Dobzhansky–Muller incompatibilities in protein evolution

Alexey S. Kondrashov[†], Shamil Sunyaev[‡§], and Fyodor A. Kondrashov[†¶]

[†]National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD 20894; and [‡]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany

Communicated by M. S. Meselson, Harvard University, Cambridge, MA, September 17, 2002 (received for review June 6, 2002)

We study fitness landscape in the space of protein sequences by relating sets of human pathogenic missense mutations in 32 proteins to amino acid substitutions that occurred in the course of evolution of these proteins. On average, ≈10% of deviations of a nonhuman protein from its human ortholog are compensated pathogenic deviations (CPDs), i.e., are caused by an amino acid substitution that, at this site, would be pathogenic to humans. Normal functioning of a CPD-containing protein must be caused by other, compensatory deviations of the nonhuman species from humans. Together, a CPD and the corresponding compensatory deviation form a Dobzhansky–Muller incompatibility that can be visualized as the corner on a fitness ridge. Thus, proteins evolve along fitness ridges which contain only ≈10 steps between successive corners. The fraction of CPDs among all deviations of a protein from its human ortholog does not increase with the evolutionary distance between the proteins, indicating that substitutions that carry evolving proteins around these corners occur in rapid succession, driven by positive selection. Data on fitness of interspecies hybrids suggest that the compensatory change that makes a CPD fit usually occurs within the same protein. Data on protein structures and on cooccurrence of amino acids at different sites of multiple orthologous proteins often make it possible to provisionally identify the substitution that compensates a particular CPD.

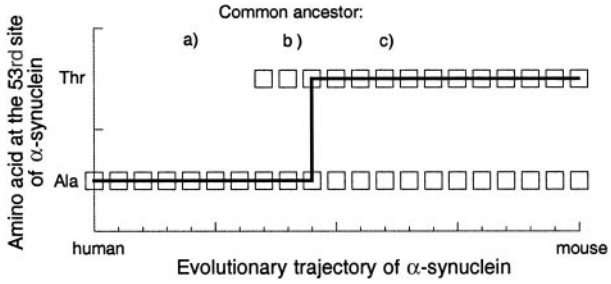

Fig. 1. Evolutionary implications of a compensated pathogenic deviation. Fit genotypes, which must provide a continuous path between human and mouse, are marked by squares, and the line shows the actual evolutionary trajectory. If we trace the evolution of α-synuclein (8) from human to mouse, at least two events will be encountered. First, the murine state, pathogenic in humans, becomes fit, because of a compensatory change(s). At this point, the fitness ridge that connects human and murine α-synucleins contains a corner, due to a DM incompatibility (11, 12), i.e., a situation in which three combinations of two binary factors are fit but the fourth one is unfit. Second, at the 53rd site of α-synuclein, Thr replaces Ala. There are three possible positions (a, b, and c) of the common ancestor of humans and mice relative to these two events. If the human state is also pathogenic in mouse (on which we have no data), the fitness ridge must contain at least two corners (not shown).

Evolution unfolds on a fitness landscape, a map that relates fitness to the genotype (1). Obviously, most of possible genotypes are always unfit, and some of rare fit genotypes must be arranged in continuous ridges (networks). This general paradigm can be applied, *inter alia*, to the evolution of proteins. "If evolution . . . is to occur, functional proteins must form a continuous network which can be traversed by unit mutational steps without passing through non-functional intermediates" (2). However, data on fitness landscapes are limited, because only a tiny fraction of all possible genotypes is actually available, and inferring fitness of currently nonexisting genotypes is difficult (3–6). Relating data on human pathogenic missense mutations, which represent unfit genotypes, to interspecies differences between homologous proteins, all of which must be fit, offers a novel opportunity to probe the fitness landscape in the space of proteins.

Human pathogenic amino acid substitutions tend to occur at less variable sites of proteins (7). Thus, an amino acid that in a nonhuman protein is different from the amino acid at the homologous site of the human ortholog would probably be benign for humans if placed into this site. Still, exceptions to this rule have been described (8, 9).

For example, the 53rd site of human α-synuclein is normally occupied by Ala, and Ala → Thr substitution at this site predisposes to Parkinson's disease. Nevertheless, healthy mice (and rats) carry Thr at the homologous site of their α-synucleins (8). We call such a situation a compensated pathogenic deviation (CPD), because high fitness of this Thr in mice must be due to some other, compensatory difference of mice from humans (10), either within or outside α-synuclein. Together, a CPD and the corresponding compensatory change form a Dobzhansky–Muller (DM) incompatibility (11, 12) (Fig. 1). At the level of intraspecies genetic variation, this phenomenon is known as genetic suppression (13) of deleterious missense mutations by compensatory second-site substitutions (14), either in the same protein (15–19) or in another molecule (20). Here, we report the results of a systematic search for compensated pathogenic deviations from human proteins.

