On the Allosteric Effect of nsSNPs and the Emerging Importance of Allosteric Polymorphism

Wei-Ven Tee [1,2], Enrico Guarnera [1] and Igor N. Berezovsky [1,2]

1 - Bioinformatics Institute (BII), *Agency for Science, Technology and Research (A*STAR), 30 Biopolis Street, #07-01, Matrix, Singapore 138671*
2 - Department of Biological Sciences (DBS), *National University of Singapore (NUS), 8 Medical Drive, Singapore 117597*

Correspondence to Igor N. Berezovsky: *Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), 30 Biopolis Street, #07-01, Matrix, Singapore 138671. igorb@bii.a-star.edu.sg*
https://doi.org/10.1016/j.jmb.2019.07.012
Edited by Anna Panchenko

Abstract

The molecular mechanisms of pathological non-synonymous single-nucleotide polymorphisms are still the object of intensive research. To this end, we explore here whether non-synonymous single-nucleotide polymorphisms can work via allosteric mechanisms. Using structure-based statistical mechanical model of allostery and analyzing energetics of the effects of mutations in a set of 27 proteins with at least 50 pathological SNPs in each molecule, we found that, indeed, some SNPs can work allosterically. We illustrate the molecular basis of disease phenotypes caused by allosteric SNPs with the case studies of human galactose 1-phosphate uridyltransferase (GALT) and glucose-6-phosphate dehydrogenase (G6PD). We also found that mutations of a number of other residues in the protein may cause modulation comparable to those observed for known pathological SNPs. In order to explain this, we propose a notion of allosteric polymorphism, which implies the presence of a number of critical positions in the protein sequence, whose mutations can allosterically disrupt the protein function and result in a disease phenotype. We conclude that the emerging importance of allosteric polymorphism calls for the development of computational framework for analyzing the allosteric effects of mutations and their role in the modulation of protein activity.

Keywords

protein dynamics
allostery
nsSNP
mutations
allosteric polymorphism

Introduction

Rapid advances in the whole-exome sequencing have provided a wealth of information on the genetic variants underlying human diseases [1,2], estimating that about 85% of disease-related mutations are located in the coding regions [3]. Moreover, comparison of two random exomes has shown about 10,000 non-synonymous single-nucleotide polymorphisms (nsSNPs), pointing to the tremendous sequence variation in individual genomes [4]. To this end, large-scale exome sequencing projects have become instrumental in identifying nsSNPs implicated in human diseases, such as cancers [5–7], cardiovascular [8], and ocular diseases [9], to name a few. At the same time, despite a massive amount of data and active research on disease-causing nsSNPs, the molecular mechanisms underlying their pathological manifestation remain mostly unclear [10–12]. There are various scenarios in which widely defined protein function can be affected or completely abolished by mutations. First, mutation of critical residue(s) in a catalytic or binding site can directly destroy the protein enzymatic or binding activity. Second, mutations can produce an impact on the overall protein stability, which in many cases can modulate different modes of the protein functional activity, for example, catalysis or binding. Third, while mutations of residues undergoing post-

