

1 **Title:**

2 Identifying allosteric hotspots with dynamics: application to inter- and intra-species  
3 conservation

4  
5 **Authors & associated information:**

6 Declan Clarke<sup>a,1</sup>, Anurag Sethi<sup>b,c,1</sup>, Shantao Li<sup>b,d</sup>, Sushant Kumar<sup>b,e</sup>, Richard W.F.  
7 Chang<sup>e</sup>, Jieming Chen<sup>b,f</sup>, and Mark Gerstein<sup>b,c,d,2</sup>

8  
9 <sup>a</sup> Department of Chemistry, Yale University, 260/266 Whitney Avenue PO Box 208114,  
10 New Haven, CT 06520 USA

11 <sup>b</sup> Program in Computational Biology and Bioinformatics, Yale University, 260/266  
12 Whitney Avenue PO Box 208114, New Haven, CT 06520, USA

13 <sup>c</sup> Department of Molecular Biophysics and Biochemistry, Yale University, 260/266  
14 Whitney Avenue PO Box 208114, New Haven, CT 06520, USA

15 <sup>d</sup> Department of Computer Science, Yale University, 260/266 Whitney Avenue PO Box  
16 208114, New Haven, CT 06520, USA

17 <sup>e</sup> Yale College, 260/266 Whitney Avenue PO Box 208114, New Haven, CT 06520, USA

18 <sup>f</sup> Integrated Graduate Program in Physical and Engineering Biology, Yale University,  
19 260/266 Whitney Avenue PO Box 208114, New Haven, CT 06520, USA

20  
21 <sup>1</sup> D.C. and A.S. contributed equally to this work.

22 <sup>2</sup> Correspondence should be addressed to M.G. ([pi@gersteinlab.org](mailto:pi@gersteinlab.org)),

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** Predicting Allosteric Hotspots Using Dynamics-Based Formalisms with Sequence Analyses Across Diverse Evolutionary Timescales - ... [1]

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** <sup>a</sup>Department

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** <sup>b</sup>Program

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** <sup>c</sup>Department

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** <sup>d</sup>Department

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** <sup>e</sup>Yale

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** <sup>f</sup>Integrated

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** ([pi@gersteinlab.org](mailto:pi@gersteinlab.org))

DECLAN CLARKE 12/13/15 2:56 PM

**Formatted:** Line spacing: 1.5 lines

54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65  
66  
67  
68  
69  
70  
71  
72  
73  
74  
75  
76

# ABSTRACT

The rapidly growing volume of data being produced by next-generation sequencing initiatives is enabling more in-depth analyses of conservation than previously possible. Deep sequencing is uncovering disease loci and regions under selective constraint, despite the fact that intuitive biophysical reasons for such constraint are sometimes ~~missing~~. Allostery may often provide the missing explanatory link. We use models of protein conformational change to identify allosteric residues by ~~finding~~ essential surface cavities ~~and~~ information flow bottlenecks, and we develop a software tool (stress.molmovdb.org) that enables users to perform this analysis on their own proteins of interest. Though fundamentally 3D-structural in nature, ~~our analysis~~ is computationally fast, thereby allowing us to run it across the PDB and to evaluate general properties of predicted allosteric residues. ~~We find that these~~ tend to be conserved over ~~diverse~~ evolutionary time scales. ~~Finally, we~~ highlight examples ~~of~~ allosteric residues ~~that~~ help explain poorly understood disease-associated variants.

- DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** unavailable
- DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** predicting
- DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** or
- DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** this software
- DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** the
- DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** , which
- DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** long and short
- DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** We
- DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** in which
- DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** can

87  
88  
89  
90

# 91 INTRODUCTION

92 The ability to sequence large numbers of human genomes is providing a much  
93 deeper view into protein evolution than previously possible. When trying to understand  
94 the evolutionary pressures on a given protein, structural biologists now have at their  
95 disposal an unprecedented breadth of data regarding patterns of conservation, both across  
96 species and between humans. As such, there are greater opportunities to take an  
97 integrated view of the context in which a protein and its residues function. This view  
98 necessarily includes structural constraints such as residue packing, protein-protein  
99 interactions, and stability. However, deep sequencing is unearthing a class of conserved  
100 residues on which no obvious structural constraints appear to be acting. The missing link  
101 in understanding these regions may be provided by studying the protein's dynamic  
102 behavior through the lens of the distinct functional and conformational states within an  
103 ensemble.

104 The underlying energetic landscape responsible for the relative distributions of  
105 alternative conformations is dynamic in nature: allosteric signals or other external  
106 changes may reconfigure and reshape the landscape, thereby shifting the relative  
107 populations of states within an ensemble (Tsai *et al.*, 1999). Landscape theory thus  
108 provides the conceptual underpinnings necessary to describe how proteins change

DECLAN CLARKE 12/13/15 2:56 PM  
Deleted: .

