

(51) International Patent Classification:
G06F 19/18 (2011.01)(21) International Application Number:
PCT/US2013/032215(22) International Filing Date:
15 March 2013 (15.03.2013)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/667,850 3 July 2012 (03.07.2012) US(71) Applicant: BAYLOR COLLEGE OF MEDICINE
[US/US]; One Baylor Plaza, BCMD 600 D, Houston, TX
77030 (US).(72) Inventor: LICHTARGE, Olivier; 4536 Beech St., Bel-
laire, TX 77401 (US).(74) Agents: WAKIMURA, Mary Lou et al.; Hamilton,
Brook, Smith & Reynolds, P.C., 530 Virginia Rd, P.O.
Box 9133, Concord, MA 01742-9133 (US).(81) Designated States (unless otherwise indicated, for every
kind of national protection available): AE, AG, AL, AM,
AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY,
BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM,DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT,
HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP,
KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD,
ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI,
NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU,
RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ,
TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA,
ZM, ZW.(84) Designated States (unless otherwise indicated, for every
kind of regional protection available): ARIPO (BW, GH,
GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ,
UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ,
TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK,
EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV,
MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM,
TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW,
ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— of inventorship (Rule 4.17(iv))

Published:

— with international search report (Art. 21(3))

— before the expiration of the time limit for amending the
claims and to be republished in the event of receipt of
amendments (Rule 48.2(h))

(54) Title: BIOLOGICAL ACTION OF MISSENSE GENOTYPE PERTURBATIONS ON PHENOTYPE

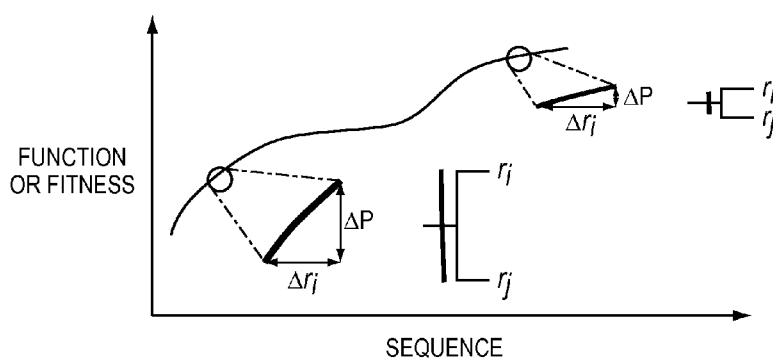


FIG. 1A

(57) Abstract: A method, computer pro-
gram product, and computer system for de-
termining or classifying a phenotypic ef-
fect of a mutation in a protein is disclosed.
An action of a protein mutation is calcula-
ted as a function of a magnitude of a
change at a protein residue and an evolu-
tionary gradient. The evolutionary gradient
measures the magnitude of an evolutionary
jump associated with the mutation at an
amino acid sequence position. The evolu-
tionary gradient also comprises the slope
of a tangent to a protein evolutionary tra-
jectory. The phenotypic effect of missense
mutations in viral, prokaryotic, and euka-
ryotic proteins is disclosed.

BIOLOGICAL ACTION OF MISSENSE GENOTYPE PERTURBATIONS ON PHENOTYPE

RELATED APPLICATION

This application claims the benefit of U.S. Provisional Application No.
5 61/667,850, filed on July 3, 2012.

The entire teachings of the above application are incorporated herein by reference.

GOVERNMENT SUPPORT

This invention was made with government support under R01 GM66099 and
10 R01 GM079656 from the National Institutes of Health and under CCF0905536 and DBI0851393 from the National Science Foundation. The government has certain rights in the invention.

BACKGROUND OF THE INVENTION

As genome sequencing becomes routine, a pressing clinical question is how
15 genotype variations affect phenotype. For protein missense mutations exact answers are elusive given the unique and complex structural and cellular context of each substitution. Here, we show a first principles equation to compute the action of infinitesimal variations of genotype on phenotype, such as the effect of point mutations on protein function and organismal fitness. In diverse proteins, this
20 computed action correlates with in vivo and in vitro data on the effect of mutations in diverse proteins, yielding better estimates than current, statistical, approaches. In the human population, an exponential distribution bias shows that human missense polymorphisms are eliminated at a rate proportional to their mutational action. And in diseases, morbidity and mortality are correlated with greater action of mutations
25 in specific genes, such as TP53 in glioblastoma or lysosomal alpha-glucosidase in the Pompe genetic disorder. These observations support an elementary, direct analytic relationship between genotype and phenotype with broad practical applications to protein design and the classification of human variations in health

and disease. Formally, the selection biases found in diseases and over the long term against mutations with greater action suggest that during evolution, protein follow trajectories that minimize action in genotype-phenotype space. This is consistent with the small number of adaptive substitutions induced by force laboratory protein evolution and with the variational principles that describe the evolution of all physical systems.

SUMMARY OF THE INVENTION

Embodiments of the present invention address the foregoing problems and shortcomings in the art. In particular, the present invention provides a method of, a computer program product, and a computer system for determining a phenotypic effect of a mutation in a protein.

In one embodiment, the invention relates to a method of determining a phenotypic effect of a mutation in a protein. The method comprises a processor calculating an action of a protein mutation in the protein as a function of (i) a magnitude of a change at a protein residue and (ii) an evolutionary gradient. Based on the calculated action, embodiments output to a user indication of a phenotypic effect of the protein mutation. In embodiments, the evolutionary gradient comprises or is a result of measuring a magnitude of an evolutionary jump associated with the mutation at an amino acid sequence position.

In another embodiment, the method of determining a phenotypic effect of a mutation in a protein calculates action of the protein mutation according to the formula:

$$\Delta P = k * \frac{\partial P}{\partial r_i} * \Delta r_i$$

wherein P is a protein, k is a scaling factor, Δr_i is a sequence displacement as the magnitude of a change at a protein residue i , and $\frac{\partial P}{\partial r_i}$ is the evolutionary gradient.

In another embodiment, the evolutionary gradient further comprises a slope of a tangent to a protein evolutionary trajectory and comprises the i^{th} component of the protein gradient and a partial derivative of the protein with respect to residue i , indicated by symbol ∂ .

In another embodiment, the scaling factor k is about 1.

In another embodiment, the sequence displacement is measured as a percent rank of all amino acid substitutions.

5 In another embodiment, the protein is a viral, eukaryotic, or prokaryotic protein.

In another embodiment, the mutation is a point mutation.

In another embodiment, the point mutation is a missense mutation.

10 In another embodiment, the method further comprises allowing the user to engineer a protein based on the calculated action and phenotypic effect of a mutation in the protein.

In one embodiment, the invention relates to a computer program product for determining a phenotypic effect of a mutation in a protein.

15 In another embodiment, the computer program product comprises a computer readable medium having program code embodied therewith. The program code is executable by a computer and configures the computer to: (a) measure a magnitude of an evolutionary jump associated with a protein mutation in a protein at an amino acid sequence position, and (b) calculate an action of the protein mutation as a function of a magnitude of (i) a change at a protein residue i and (ii) an evolutionary gradient, and (c) display to a user an indication of phenotypic effect of the protein mutation based on (corresponding to) the calculated action. In
20 embodiments, the evolutionary gradient utilizes the measured magnitude.

In another embodiment, the computer program product for determining a phenotypic effect of a mutation in a protein calculates action of the protein mutation according to the formula:

$$\Delta P = k * \frac{\partial P}{\partial r_i} * \Delta r_i$$

25 wherein P is a protein, k is a scaling factor, Δr_i is a sequence displacement as the magnitude of a change at a protein residue i , and $\frac{\partial P}{\partial r_i}$ is the evolutionary gradient.

In another embodiment of the computer program product, the evolutionary gradient further comprises a slope of a tangent to a protein evolutionary trajectory

and comprises the i^{th} component of the protein gradient and a partial derivative of the protein with respect to residue i , indicated by symbol ∂ .

In other embodiments of the computer program product, the scaling factor k is about 1 and the sequence displacement may be measured as a percent rank of all amino acid substitutions.

In another computer program product embodiment, the protein is a viral, eukaryotic, or prokaryotic protein. The mutation may be a point mutation. And, the point mutation may be a missense mutation.

In one embodiment, the invention relates to a computer system for classifying a phenotypic effect of a mutation in a protein.

In another embodiment, the computer system comprises a protein mutation execution engine (module/subassembly) that calculates an action of a protein mutation in the protein P as a function of (i) a magnitude of a change at a protein residue and (ii) an evolutionary gradient. Based on the calculated action, the computer system outputs to a user indications of phenotypic effect of the protein mutation. The evolutionary gradient comprises measuring a magnitude of an evolutionary jump associated with the mutation at a amino acid sequence position, and.

In another embodiment, the protein mutation execution engine (module) calculates the action of a protein mutation according to the formula:

$$\Delta P = k * \frac{\partial P}{\partial r_i} * \Delta r_i$$

wherein P is a protein, k is a scaling factor, Δr_i is a sequence displacement as the magnitude of a change at a protein residue i , and $\frac{\partial P}{\partial r_i}$ is the evolutionary gradient.

In another computer system embodiment, the evolutionary gradient further comprises a slope of a tangent to a protein evolutionary trajectory and comprises the i^{th} component of the protein gradient and a partial derivative of the protein with respect to residue i , indicated by symbol ∂ .

In other embodiments of the computer system, the scaling factor k is about 1. And the sequence displacement may be measured as a percent rank of all amino acid substitutions.

In another computer system embodiment, the protein is a viral, eukaryotic, or prokaryotic protein. The mutation may be a point mutation. In another embodiment, the point mutation is a missense mutation.

In another embodiment, the computer system further comprises allowing the user to engineer a protein based on the calculated action and phenotypic effect of a mutation in the protein.

The following describes methods, computer systems and computer program products to determine a phenotypic effect of a mutation in a protein.

BRIEF DESCRIPTION OF THE DRAWINGS

The patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawing(s) will be provided by the Office upon request and payment of the necessary fee.

The foregoing will be apparent from the following more particular description of example embodiments of the invention, as illustrated in the accompanying drawings in which like reference characters refer to the same parts throughout the different views. The drawings are not necessarily to scale, emphasis instead being placed upon illustrating embodiments of the present invention.

Fig. 1 shows a graph of the function of fitness versus sequence (panel a), a pictorial representation of Evolutionary Tracing (panel b), an illustration of all possible alanine substitutions (panel c), and the substitution odds from alanine to valine, threonine, and arginine (panel d).

Fig. 2 shows the substitution odds on the solvent accessibility and the secondary protein structure element.

Fig. 3 are graphs showing the action of a protein mutation is proportional to the experimental impact of mutations in lac repressor, T4 lysozyme, HIV-1 protease, and p53.

Fig. 4 are graphs of receiver operating characteristic (ROC) curves of the sensitivity and specificity of various methods.

Fig. 5 are graphs of the ROC curve for T4 lysozyme (panel a), loss of recombination activity versus evolutionary action for RecA (panel b), and the

distribution of disease-associated mutations and polymorphisms in human genes (panel c).

Fig. 6 are graphs showing the relationship between the fraction of deleterious mutations and evolutionary action.

5 Fig. 7 are graphs showing the relationship between fraction of polymorphisms and evolutionary action (panel a), the relationship of polymorphisms and minor allele frequency (MAF) (panel b), somatic p53 mutations in tumors and evolutionary action score (panel c), sporadic p53 mutations and evolutionary action score (panel d), survival rate of glioblastoma patients (panel e), and the evolutionary
10 action of mutations in the GAA gene (panel f).

Fig. 8 is a block diagram of computer nodes or devices in the computer network of Fig. 9.

Fig. 9 is a schematic view of a computer network environment in which embodiments of the present invention may be deployed.

15 Fig. 10 is a flow diagram of one computer-based embodiment of the present invention.

DETAILED DESCRIPTION OF THE INVENTION

A description of example embodiments of the invention follows.

20 Panel a in Fig. 1 shows the evolutionary trajectory of a protein and how changes in sequence over time, along the x-axis, translate into changes in structure, function and fitness, along the y-axis. Each elementary step of this trajectory is a mutation, denoted Δr_i , associated with a small change in phenotype, denoted ΔP_i . Compared to long, evolutionary time-scales, both Δr_i and ΔP_i are infinitesimal so
25 that the trajectory appears continuous and smooth. Their ratio (Δr_i to ΔP_i), however, is fixed at any given time and defines the slope of the trajectory curve (at that time), which is the first derivative or gradient of the protein with respect to its sequence variables. This slope will be small, or large, if a mutation is associated with relatively smaller, or larger evolutionary steps.

In panel b in Fig. 1, this slope criterion is precisely the basis for ranking the relative importance of sequence positions with Evolutionary Tracing (ET), suggesting that ET approximates evolutionary gradients.

Panel c in Fig. 1 is an illustration of the evolutionary gradient as an important determinant of substitution patterns. This is illustrated for all possible substitutions of alanine to any of the other amino acids. The substitutions were tallied over 70,000 sequence alignments and their log-odds were binned by evolutionary gradient decile. The red to green spectrum represents a shift from least likely (log odds are -6) to most likely (log odds are 2) substitutions. Each column is gradually different from its neighbor leading to significant difference of the whole spectrum of evolutionary importance and compared to the BLOSUM62 log odds, as a reference. The amino acids labels follow standard definitions.

Panel d in Fig. 1 shows the different patterns of rise and fall of substitution odds from alanine (A) to various valine (V), threonine (T) and arginine (D) illustrate the complex and amino acid specific dependence of these odds as a function of the evolutionary gradient. The odds of BLOSUM62 are shown as straight lines, and are close to the overall average.