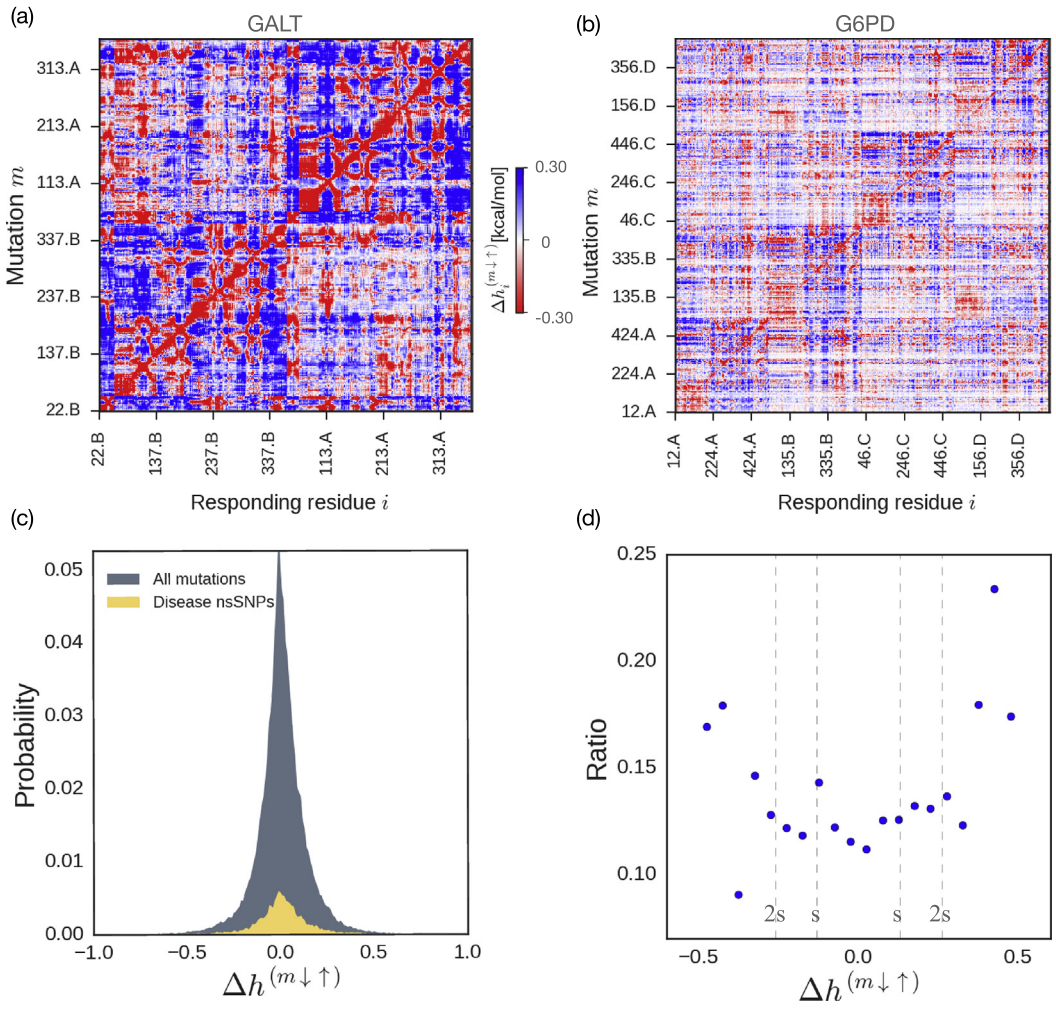

Fig. 1. ASMs, distribution of modulation ranges in functional sites of 27 proteins and its analysis. (a–b) ASMs for GALT and G6PD. (c) The probability distribution of $\Delta h^{(m\downarrow\uparrow)}$ in functional sites of 27 proteins with more than 50 nsSNPs as a result of all distal amino acid substitutions (separated from all residues of a site by more than 11 Å, which is the distance cutoff for directly interacting residues in the model [21]) is colored in gray, and distribution of the effects of substitutions caused by positions hosting pathological nsSNPs is colored in yellow. A list of ligand binding sites is tabulated in Suppl. Table 1. Pathological nsSNPs are collected from the human polymorphisms and disease mutations index (humsavar) from UniProtKB/Swiss-Prot and from the literature. (d) The ratio of the number of position with pathological nsSNPs to the number of all considered distal mutations with the same modulation strength.

translational modifications can affect its activity by changing the protein structure and stability, they can also disrupt protein–protein interactions facilitated by post-translational modifications in the protein signaling network [13]. Finally, some mutations can affect distal functional sites via allosteric signaling, which remotely changes catalytic or binding activity as a result of the altered dynamics of the whole structure caused by the mutation [14–16].

The abundance of disease-causing nsSNPs and, at the same time, multiple indications of the involvement of allosteric mutations in different pathologies [17–19] raise a question if some SNPs work allosterically. It has been shown that consideration of specific mutational aberrations compels the paradigm shift in what is called “precision oncology” in which therapeutic selection will be determined by the genomic sequence of an individual and its mutations rather than by the cancer type and tissue/organ location [12]. At the same time, only genomic sequence data may not provide sufficient information, because statistics of mutations do not necessarily reflect changes in the protein conformation. Therefore, rare mutations or so-called latent drivers (passenger mutations that can transform into drivers) can be underestimated or even ignored in the sequence-based analysis. Our goal here is to quantify how effects of mutations are manifested in the changes of protein conformational ensemble, which may lead to a distortion of the protein