Methods

We were able to identify 32 human proteins for each of which (*i*) at least 50 different pathogenic missense mutations are known, and (*ii*) sequences of orthologous proteins with identity >50% are known from at least three other species.

Pathogenic missense mutations were taken from locus-specific databases (see Table 2, which is published as supporting information on the PNAS web site, www.pnas.org) or from reviews on CYBB (21) and HPRT1 (22). For a locus, let Tm be the number of all amino acid substitutions possible because of a single nucleotide substitution, Pm be the total number of pathogenic substitutions among them, Sm be the number of different amino acid substitutions among all known pathogenic mutations, and Tn, Pn, and Sn be the corresponding numbers of single-nucleotide nonsense substitutions. We estimated Pm as Tn(Sm/Sn), which is unbiased as long as (*i*) all nonsense

Abbreviations: CPD, compensated pathogenic deviation; DM, Dobzhansky–Muller.

[§]Present address: Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Thorn Building, Room 1019, 20 Shattuck Street, Boston, MA 02115.

[¶]To whom correspondence should be addressed. E-mail: fkondras@ncbi.nlm.nih.gov.

www.pnas.org/cgi/doi/10.1073/pnas.232565499

mutations are pathogenic (Tn = Pn) and (*ii*) pathogenic missense and nonsense mutations are ascertained with the same probability. Tn and Tm were deduced from the coding sequence. For PMM2, this estimate of Pm exceeds Tm, and we accepted Pm = Tm. For five loci where nonsense mutations are not ascertained because they are either early lethal or not obviously pathogenic (G6PD, KCNQ1, MYH7, RHO, and TTR), we assumed Pm = 0.5Tm, as it is the average over the data set.

For each human protein, orthologs were found in the nr (nonredundant) database by BLAST (23), a multiple alignment was created by CLUSTALW (24), and the number of CPDs, N, was recorded for each nonhuman protein (see Table 2 and supporting text, which is published as supporting information on the PNAS web site). We used only CPDs at sites with unambiguous local homology to the human ortholog (no gaps closer than 10 sites to a CPD, and at least four exact matches at 10 sites before and 10 sites after it). This requirement led to rejection of ≈20% of unreliable CPDs, mostly in more distant proteins.

For each nonhuman protein, we calculated D, the number of amino acids that deviate from homologous human amino acids in such a way that a deviating amino acid can appear in the human protein after a single nucleotide substitution (in most proteins, >70% of all deviations met this criterion). The fraction of CPDs from all such deviations of a nonhuman protein was estimated as $F = N^*(\mathrm{Pm}/\mathrm{Sm})/D$, which is valid as long as ascertainment of human pathogenic mutations is independent of whether they coincide with CPDs in other species.

Identification of candidate compensatory second-site substitutions that ensure high fitness of CPD-carrying proteins was performed as follows. First, we identified all amino acid residues that directly interact with a CPD residue, requiring that the distance between their closest atoms does not exceed 4 Å. For each site occupied by such an amino acid, we used multiple alignment of orthologous proteins to check whether an amino acid that is pathogenic in humans can cooccur with the human amino acid at this site. If the cooccurrence was never observed, and if all nonhuman proteins that carry the CPD also carry a particular nonhuman amino acid at the second site, we hypothesized that this second-site substitution is compensatory for the CPD.

Results

We detected 608 CPDs in orthologs of 32 human proteins responsible for Mendelian diseases (Tables 1 and 2, Fig. 2). These CPDs cannot all be due to errors in the data, because proteins from several species often contain identical CPDs, and frequently a CPD corresponds to a pathogenic mutation known from several patients.

Table 1. Summary of data on pathogenic mutations and CPDs for individual human loci