Figure 2 shows the dependence of the substitution odds on the solvent accessibility and the secondary structure element, when alanines have small (0~10), in the upper panel, and large (90~100), in the lower panel, evolutionary gradient. The odds of helical solvent exposed alanines are compared when the alanines become either solvent inaccessible or on strands, keeping the other features same.

In Fig. 3, the action of a protein mutation is proportional to the experimental impact of mutations. Panel a in Fig. 3 shows the experimental impact of missense mutations in four proteins (lac repressor, T4 lysozyme, HIV-1 protease, and p53) is a near linear function of action computed with equation:

$$\Delta P = k * \frac{\partial P}{\partial r_i} * \Delta r_i$$

wherein P is a protein, k is a scaling factor, Δr_i is a sequence displacement as the magnitude of a change at a protein residue i , and $\frac{\partial P}{\partial r_i}$ is the evolutionary gradient.

ET rank and ET-dependent substitution log odds were used to approximate the evolutionary gradient and the sequence displacement, respectively. In *E. coli* lac repressor, T4 lysozyme and HIV-1, fitness was measured as the fraction of mutations that were judged to be deleterious. In human p53 protein the functional impact was measured in yeast as the average loss of p53 transactivation activity over eight response-elements. The scaling factor k was $k_{lac} = 1.00$, $k_{T4} = 0.63$, $k_{HIV} = 1.10$ and $k_{p53} = 0.76$, respectively.

In Panel b, there is linear correlation between experimental loss of fitness and the evolutionary gradient ($\partial P / \partial r_i$).

In Panel c, the magnitude of change (Δr_i), when the other term is held fixed, is shown using the pooled data over all T4 lysozyme and lac repressor mutations, binned into low (0-40%), mild (40-70%) or high (70-100%) scores for the fixed variable.

In Figure 4, the Receiver Operating Characteristic (ROC) curves show the relative sensitivity and specificity of various methods to separate harmful from harmless mutations. The action always achieves the greatest Area Under the Curve (AUC) compared to Polyphen2, SIFT, MAPP and a control of the action equation that uses entropy and BLOSUM62 substitution log odds in each of the proteins. Panel a represents 2015 T4 lysozyme mutations were assayed by the amount of formed plaque, due to lysozyme's break up of the host cell walls. Panel b represents 4,041 *E. coli* lac repressor mutants to repress beta-galactosidase more than 20 fold. Panel c represents 336 HIV-1 protease mutants to cleave the Gag and Gag-Pol precursor proteins. Polyphen2 returned no predictions for the HIV-1 protease mutations. Panel d represents 2,314 human p53 mutants to transactivate 8 p53 response-elements in yeast. Finally, Panel e represents the ROC curves of predictions by the evolutionary action and by Polyphen2, SIFT and MAPP methods for the association of mutations with disease or not, in a dataset that involves 218 genes (see methods).

In Fig. 5, panel a, the Receiver Operating Characteristic (ROC) curves show the relative sensitivity and specificity of various methods to separate harmful from harmless mutations. The action achieves the greatest Area Under the Curve (AUC)

compared to Polyphen2, SIFT and MAPP methods in 2,015 bacteriophage T4 lysozyme mutants to break the host cell walls. As a control, the action based on less accurate measures for the gradient and the sequence displacement performs poorly. In panel b, action correlates with a gradual loss of activity caused by targeted mutations in RecA. The loss of recombination activity of RecA as a function of the action is shown for individual mutants, in grey, and binned, in black. The latter has a correlation factor R^2 of 0.89. Occasional mutations, in red, have smaller loss of recombination than expected, but they do impact other RecA functions, such as mediating self-cleavage in LexA. In panel C, there is marked difference between the distribution of 8,553 disease-associated mutations in 218 human genes obtained from the Uniprot database, in black, and the distribution of 794 polymorphisms in the same proteins, in grey, as a function of the action. This suggests that action should help distinguish harmful from harmless mutations among all these proteins.

In Figure 6, the linear dependencies of the functional impact on the action (Panel a), the gradient (Panel b), and the magnitude of change (Panel c) are similar to results shown in Fig. 3 for the entropy-BLOSUM62 control of the action equation. The dependencies are also nearly linear, but have lower correlation coefficients and smaller slopes.

In Fig. 7, the number of human protein coding polymorphisms decays exponentially with the action of their missense mutations in Panel a. This is shown for 31,803 human missense polymorphisms taken from the dbSNP database. This suggests that the purification of polymorphism is a random Poisson process parameterized by the action. Panel b shows that the exponential decay of polymorphisms becomes steeper for high minor allele frequency (MAF). This suggests that older polymorphisms, with larger MAF, have been subjected to higher selection. The distribution of 343 highly recurrent (>10 times) somatic p53 mutations found in 26,597 tumor samples of the IARC database, rises exponentially with evolutionary action shown in Panel c. However, in Panel d, 1,026 sporadic p53 mutations in the same tumor samples have no bias with respect to action. The mutations in that group that have been shown to cause loss of p53 activity in yeast assays recover the distribution bias of Panel c. Panel e shows the survival rate of glioblastoma patients with somatic TP53 mutations stratifies as a function of their

action. Patients from TCGA were selected for having TP53 mutations with deleterious evolutionary action (more than 70%) and divided into three groups. Black, grey and light grey colors denote the highest, high and milder deleterious evolutionary action, respectively. Panel f shows the evolutionary action distribution of 135 mutations in the GAA gene that were rated in five classes of Pompe disease severity. Class B is the most severe for missense mutations, and the severity gradually decreases for classes C, D, E. Class F contains mutations rated as non-pathogenic. The horizontal bars represent the median evolutionary action and the numbers the quantity of mutations for each class type.

Fig. 9 illustrates a computer network or similar digital processing environment in which the present invention may be implemented.

Client computer(s)/devices 50 and server computer(s) 60 provide processing, storage, and input/output devices executing application programs and the like. Client computer(s)/devices 50 can also be linked through communications network 70 to other computing devices, including other client devices/processes 50 and server computer(s) 60. Communications network 70 can be part of a remote access network, a global network (e.g., the Internet), a worldwide collection of computers, Local area or Wide area networks, and gateways that currently use respective protocols (TCP/IP, Bluetooth, etc.) to communicate with one another. Other electronic device/computer network architectures are suitable.

Fig. 8 is a diagram of the internal structure of a computer (e.g., client processor/device 50 or server computers 60) in the computer system 100 of Fig. 9 embody the present invention. Each computer 50, 60 contains system bus 79, where a bus is a set of hardware lines used for data transfer among the components of a computer or processing system. Bus 79 is essentially a shared conduit that connects different elements of a computer system (e.g., processor, disk storage, memory, input/output ports, network ports, etc.) that enables the transfer of information between the elements. Attached to system bus 79 is I/O device interface 82 for connecting various input and output devices (e.g., keyboard, mouse, displays, printers, speakers, etc.) to the computer 50, 60. Network interface 86 allows the computer to connect to various other devices attached to a network (e.g., network 70 of Fig. 9). Memory 90 provides volatile storage for computer software instructions

92 and data 94 used to implement an embodiment of the present invention (e.g., Evolutionary Action calculation, evolutionary gradient determination and phenotypic effect indicator detailed above and below). Disk storage 95 provides nonvolatile storage for computer software instructions 92 and data 94 used to
5 implement an embodiment of the present invention. Central processor unit 84 is also attached to system bus 79 and provides for the execution of computer instructions.

In one embodiment, the processor routines 92 and data 94 are a computer program product (generally referenced 92), including a computer readable medium (e.g., a removable storage medium such as one or more DVD-ROM's, CD-ROM's,
10 diskettes, tapes, etc.) that provides at least a portion of the software instructions for the invention system. Computer program product 92 can be installed by any suitable software installation procedure, as is well known in the art. In another embodiment, at least a portion of the software instructions may also be downloaded over a cable, communication and/or wireless connection. In other embodiments, the invention
15 programs are a computer program propagated signal product 107 embodied on a propagated signal on a propagation medium (e.g., a radio wave, an infrared wave, a laser wave, a sound wave, or an electrical wave propagated over a global network such as the Internet, or other network(s)). Such carrier medium or signals provide at least a portion of the software instructions for the present invention
20 routines/program 92.

In alternate embodiments, the propagated signal is an analog carrier wave or digital signal carried on the propagated medium. For example, the propagated signal may be a digitized signal propagated over a global network (e.g., the Internet), a telecommunications network, or other network. In one embodiment, the propagated
25 signal is a signal that is transmitted over the propagation medium over a period of time, such as the instructions for a software application sent in packets over a network over a period of milliseconds, seconds, minutes, or longer. In another embodiment, the computer readable medium of computer program product 92 is a propagation medium that the computer system 50 may receive and read, such as by
30 receiving the propagation medium and identifying a propagated signal embodied in the propagation medium, as described above for computer program propagated signal product.

Generally speaking, the term “carrier medium” or transient carrier encompasses the foregoing transient signals, propagated signals, propagated medium, other mediums and the like.

In particular, embodiments provide a computer based system or apparatus
5 100 programmed or otherwise configured to carry out the follow steps outlined in Fig. 10.

A protein mutation module (at 192) receives as input an evolutionary trajectory of a subject protein P. The evolutionary trajectory is formed by a time series of sequence data of a sample of protein P.

10 From the input ET, the protein mutation module 192 calculates, for the subject protein P, action ΔP of protein mutation in the protein. To accomplish this, the protein mutation module 192 measures a magnitude of an evolutionary jump associated with a mutation at an amino acid sequence position. Based on this measurement, the protein mutation module 192 determines an evolutionary gradient
15 for the subject protein. Next, the protein mutation module 192 calculates magnitude of a change at a protein residue (Δr_i).

Then as a function of this calculated magnitude of change (Δr_i) and determined evolutionary gradient, module 192 calculates the action ΔP of the protein mutation. In particular, module 192 computes Equation 1 further detailed below.
20 This calculated action ΔP of the protein mutation represents the phenotypic response to the genotype perturbation.

An output 195, system 100 produces indications of the phenotypic effect of the protein mutation.

25 **Example 1.**

Each human birth introduces about 70 new genetic variations that may affect health and disease. At least 40% of these are lost, while the rest persist and lead, on average, to four million DNA differences between random individuals. These differences include insertions, deletions, and copy number variations. Up to 80%
30 single nucleotide substitutions translate to ten thousand single amino acid variations per individual. A classification of those that impact our susceptibility to diseases

could guide surveillance, prognosis, and therapeutic choices. With the advent of personal exome sequencing, association studies have linked over 2,500 diseases to genetic variations. These variations, however, explain only a small fraction of each disorder leaving many other contributing genetic variations cryptic. Moreover, 15% percent of rare or personal mutations cannot be associated to disease by these methods. Computational alternatives to evaluate harmful mutations in proteins are therefore of wide interest. They are based on statistical and biophysical models of evolutionary conservation and of structure and function; or on their combination through machine learning. The degree of agreement is variable and suggests overall limitations in these models of the sequence, structure and function relationship.

Here we sought instead to evaluate how missense mutations affect proteins from first principles, and assuming that evolution is a smooth process. A protein, P , is viewed as a function of n residues, $P = P(r_1, r_2, \dots, r_n)$. It describes a trajectory in sequence-function space or, more broadly, in phenotype-genotype space during evolution, as depicted in Fig. 1, panel a. If P is differentiable, then a mutation at residue i will induce a first order response:

Equation 1:
$$\Delta P = k * \frac{\partial P}{\partial r_i} * \Delta r_i$$

where k is a scaling factor; Δr_i is the magnitude of the change at residue i ; and $\frac{\partial P}{\partial r_i}$ is the i^{th} component of the protein gradient, namely, the partial derivative of the protein with respect to residue i , indicated by the symbol ∂ . This gradient term is the sensitivity of the protein to a change of residue i . Equation 1 translates a genotype perturbation into a phenotype response, via the evolutionary gradient, and it yields the action of a protein mutation. We therefore call it the action equation.

For the action equation to be meaningful, we must also show how to compute Δr_i and $\frac{\partial P}{\partial r_i}$. The evolutionary gradient, $\frac{\partial P}{\partial r_i}$, is the slope of the tangent to the protein evolutionary trajectory (Fig. 1, panel a), which is large (or small) if the trajectory changes by a large (or small) amount when residue i mutates. Hence the gradient measures the magnitude of the evolutionary jumps associated with

mutations at sequence position i . This association can be measured directly by comparing all the mutations at position i seen in a sequence alignment, with the evolutionary jumps for each one seen in an evolutionary divergence tree, depicted in Fig. 1, panel b. A residue that only varies between large and distant evolutionary branches will have a large gradient, and one that varies between small and recently diverged branches will have a small gradient. This matches precisely the definition of evolutionary importance ranks of sequence positions by Evolutionary Trace (ET). Thus, evolutionary gradients and ET ranks are a measure of the gradients and these terms will be used interchangeably thereafter.

In order to evaluate the magnitude of a residue mutation, Δr_i , we reason that larger mutations will impact function more and be selected against accordingly. Therefore substitution odds should be inversely related to the size of a mutation. A tally of these odds over 67,000 protein families of known structures reveals, however, that they also depend on ET ranks, as shown in Fig. 1, panel c, for alanine. The average of substitution odds over all Evolutionary Trace is close to typical (BLOSUM62) values, but deviations are the norm for any given ET rank. These are especially large for extreme ET ranks and their exact dependencies are complex and substitution specific. In Fig. 1, panel d, alanine to valine mutations are a bell-shaped function of the evolutionary importance, alanine to threonine mutations are initially flat then tail off, and alanine to arginine mutations fall exponentially. These data show that amino acid substitutions odds are different for evolutionary important residues than for ones less so, and this effect is distinct from the effect of structural context, represented in Fig. 2. Therefore we chose to approximate mutational size through ET-dependent substitution matrices.