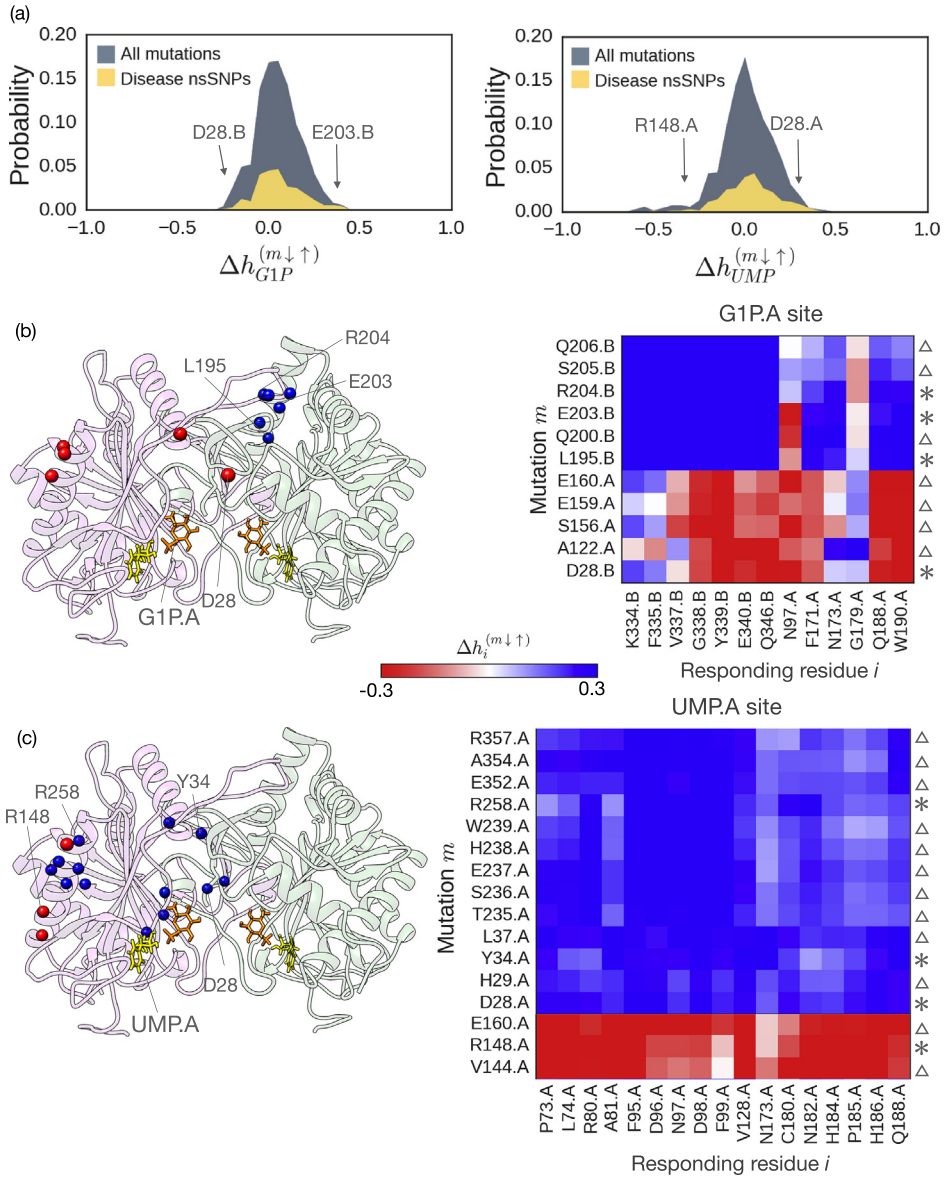

Fig. 2. Analysis of known galactosemia-associated nsSNPs in galactose 1-phosphate uridyltransferase (GALT) and detection of additional positions with potentially harmful allosteric effects. (a) The probability distributions of the modulation ranges $\Delta h_{\text{UMP}}^{(m\downarrow\uparrow)}$ and $\Delta h_{\text{G1P}}^{(m\downarrow\uparrow)}$ obtained as a result of distal amino acid substitutions in GALT. The residue index is followed by the protein chain identifier. Residues in the UMP and G1P sites in GALT are identified based on a contact distance cutoff of 4.5 Å with corresponding ligands in the crystal structure (PDB: 5in3), in addition to residues that are known to interact with the ligands (literature-based). The fraction of residue positions associated with diseases that can cause modulation ranges greater than one standard deviation (s= 0.12 kcal/mol) is 0.38. (b–c) Examples of residues are shown as spheres on the structures (left), and those implicated in disease-causing nsSNPs are labeled. The spheres are colored according to the sign of the modulation ranges resulted in the corresponding sites—positive and negative $\Delta h_{site}^{m\downarrow\uparrow}$ values are indicated in blue and red, respectively. Chains A and B of the GALT dimer are colored in light purple and green, respectively. G1P ligand is colored in orange, and UMP ligand - in yellow. On the right, the per-residue modulation ranges in the functional sites are obtained from the relevant parts of the complete ASM. Residues that are associated with disease-causing nsSNPs are marked by asterisks, while other positions that produce allosteric modulation similar to that by the positions hosting pathological nsSNPs are marked by triangles. The negative-to-positive variation of the modulation range $\Delta h_{\text{site}}^{(m\downarrow\uparrow)}$ is represented by the red-to-blue color gradient (same in Fig. 3).

activity via allosteric mechanisms. We investigate the effects of mutations, starting from the analysis of allosteric modulation that is originated by residue positions hosting known nsSNPs, then we perform a comprehensive per-residue analysis of the allosteric signaling in 27 proteins with multiple nsSNPs (more than 50 in each protein). We use here previously introduced structure-based statistical mechanical model of allostery (SBSMMA), which allows to estimate the energetics of allosteric communication caused by the ligand binding perturbations [20], to identify allosteric sites by reversing the allosteric communication via perturbation of functional sites [21], and to observe allosteric effects of mutations [14,15]. The SBSMMA was implemented in the AlloSigMA webserver [22] and was used for building the AlloMAPS database, which contains comprehensive Allosteric Signaling Maps (ASMs) for about 2000 proteins and protein chains [23].

Statistical mechanical model for the analysis of allosteric effect of mutations

In the framework of SBSMMA [14,20], a single crystal structure is used to build the Cα harmonic models for the native and the perturbed states of the protein. For the native system, the energy function is

$$E^{(0)}(\delta \boldsymbol{r}) = \sum_{i<j} k_{ij}\left(d_{ij}-d_{ij}^{0}\right)^{2} \qquad (1)$$

where $\delta \boldsymbol{r}$ is the 3N-dimensional vector of displacements of the Cα atoms with respect to the reference structure. The distance between a pair of Cα atoms i and j is denoted by $d_{ij,}$ and the corresponding distance in the reference structure is d_{ij}^{0}. The k_{ij} is a distance-dependent force constant that decays as $(1/d_{ij}^{0})^{6}$, and the global distance cutoff is 25 Å [24]. The energy function for the perturbed (mutated) state is

$$E^{(P)}(\delta \boldsymbol{r}, m) = \sum_{i<j, i\neq m} k_{ij}\left(d_{ij}-d_{ij}^{0}\right)^{2} + \alpha \sum_{j} k_{mj}\left(d_{mj}-d_{mj}^{0}\right)^{2} \qquad (2)$$

where in the second term the alteration of the force constants that couple the mutated residue m with all neighboring residues is modeled via the scaling parameter α. Two types of residue mutations are considered here: substitution to bulky amino acids, strongly interacting with the environment (stabilizing, UP mutation, $m\uparrow$ with $\alpha = 10^{2}$), and to small ones, less interacting, such as glycine or alanine (destabilizing DOWN mutation, $m\downarrow$ with $\alpha = 10^{-2}$). The configurational ensembles in the native and perturbed states are estimated from the two sets of normal modes $\boldsymbol{e}_{\mu}^{(0)}$ and $\boldsymbol{e}_{\mu}^{(P)}$ obtained from the Hessian matrices of the energy functions in Eqs. (1) and (2), respectively. The energetic impact of allosteric signaling is evaluated for each state using the allosteric potential, which measures the total elastic work applied on a residue as a result of the change in displacements of all neighbors

$$U_{i}(\sigma) = \frac{1}{2}\sum_{\mu} \varepsilon_{\mu,i}\sigma_{\mu}^{2} \qquad (3)$$

where $\varepsilon_{\mu,i} = \sum |\boldsymbol{e}_{\mu,i}-\boldsymbol{e}_{\mu,j}|^{2}$ are parameters obtained from the normal modes of corresponding states (native or mutated), and σ is a vector of Gaussian distributed amplitudes $\sigma = (\sigma_1, \ldots, \sigma_{\mu}, \ldots)$ with variance $1/\varepsilon_{\mu,\ i}$ [20]. Since the generic displacement of a residue i is $\delta \boldsymbol{r}_{i}(\sigma) = \sum_{\mu} \sigma_{\mu}\boldsymbol{e}_{\mu,i}$, the vector σ identifies a configurational state. Thus, by integrating the allosteric potential in Eq. 3 over all possible configurations σ of a residue in the native and perturbed (mutated) states the corresponding partition functions can be calculated, providing the free energies of the native and mutated states [14,20]. Finally, the per-residue free energy change upon the perturbation is given by

$$\Delta g_{i}^{(P)} = \frac{1}{2}k_{B}T\sum_{\mu} \ln\frac{\varepsilon_{\mu,i}^{(P)}}{\varepsilon_{\mu,i}^{(0)}} \qquad (4)$$

Since, in principle, any mutation induces an allosteric signal to all distal residues of a protein, we consider the background-free allosteric modulation on a residue i, showing the extent at which obtained allosteric signaling differs from the global effect. To this end, we estimate a deviation of the free energy change from its mean over all residues in the protein chain

$$\Delta h_{i}^{(P)} = \Delta g_{i}^{(P)} - \left\langle \Delta g_{i}^{(P)} \right\rangle_{Prot} \qquad (5)$$

A positive allosteric modulation indicates an increase of configurational work applied on the residue that may cause local conformational changes due to the changes in the configurational ensemble of its neighbors. On the other hand, a negative allosteric modulation leads to the opposite effect, which results in preventing the conformational change. The allosteric modulation at the site of interest is calculated by averaging the per-residue allosteric modulations of all residues that fall within a cutoff contact distance of 4.5 Å from the corresponding substrate, catalytic product, or cofactor molecule in the crystal structure.

