain mutations can be rescued or compensated by an amino acid change will also allow us to identify types of mutations that, in general, can be more easily rescued leading us towards the possibility of drugs that can rescue protein function. Consequently studying CPDs is not only of interest in understanding evolution, but is also important in developing future drugs.

# Highlights

- Compensation of deleterious mutations through epistatic protein mutations is a very common effect.
- The frequency of compensated mutations depends on the time elapsed from the common ancestor — more diverged sequences are more likely to show compensatory events
- Study of RNA molecules and *in silico* models of RNA evolution clearly show one-on-one compensation.
- Compensation in proteins is more likely to involve multiple compensatory events, but there are also several examples of one-on-one compensation.
- 'Cancer rescue mutations' in p53 are an example of one-on-one compensation.
- CPDs are more likely to occur on the protein surface, be more conservative in nature and be less damaging in structural terms than PDs.
- Prediction of compensatable mutations may allow design of drugs that are able to compensate for the effects of a damaging mutation.

# Acknowledgments

AB was supported by an UK Overseas Research Scholarship and by a UCL Graduate School Research Scholarship.

# References

1. Meer, M. V., Kondrashov, A. S., Artzy-Randrup, Y., and Kondrashov, F. A. Compensatory evolution in mitochondrial tRNAs navigates valleys of low fitness. Nature (London) 2010;464:279–282.

2. Cowperthwaite, M. C., Bull, J. J., and Meyers, L. A. From bad to Good: Fitness Reversals and the Ascent of Deleterious Mutations. PLoS Comput. Biol. 2006;2:e141.

3. Cowperthwaite, M. C. and Meyers, L. A. How mutational networks shape evolution: Lessons from RNA models. Annu. Rev. Ecol. Evol. and Systematics 2007;38:203–230.

4. Kimura, M. The role of compensatory neutral mutations in molecular evolution. J. Genet. 1985;64:7–19.

5. Wright, S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. Proc. VI Intl. Cong. Genet. 1932;1:356–366.

6. Dawkins, R. "Climbing Mount Improbable". W.W. Norton, 1996.

7. Hurst, J. M., McMillan, L. E. M., Porter, C. T., Allen, J., Fakorede, A., and Martin, A. C. R. The SAAPdb web resource: a large-scale structural analysis of mutant proteins. Human Mut. 2009;30:616–624.

8. Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29:308–311.

9. Zhang, G., Pei, Z., Krawczak, M., Ball, E. V., Mort, M., Kehrer-Sawatzki, H., and Cooper, D. N. Triangulation of the human, chimpanzee, and neanderthal genome sequences identifies potentially compensated mutations. Human Mutation 2010;31:1286–1293.

10. Barešić, A., Hopcroft, L. E. M., Rogers, H. H., Hurst, J. M., and Martin, A. C. R. Compensated pathogenic deviations: Analysis of structural effects. J. Mol. Biol. 2010;396:19–30.

11. Kondrashov, A. S., Sunyaev, S., and Kondrashov, F. A. Dobzhansky-Muller incompatibilities in protein evolution. Proc. Natl. Acad. Sci. USA 2002;99:14878–14883.

12. McMillan, L. E. M. and Martin, A. C. R. Automatically extracting functionally equivalent proteins from SwissProt. BMC Bioinf. 2008;9:418–418.

13. Fitch, W. M. Homology a personal view on some of the problems. Trends Genet. 2000;16:227–231.

14. Shibata, S., Sasaki, M., Miki, T., Shimamoto, A., Furuichi, Y., Katahira, J., and Yoneda, Y. Exportin-5 orthologues are functionally divergent among species. Nucleic Acids Res. 2006;34:4711–4721.

15. Pazos, F. and Valencia, A. Protein co-evolution, co-adaptation and interactions. EMBO J. 2008;27:2648–2655.

16. Cordell, H. J. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Human Molec. Genet. 2002;11:2463–2468.

17. Zuker, M. and Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. 1981;9:133–148.

18. Zuker, M. On finding all suboptimal foldings of an RNA molecule. Science 1989;244:48–52.

19. Hofacker, I., Fontana, W., Stadler, P., Bonhoeffer, L., Tacker, M., and Schuster, P. Fast folding and comparison of RNA secondary structures. Monatsh. Chem. 1994;125:167–188.

20. Bernhart, S. H., Tafer, H., Mückstein, U., Flamm, C., Stadler, P. F., and Hofacker, I. L. Partition function and base pairing probabilities of RNA heterodimers. Algorithms Mol. Biol. 2006;1:3–3.

21. Mathews, D. Revolutions in RNA secondary structure prediction. J. Mol. Biol. 2006;359:526–532.

22. Mathews, D. and Turner, D. Prediction of RNA secondary structure by free energy minimization. Curr. Opin. Struct. Biol. 2006;16:270–278.