With its factors defined, we can test the action equation against the impact of mutations in diverse proteins. Our comparisons included the effect of 4,041 lac repressor mutations in *E. coli* on β -galactosidase hydrolysis, 2,015 lysozyme mutations in bacteriophage T4 on plaque formation 34, 336 HIV-1 protease mutations on cleavage products 35, and 2,314 p53 mutants in yeast on the loss of transactivation over eight response-elements 36. In every case, the action computed with equation 1 correlated linearly with experimental results. R^2 was 0.94 - 0.96 for the lac repressor, HIV-1 protease, and p53 data, and 0.73 for the T4 lysozyme data,

shown in Fig. 3, panel a. Also in accord with the action equation, the experimental impact of mutations was linear with when was binned into at low, medium and high values, in Fig. 3, panel b, and vice-versa in panel c. These data from viral, eukaryotic and prokaryotic proteins assayed in vivo or in vitro, show that the experimental impact of mutations is proportional to the product of the evolutionary gradient with the size of the mutation.

As a second test we also compared the action equation with popular, state-of-the-art methods, such as the SIFT, MAPP and Polyphen2, to measure the deleterious impact of residue variations. Performance was the aggregate sensitivity and specificity of ranking mutations from most to least deleterious, measured by the area under a Receiver-Operator Characteristic curve (AUC). The ET-based action equation always yielded the greatest AUC, between 0.86 and 0.89, improving by 5% on Polyphen2, 5% on MAPP, and 15% on SIFT, shown in Fig. 4, panel a. These data show that the action equation ranks the relative impact of protein mutations more accurately than statistical approaches when the evolutionary gradient and mutation size are based on ET.

In a third test, we asked whether the action equation also anticipated the effect of engineered mutations in a complex system. The prokaryotic RecA protein is a central component of bacterial DNA repair, normally mediating error-free DNA repair but also triggering error-prone repair and drug resistance, when it binds LexA in response to genotoxic stress. In order to identify its functional sites, a total of 31 mutations were directed to surface amino acids ranked as important by ET, with some control residues that were ranked as unimportant. An assay, based on the integration of a selectable GFP-Chloramphenicol resistance fusion gene into the chromosome, using P1 phage mediated transduction, quantified the decrease in recombination activity. As seen before in large-scale tests, the action predicted by equation 1 correlated linearly with the average loss of function with an R^2 of 0.89, in Fig. 5, panel b. RecA is a multifunctional protein, however, so these mutations could perturb functions other than recombination. For example, in panel b, the points in red also inhibit the interaction with LexA, which may explain why they appear to be outliers for the correlation of recombination function with the action.

These data show the potential of the action equation to guide substitutions that tune the functional impact of mutants.

Next, we asked whether the action equation could discriminate harmful genetic polymorphisms from benign ones. We collected from the Uniprot database 218 genes with annotated with 8,553 disease-associated mutations (8,553) and also with 794 other mutations that had no known disease association. In Fig. 5, panel c, the distribution of action was significantly different between these two groups (p -value $< 10^{-16}$). Moreover, even in this mixed group of proteins, the action equation ranked harmful mutations with greater sensitivity and specificity than other methods. The AUC value was 0.85 over all cases, and rose to 91% when only polymorphisms with action above the 80th percentile or below the 20th percentile were considered in Fig. 4, panel e. These data show that the distribution of harmful genetic mutations is skewed to large action values, and conversely benign mutations are skewed to small action.

This distribution bias of neutral mutations was also shown on a large scale, with 31,803 distinct single residue human polymorphisms from 10,144 genes in the dbSNP database. The number of these polymorphisms decays exponentially as their action rises as shown in Fig. 7, panel a. This trend suggests that regardless of important individual features, such as zygosity and dominance, some polymorphisms are lost following a typical Poisson process with a rate parameter proportional to the action. In Fig. 7, panel b, polymorphisms are binned by minor allele frequency (MAF) the exponential decay persists in every bin, and, strikingly, the rate increases in proportion to the allele frequency. This indicates that as the surviving polymorphisms propagate and reach greater MAF, the selection against those with greater action grows ever higher relative to those with lesser action. Hence, polymorphisms with greater action have a selection disadvantage relative to those with lesser action, and this difference grows greater even as the polymorphisms spread throughout the population. Conversely, polymorphisms with the least action have the greatest chance of fixation over long time scales, a fact reflected by the J curve in Fig. 7, panel b for low action polymorphisms as a function of MAF.

Next, we asked whether strong selection biases could be witnessed on short timescales. We focused on a subset of p53 mutations seen in a collection of 26,597 human tumors. In Fig. 7, panel c, 85% of the (343) mutations that were thought to play causative roles in cancer, because they were seen frequently, had action values in the top half ($p\text{-value} = 9 \times 10^{-34}$), and were non-functional in the yeast transactivation assays ($p\text{-value} = 4 \times 10^{-58}$). In contrast, the (1,026) sporadic p53 mutations with a more uncertain role in pathogenesis, had a random distribution with respect to action. However, the third of these that were non-functional in yeast assays, shown in Fig. 7, panel d, were also biased towards large evolutionary action (79% of them, $p\text{-value} = 2 \times 10^{-47}$). These data show that the p53 mutations that are most likely to play a role in cancer are strongly biased towards large mutational action. This is not true of mutations not clearly associated with cancer, except for a subgroup that have *in vitro* functional defects, and which may therefore play an unrecognized role in the disease.

These data suggest that the action of mutations in genes associated with disease may stratify morbidity and mortality. Glioblastoma survival is affected by various mutations and impairment of the tumor suppressor function of TP53 generally worsens the prognosis. Since neutral mutations or wild type p53 instances are often accompanied by amplifications, deletion and other driver mutations in many other genes that affect p53 expression, and glioblastoma progression, we focused only on p53 missense mutations with evolutionary action greater than 70-80%, 80-90% and 90-100% assuming that these three subsets were most directly linked to disease outcomes. Although these groups are small, with only 10, 13 and 5 patients, respectively, their mortality rates are significantly different (overall log-rank $p\text{-value}$ is 0.026) seen in Fig. 7, panel e, such that two years after diagnosis the fraction of patients still alive was 60%, 30% and 0%, respectively. These data show that the mutational action in a known disease-associated gene, such as p53, may be a useful marker of disease in cancer.

In order to evaluate the same hypothesis in a genetic disease, we turned to lysosomal glycogen storage type II, or Pompe disease. This is a clinically heterogeneous inherited condition that occurs once in every 40,000 births due to a deficiency of the acid alpha-glucosidase enzyme encoded by the GAA gene. The

135 missense mutations of GAA have been rated by decreasing clinical severity into types B, C, D and E, and ending with type F that is thought to be entirely non-pathogenic. In Fig. 7, panel f, the evolutionary action of the pathogenic types B-E was in the top half, and it was in the bottom half for the non-pathogenic type F.

- 5 Moreover, the median action rose with the severity type, which is 80% for B and C, 60% for D and E and 20% for F types, and these differences were significant (p-value = 5×10^{-6}). These data show that evolutionary action is clinically relevant to the classification of relative health risk in a Mendelian disease, and together with the p53 data they illustrate that mutations with larger action undergo more rapid
10 elimination.

The current evaluation of the action equation has a number of limitations. The perturbation and the sensitivity to it, Δr_i and $\frac{\partial P}{\partial r_i}$, should be independent of each other but are not. As computed, Δr_i is approximated with ET-dependent substitution odds. Moreover, the ET algorithm to approximate $\frac{\partial P}{\partial r_i}$ is not necessarily optimal, and
15 the substitution odds may be refined for specific cases (e.g eukaryotes or prokaryotes, membrane or globular proteins, structured or unstructured). At a deeper level, these terms are given as normalized ranks instead of absolute values. This rescaling introduces errors when comparing mutational impact from different proteins and it explains the need for a scaling factor, k , which normally should be 1
20 but that deviates from it protein by protein as seen in the unequal slopes of Fig. 3. Most broadly, the threshold at which a mutation causes harm will differ among proteins, in part because impairment in proteins does not translate uniformly into harm to the host. Surprisingly, despite these caveat, the action classifies harmful mutations better than current techniques in all of our benchmarks, even when
25 pooling them together shown in Fig. 4, panel e.

It is possible therefore that other approximations for Δr_i and $\frac{\partial P}{\partial r_i}$, would solve these issues. For example, conservation entropy could approximate the gradient and BLOSUM62 substitution log-odds could approximate the size of the mutation. This entropy-BLOSUM62 model applied to the action equation also yields
30 impressive linear correlations from 0.80 to 0.97. These correlations are flatter than

for the original ET-based model, however, and they discriminate deleterious mutations with a 10% loss of performance shown in Fig. 6. This alternative model shows that the laws of proportionality in the action equation are robust but that accurate evaluation of the action requires better measurement of the evolutionary
 5 gradients and perturbation, such as provided by ET, and ET-dependent substitution matrices.

This work describes the impact of a mutation on a protein through an analytic action equation that links genotype variations to phenotype responses. The agreement between this action and experiments supports the fundamental hypothesis
 10 that evolution is a differentiable process and the ET-based approximations for the evolutionary gradient and the magnitude of sequence mutations. In practice, this opens molecular evolution, protein redesign, and the interpretation of human allelic variations to an analytic view of the trajectories that proteins follow as they evolve in genotype-phenotype space.

15 Interestingly, the data show that in disease, as well as in health, there is an exponential selection bias against mutations with greater action. Conversely, the retention of mutations that minimize evolutionary action is systematically favored. This suggests a least action conjecture: proteins evolve along a genotype-phenotype evolutionary trajectory that minimizes the sum of the action

$$\sum_i^{neighbors} \frac{\partial P}{\partial r_i} * \Delta r_i$$

20 over closest evolutionary neighbors. This conjecture would be consistent with evolutionary experiment in single proteins that find narrow adaptive paths limited to a few residue mutations in a defined order; it is also closely related to the principle of parsimony, which is the basis for sequence alignments and already states that nearest evolutionary neighbors have the fewest and most conservative sequence
 25 substitutions, and it would have the advantage of placing protein evolution in a variational framework consistent with description of evolution in physical systems.

Methods

Calculation of the Evolutionary Action

The evolutionary action is calculated by equation 1 as the product of the evolutionary gradient, $\partial P / \partial r_i$ and the sequence displacement Δr_i . The gradient is measured by the Evolutionary Trace method and the displacement by amino acid substitution odds, as detailed below. Using percent ranks instead of absolute values for both of these measures results to action in dimensionless scale. Because this is hard to interpret, results were normalized become percent ranks out of all possible amino acid substitutions of each protein. Small or large coverage values indicate neutral or deleterious predictions, respectively, such that an action of 68 corresponds to mutations with higher impact than 68% of all possible substitutions in a protein.

Calculate the Evolutionary Trace ranks

The Evolutionary Gradient is measured as the percent rank of each protein residue as calculated by the Evolutionary Trace (ET) method. ET analyses multiple sequence alignments, such that it generates the phylogenetic tree of the sequences and it correlates variations with phylogenetic branching. Residues that vary within smaller branches are ranked as less important, and vice versa. These ranks are normalized for the sequence length to become percent ranks out of all protein residues. Here, we generated three alignments by blasting the NCBI nr, the Uniprot100 and the Uniprot90 database, while the sequences with gaps at the ranked position were excluded. Then, the relative importance of each residue was averaged and normalized to become percent ranks.

Calculate the substitution log-odds

The sequence displacement is measured as the percent rank of all distinct amino acid substitutions. Since we noted that the substitution log-odds depend strongly on the evolutionary gradient and the structure features, we modified the methodology for obtaining the BLOSUM matrices, to derive class-specific log-odds. Therefore, we used sequence alignment substitutions of more than 67,000 protein chains, available in the PDB database, to calculate matrices for each decile of evolutionary gradient. Additionally, we subdivided these matrices into three classes of solvent accessibility, defined by 10 and 50 Å², and in secondary structure

elements of helix, strand and coil. Therefore, the substitution matrix consists of either 10 or 90 specific classes.

For each column of a multiple sequence alignment, we referred to the most frequent amino acid, i , and we counted the number of all mismatches. Substitutions to alignment gaps were ignored. Also, columns with more gaps than amino acids were excluded. Let the total number of amino acid pairs i, j ($1 \leq j \leq 20, 1 \leq i \leq 20, i \neq j$) of class c ($1 \leq c \leq 10$ or $1 \leq c \leq 90$), be f_{ijc} , then the observed frequency for the amino acid i to be substituted by j for the specific class c is

$$q_{ijc} = \frac{f_{ijc}}{\sum_j f_{ijc}}$$

The probability of occurrence of the amino acid pair i, j in the dataset is

$$e_j = \frac{\sum_i \sum_c f_{ijc}}{\sum_i \sum_j \sum_c f_{ijc}}$$

The log-odds matrices are then calculated with entries

$$s_{ijc} = \log_2\left(\frac{q_{ijc}}{e_j}\right)$$

No rounding to the nearest integer was made and so the scaling factor assumed to equal unit. Instead, a factor c was used to adjust the distance between the classes (see below). The log-odds were normalized to become percent ranks over all distinct amino acid substitutions.

Generate a multiple sequence alignment

Up to 5,000 sequence homologs were retrieved from a blastp search over the target databases of protein sequences. The default blast parameters were used, except the e-value cutoff that was set to 10^{-5} and the minimum sequence identity that was set to 0.3. Only up to two (lowest e-value) homologs with the same sequence identity to the query sequence were selected. The selected sequences were aligned using muscle. All columns with gap in the query sequence were removed.