Locus	Known missense	Known nonsense	All missense	All nonsense	Pathogenic missense	Orthologs	CPD
ABCD1	124	30	4,415	236	0.22	3	0
ALPL	83	5	3,147	156	0.82	6	1
AR	212	28	5,446	346	0.48	10	10
ATP7B	118	16	8,829	421	0.35	3	1
BTK	135	82	3,944	297	0.12	3	2
CASR	52	3	6,477	360	0.96	5	0
CBS	74	4	3,295	166	0.93	4	3
CFTR	446	133	8,733	650	0.25	15	70
CYBB	80	47	3,398	218	0.11	7	0
F7	82	6	2,778	166	0.82	3	1
F8	354	74	13,995	996	0.34	3	4
F9	419	55	2,759	211	0.58	9	3
G6PD	103	—	3,114	169	0.5*	15	17
GALT	99	11	2,247	139	0.56	6	7
GBA	94	11	3,060	173	0.48	3	0
GJB1	188	19	1,696	88	0.51	6	4
HBB	152	13	877	34	0.45	218	109
HPRT1	90	6	1,307	81	0.93	8	1
IL2RG	51	27	2,177	148	0.13	3	0
KCNH2	72	10	6,919	282	0.29	4	0
KCNQ1	69	—	4,022	186	0.5*	4	0
L1CAM	52	17	7,524	423	0.17	3	0
LDLR	273	67	5,242	297	0.23	8	15
MPZ	52	6	1,519	85	0.48	6	1
MYH7	62	—	11,620	755	0.5*	26	14
OCA1	71	13	3,155	226	0.39	106	31
PAH	272	23	2,699	193	0.85	6	8
PMM2	53	1	1,473	100	1.00	6	5
RHO	64	—	2,102	105	0.5*	116	10
TP53	80	7	2,345	129	0.63	29	5
TTR	74	—	870	41	0.5*	19	13
VWF	122	8	2,507	122	0.74	125	215

The columns are (left to right): name of a locus according to Online Mendelian Inheritance in Man; number of known different pathogenic missense mutations; number of known different pathogenic nonsense mutations (—, such mutations are not routinely ascertained); number of all possible missense mutations; number of all possible nonsense mutations; estimated fraction of pathogenic mutations among all possible missense mutations (*, absence of direct data); number of analyzed nonhuman orthologs; and total number of reliable CPDs detected in these orthologs.

```
Hs  1   mvhltpeeksavtalwgkvnvdevggealgrllvvypwtqrffesfgdlstpdavmgnpkvkahgkkvlgafsd 74
Sf      -....g......t.......e.............................s.....................
Hg      -tf.....ngh..s.....d.ek.............................s.i..............s...e
Ef      .tl.sa..nah..s.....d.ek.............................s.s..............s...e
Oc      -.....d..n..c........e.....................d......s.s..............s...e
Ba      -....a..........a....e.....................a......a....k............as...
Cg      -...sa...a...g.......e.....................................saa..............ts..e
Rn      .....da..a..n.......p.d..................Y.d......sas.i.............in..n.
Mm      -....a.......s.............................Y.d.........................ns..e
Ee      -....a...al..g.....k.e.F..................d......sa.............a...qsmg.
Bt      -...add..a..s...n..h.e.F..............s..........sa...fs.a..........ts.ge
Ru      -.d..a...a..l.......E.....................d........a..l..a..........hs.g.
Tr      -.e..g...a..l...d..dE.k...................d.......aa................hs.g.
Tg      -...sg...a...g.....dlek...qS..S..i.........d......s.s...s............ts...
Sc      ....sa...ghin.i.s..S..qt.a.......i.....s...dh.....sakg....a..qg..a...ts.g.

Hs  74  glahldnlkgtfatlselhcdklhvdpenfrllgnvlvcvlahhfgkeftppvqaayqkvvagvanalahkyh 147
Sf      .............q......N.................................q.....................
Hg      ..h..........q.............Q..t.e.....i...e...na.s......f....t...........
Ef      ..h..........q.............Q..t.......v...e...na.s.a....f.................
Oc      ..n..........k.............................v........d...q.............t.......
Ba      ..k...d.....................................i...r........el..................
Cg      ..s..........k.....................Mi.it....y.p..g.qt...f.................
Rn      ..k..........h.....................Mi.i..g..l......ca...f........s.......
Mm      ..kn.........k................k.....................q.....................
Ee      .ikn........sk.............................r....d...aa...f............a...
Bt      ..k...d....y.h.................k.......i...r........ql..s....tt..st.......
Ru      .vhn.......y.a.....................v...q...q....el....................
Tr      .vh...d..v...q.....................v...qq...a...el....................
Tg      ..n..........k.....................r...cn..p....q....f.................
Sc      avknm........k.....................i..ic..e....d...e....w..l.....t.......
```

Fig. 2. Alignment of the human β-hemoglobin and a few of the known mammalian orthologs. The organisms are: Hs, *Homo sapiens*; Sf, *Saguinus fuscicollis*; Hg, *Hapalemur griseus*; Ef, *Eulemur fulvus fulvus*; Oc, *Otolemur crassicaudatus*; Ba, *Balaenoptera acutorostrata*; Cg, *Ctenodactylus gundi*; Rn, *Rattus norvegicus*; Mm, *Meles meles*; Ee, *Erinaceus europaeus*; Bt, *Bradypus tridactylus*; Ru, *Rhinoceros unicornis*; Tr, *Tapirus terrestris*; Tg, *Tupaia glis*; and Sc, *Sminthopsis crassicaudata*. CPDs are in uppercase.