In order to have a generic description of a strength of the allosteric effect on a protein residue

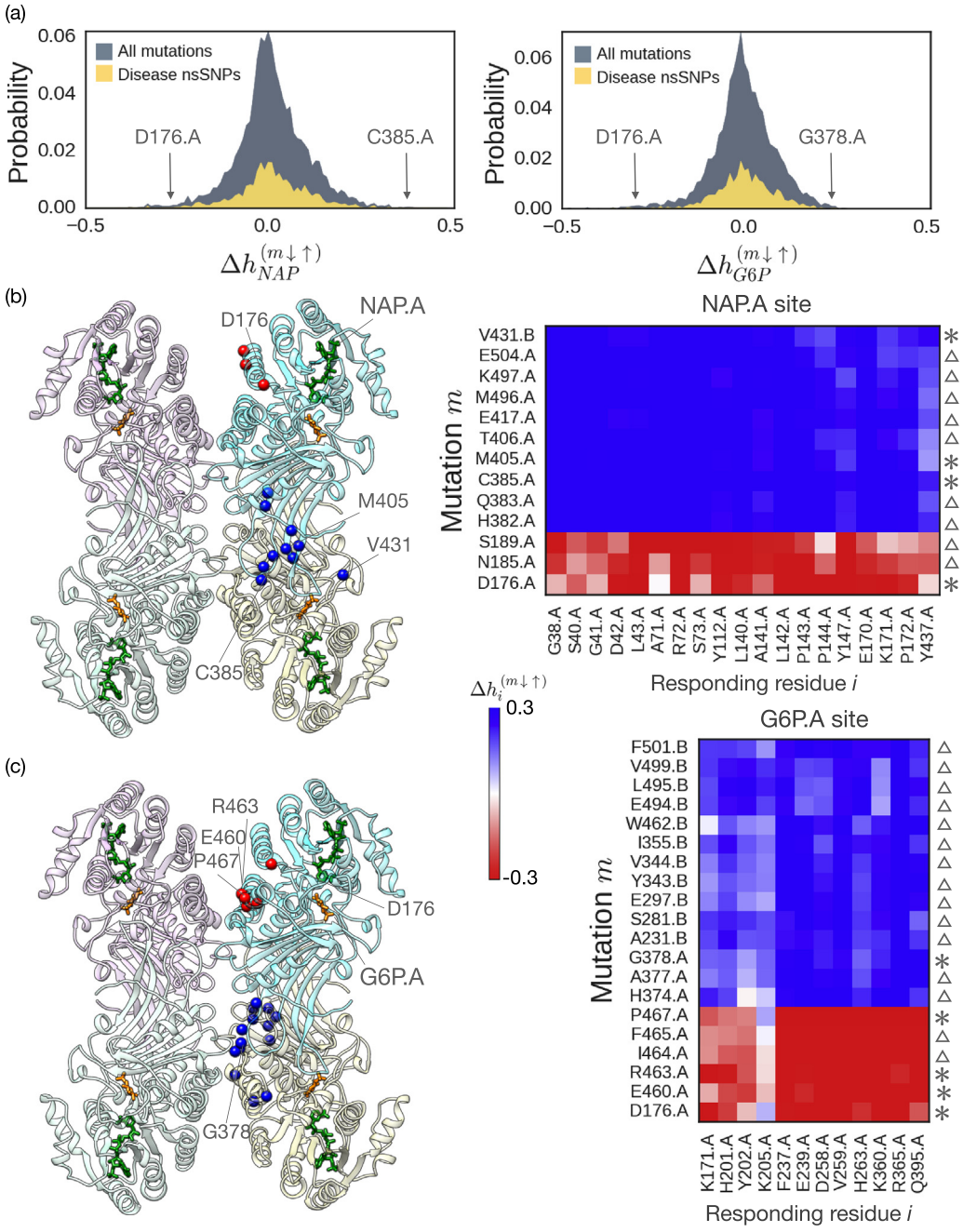

Fig. 3. Analysis of known nsSNPs associated with G6PD deficiency in glucose-6-phosphate dehydrogenase (G6PD) and detection of additional positions with potentially harmful allosteric effects. (a) $\Delta h_{\mathrm{NAP}}^{(m\downarrow\uparrow)}$ and $\Delta h_{\mathrm{G6P}}^{(m\downarrow\uparrow)}$ probability distributions obtained from the G6PD ASM (Fig. 1b). The NAP and G6P sites contain residues within a distance cutoff of 4.5 Å from the corresponding ligands in the crystal structure (PDB: 1qki). The fraction of disease-related residue positions that result in modulation ranges larger than one standard deviation (s= 0.09 kcal/mol) is 0.27. (b–c) Examples of residues mapped on the structure and relevant fragments of ASMs are presented (see also legend for Fig. 2). For clarity, only mutations in chains A and B are shown. Chains A, B, C, and D are colored in light cyan, yellow, green, and magenta, respectively. NAP cofactor is colored in green, and G6P ligand is colored in orange.

i due to a substitution at protein position m irrespective of the original amino acid, we characterize each affected protein position via the allosteric modulation range. We define the allosteric modulation range on residue *i* as the difference between the allosteric modulations induced by the

stabilizing (UP) and destabilizing (DOWN) mutations of the residue *m*

$$\Delta h_i^{(m\downarrow\uparrow)} = \Delta h_i^{(m\uparrow)} - \Delta h_i^{(m\downarrow)} \quad (6)$$

Performing an exhaustive mutational scanning, we obtain a complete data on the allosteric signaling expressed as the modulation range at residue *i* upon perturbation at residue *m*. These data are represented in the form of the ASM of the protein [14], which can be instrumental in quantifying the allosteric modulation of the protein activity, finding alternative ways of signaling, designing new allosteric sites and tuning them using additional mutations, and exploring "latent drivers" that expand the cancer mutational landscapes [19]. Figure 1a and b contains the ASMs for two case study proteins, human galactose 1-phosphate uridyltransferase (GALT) and glucose-6-phosphate dehydrogenase (G6PD), analyzed in this work.