23. Doudna, J. A. Structural genomics of RNA. Nature: Struct. Biol. 2000;7 Suppl:954–956.

24. Poon, A., Davis, B. H., and Chao, L. The coupon collector and the suppressor mutation: Estimating the number of compensatory mutations by maximum likelihood. Genetics 2005;170:1323–1332.

25. Dawson, K., Thorpe, R. S., and Malhotra, A. Estimating genetic variability in non-model taxa: a general procedure for discriminating sequence errors from actual variation. PLoS One 2010;5:e15204–e15204.

26. Kulathinal, R. J., Bettencourt, B. R., and Hartl, D. L. Compensated deleterious mutations in insect genomes. Science 2004;306:1553–1554.

27. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. J. Mol. Biol. 1990;215:403–410.

28. Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., and Bateman, A. The Pfam protein families database. Nucleic Acids Research 2010;38:D211–D222.

29. McMillan, L. E. M. and Martin, A. C. R. Automatically extracting functionally equivalent proteins from SwissProt. BMC Bioinf. 2008;9:418.

30. DePristo, M. A., Weinreich, D. M., and Hartl, D. L. Missense meanderings in sequence space: a biophysical view of protein evolution. Nat. Rev. Genet. 2005;6:678–687.

31. Henikoff, S. and Henikoff, J. G. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA 1992;89:10915–10919.

32. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. J. Mol. Biol. 1990;215:403–410.

33. Ferrer-Costa, C., Orozco, M., and de la Cruz, X. Characterization of compensated mutations in terms of structural and physico-chemical properties. J. Mol. Biol. 2007;365:249–256.

34. Amberger, J., Bocchini, C. A., Scott, A. F., and Hamosh, A. McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res. 2009;37:D793–D796.

35. McKusick, V. A. Online Mendelian Inheritance in Man (OMIM) (TM). McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD) 2011;http://www.ncbi.nlm.nih.gov/omim/.

36. Povolotskaya, I. S. and Kondrashov, F. A. Sequence space and the ongoing expansion of the protein universe. Nature 2010;465:922–926.

37. Bentley, D. R. Whole-genome re-sequencing. Curr. Opin. Genet. Dev. 2006;16:545–552.

38. Yue, P., Melamud, E., and Moult, J. SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinf. 2006;7:166–166.

39. Uzun, A., Leslin, C. M., Abyzov, A., and Ilyin, V. Structure SNP (StSNP): a web server for mapping and modeling nsSNPs on protein structures with linkage to metabolic pathways. Nucleic Acids Res. 2007;35:W384–W392.

40. Yip, Y. L., Scheib, H., Diemand, A. V., Gattiker, A., Famiglietti, L. M., Gasteiger, E., and Bairoch, A. The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. Human Mut. 2004;23:464–470.

41. Dantzer, J., Moad, C., Heiland, R., and Mooney, S. MutDB services: Interactive structural analysis of mutation data. Nucleic Acids Res. 2005;33:W311–W314.

42. Karchin, R., Diekhans, M., Kelly, L., Thomas, D. J., Pieper, U., Eswar, N., Haussler, D., and Sali, A. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics 2005;21:2814–2820.

43. Stitziel, N. O., Binkowski, T. A., Tseng, Y. Y., Kasif, S., and Liang, J. topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. Nucleic Acids Res. 2004;32:D520–D522.

44. Ng, P. C. and Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31:3812–3814.

45. Reumers, J., Maurer-Stroh, S., Schymkowitz, J., and Rousseau, F. SNPeffect v2.0: a new step in investigating the molecular phenotypic effects of human non-synonymous SNPs. Bioinformatics 2006;22:2183–2185.

46. Ramensky, V., Bork, P., and Sunyaev, S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 2002;30:3894–3900.

47. Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. A method and server for predicting damaging missense mutations. Nature: Methods 2010;7:248–249.

48. Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 2003;13:2129–2141.

49. Bao, L., Zhou, M., and Cui, Y. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res. 2005;33:W480–W482.

50. Callaway, E. Mutation-prediction software rewarded, 2010. Nature News, doi:10.1038/news.2010.679.

51. Vogelstein, B., Lane, D., and Levine, A. J. Surfing the p53 network. Nature (London) 2000;408:307–310.

52. Lakin, N. D. and Jackson, S. P. Regulation of p53 in response to DNA damage. Oncogene 1999;18:7644–7655.

53. Chao, C., Saito, S., Kang, J., Anderson, C. W., Appella, E., and Xu, Y. p53 transcriptional activity is essential for p53-dependent apoptosis following DNA damage. EMBO J. 2000;19:4967–4975.

54. Nikolova, P. V., Wong, K. B., DeDecker, B., Henckel, J., and Fersht, A. R. Mechanism of rescue of common p53 cancer mutations by second-site suppressor mutations. EMBO J. 2000;19:370–378.

55. Joerger, A. C., Ang, H. C., Veprintsev, D. B., Blair, C. M., and Fersht, A. R. Structures of p53 cancer mutants and mechanism of rescue by second-site suppressor mutations. J. Mol. Biol. 2005;280:16030–16037.