For each of the 32 human proteins, we determined, by comparing data on missense and nonsense mutations, the number of all possible pathogenic missense mutations (on average, 51% of all possible missense mutations, Table 1). We used these numbers to estimate F, the fraction of CPDs among all amino acid deviations of a nonhuman protein from its human ortholog, to be ≈10% (Fig. 3). Thus, although pathogenic substitutions are, indeed, overrepresented at slowly evolving sites (Fig. 4), a deviation of a nonhuman protein from its human ortholog would be pathogenic in humans with probability that is only ≈5 times lower than if proteins evolved regardless of selection within humans.

We looked for possible compensatory substitutions only for CPDs that are present in mammalian proteins, where the number of all deviations from their human orthologs is not too large. Thus, our analysis requires knowledge of the spatial structure of the mammalian proteins, at least around a CPD site, and of a substantial number of orthologous mammalian protein sequences. These data currently exist for only three of the 32 proteins, β-hemoglobin, von Willebrand factor, and transtherytin.

Known mammalian orthologs of these three proteins together carry 20 different CPDs. For one of these CPDs (27 B A → S in β-hemoglobin; Protein Data Bank ID 4hhb), the pathology of the corresponding human mutation is caused by its impact on RNA (25), and, therefore, its compensation also probably happens at the RNA level. For two CPDs, no amino acid site involved in direct interaction with them carried a candidate compensatory substitution. These CPDs might be compensated by substitutions involved in long-range interactions or by substitutions in other proteins/DNA. For each of 13 CPDs a particular, nonhuman amino acid at a single interacting site is present in all orthologs that carry the CPD (10 cases), or in all orthologs except one in which the structural neighborhood of the CPD site contains numerous deviations from the human protein (three cases). We hypothesize that a nonhuman amino acid that (almost) always accompanies such a CPD represents a compensatory substitution for it. Finally, for four CPDs there were more than one candidate compensatory substitutions.

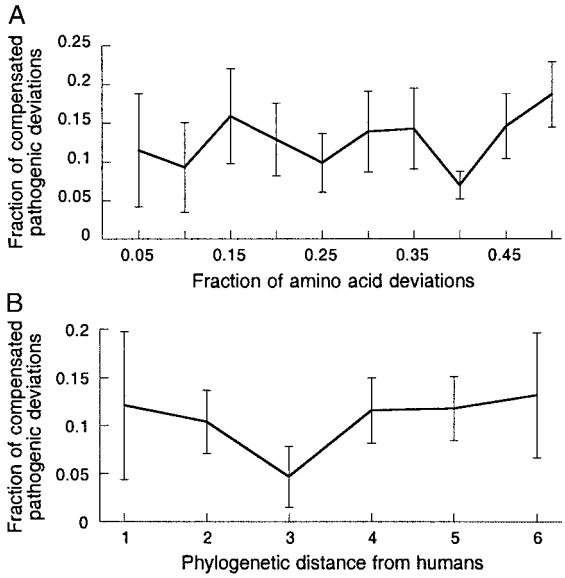

Fig. 3. The fraction of CPDs among all deviations of nonhuman proteins from their human orthologs, plotted against the fraction of amino acids that are different between a nonhuman protein and its human ortholog (*A*) or phylogenetic distance of the nonhuman species from humans (*B*) (1, nonhuman primates; 2, nonprimate mammals; 3, nonmammalian amniotes; 4, anamniote tetrapods; 5, fishes; 6, invertebrates). For each locus, the average F was calculated for each bin, and the average of these averages among all 32 loci is presented for each bin.

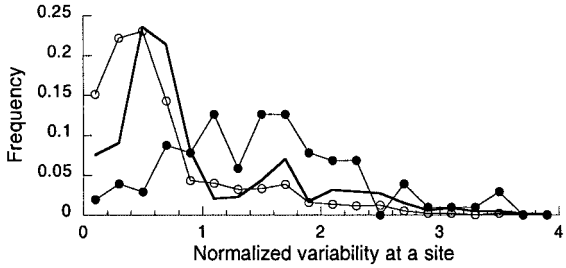

Fig. 4. Distributions of normalized variability (entropy of amino acid frequencies at a site normalized by the average entropy at all sites within the multiple alignment) at all sites (thick line), at sites where at least one pathogenic mutation is known (open circles), and at sites where at least one CPD is known (filled circles). Averages of these distributions are 1.00, 0.70, and 1.52, respectively.