Calculations of the adjustment factor c

A factor, c, adjusts the distance of class dependent matrices so that the impact of any substitution depends monotonically on the importance. For the substitution of amino acid i to j of the evolutionary class e ($2 \leq e \leq 10$) and structural class s ($1 \leq s \leq 9$) the adjustment factor is

$$c_{i,e,s} = \frac{1}{\sum_j \delta_{i,j,e,s}} \sum_{j \neq i} \frac{\delta_{i,j,e,s} * (p_{i,j,e,s} - p_{i,j,e-1,s})}{p_{i,j,e,s}}$$

where $p_{i,j,e,s}$ represents the product of ET rank and log-odds rank and the function $\delta_{i,j,e,s}$ selects for drops and is

$$\delta_{i,j,e,s} = \begin{cases} 1 & \text{if } (p_{i,j,e,s} - p_{i,j,e-1,s}) < 0 \\ 0 & \text{if } (p_{i,j,e,s} - p_{i,j,e-1,s}) \geq 0 \end{cases}$$

For the residues of the highest evolutionary importance class the $c_{i,e=1,s}$ equals unit.

Use SIFT, MAPP, Polyphen and Polyphen2

SIFT

The SIFT predictions were obtained using the “SIFT BLink” and the “SIFT Aligned Sequences” tools from the web server (www.sift.jcvi.org). The default parameters were used, such as select sequences include best BLAST hit to each organism and remove sequences more than 90 percent identical to query for BLink and remove sequences more than 100 percent identical to query for Aligned Sequences.

MAPP

The MAPP software was downloaded from the web page (www.mendel.stanford.edu/SidowLab/downloads/MAPP/) and run using the Uniprot90 alignment and the phylogenetic tree file generated by the ET. The “p-value interpretations of the MAPP scores” were used as the ultimate predicted impact of each amino acid variant.

Polyphen2

The Polyphen2 predictions were obtained using batch query tab from the web page (www.genetics.bwh.harvard.edu/pph2/). The default parameters were used, such as classifier model is HumDiv, genome assembly is GRCh37/hg19 and
5 transcripts are canonical.

Datasets

bacteriophage T4 lysozyme

The set of 2015 T4 lysozyme mutations were assayed by the amount of
10 formed plaque, due to lysozyme's break up of the host cell walls. Mutants with no (-) and difficult to discern (-+) plaque formation were considered as deleterious, while mutants with normal (+) and small plaque formation (+-) were considered as neutral.

15 E. coli lac repressor

The set of 4041 lac repressor mutations were assayed by the protein's repression activity. Mutations with phenotypes less than 4-fold (-) and 4- to 20- fold (-+) repression activity were considered as deleterious, while mutants with more than 200- fold (+) and 20- to 200- fold (+-) repression activity were considered as
20 neutral.

HIV-1 protease

The set of 336 HIV-1 protease mutations were assayed by the amount of cleavage products of Gag and Gag-Pol precursor proteins. Mutants with no (-) and
25 some (-+) product were considered as deleterious, while mutants with normal (+) function were considered as neutral.

P53

The set of 2314 p53 mutations were assayed in yeast for transactivation on 8
30 p53 response-elements. Since predictions assume that mutations result in loss of function, we treated single assay values greater than 100% as equal to 100%. Then, we calculated the average transactivation of each mutant over all elements and we

considered as deleterious those mutants with less than 50% of the wild-type activity and the rest as neutral.

The set of 26,597 p53 mutations was obtained from the IARC p53 database (p53 somatic mutations version R14). We divided it in recurrent mutations in cancer patients (at least 10 times) and non-recurrent (9 times or less). The recurrent set consists of 342 SNPs and the non-recurrent of 1,023 SNPs. Some of these SNPs were not functionally annotated in yeast.

Uniprot Dataset

We retrieved SNP phenotypes for all human proteins from the Uniprot database (www.uniprot.org/) and classified them as neutral if they match the keywords “dbSNP”, “polymorphism” and “VAR_” (variant) or as deleterious otherwise. From the 20,343 human proteins found the 11,995 (70%) had at least one SNP reported and only 3,023 (15%) had at least one deleterious phenotype. To select disease-causing proteins with higher confidence we selected only those that have at least 10 SNPs associated with the same phenotype and at most 10 SNPs associated by “Uncertain pathogenicity”. The resulting SNPs were manually reviewed to correct for misclassification due to the unmatched keyword search. Also, we ignored SNPs with uncertain pathogenicity and those with cancer related phenotypes when cancer was a minority of the disease phenotypes (less than 3 SNPs per protein). The p53 protein was excluded, since it was studied separately. The final dataset consists of 9,347 SNPs coming from 218 proteins.

Linear dependences

The scores of $\partial P / \partial r_i$ and Δr_i were calculated for each of the 6,056 mutations of T4 lysozyme and lac repressor and the 2,314 p53 mutations. The $\partial P / \partial r_i$ was approximated by the coverage of either the ET ranks or the Entropy of conservation, while the Δr_i was approximated by the coverage of either the ET-dependent or the BLOSUM62 log-odds. Either term of $\partial P / \partial r_i$ or Δr_i was classified into three bins of 0-40, 40-70 and 70-100 scores that defined the fixed groups, while the other term

was binned into deciles and was used as a variable. The trend lines were weighted by the number of mutations used for each plotted point.

Minor Allele Frequency

5 The Minor Allele Frequency (MAF) data were obtained by dbSNP using the site: <https://cgscmd.isi.edu/dbsnpq/downloads.php>. The entries of the _loc_maf and _loc_snp_gene_list_ref files were merged for the same SNP IDs. Since more than one frequency entries were found per SNP, we removed those with AF=1, and picked the maximum AF for each SNP. This yielded in 31,803 amino acid
10 variations. The average MAF was calculated for 50 equidistant bins of EGradient scores.

Statistical significance

 The statistical significance was calculated as follows. The Fisher's exact test
15 was used to rate the overlapping of evolutionary action with clinical association and yeast assay data. The t-test was used to compare the distributions of disease and polymorphism SNPs, and of the Pompe mutations. The log-rank test was used to assess the hazard of mutations found in glioblastoma tumors.

20 *Applications of action score*

 One of ordinary skill in the art will appreciate the biological, biochemical, and engineering applications of the ability to determine a phenotypic effect of a mutation in a protein.

 For example, it has been shown by the teachings above that there is a linear
25 relationship between the change in function of a protein and protein activity (Fig. 3 and 6). Using the action equation, as described above, one of ordinary skill in the art could engineer a protein with a selected mutation to have any desired level of protein function. The desired level of protein function could be anything from 0 to 100% activity. It is usually the case that one would want to selectively decrease
30 protein activity. By determining the location and type of amino acid residue to mutate using the action equation, one of ordinary skill in the art could easily and

accurately predict the resulting mutated protein have a certain level of protein activity.

Furthermore, titration or customization of protein activity could also relate to or equate to enzymatic activity, catalytic activity, binding activity, other biochemical activity, *in vivo* activity, *in vitro* activity, or overall organismal fitness. In the context of synthetic biology, existing proteins can be redesigned or reconstructed to have a specified level of phenotypic activity. Also, novel proteins can be designed and constructed for the same purpose.

Also, identified mutations (in any coding sequence) could easily and accurately be assessed for those with a higher probability of potentially causing disease and/or clinical symptoms in a patient or organism. Single nucleotide variations can result in single amino acid substitutions, and can have a deleterious effect on the protein, cellular, and/or organismal level.

As described in detail above, this process allows one to calculate the impact of coding mutations in proteins. This information can then be used to interpret mutations in patients, or other organisms. This process also allows for the computation of the evolutionary gradient, designated as $\partial P / \partial r_i$. This process also allows for the computation of the magnitude of a mutation in a sequence, designated as Δr_i . Finally, this process allows for the translation of the magnitude of the mutation in a sequence into a perturbation magnitude in function, phenotype, or fitness space.

The teachings of all patents, published applications and references cited herein are incorporated by reference in their entirety.

While this invention has been particularly shown and described with references to example embodiments thereof, it will be understood by those skilled in the art that various changes in form and details may be made therein without departing from the scope of the invention encompassed by the appended claims.

CLAIMS

What is claimed is:

1. A method of determining a phenotypic effect of a mutation in a protein, the method comprising:
 - 5 in a processor calculating an action of a protein mutation in the protein as a function of a magnitude of a change at a protein residue and an evolutionary gradient;
 - wherein the evolutionary gradient comprises measuring a magnitude of an evolutionary jump associated with the mutation at an amino acid
 - 10 sequence position; and
 - based on the calculated action, outputting to a user indication of a phenotypic effect of the protein mutation.

2. The method of claim 1, wherein calculating the action of the protein mutation is according to the formula:

$$\Delta P = k * \frac{\partial P}{\partial r_i} * \Delta r_i$$

wherein P is a protein, k is a scaling factor, Δr_i is a sequence displacement as the magnitude of a change at a protein residue i , and $\frac{\partial P}{\partial r_i}$ is the evolutionary gradient.

3. The method of claim 2, wherein the evolutionary gradient further comprises a slope of a tangent to a protein evolutionary trajectory and comprises the i^{th} component of the protein gradient and a partial derivative of the protein with respect to residue i , indicated by symbol ∂ .
4. The method of claim 2, wherein the scaling factor k is about 1.

5. The method of claim 2, wherein the sequence displacement is measured as a percent rank of all amino acid substitutions.
6. The method of claim 1, wherein the protein is a viral, eukaryotic, or
5 prokaryotic protein.
7. The method of claim 1, wherein the mutation is a point mutation.
8. The method of claim 7, wherein the point mutation is a missense mutation.
10
9. The method of claim 8, further comprising allowing the user to engineer a protein based on the calculated action and phenotypic effect of a mutation in the protein.
- 15 10. A computer program product for determining a phenotypic effect of a mutation in a protein comprising:
a computer readable medium having program code embodied, the program code executable by a computer to
measure a magnitude of an evolutionary jump associated with a
20 protein mutation in a protein at an amino acid sequence position;
calculate an action of the protein mutation as a function of a magnitude of a change at a protein residue i and an evolutionary gradient, wherein the evolutionary gradient utilizes the measured magnitude; and
display to a user an indication of phenotypic effect of the protein
25 mutation based on (corresponding to) the calculated action.
11. A computer system for classifying a phenotypic effect of a mutation in a protein, the system comprising:

a protein mutation execution engine (module/subassembly) to calculate an action of a protein mutation in the protein as a function of a magnitude of a change at a protein residue and an evolutionary gradient;

5 wherein the evolutionary gradient comprises measuring a magnitude of an evolutionary jump associated with the mutation at a amino acid sequence position; and

based on the calculated action, outputting to a user indication a phenotypic effect of the protein mutation.

- 10 12. The computer system of claim 11, wherein the protein mutation execution engine (module) calculates the action of a protein mutation according to the formula:

$$\Delta P = k * \frac{\partial P}{\partial r_i} * \Delta r_i$$

- 15 wherein P is a protein, k is a scaling factor, Δr_i is a sequence displacement as the magnitude of a change at a protein residue i , and $\frac{\partial P}{\partial r_i}$ is the evolutionary gradient.

13. The computer system of claim 12, wherein the evolutionary gradient further comprises a slope of a tangent to a protein evolutionary trajectory and comprises the i^{th} component of the protein gradient and a partial derivative of the protein with respect to residue i , indicated by symbol ∂ .
- 20

14. The computer system of claim 12, wherein the scaling factor k is about 1.

15. The computer system of claim 12, wherein the sequence displacement is measured as a percent rank of all amino acid substitutions.
- 25

16. The computer system of claim 11, wherein the protein is a viral, eukaryotic, or prokaryotic protein.
17. The computer system of claim 11, wherein the mutation is a point mutation.
18. The computer system of claim 17, wherein the point mutation is a missense mutation.
19. The computer system of claim 11, further comprising allowing the user to engineer a protein based on the calculated action and phenotypic effect of a mutation in the protein.