DECLAN CLARKE 12/13/15 2:56 PM  
Deleted: amongst  
DECLAN CLARKE 12/13/15 2:56 PM  
Deleted: a more  
DECLAN CLARKE 12/13/15 2:56 PM  
Deleted: integrated

DECLAN CLARKE 12/13/15 2:56 PM  
Deleted: often  
DECLAN CLARKE 12/13/15 2:56 PM  
Deleted: considering  
DECLAN CLARKE 12/13/15 2:56 PM  
Deleted: and

DECLAN CLARKE 12/13/15 2:56 PM  
Formatted: Font:Italic, Check spelling and grammar

116 behavior and shape under changing conditions. A primary driving force behind the  
117 evolution of these landscapes is the need to efficiently regulate activity in response to  
118 changing cellular contexts, thereby making allostery and conformational change essential  
119 components of protein evolution.

120 | Given the importance of allosteric regulation, as well as its role in imparting  
121 | efficient functionality, several methods have been devised for the identification of likely  
122 | allosteric residues. Conservation itself has been used, either in the context of conserved  
123 | residues (Panjkovich and Daura, 2012), networks of co-evolving residues (Halabi *et al.*,  
124 | 2009; Lee *et al.*, 2008; Lockless *et al.*, 1999; Reynolds *et al.*, 2011; Shulman *et al.*, 2004;  
125 | Süel *et al.*, 2003), or local conservation in structure (Panjkovich and Daura, 2010). In  
126 | related studies, both conservation and geometric-based searches for allosteric sites have  
127 | been successfully applied to several systems (Capra *et al.*, 2009).

128 | The concept of ‘protein quakes’ has been introduced to explain local  
129 | conformational changes that are essential for global conformation transitions of  
130 | functional importance (Ansari *et al.*, 1985; Miyashita *et al.*, 2003). These local changes  
131 | cause strain within the protein that is relieved by subsequent relaxations (which are also  
132 | termed functionally important motions) that terminate when the protein reaches the  
133 | second equilibrium state. Such local perturbations often end with large conformational  
134 | changes at the focal points of allosteric regulation, and these motions may be identified in  
135 | a number of ways, including modified normal modes analysis (Miyashita *et al.*, 2003) or  
136 | time-resolved X-ray scattering (Arnlund *et al.*, 2014).

137 | In addition to conservation and geometry, protein dynamics have also been used  
138 | to predict allosteric residues. Normal modes analysis has been used to examine the extent

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** the

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** of allostery

DECLAN CLARKE 12/13/15 2:56 PM

**Formatted:** Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM

**Formatted:** Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM

**Formatted:** Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM

**Formatted:** Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM

**Formatted:** Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM

**Formatted:** Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM

**Formatted:** Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** A number of methods employing support vector machines have also been described (Huang and Schroeder, 2006; Huang *et al.*, 2013). Normal modes analysis, coupled with ligands of varying size, have

DECLAN CLARKE 12/13/15 2:56 PM

**Moved down [1]:** been used to examine the extent to which bound ligands interfere with low-frequency motions, thereby identifying potentially important residues at the surface (Ming and Wall, 2005; Mitternacht and Berezovsky, 2011; Panjkovich and Daura, 2012).

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** The concept of ‘protein quakes’ has been introduced to explain local regions of proteins that are essential for conformation transitions (Miyashita *et al.*, 2003). A protein may relieve the strain of a high-energy configuration by local structural changes. Such local changes often occur at the focal points of allosteric regulation, and these regions may be identified in a number of ways, including modified normal modes analysis (Miyashita *et al.*, 2003) or time-resolved X-ray scattering (Arnlund *et al.*, 2014).

DECLAN CLARKE 12/13/15 2:56 PM

**Moved (insertion) [1]**

165 | to which bound ligands interfere with low-frequency motions, thereby identifying  
166 | potentially important residues at the surface (Ming and Wall, 2005; Mitternacht and  
167 | Berezovsky, 2011; Panjkovich and Daura, 2012). Normal modes have also been used by  
168 | the Bahar group to identify important subunits that act in a coherent manner for specific  
169 | proteins (Chennubhotla and Bahar, 2006; Yang and Bahar, 2005). Rodgers *et al.* have  
170 | applied normal modes to identify key residues in CRP/FNR transcription factors  
171 | (Rodgers *et al.*, 2013).

172 | With the objective of identifying allosteric residues within the interior, molecular,  
173 | dynamics (MD) simulations and network analyses have been used to identify residues  
174 | that may function as internal allosteric bottlenecks (Csermely *et al.*, 2013; Gasper *et al.*,  
175 | 2012; Rousseau and Schymkowitz, 2005; Sethi *et al.*, 2009; Vanwart *et al.*, 2012). Ghosh  
176 | *et al.* (2008) have taken a novel approach of combining MD and network principles to  
177 | characterize allosterically important communication between domains in methionyl  
178 | tRNA synthetase. In conjunction with NMR, Rivalta *et al.* have use MD and network  
179 | analysis to identify important regions in imidazole glycerol phosphate synthase (Rivalta  
180 | *et al.*, 2012).