Several examples of CPDs together with provisional compensatory substitutions are shown in Fig. 5*A*. In one case, where 3D structures of both the human protein and a CPD-carrying mammalian ortholog are known, we know the exact positions of side chains in both structures and can observe the interactions probably involved in compensation (Fig. 5*B*). Details on the analysis of all individual CPDs and their provisional compensatory substitutions are provided in Table 3, which is published as supporting information on the PNAS web site.

Discussion

Selection Acting on CPDs. Our analysis shows that a remarkably high fraction ($F \approx 10\%$) of all deviations of nonhuman proteins from their human orthologs is compensated pathogenic deviations. Thus, 1 in every 10 amino acid substitutions accepted by evolving proteins becomes benign only after a compensatory substitution is accepted. Geometrically, proteins evolve along fitness ridges (Fig. 1), which contain only ≈10 steps between successive corners.

Even more surprisingly, F is approximately invariant for a wide variety of nonhuman proteins and species (Fig. 3). Indeed, F must approach zero (or nearly so) when we consider organisms that are increasingly close to humans, because suppressor (compensatory) substitutions rarely segregate in human populations, and most mutations causing Mendelian diseases are pathogenic unconditionally (26). Almost certainly, a mutation that is pathogenic to us was also pathogenic to Neanderthals. Nevertheless, F is already high even in proteins from the closest living relatives of humans, as well as in proteins that individually are most similar to their human orthologs, and does not grow further when more distant species and proteins are considered (Fig. 3).

To understand this paradox, let us consider, starting from modern humans, events in a lineage with a CPD (Fig. 1). If, after a single compensatory substitution, the CPD can appear over a long period, F must grow linearly with evolutionary distance of the nonhuman protein from its human ortholog (11, 12), and even more rapid growth of F is expected if several compensatory

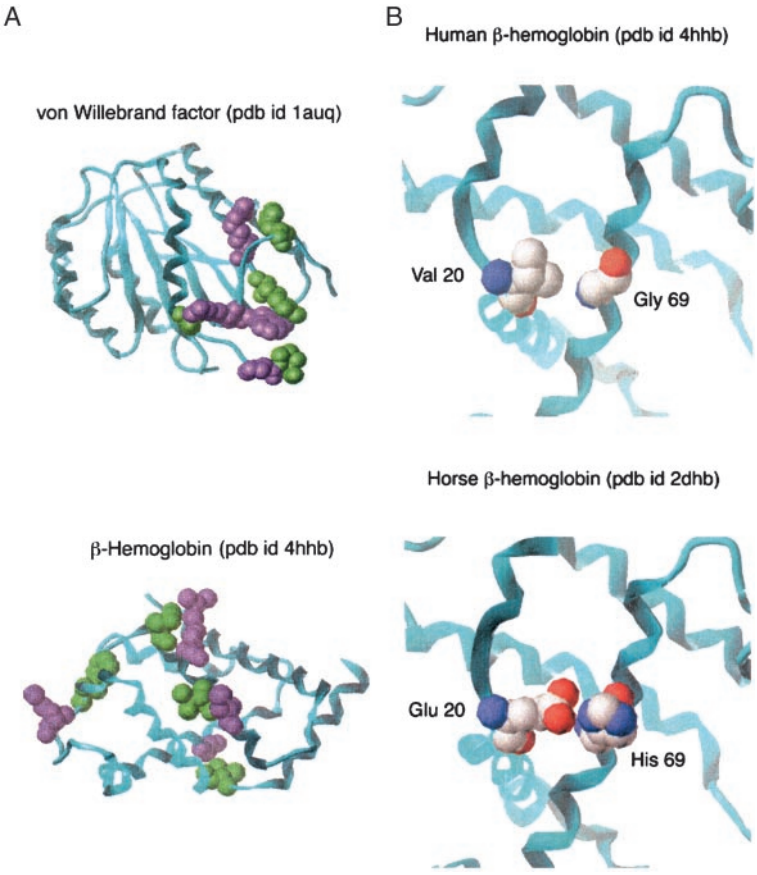

Fig. 5. Examples of CPDs and inferred compensatory substitutions. (*A*) Inferences based on the 3D structure of human protein for von Willebrand factor and β-hemoglobin. CPD sites are green and sites of possible compensatory substitutions are purple. (*B*) An inference based on the 3D structures of human and nonhuman protein. Val → Glu substitution at site 20 of human β-hemoglobin is pathogenic. Glu at this site is present in several β-hemoglobins, including that of horse, for which the 3D structure is known. Glu at site 20 is always accompanied by His at site 69, whereas human β-hemoglobin contains Gly at site 69. His-69 in horse β-hemoglobin interacts with Glu-20, possibly forming a hydrogen bond, whereas Val-20 and Gly-69 form a van der Waals interaction in human β-hemoglobin. Thus 69 Gly → His may compensate the deleterious effect of 20 Val → Glu.

substitutions are required to make the CPD fit (12). In contrast, if the CPD always appears very soon after a single compensatory substitution, F will be constant, as observed (Fig. 3). Simultaneous occurrence, caused by a complex mutational event (27), of a CPD and the corresponding compensatory substitution would, of course, also lead to constant F.