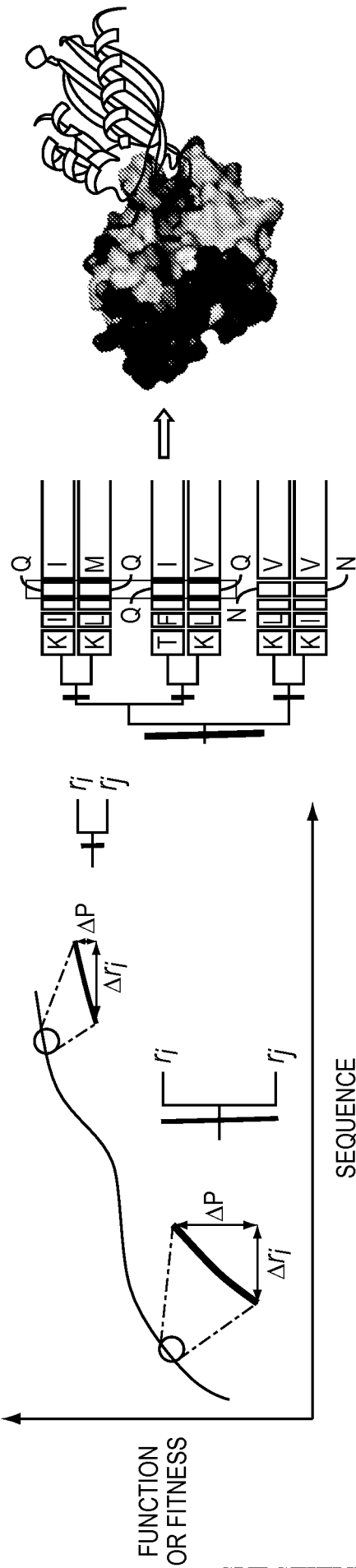
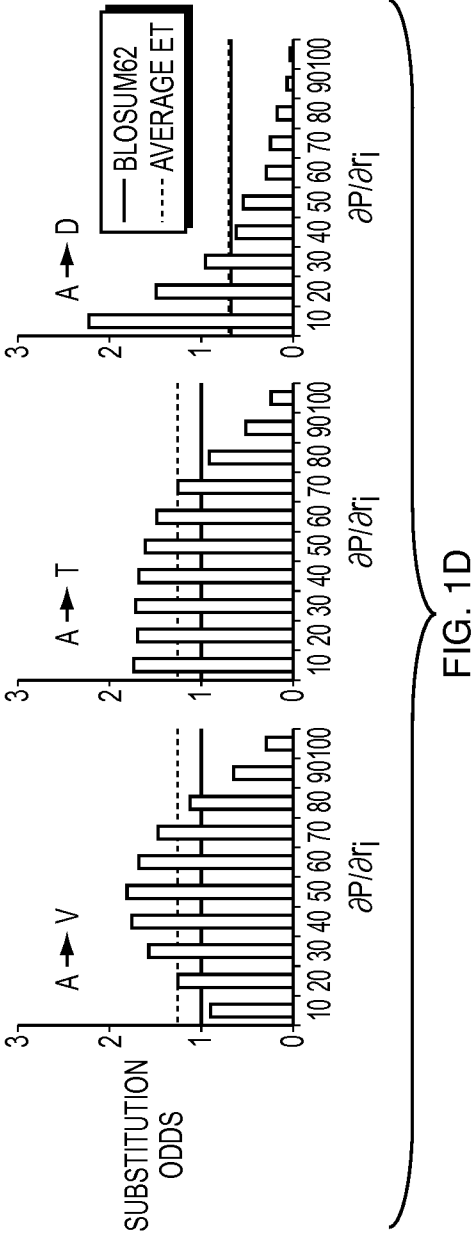
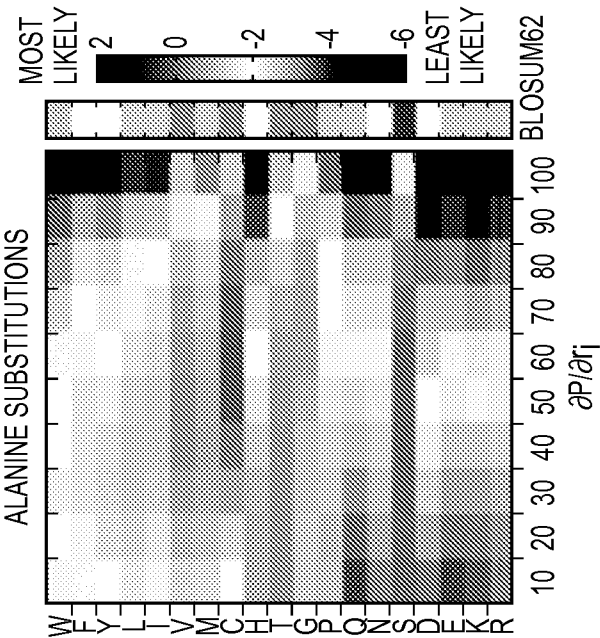


FIG. 1B



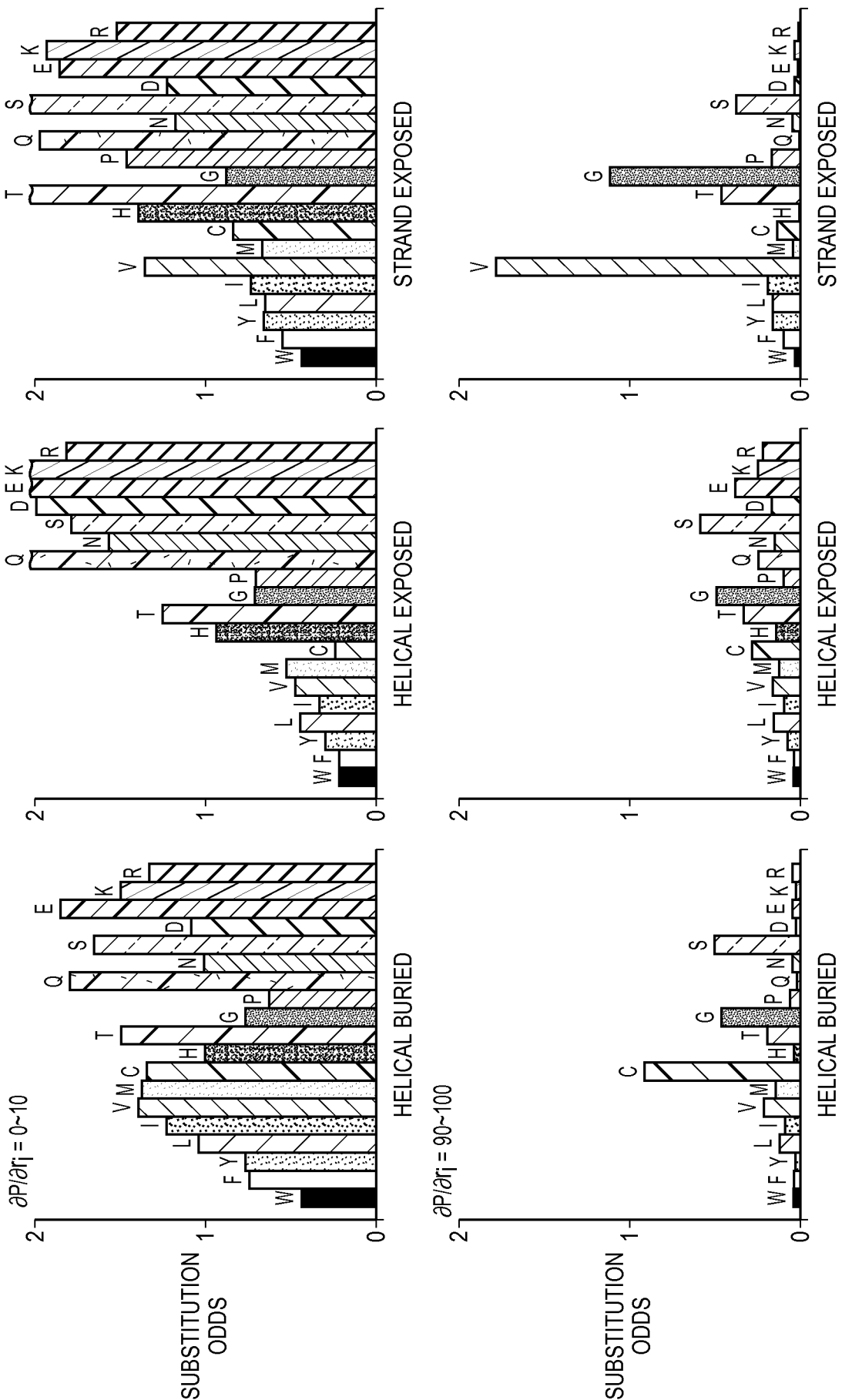


FIG. 2

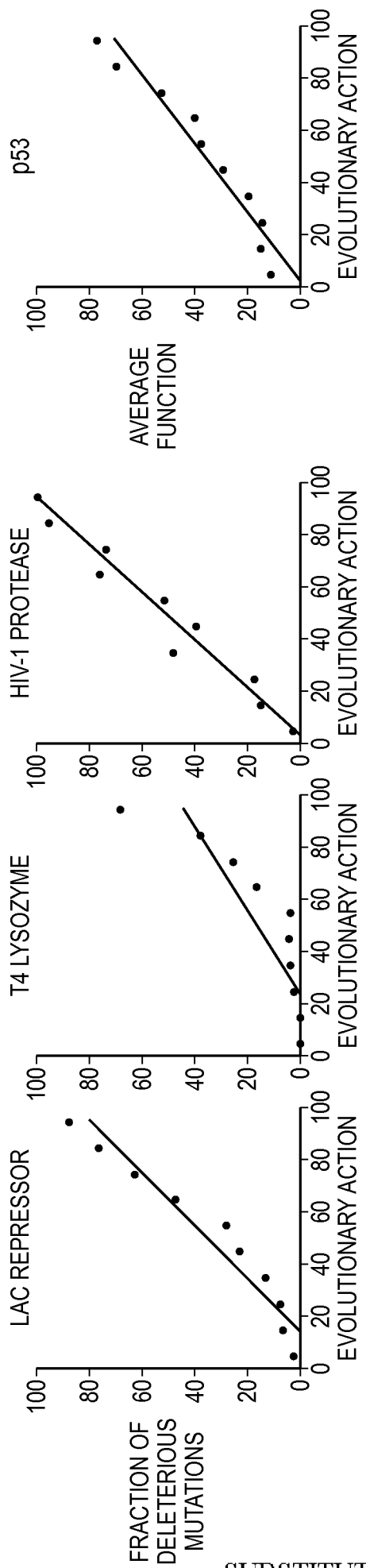


FIG. 3A

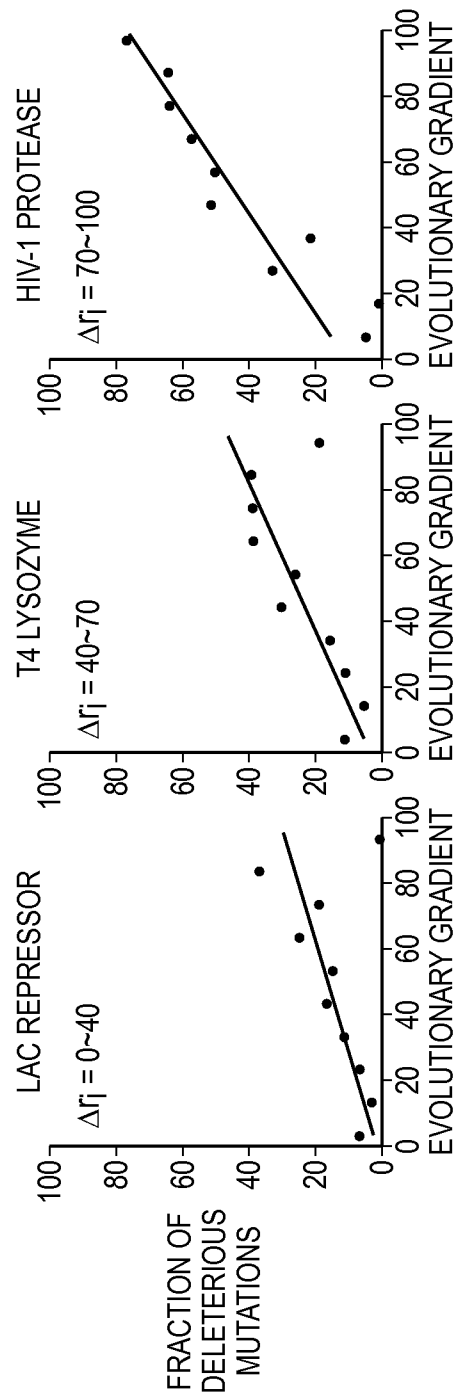
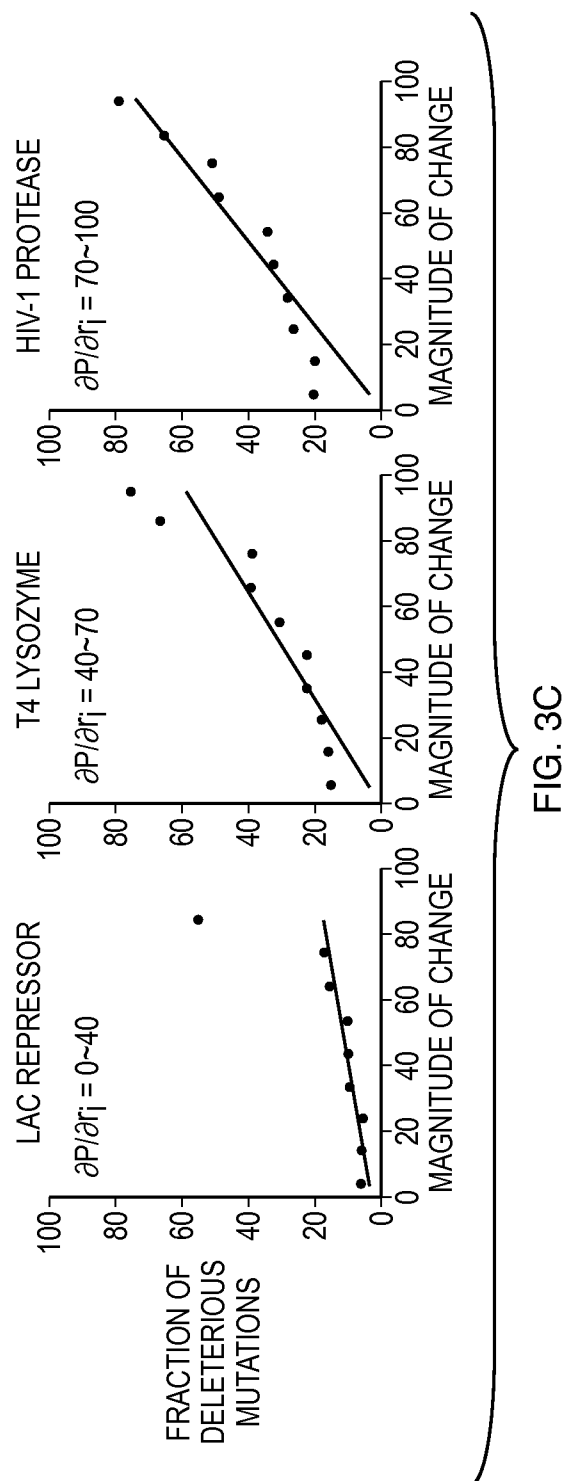