Exhaustive analysis of potential allosteric effect of mutations

In order to explore a potential role of allosteric mechanisms in the action of pathological nsSNPs, we analyzed a set of 27 proteins (Supplementary Table 1) with at least 50 pathological nsSNPs mapped on the sequence and structure of each protein. On the basis of ASMs obtained for these proteins [see examples of GALT (chart a) and G6PD (b) ASMs in Fig. 1], we derived the distribution of modulation ranges $\Delta h^{(m\downarrow\uparrow)}$ observed on their functional sites as a result of the exhaustive single-residue mutagenesis (single-residue perturbations, Fig. 1c). The probability distribution of $\Delta h^{(m\downarrow\uparrow)}$ values in the protein set (Fig. 1c) shows that effect of mutations in distal amino acid positions (more than 11 Å from affected residue [21]) including disease-causing nsSNPs can vary in magnitude and sign of the modulation (see also $\Delta h_{\text{site}}^{(m\downarrow\uparrow)}$ distributions for individual proteins in Supplementary Fig. 1), and a significant proportion of distal perturbations can cause relatively strong allosteric modulation on the protein functional sites. For example, fractions of positions with disease-related SNPs with allosteric modulation above one (s= 0.13 kcal/mol) and two standard deviations are 25% and 7% (yellow distribution in Fig. 1c), respectively. Distribution of the allosteric modulation ranges obtained for all positions in analyzed proteins (gray distribution, Fig. 1c) yields similar fractions of residues that produce corresponding allosteric effect: 23% for the modulation above one standard deviation (s) and 6% for modulations above 2s. Noteworthy, the ratio of the number of positions with pathological SNPs (yellow distribution, Fig. 1c) that strongly modulate protein function to the number of all considered mutations with corresponding allosteric effects (gray distribution, Fig. 1c) increases with the strength of the modulation (Fig. 1d). These ratios are about 0.13 and 0.15 for the modulation strengths above s and 2s, respectively (Fig. 1d). At the same time, the comparison between distributions of allosteric modulation caused by SNPs and all residues (yellow and gray distributions, respectively, Fig. 1c) shows that, in addition to known SNPs, there are many protein positions that can originate strong allosteric modulation on the functional sites of corresponding proteins upon mutations. Therefore, comprehensive analysis should be performed in order to obtain sets of these residues, to explore their potential involvement in diseases characterized by known SNPs with similar allosteric modulation, and to search for new, yet undetected or overlooked mutations in these positions that may cause harmful allosteric effects. Below, we analyze in detail strongly modulating nsSNPs in two case studies of the human galactose 1-phosphate uridyltransferase (GALT) and glucose-6-phosphate dehydrogenase (G6PD), following up with exploring other positions that may originate comparable allosteric modulation in these proteins.

Human galactose 1-phosphate uridyltransferase (GALT)

Type I galactosemia or classic galactosemia is caused by toxic accumulation of galactose 1-phosphate in blood as a result of the impaired galactose metabolism, resulting in cognitive impairment and poor bone health in most patients [25]. The disease is triggered by severely diminished GALT activity, which is responsible for the reversible conversion of galactose 1-phosphate and uridine diphosphate glucose (UDP-Glc) to glucose 1-phosphate (G1P) and UDP galactose [26]. The human GALT is an obligate dimer with the active sites situated in the interface of its subunits (PDB: 5in3, [27]). Experimental and clinical data contain a large number of disease-causing nsSNPs, 97 of which can be mapped to the protein structure. While the most frequently observed Q188R mutation that severely hampers uridylylation of the protein [27] is in the catalytic site, the molecular mechanisms of protein dysfunction due to other deleterious nsSNPs are largely unknown.

Using the ASM of GALT (Fig. 1a), we obtained $\Delta h_{\text{UMP}}^{(m\downarrow\uparrow)}$ and $\Delta h_{\text{G1P}}^{(m\downarrow\uparrow)}$ distributions of the allosteric effects (expressed via modulation range) on residues of the UMP (uridine monophosphate) and G1P sites upon all distal mutations (Fig. 2a). The probability distributions of modulation ranges show that similar to the general observation on the basis of corresponding distributions for 27 proteins (Fig. 1c), amino acid substitutions can induce allosteric response of various signs and strength in the UMP and G1P sites of GALT (Fig. 2a). For example,

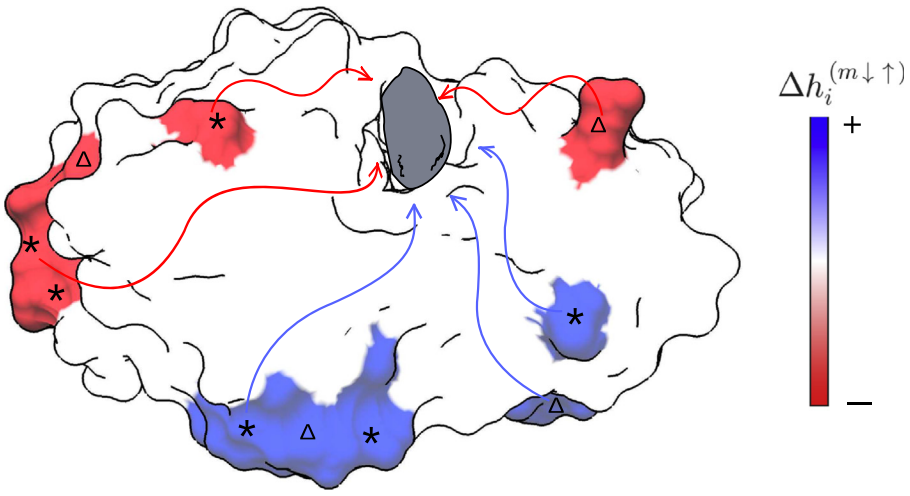

Fig. 4. Illustration of the notion of allosteric polymorphism. An illustration of allosteric polymorphism shows that mutations at different locations in a protein can exert similar allosteric modulation in the functional site. Amino acid positions with known pathological effects upon mutation are marked by asterisks, whereas those that have yet to be characterized are indicated by triangles. The functional site of the protein is indicated by a bound ligand in gray. One can consider residues that can induce allosteric modulation similar to one from the protein positions with known SNPs as latent allosteric triggers.

positions Asp28 and Glu203 in chain B cause negative and positive modulations in the G1P site of the chain A ($\Delta h_{G1P}^{(Asp28\downarrow\uparrow)} = -0.18$ kcal/mol and $\Delta h_{G1P}^{(Glu203\downarrow\uparrow)} = 0.33$ kcal/mol), respectively. On the other hand, positions Asp28 and Arg148 in chain A show modulation ranges of 0.24 and −0.36 kcal/mol in the UMP site of the same monomer (chain A). The allosteric signaling from these positions is confirmed by the experimentally observed very low enzymatic activity of GALT as a result of pathological mutations Asp28His, Asp28Tyr, Arg148Gly, Arg148Gln, Arg148Trp, and Glu203Lys found in patients with galactosemia.