Such dynamics must be caused by epistatic selection (28). If both events shown in Fig. 1 happened in the nonhuman lineage (case **a**), the compensatory substitution must make the CPD not just acceptable but advantageous, causing its rapid fixation. Alternatively, if both events happened in the human lineage (case **c**), fixation of the human state (reversal of CPD) must trigger reversal of the compensatory substitution (aggravating substitution), which makes the nonhuman state unfit. Obviously, if the two events were almost simultaneous, they are unlikely to occur in different lineages (case **b**).

Such selection may appear if a compensatory substitution without the corresponding CPD is mildly deleterious. In this case, the highest fitness is achieved either by combination of the compensatory substitution and the CPD, or if both are absent. However, reduction of fitness due to the compensatory susbsitution alone, in contrast to that due to the CPD alone, is probably small, as otherwise the transition between the two most fit states would have been impossible (2), at least in the absence of complex mutational events.

Dynamics of this type have been detected in evolution of core rRNA sequences, where "compensatory substitutions can be fixed so rapidly as to appear to be instantaneous" (28). Known cases of epistatic selection in rRNAs and mRNAs are due to one obvious mechanism, Watson–Crick pairing of nucleotides that maintains secondary structure (28–30). In contrast, studies of second-site intragenic suppression of missense mutations revealed a variety of mechanisms of molecular compensation in proteins (15–19).

Thus, molecules that fold into the same structure (as a human protein, its orthologs in other eukaryotes, and all of the evolutionary intermediates certainly do), nevertheless, do not always have the same fitness, as it is often assumed in studies of protein (3) and RNA (5, 6) evolution. Evolving proteins do not just wander randomly on neutral manifolds or flat holey fitness landscapes (31) but experience selection caused by quantitative differences in fitnesses of fit sequences, all having a particular structure (28). Such differences may involve mild DM incompatibilities, with some combinations of states having reduced but nonzero fitness. For selection acting on RNAs, such DM incompatibilities can occur even between polymorphic sites within a population (30). However, here we considered mostly lethal DM incompatibilities, because almost all pathogenic mutations causing Mendelian diseases would be lethal to a human under natural conditions, at least when homozygous.

Preponderance of Intramolecular Compensation. A CPD in a protein that works in interaction with other molecules can be compensated by a change within the same protein or within a different molecule. In contrast, a CPD in a protein that can function in isolation must be compensated by a substitution within the same protein. Apparently, 11 of the 32 proteins considered here can still perform their functions after being isolated, and F in these proteins is not significantly different from F in the whole set. This observation suggests that many compensatory changes occur within the same protein, although both intraprotein (32, 33) and interprotein (34) DM incompatibilities have been found by interspecies comparisons. Within a species, compensatory mutations leading to genetic suppression also can be both intragenic and intergenic (15).

Hybrids between individuals from isolated populations become sterile or inviable after these populations turn, in the course of their independent evolution, different corners on fitness ridges and thus accumulate DM incompatibilities (11, 12, 31–33). In *Drosophila* (11, 38) and mammals (39, 40) many pairs of species that accumulated differences at $\approx 10^5$ or more amino acid sites can produce viable or even fertile hybrids. This implies that between closely related species $F < 10^{-5}$, in an apparent discrepancy with our estimate of $F \approx 10^{-1}$. This discrepancy strongly suggests that a vast majority of CPDs described here are caused by intraprotein interactions. Indeed, intraprotein DM incompatibilities would not cause immediate inviability of hybrids because they remain silent until incompatible amino acids are brought into the *cis* state by intragenic recombination.

As long as intraprotein DM incompatibilities are common, different sites of a protein cannot evolve independently, so that covarion or related models are necessary to describe protein evolution (41–43). Also, corners on fitness ridges that reflect intragenic interactions are compatible with Fisher–Muller advantage of sex (44). Instead of harming hybrids instantly, intraprotein *cis* DM incompatibilities will lead to slow hybrid breakdown in later generations. With $F \approx 10\%$, recombination between two alleles encoding proteins that are different at just 1% of sites will create an unfit allele with a substantial probability, as long as incompatible amino acid sites are not always located very close to each other; this may be detectable in populations derived from interspecies crosses (38).