FIG. 3B



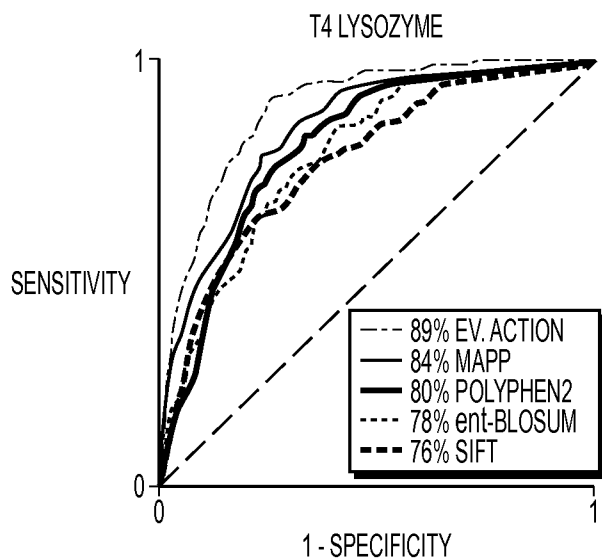


FIG. 4A

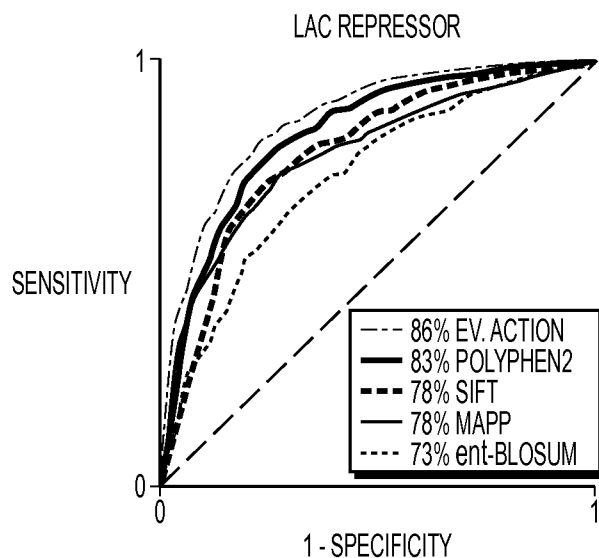


FIG. 4B

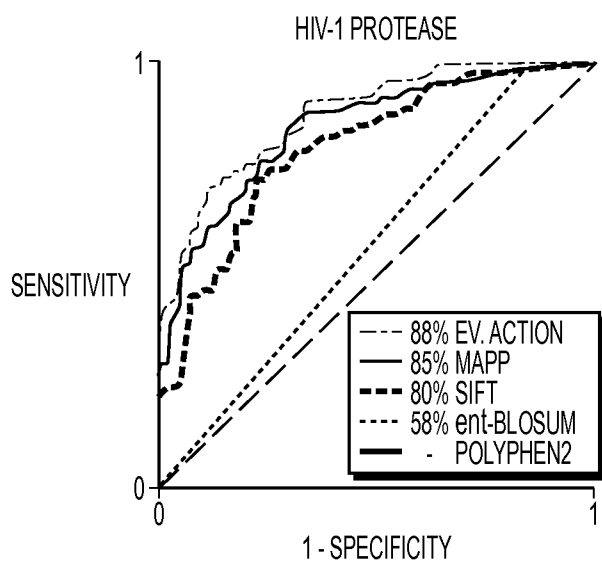


FIG. 4C

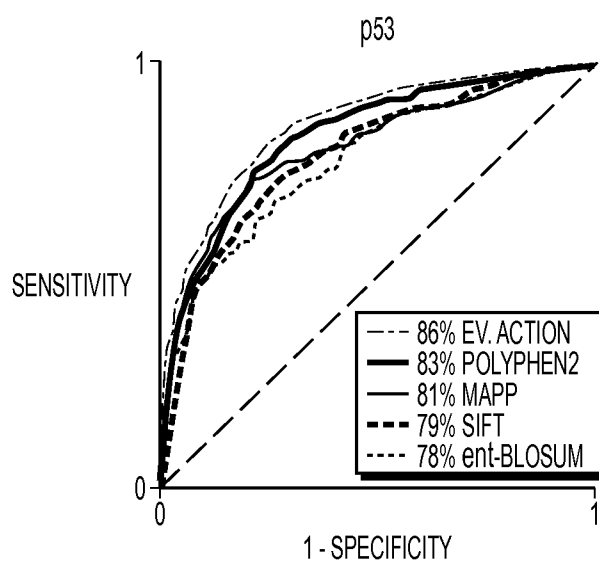


FIG. 4D

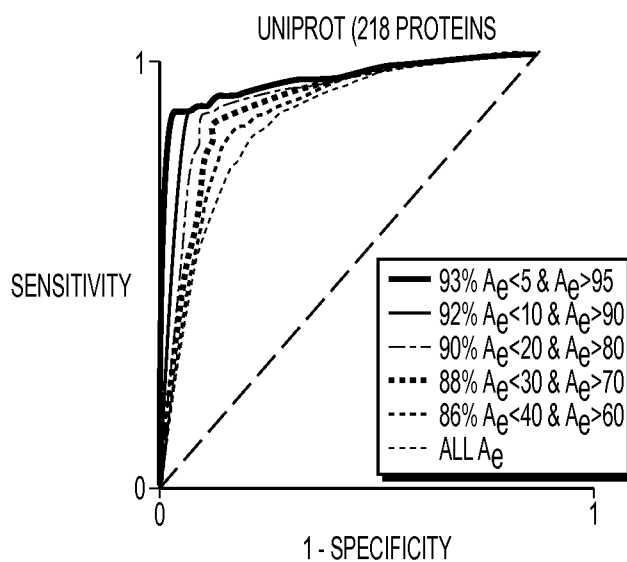


FIG. 4E

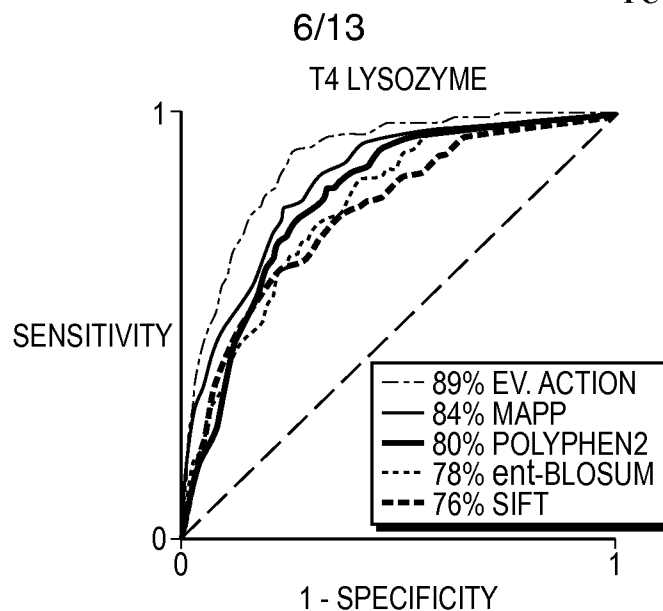


FIG. 5A

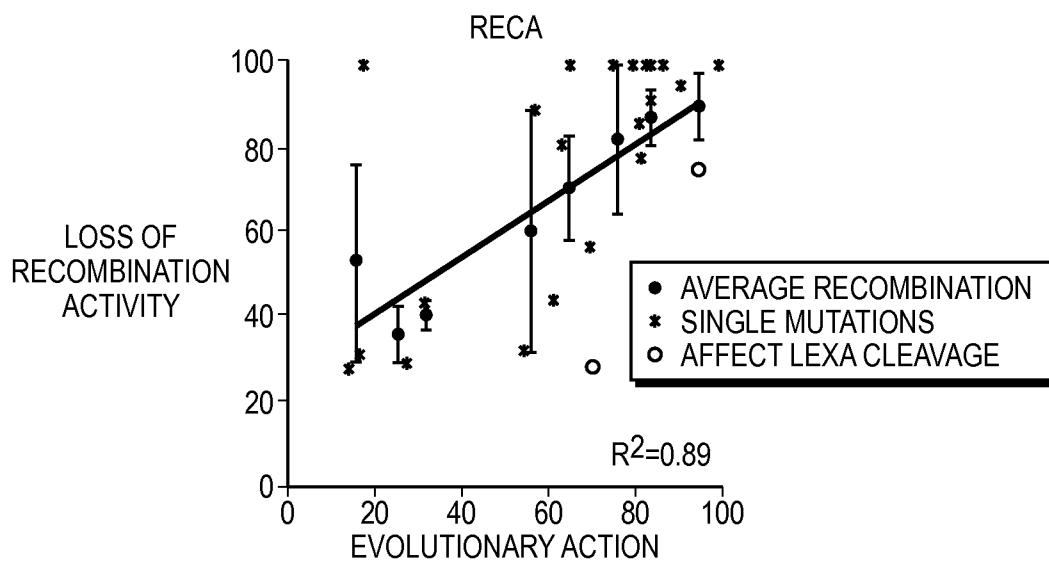


FIG. 5B

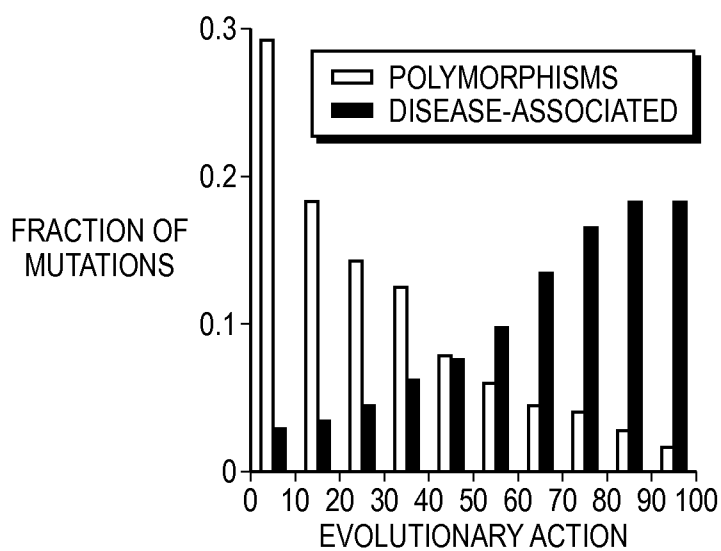


FIG. 5C

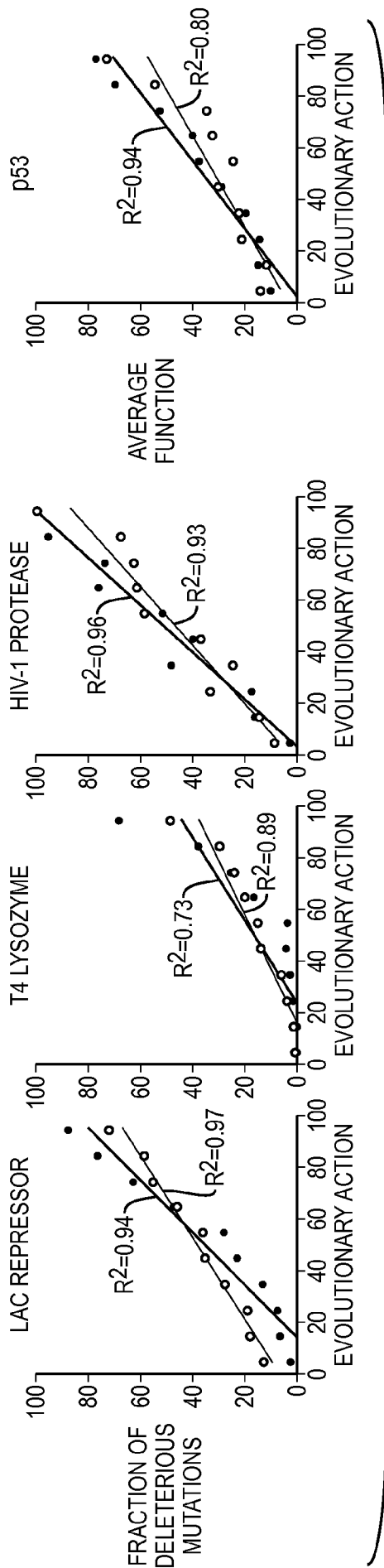


FIG. 6A

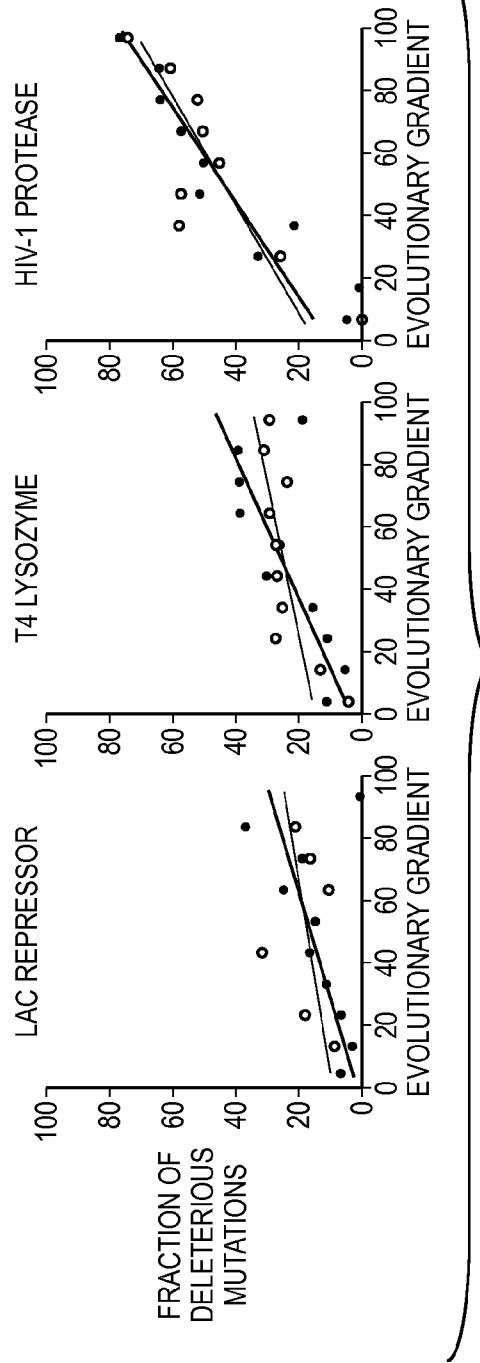
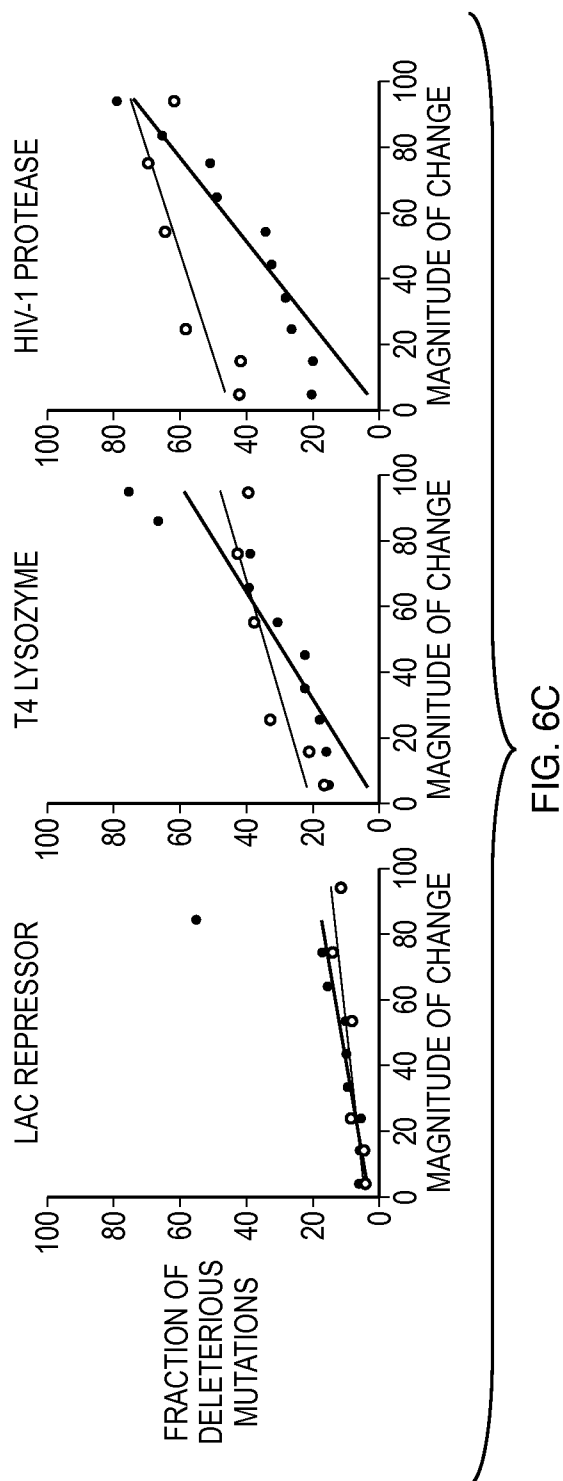