56. Danziger, S. A., Swamidass, S. J., Zeng, J., Dearth, L. R., Lu, Q., Chen, J. H., Cheng, J., Hoang, V. P., Saigo, H., Luo, R., Baldi, P., Brachmann, R. K., and Lathrop, R. H. Functional census of mutation sequence spaces: the example of p53 cancer rescue mutants. IEEE/ACM Trans. Comput. Biol. Bioinform. 2006;3:114–125.

57. Baronio, R., Danziger, S. A., Hall, L. V., Salmon, K., Hatfield, G. W., Lathrop, R. H., and Kaiser, P. All-codon scanning identifies p53 cancer rescue mutations. Nucleic Acids Res. 2010;38:7079–7088.

58. Danziger, S. A., Zeng, J., Wang, Y., Brachmann, R. K., and Lathrop, R. H. Choosing where to look next in a mutation sequence space: Active Learning of informative p53 cancer rescue mutants. Bioinformatics 2007;23:i104–i114.

59. Danziger, S. A., Baronio, R., Ho, L., Hall, L., Salmon, K., Hatfield, G. W., Kaiser, P., and Lathrop, R. H. Predicting positive p53 cancer rescue regions using Most Informative Positive (MIP) active learning. PLoS Comput. Biol. 2009;5:e1000498–e1000498.

60. Friedler, A., Hansson, L. O., Veprintsev, D. B., Freund, S. M. V., Rippin, T. M., Nikolova, P. V., Proctor, M. R., Rüdiger, S., and Fersht, A. R. A peptide that binds and stabilizes p53 core domain: Chaperone strategy for rescue of oncogenic mutants. Proc. Natl. Acad. Sci. USA 2002;99:937–942.

61. Friedler, A., DeDecker, B. S., Freund, S. M. V., Blair, C., Rüdiger, S., and Fersht, A. R. Structural distortion of p53 by the mutation R249S and its rescue by a designed peptide: Implications for "mutant conformation". J. Mol. Biol. 2004;336:187–196.

62. Boeckler, F. M., Joerger, A. C., Jaggi, G., Rutherford, T. J., Veprintsev, D. B., and Fersht, A. R. Targeted rescue of a destabilized mutant of p53 by an in silico screened drug. Proc. Natl. Acad. Sci. USA 2008;105:10360–10365.

63. Girardini, J. E. and Del Sal, G. Improving pharmacological rescue of p53 function: RITA targets mutant p53. Cell Cycle 2010;9:2059–2062.

64. Zhao, C. Y., Grinkevich, V. V., Nikulenkov, F., Bao, W., and Selivanova, G. Rescue of the apoptotic-inducing function of mutant p53 by small molecule RITA. Cell Cycle 2010;9:1847–1855.

65. Zhao, C. Y., Szekely, L., Bao, W., and Selivanova, G. Rescue of p53 function by small-molecule RITA in cervical carcinoma by blocking E6-mediated degradation. Cancer Res. 2010;70:3372–3381.

66. Auerbach, G., Herrmann, A., Bracher, A., Bader, G., Gutlich, M., Fischer, M., Neukamm, M., Garrido-Franco, M., Richardson, J., Nar, H., Huber, R., and Bacher, A. Zinc plays a key role in human and bacterial GTP cyclohydrolase I. Proc. Natl. Acad. Sci. USA 2000;97:13567–13572.

67. Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D., and Zardecki, C. The Protein Data Bank. Acta Crystallogr. D Biol. Crystallogr. 2002;58:899–907.

68. Shih, H. L., Brady, J., and Karplus, M. Structure of proteins with single-site mutations: A minimum perturbation approach. Proc. Natl. Acad. Sci. USA 1985;82:1697–1700.

69. Martin, A. C. R., Facchiano, A. M., Cuff, A. L., Hernandez-Boussard, T., Olivier, M., Hainaut, P., and Thornton, J. M. Integrating mutation data and structural analysis of the tp53 tumor-suppressor protein. Human Mut. 2002;19:149–164.

70. Krissinel, E. and Henrick, K. Inference of macromolecular assemblies from crystalline state. J. Mol. Biol. 2007;372:774–797.

71. Shi, D., Morizono, H., Ha, Y., Aoyagi, M., Tuchman, M., and Allewell, N. M. 1.85-A resolution crystal structure of human ornithine transcarbamoylase complexed with N-phosphonacetyl-L-ornithine. Catalytic mechanism and correlation with inherited deficiency. J. Mol. Biol. 1998;273:34247–34254.

72. Suriano, G., Azevedo, L., Novais, M., Boscolo, B., Seruca, R., Amorim, A., and Ghibaudi, E. M. In vitro demonstration of intra-locus compensation using the ornithine transcarbamylase protein as model. Human Molec. Genet. 2007;16:2209–2214.

73. Azevedo, L., Suriano, G., van Asch, B., Harding, R. M., and Amorim, A. Epistatic interactions: how strong in disease and evolution? Trends Genet. 2006;22:581–585.