Detecting Compensatory Substitutions. In studies of intragenic suppression of missense mutations, compensatory substitutions are usually detected directly, by assaying function of proteins carrying individual second-site substitutions together with a damaging mutation. In contrast, it may be difficult to identify the compensatory substitution that corresponds to a particular CPD among many deviations of a nonhuman protein from its human ortholog. Still, consideration of protein structure and of the multiple sequence alignment can often lead to plausible hypotheses.

We were looking only for compensatory substitutions of residues that interact directly with the CPD site. Such substitutions would replace one type of stabilizing residue–residue interaction with a different type of stabilizing interaction, whereas a pathogenic deviation without the compensation would destabilize a protein or cause conformational changes and destroy the function. Short-range compensatory substitutions are likely to be responsible for most of CPDs, because a majority of pathogenic mutations affect protein stability through changing hydrophobic packing, hydrogen bonding, salt bridges, and other types of short-range interactions (45).

Of course, long-range compensation caused by interactions between distant residues through other residues forming densely packed clusters in the folded protein (46) or caused by interactions in the course of protein folding (47) is also possible. Distant substitutions may simply have opposite effects on protein stability and thus compensate each other without interacting physically. Intermolecular compensation also happens (48), although a majority of CPDs observed here apparently are compensated intramolecularly. Still, short-range compensation is probably the most common, and identifying distant compensatory substitutions without experimental evidence is currently impossible.

About one-third of provisional compensatory substitutions involve charged amino acids. Replacement of an electrostatic interaction by a van der Waals contact or vice versa appears to be a common mechanism of compensation (e.g., Fig. 5*B*). We also observed a likely replacement of a possible stacking interaction between an aromatic residue and a positively charged residue for the same type of interaction, whereas the aromatic and the charged residue swapped (interaction between Arg-545 and Trp-550 shown in Fig. 5*A* is substituted by the interaction of His-545 and Arg-550). However, we did not observe a single case of swapping of positively and negatively charged residues. Evolutionarily, swapping of an

aromatic and a positively charged residue can go through a fit intermediate state of stacking interaction between two aromatic residues. In contrast, swapping positively and negatively charged residue requires going either through a single state with destabilizing interaction or through multiple steps.

Our analysis adds a previously undescribed dimension to earlier studies of correlated mutations in multiple sequence alignments (46, 49, 50). An alignment represents a sample of amino acids allowed by natural selection, perhaps conditionally on amino acids at other sites of the protein. Data on pathogenic mutations provide important information on amino acids that are forbidden. Together with structural data, this helps to reveal the molecular basis of compensatory substitutions.

Obviously, our predictions of the identities of compensatory substitutions are only hypotheses that must be tested by experimental data, similar to those used for double-mutant cycle analysis (51, 52).