FIG. 6B



9/13

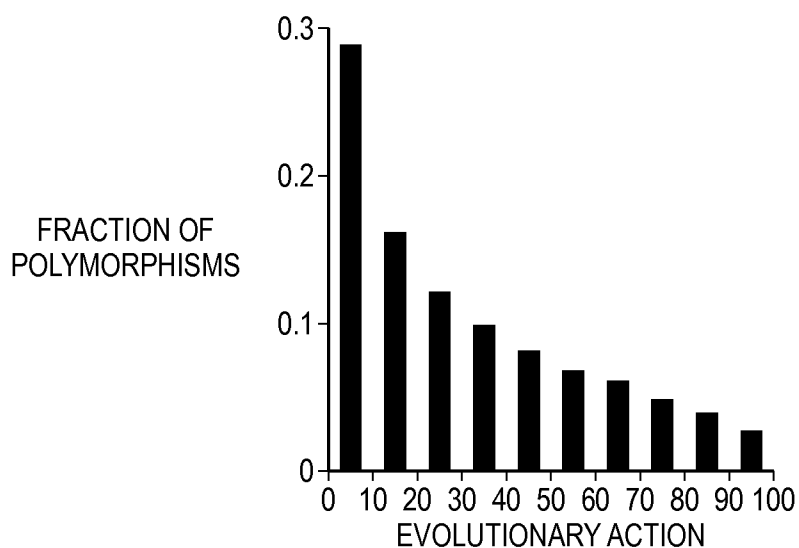


FIG. 7A

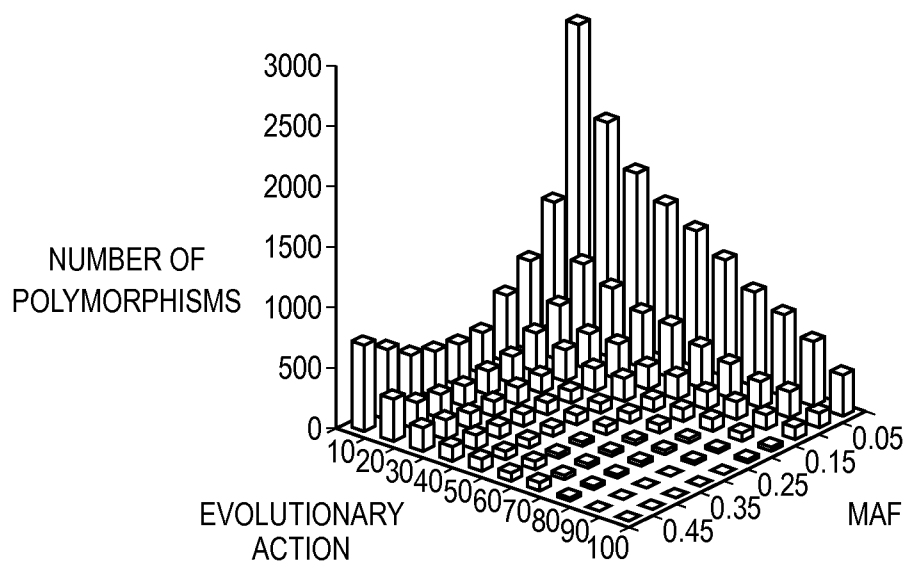


FIG. 7B

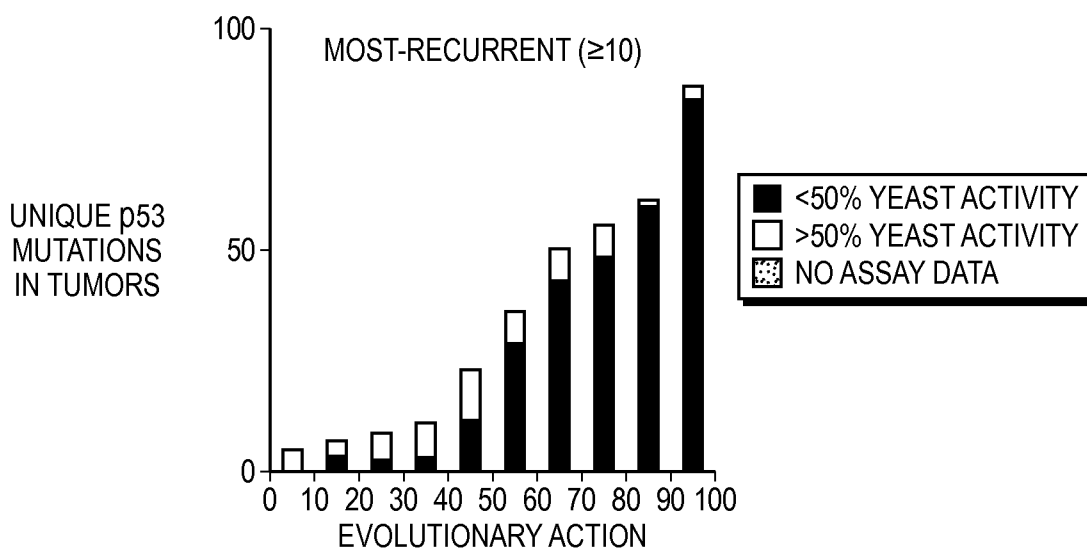


FIG. 7C

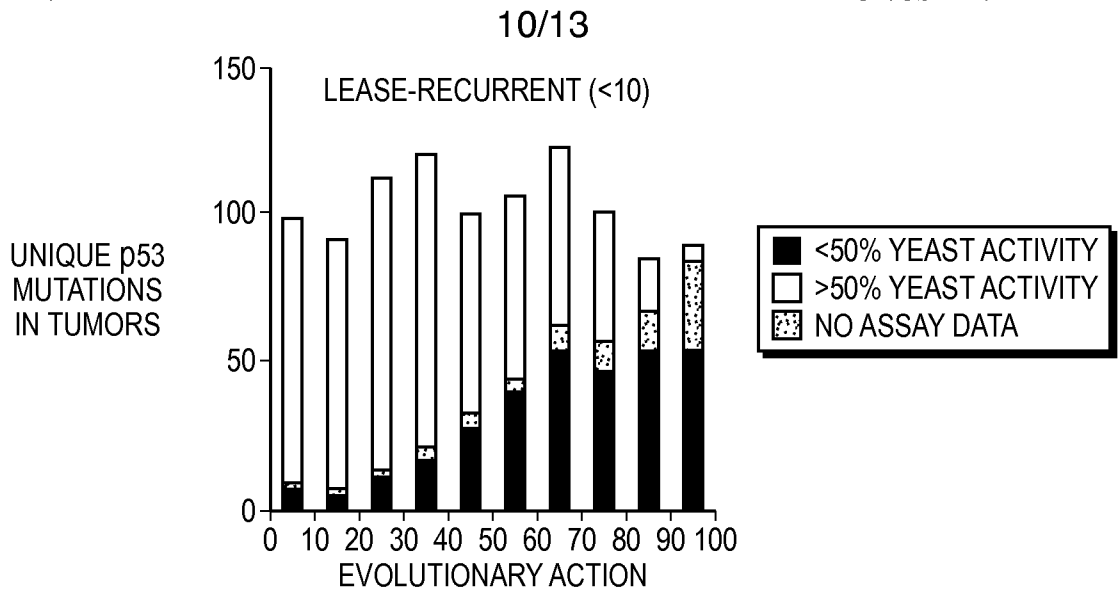


FIG. 7D

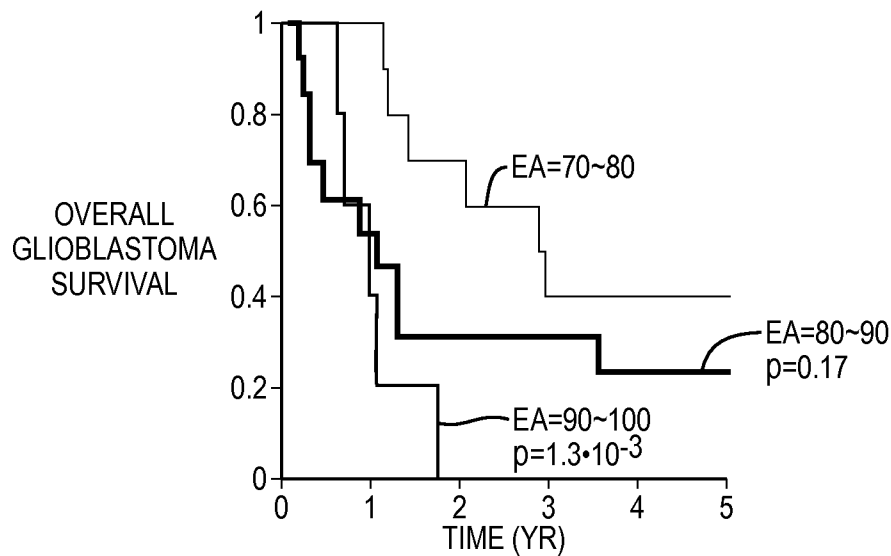


FIG. 7E

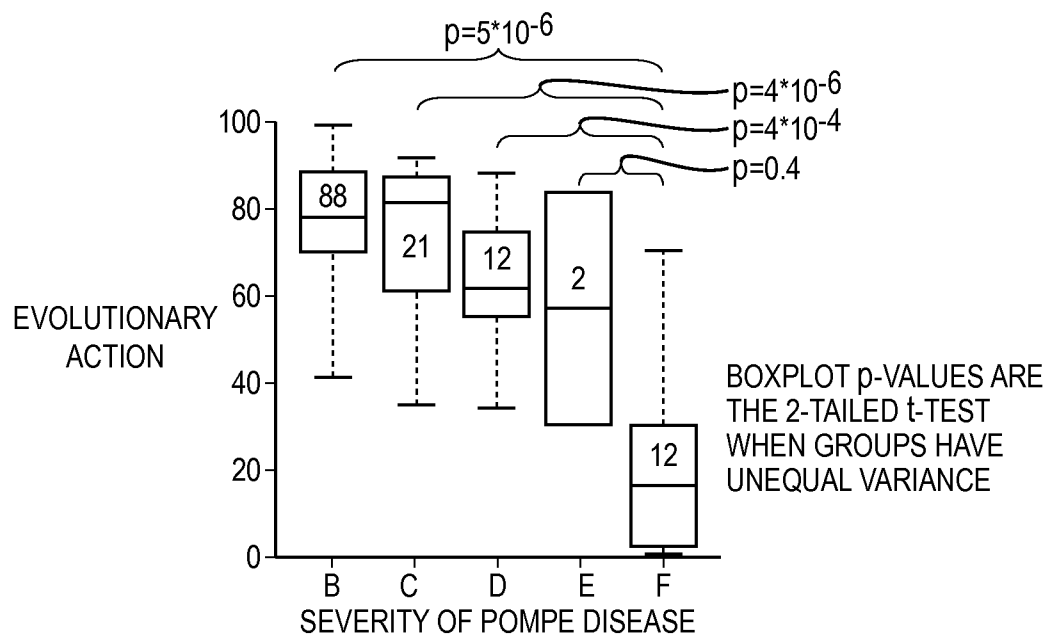


FIG. 7F

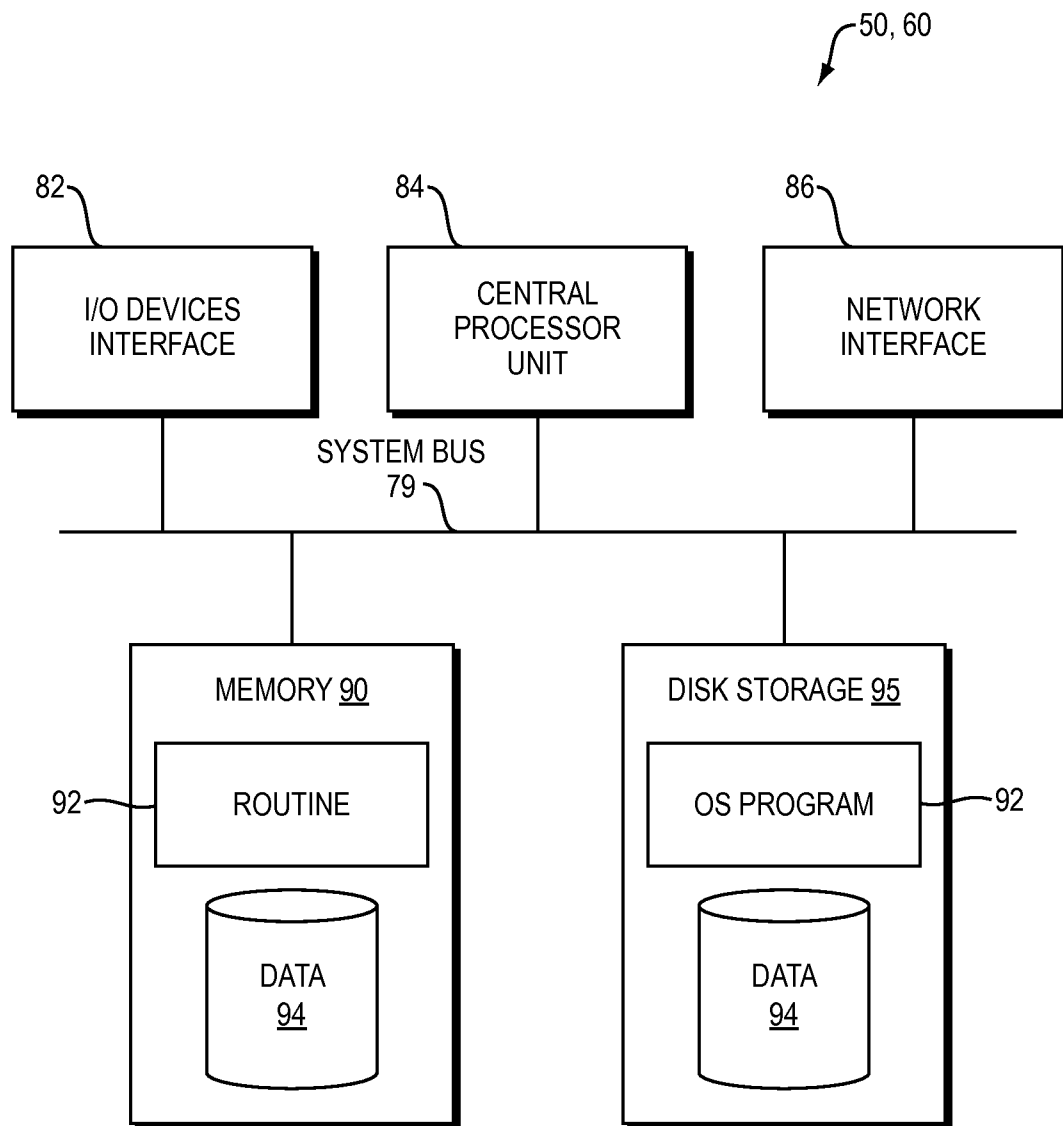


FIG. 8

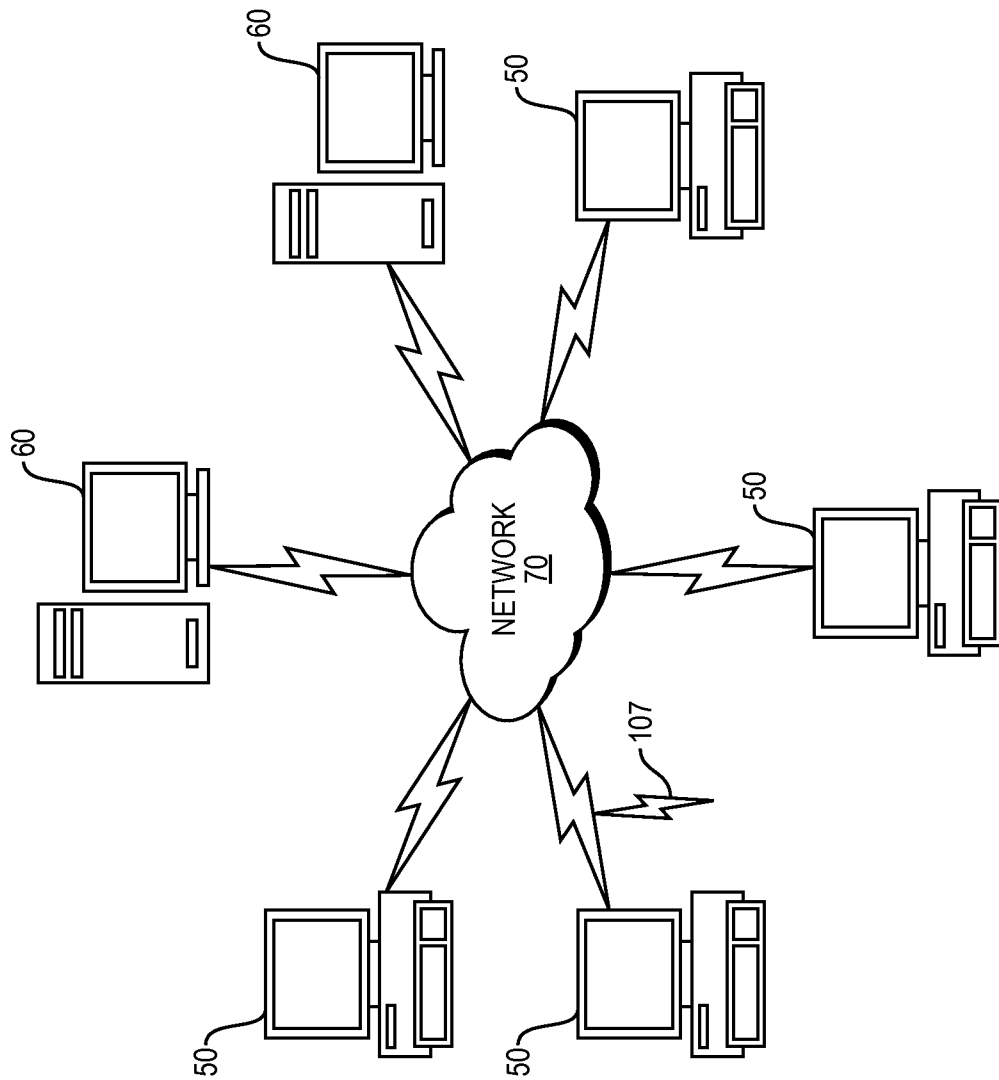


FIG. 9

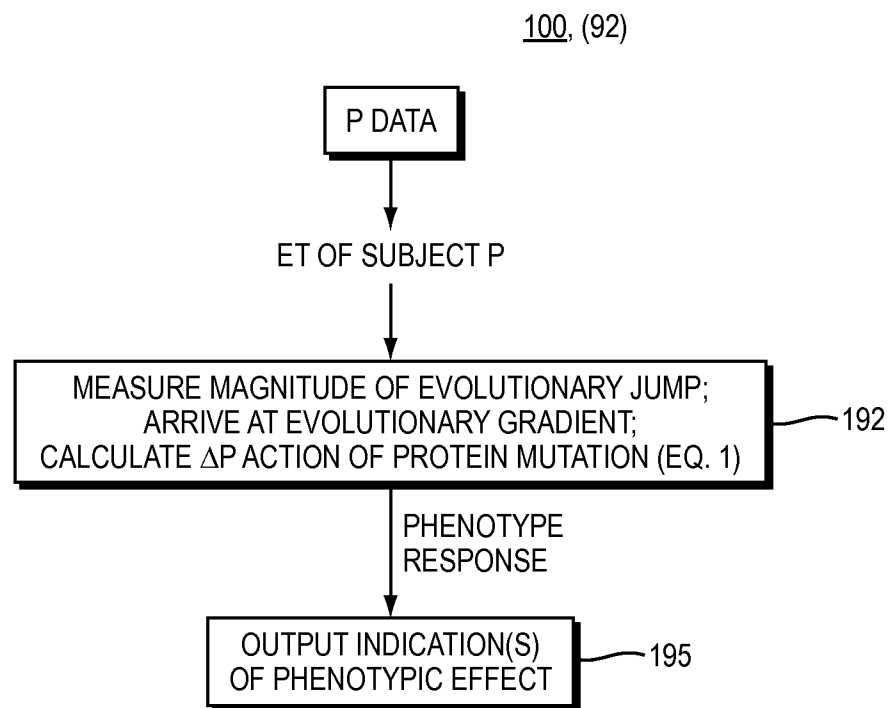


FIG. 10

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2013/032215

A. CLASSIFICATION OF SUBJECT MATTER

INV. G06F19/18
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EP0-Internal, BIOSIS, EMBASE

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	ERDIN S ET AL: "Evolutionary Trace Annotation of Protein Function in the Structural Proteome", JOURNAL OF MOLECULAR BIOLOGY, ACADEMIC PRESS, UNITED KINGDOM, vol. 396, no. 5, 12 March 2010 (2010-03-12), pages 1451-1473, XP026906470, ISSN: 0022-2836, DOI: 10.1016/J.JMB.2009.12.037 [retrieved on 2010-02-12] abstract page 1452, left-hand column, last paragraph - page 1456, left-hand column, paragraph 1 page 1462, left-hand column, last paragraph - page 1470, right-hand column, last paragraph figures 1-3 -/--	1-19



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

25 October 2013

Date of mailing of the international search report

04/11/2013

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Hilbig, Matthias

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2013/032215