Mapping of strongly modulating amino acids from the distributions to the structure reveals two clusters of residues—on the α1 helix in the 2-layer sandwich well separated from the UMP and G1P sites, and near the N-terminal end of the α2 helix (Supplementary Fig. 2). This observation is consistent with patterns of signaling in the GALT ASM (Fig. 1a), which suggest that mutations of amino acids close to positions of SNPs can cause similar allosteric signaling. Figure 2b and c illustrates two examples of allosteric signaling to G1P and UMP sites of chain A, showing the GALT structure with marked positions of mutated residues (left) and fragments of ASM with modulation ranges ($\Delta h_i^{(m\downarrow\uparrow)}$) obtained on the residues of responding UMP and G1P sites. Amino acid substitutions associated with known pathological nsSNPs are marked by asterisks, whereas those with uncharacterized mutational effect are marked by triangles. We observed that positions 122, 156, 159, and 160 in chain A cause a negative modulation on most of the responding residues of G1P site of chain A (note that the site is actually located in the interface, containing residues from both chains), similar to the modulation yielded by SNP-containing position 28 in chain B (Fig. 2b). At the same time, positions 200, 205, and 206 can induce in most of the responding positions a positive modulation, similar to the one caused by SNP-positions 195, 203, and 204 in chain B. In the case of UMP site in chain A, positions 28, 34, and 258 with documented pathological mutations involved in galactosemia yield positive modulation ranges on the site (e.g., $\Delta h_{UMP}^{(Asp28\downarrow\uparrow)} = 0.24$ kcal/mol), whereas galactosemia-associated position 148—negative modulation ($\Delta h_{UMP}^{(Arg148\downarrow\uparrow)} = -0.36$ kcal/mol).The GALT ASM also reveals additional residues that can elicit modulation comparable to that caused by known galactosemia nsSNPs: positive caused by positions 29, 37, 235–239, 352, 354, and 357; negative—by positions 144 and 160 (Fig. 2c). These observations suggest a potential deleterious effect of mutations in above positions in addition to known SNPs. Because of the GALT dimer symmetry, the allosteric modulation on the G1P and UMP sites in chain B is similar to that in chain A (Fig. 1a).

Human glucose-6-phosphate dehydrogenase (G6PD)

The G6PD is a key enzyme in the pentose–phosphate pathway, which catalyzes the rate-limiting step of the glucose-6-phosphate (G6P) oxidation and the concomitant reduction of $NADP^+$ cofactor to NADPH. The G6PD deficiency, which affects approximately 7% of the world population

[28], can lead to hemolytic anemia after an acute oxidative stress. The severity of G6PD deficiency ranges from the most acute class I, characterized by chronic hemolytic anemia and <10% of G6PD activity *in vitro*, to class IV with above 60% activity and no clinical manifestations [29]. Despite more than 160 disease-causing nsSNPs identified to date [30], the underlying mechanisms of their pathological effects remain unknown in most of the cases.

We obtained the $\Delta h_{G6P}^{(m\downarrow\uparrow)}$ and $\Delta h_{NAP}^{(m\downarrow\uparrow)}$ distributions for the substrate G6P and the cofactor $NADP^+$ binding sites (Fig. 3a) from the ASM (Fig. 1b) of G6PD (PDB:1qki). Similar to GALT, the probability distributions of modulation ranges reveal a number of amino acid substitutions that can produce significant allosteric signal on the functional sites of G6PD (Supplementary Fig. 3). For instance, mutation of Asp176 in chain A leads to a negative modulation on the cofactor (NAP) and the substrate (G6P) binding sites in the monomer ($\Delta h_{NAP}^{(Asp176\downarrow\uparrow)} = -0.27$ kcal/mol and $\Delta h_{G6P}^{(Asp176\downarrow\uparrow)} = -0.31$ kcal/mol, Fig. 3a). On the other hand, amino acid substitution at Cys385 results in a positive modulation on the cofactor site ($\Delta h_{NAP}^{(Cys385\downarrow\uparrow)} = 0.34$ kcal/mol), whereas mutation at Gly378 can modulate the G6P site with an increase of configurational work ($\Delta h_{G6P}^{(Gly378\downarrow\uparrow)} = 0.23$ kcal/mol). The allosteric modulation observed for these mutations is consistent with a drastic decline in the enzymatic activity caused by mutations Asp176Gly, Gly378Ser, Cys385Arg, Cys385Phe, and Cys385Trp in patients with symptoms of the class I G6PD deficiency. Mutations Met405Ile and Val431Gly are associated with pathological nsSNPs responsible for class I G6PD deficiency, while Val431Met is known to cause class II G6PD deficiency (Fig. 3b). The modulation range of positions 385, 405, and 431 clearly indicates their potential in positive allosteric modulation of the cofactor-binding site as the basis for the G6PD deficiency. In addition to Asp176Gly, mutations Glu460Asp, Arg463Ser, Arg463Cys, Arg463His, and Pro467Arg cause the class II and III G6PD deficiency. According to ASM data, these positions hosting pathological mutations at the tetramer interface affect the protein activity via negative allosteric modulation on the substrate binding site. Several other positions, for example, 464 and 465, can also cause negative modulation in the G6P site in chain A (Fig. 3c). Overall, the G6PD ASM reveals a number of amino acid positions (Fig. 3b, c), whose mutations can result in the negative/positive allosterically induced work, similar to those originated by positions with known pathological SNPs. These protein positions are spread across different monomers in the tetrameric structure, and residue substitutions at these positions may lead to allosteric modulation on the cofactor-/substrate-binding sites of G6PD (Fig. 3b, c).

Definition and implications of the allosteric polymorphism

Figures 2 and 3 contain several examples of the allosteric modulation ($\Delta h_i^{(m\downarrow\uparrow)}$) caused by SNP-containing positions, similar to that observed for other protein positions upon perturbation (UP/DOWN mutations). One can also consider position Leu37 in GALT, yielding uniformly large positive modulation ($\Delta h_i^{(m\downarrow\uparrow)}$) in all residues of the UMP site of the same monomer. As allosteric modulation range produced by position 37 is very similar to the one originated by position 28 that causes galactosemia upon substitution to histidine or tyrosine, it can be suggestive of the potential pathological development by mutating Leu37 (Fig. 2c). Similarly, mutations at His374 and Ala377 in the G6PD can potentially cause the class I G6PD deficiency as they result in the almost identical $\Delta h_i^{(m\downarrow\uparrow)}$ in the G6P site as the pathological Gly378Ser mutation (Fig. 3c).