1. Wright, S. (1932) *Proc. Sixth Int. Congr. Genet.* **1,** 356–366.
2. Maynard Smith, J. (1970) *Nature* **225,** 563–564.
3. Lipman, D. J. & Wilbur, W. J. (1991) *Proc. R. Soc. London Ser. B* **245,** 7–11.
4. Govindarajan, S. & Goldstein, R. A. (1997) *Biopolymers* **42,** 427–438.
5. Huynen, M. A., Stadler, P. F. & Fontana, W. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 397–401.
6. Huynen, M. A. (1996) *J. Mol. Evol.* **43,** 165–169.
7. Sunyaev, S., Ramensky, V. & Bork, P. (2000) *Trends Genet.* **16,** 198–200.
8. Polymeropoulos, M. H., Lavedan, C., Leroy, E., Ide, S. E., Dehejia, A., Dutra, A., Pike, B., Root, H., Rubenstein, J., Boyer, R., *et al.* (1997) *Science* **276,** 2045–2047.
9. Schaner, P., Richards, N., Wadhwa, A., Aksentijevich, I., Kastner, D., Tucker, P. & Gumucio, D. (2001) *Nat. Genet.* **27,** 318–321.
10. Afonnikov, D. A., Oshchepkov, D. Y. & Kolchanov, N. A. (2001) *Bioinformatics* **17,** 1035–1046.
11. Orr, H. A. & Turelli, M. (2001) *Evolution (Lawrence, Kans.)* **55,** 1085–1094.
12. Orr, H. A. (1995) *Genetics* **139,** 1805–1813.
13. Benzer, S. (1955) *Proc. Natl. Acad. Sci. USA* **41,** 344–354.
14. Feynman, R. P. (1985) *Surely You're Joking, Mr. Feynman!* (W. W. Norton, New York).
15. Davis, J. E., Voisine, C. & Craig, E. A. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 9269–9276.
16. Schulein, R., Zuhlke, K., Krause, G. & Rosenthal, W. (2001) *J. Biol. Chem.* **276,** 8384–8392.
17. Brasseur, G., Di Rago, J. P., Slonimski, P. P. & Lemesle-Meunier, D. (2001) *Biochim. Biophys. Acta* **1506,** 89–102.
18. Izumi, T., Malecki, J., Chaudhry, M. A., Weinfeld, M., Hill, J. H., Lee, J. C. & Mitra, S. (1999) *J. Mol. Biol.* **287,** 47–57.
19. Chen, R. D., Grobler, J. A., Hurley, J. H. & Dean, A. M. (1996) *Protein Sci.* **5,** 287–295.
20. Klein, G. & Georgopoulos, C. (2001) *Genetics* **158,** 507–517.
21. Heyworth, P. G., Curnutte, J. T., Rae, J., Noack, D., Roos, D., van Koppen, E. & Cross, A. R. (2001) *Blood Cells Mol. Dis.* **27,** 16–26.
22. Jinnah, H. A., De Gregorio, L., Harris, J. C., Nyhan, W. L. & O'Neill, J. P. (2000) *Mutat. Res.* **463,** 309–326.
23. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
24. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680.
25. Orkin, S. H., Antonarakis, S. E. & Loukopoulos, D. (1984) *Blood* **64,** 311–313.
26. Krawczak, M., Ball, E. V., Fenton, I., Stenson, P. D., Abeysinghe, S., Thomas, N. & Cooper, D. N. (2000) *Hum. Mutat.* **15,** 45–51.
27. Ninio, J. (1996) *Mol. Gen. Genet.* **251,** 503–508.
28. Chen, Y., Carlini, D. B., Parsch, J., Braverman, J. M., Tanda, S. & Stephan, W. (1999) *Genes Genet. Syst.* **74,** 271–286.
29. Tillier, E. R. M. & Collins, R. A. (1998) *Genetics* **148,** 1993–2002.
30. Parsch, J., Tanda, S. & Stephan, W. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 928–933.
31. Gavrilets, S. (1999) *Am. Nat.* **154,** 1–22.
32. Mateu, M. G. & Fersht, A. R. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 3595–3599.
33. Peixoto, A. A., Hennessy, J. M., Townson, I., Hasan, G., Rosbash, M., Costa, R. & Kyriacou, C. P. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 4475–4480.
34. Liu, J.-C., Makova, K. D., Adkins, R. M., Gibson, S. & Li, W.-H. (2001) *Mol. Biol. Evol.* **18,** 945–953.
35. Wu, C.-I. (2001) *J. Evol. Biol.* **14,** 851–865.
36. Ting, C. T., Takahashi, A. & Wu, C.-I. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 6709–6713.
37. Fishman, L. & Willis, J. H. (2001) *Evolution (Lawrence, Kans.)* **55,** 1932–1942.
38. Coyne, J. A. & Orr, H. A. (1998) *Philos. Trans. R. Soc. London B* **353,** 287–305.
39. Short, R. V. (1997) *J. Hered.* **88,** 355–357.
40. Skidmore, J. A., Billah, M., Binns, M., Short, R. V. & Allen, W. R. (1999) *Proc. R. Soc. London Ser. B* **266,** 649–656.
41. Fitch, W. M. & Markowitz, E. (1970) *Biochem. Genet.* **4,** 579–593.
42. Miyamoto, M. M. & Fitch, W. M. (1995) *Mol. Biol. Evol.* **12,** 503–513.
43. Gaucher, E. A., Miyamoto, M. M. & Benner, S. A. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 548–552.
44. Kondrashov, F. A. & Kondrashov, A. S. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 12089–12092.
45. Wang, Z. & Moult, J. (2001) *Hum. Mutat.* **17,** 263–270.
46. Lockless, S. W. & Ranganathan, R. (1999) *Science* **286,** 295–299.
47. Northey, J. G., Di Nardo, A. A. & Davidson, A. R. (2002) *Nat. Struct. Biol.* **9,** 126–130.
48. Pazos, F. & Valencia, A. (2002) *Proteins* **47,** 219–227.
49. Fariselli, P., Olmea, O., Valencia, A. & Casadio, R. (2001) *Proteins* **5,** Suppl., 157–162.
50. Shindyalov, I. N., Kolchanov, N. A. & Sander, C. (1994) *Protein Eng.* **7,** 349–358.
51. Horovitz, A., Bochkareva, E. S., Yifrach, O. & Girshovich, A. S. (1994) *J. Mol. Biol.* **238,** 133–138.
52. Germain, N., Merienne, K., Zinn-Justin, S., Boulain, J. C., Ducancel, F. & Menez, A. (2000) *J. Biol. Chem.* **275,** 21578–21586.