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	----- MARIMUTHU PARTHIBAN ET AL: "In silico point mutation and evolutionary trace analysis applied to nicotinic acetylcholine receptors in deciphering ligand-binding surfaces", JOURNAL OF MOLECULAR MODELING, SPRINGER VERLAG, DE, vol. 16, no. 10, 5 March 2010 (2010-03-05), pages 1651-1670, XP019854504, ISSN: 0948-5023 abstract page 1653, left-hand column, paragraph 1 - page 1656, right-hand column, paragraph 1 page 1660, left-hand column, paragraph 2 - page 1669, right-hand column, last paragraph figure 9	1-19
X	----- LICHTARGE OLIVIER ET AL: "An evolutionary trace method defines binding surfaces common to protein families", JOURNAL OF MOLECULAR BIOLOGY, ACADEMIC PRESS, UNITED KINGDOM, vol. 257, no. 2, 1 January 1996 (1996-01-01), pages 342-358, XP002252415, ISSN: 0022-2836, DOI: 10.1006/JMBI.1996.0167 abstract page 342, right-hand column, last paragraph - page 346, left-hand column, paragraph 3 page 353, left-hand column, last paragraph - page 357, left-hand column, last paragraph figures 1-3	1-19
X	----- MIHALEK I ET AL: "A Family of Evolution-Entropy Hybrid Methods for Ranking Protein Residues by Importance", JOURNAL OF MOLECULAR BIOLOGY, ACADEMIC PRESS, UNITED KINGDOM, vol. 336, no. 5, 5 March 2004 (2004-03-05), pages 1265-1282, XP004490180, ISSN: 0022-2836, DOI: 10.1016/J.JMB.2003.12.078 abstract page 1265, left-hand column, paragraph 1 - page 1271, right-hand column, paragraph 1 page 1280, left-hand column, paragraph 2 - page 1281, right-hand column, paragraph 1 figure 1 -----	1-19