We found that positions in GALT and G6PD with disease-causing SNPs comprise 26% and 13% of all positions (Figs. 2a and 3a) that originate allosteric modulation range with the strength above one standard deviation in corresponding distributions. At the same time, our data show that there is always a number of protein positions that can produce an allosteric modulation on the residues of protein functional sites similar to that observed from positions with known disease-causing SNPs. Specifically, Figure 1c and d shows that only 13% and 15% of positions that produce allosteric modulation range above one and two standard deviations, respectively, are known to contain disease-causing SNPs, leaving majority of strongly modulating positions as candidates for being allosteric modulators upon certain mutations. One can infer therefore a phenomenon of the allosteric polymorphism (Fig. 4), which is based on a number of residue positions in the protein structure, whose mutations may lead to the allosteric modulation/disruption of the protein function. As a consequence, pathological developments similar to those caused by known allosterically acting nsSNPs or with different clinical phenotypes can be triggered. The allosteric polymorphism is rather independent or complementary to the nsSNP, which considers an effect of the amino acid substitutions in a certain protein position determined by the non-synonymous nucleotide substitutions. The fundamental difference lies in the very mechanism of the protein function modulation, which is the result of the work produced on the protein functional sites because of the variation in the protein conformational ensemble. In case of the allosteric polymorphism, distal (allosteric) mutations of different nature, including nsSNPs, play a role of the perturbation that initiates a change in the protein activity.

Discussion

Rapidly accumulating protein structural data have been immensely helpful in linking disease-related nsSNPs to protein dysfunctions that result in different pathological developments. Among traditionally considered scenarios of the damaging action of mutations are alteration of the catalytic/binding function by removing the active/binding site residues, aberrations in protein–protein interactions, and change of the overall protein stability. The mutagenesis experiments coupled with functional assays have been the traditional method for investigating a small number of mutations suspected to be harmful for the protein function. High-throughput mutagenesis and functional assays show that some of these deleterious mutations occur at a distance from the functional sites, suggesting that they may affect the protein activity allosterically. Recent advances in understanding of allostery allow one to investigate its molecular mechanisms at per-residue resolution in a wide diversity of allosteric systems. It becomes especially important because of the mounting evidences of the role of allosteric effects of mutations, which were shown to be involved in cancerogenesis [18,19], gain- or loss-of function in GPCRs, and regulation in ligand- and voltage-gated channels [31], to name a few. Allosteric effects caused by single-residue mutations were instrumental in activation of insulin-degrading enzyme against amyloid Aβ peptide [14,15].

Using our computational framework for the comprehensive analysis of allosteric effects of mutations, we asked here two questions: (i) Is there an allosteric component in the action of disease-causing SNPs [14,16]? and (ii) Can some protein positions work as "latent allosteric triggers," which may affect widely defined function of the protein by allosteric modulation on corresponding sites (binding, catalytic, post-translational modification, etc.) upon certain mutations? We performed here an exhaustive scanning of allosteric effects of mutations and obtained ASMs for 27 proteins with a large number (more than 50 in each protein) annotated disease-causing nsSNPs. The modulation produced on functional sites, of which majority are catalytic or substrate/cofactor-binding sites with only two representing protein–protein interactions (Supplementary Table S1), varies in strength and can be of negative or positive value, indicating that work applied to regulated sites can prevent or produce their conformational changes. We found that many positions that host disease-causing nsSNPs produce strong allosteric modulation on the protein functional sites (Fig. 1c). Moreover, the ratio between the SNP-containing and all positions yielding the same allosteric modulation increases with the increase of the modulation value (Fig. 1d). Two case studies of human galactose 1-phosphate uridyltransferase (GALT) and glucose-6-phosphate dehydrogenase (G6PD) with many well-documented disease-causing SNPs illustrate an apparent relevance of allosteric mechanisms to the action of pathological nsSNPs that strongly affect their functional sites, resulting in a drastic reduce of GALT catalytic activity *in vitro* and severe class I galactosemia in case of G6PD mutations.

The analysis of ASMs also reveals that in addition to protein positions with annotated pathological SNPs, there are many other residues producing similar allosteric modulation on the protein activity. We therefore proposed here a notion of the allosteric polymorphism, according to which a variety of residues in a protein may allosterically modulate its function. The allosteric polymorphism is based on the change in the protein conformational ensemble caused by the different distal (allosteric) mutations, among which nsSNPs can also be present. By the analogy with the function of so-called latent drivers in cancerogenesis [19], allosteric polymorphism hints on the potential presence of "latent allosteric triggers" that might be harmful for the protein function upon certain mutations.

To conclude, apparently sheer number of residues that can allosterically modulate protein activity prompts us to develop a computational framework for investigating the regulation of protein function via allosteric mutations. We propose to use comprehensive ASMs obtained from the exhaustive mutational scanning, which allow to identify the latent pathological mutations in addition to the nsSNPs that affect the protein function via allosteric mechanisms. ASMs can be also instrumental in delineating the so-called latent drivers that are potential sources of expanding the cancer landscape and in investigating possible transformation of combinations of passenger mutations into the drivers of different diseases. Moving from understanding the allosteric basis of loss-of-function mutations, the ASMs can be also used in detecting the compensatory mutations that would neutralize the damage from the disease-causing ones and/or can be used for engineering new gain-of-function mutations that would induce the required strength and mode of allosteric signaling to any site of interest in a protein in order to change its activity.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2019.07.012.

Acknowledgments

This work was funded by the Biomedical Research Council (BMRC) of the Agency for Science Technology, and Research (A*STAR), Singapore.

Received 27 March 2019;
Received in revised form 11 June 2019;
Accepted 4 July 2019
Available online 12 July 2019

Abbreviations:
nsSNP, non-synonymous single-nucleotide polymorphism; SBSMMA, structure-based statistical mechanical model of allostery; ASM, Allosteric Signaling Map.

References

[1] J.K. Teer, J.C. Mullikin, Exome sequencing: the sweet spot before whole genomes, Hum. Mol. Genet. 19 (2010) R145–R151.
[2] B. Rabbani, M. Tekin, N. Mahdieh, The promise of whole-exome sequencing in medical genetics, J. Hum. Genet. 59 (2014) 5–15.
[3] D. Botstein, N. Risch, Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease, Nat. Genet. 33 (2003) 228–237 Suppl.
[4] P.C. Ng, S. Levy, J. Huang, T.B. Stockwell, B.P. Walenz, K. Li, et al., Genetic variation in an individual human exome, PLoS Genet. 4 (2008), e1000160.
[5] N. Agrawal, M.J. Frederick, C.R. Pickering, C. Bettegowda, K. Chang, R.J. Li, et al., Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1, Science. 333 (2011) 1154–1157.
[6] C. Bettegowda, N. Agrawal, Y. Jiao, M. Sausen, L.D. Wood, R.H. Hruban, et al., Mutations in CIC and FUBP1 contribute to human oligodendroglioma, Science. 333 (2011) 1453–1455.
[7] G. Guo, J. Chmielecki, C. Goparaju, A. Heguy, I. Dolgalev, M. Carbone, et al., Whole-exome sequencing reveals frequent genetic alterations in BAP1, NF2, CDKN2A, and CUL1 in malignant pleural mesothelioma, Cancer Res. 75 (2015) 264–269.
[8] S.B. Seidelmann, E. Smith, L. Subrahmanyan, D. Dykas, M. D. Abou Ziki, B. Azari, et al., Application of whole exome sequencing in the clinical diagnosis and management of inherited cardiovascular diseases in adults, Circ. Cardiovasc. Genet. 10 (2017).
[9] S. Gupta, S. Chatterjee, A. Mukherjee, M. Mutsuddi, Whole exome sequencing: uncovering causal genetic variants for ocular diseases, Exp. Eye Res. 164 (2017) 139–150.
[10] S. Stefl, H. Nishi, M. Petukh, A.R. Panchenko, E. Alexov, Molecular mechanisms of disease-causing missense mutations, J. Mol. Biol. 425 (2013) 3919–3936.
[11] P. Katsonis, A. Koire, S.J. Wilson, T.K. Hsu, R.C. Lua, A.D. Wilkins, et al., Single nucleotide variations: biological impact and theoretical interpretation, Protein Sci. 23 (2014) 1650–1666.
[12] R. Nussinov, H. Jang, C.J. Tsai, F. Cheng, Review: precision medicine and driver mutations: computational methods, functional assays and conformational principles for interpreting cancer drivers, PLoS Comput. Biol. 15 (2019), e1006658.
[13] I.N. Berezovsky, E. Guarnera, Z. Zheng, B. Eisenhaber, F. Eisenhaber, Protein function machinery: from basic structural units to modulation of activity, Curr. Opin. Struct. Biol. 42 (2017) 67–74.
[14] E. Guarnera, I.N. Berezovsky, Toward comprehensive allosteric control over protein activity, Structure. 27 (2019) 866–878.
[15] I.V. Kurochkin, E. Guarnera, J.H. Wong, F. Eisenhaber, I.N. Berezovsky, Toward allosterically increased catalytic activity of insulin-degrading enzyme against amyloid peptides, Biochemistry. 56 (2017) 228–239.
[16] E. Guarnera, I.N. Berezovsky, On the perturbation nature of allostery: sites, mutations, and signal modulation, Curr. Opin. Struct. Biol. 56 (2019) 18–27.
[17] M.R. Williams, S.J. Lehman, J.C. Tardiff, S.D. Schwartz, Atomic resolution probe for allostery in the regulatory thin filament, Proc. Natl. Acad. Sci. U. S. A. 113 (2016) 3257–3262.
[18] M. Li, S.C. Kales, K. Ma, B.A. Shoemaker, J. Crespo-Barreto, A.L. Cangelosi, et al., Balancing protein stability and activity in cancer: a new approach for identifying driver mutations affecting CBL ubiquitin ligase activation, Cancer Res. 76 (2016) 561–571.
[19] R. Nussinov, C.J. Tsai, Latent drivers' expand the cancer mutational landscape, Curr. Opin. Struct. Biol. 32 (2015) 25–32.
[20] E. Guarnera, I.N. Berezovsky, Structure-based statistical mechanical model accounts for the causality and energetics of allosteric communication, PLoS Comput. Biol. 12 (2016), e1004678.
[21] W.V. Tee, E. Guarnera, I.N. Berezovsky, Reversing allosteric communication: from detecting allosteric sites to inducing and tuning targeted allosteric response, PLoS Comput. Biol. 14 (2018), e1006228.
[22] E. Guarnera, Z.W. Tan, Z. Zheng, I.N. Berezovsky, AlloSigMA: allosteric signaling and mutation analysis server, Bioinformatics. 33 (2017) 3996–3998.
[23] Tan ZW, Tee WV, Guarnera E, Booth L, Berezovsky IN. AlloMAPS: allosteric mutation analysis and polymorphism of signaling database. Nucleic Acids Res. 2018;47:D265-D70.
[24] K. Hinsen, Analysis of domain motions by approximate normal mode calculations, Proteins. 33 (1998) 417–429.
[25] A.I. Coelho, M.E. Rubio-Gozalbo, J.B. Vicente, I. Rivera, Sweet and sour: an update on classic galactosemia, J. Inherit. Metab. Dis. 40 (2017) 325–342.
[26] L.J. Elsas 2nd, K. Lai, The molecular biology of galactosemia, Genet Med. 1 (1998) 40–48.
[27] T.J. McCorvie, J. Kopec, A.L. Pey, F. Fitzpatrick, D. Patel, R. Chalk, et al., Molecular basis of classic galactosemia from the structure of human galactose 1-phosphate uridylyltransferase, Hum. Mol. Genet. 25 (2016) 2234–2244.
[28] T. Vulliamy, P. Mason, L. Luzzatto, The molecular basis of glucose-6-phosphate dehydrogenase deficiency, Trends Genet. 8 (1992) 138–143.
[29] Standardization of procedures for the study of glucose-6-phosphate dehydrogenase. Report of a WHO Scientific Group, World Health Organ Tech Rep Ser. 366 (1967) 1–53.
[30] A. Minucci, K. Moradkhani, M.J. Hwang, C. Zuppi, B. Giardina, E. Capoluongo, Glucose-6-phosphate dehydrogenase (G6PD) mutations database: review of the "old" and update of the new mutations, Blood Cells Mol. Dis. 48 (2012) 154–165.
[31] J.P. Changeux, A. Christopoulos, Allosteric modulation as a unifying mechanism for receptor function and regulation, Cell. 166 (2016) 1084–1102.