181 | Though having provided valuable insights, many of these approaches have been,  
182 | limited in terms of scale (the numbers of proteins which may feasibly be investigated),  
183 | computational demands, or the class of residues to which the method is tailored (surface  
184 | or interior). Here, we use, models of protein conformational change to, identify both  
185 | surface and interior residues that may act as essential allosteric hotspots, in a  
186 | computationally tractable manner, thereby enabling high-throughput analysis. This  
187 | framework directly incorporates information regarding 3D protein structure and

DECLAN CLARKE 12/13/15 2:56 PM  
**Formatted:** Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** Molecular

DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** interior

DECLAN CLARKE 12/13/15 2:56 PM  
**Formatted:** Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM  
**Formatted:** Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** Along similar lines,

DECLAN CLARKE 12/13/15 2:56 PM  
**Formatted:** Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM  
**Formatted:** Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM  
**Formatted:** Font:Italic

DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** inter-domain

DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** (Ghosh et al., 2008).

DECLAN CLARKE 12/13/15 2:56 PM  
**Formatted:** Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** may be

DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** Using

DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** , we

DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** regions

197 dynamics, and it can be applied on a PDB-wide scale to proteins that exhibit  
198 conformational change (Berman *et al.*, 2000). The residues identified tend to be  
199 conserved both across species and between humans, and they may help to elucidate many  
200 of the otherwise poorly understood regions in proteins. In a similar vein, several of our  
201 identified sites correspond to human disease loci for which no clear mechanism for  
202 pathogenesis had previously been proposed. Finally, we make the software associated  
203 with this framework (termed STRESS, for STRucturally-identified ESSential residues)  
204 publically available through a tool to enable users to submit their own structures for  
205 analysis.  
206

## 207 RESULTS

### 208 Identifying Potential Allosteric Residues

209 Allosteric residues at the surface generally play a regulatory role that is  
210 fundamentally distinct from that of allosteric residues within the protein interior. While  
211 surface residues may often constitute the sources or sinks of allosteric signals, interior  
212 residues act to transmit such signals. We use models of protein conformational change to  
213 identify both classes of residues (Figure 1). Throughout, we term these potential allosteric  
214 residues at the surface and interior “surface-critical” and “interior-critical” residues,  
215 respectively.

216 Critical residues are identified and analyzed within a set of 12 well-studied  
217 canonical systems (see Figure S1, as well as Table S1 for rationale), and they are then

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: is

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: throughout the PDB (Berman et al., 2000)

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: .

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: relatively greater conservation of the

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: (

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: amongst

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: )

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: our

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: is made

DECLAN CLARKE 12/13/15 2:56 PM

Formatted: Line spacing: 1.5 lines, Tabs: 2.3", Left

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: in an attempt

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: first

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: in

232 [investigated](#) on a large scale across hundreds of proteins [for which crystal structures of](#)  
233 [alternative conformations are available](#).

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: identified

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: distinct

### 235 Identifying Surface-Critical Residues

236 Allosteric ligands often act by binding to surface cavities and modulating protein  
237 conformational dynamics. The surface-critical residues, some of which may act as latent  
238 ligand binding sites and active sites, are first identified by finding cavities using Monte  
239 Carlo simulations to probe the surface with a flexible ligand (Figure 1A, top-left). The  
240 degree to which cavity occlusion by the ligand disrupts large-scale conformational  
241 change is used to assign a score to each cavity – sites at which ligand occlusion strongly  
242 [interferes](#) with conformational change earn high scores (Figure 1A, top-right), whereas  
243 shallow pockets (Figure 1A, bottom-left) or sites at which large-scale motions are largely  
244 unaffected (Figure 1A, bottom-right) earn lower scores. Further details are provided in SI  
245 Methods section 3.1-a.

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: interfere

246 This approach is a modified version of the binding leverage framework  
247 introduced by Mitternacht and Berezovsky (Mitternacht and Berezovsky, 2011). The  
248 main modifications [implemented here](#) include the use of heavy atoms in the protein  
249 during the Monte Carlo search, in addition to an automated means of thresholding the list  
250 of ranked scores. These modifications were implemented to provide a more selective set  
251 of sites; [without](#) them, an exceedingly large fraction of the protein surface would be  
252 captured (Figure 2C). [Within our dataset of proteins exhibiting alternative conformations,](#)  
253 [we](#) find that this modified approach results in an average of ~2 distinct sites per domain  
254 (Figure 2A; [see Figure 2B for the](#) distribution for distinct sites within entire complexes).

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: . Without

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: We

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: ). The

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: is given in Figure 2B.

262 Within the canonical set of 12 proteins, we positively identify an average of 56%  
263 of the sites known to be directly involved in ligand or substrate binding (see Table 1,  
264 Figure S1, and SI Methods section 3.1-a-iv). Some of the sites identified do not directly  
265 overlap with known binding regions, but we often find that these “false positives”  
266 nevertheless exhibit some degree of overlap with binding sites (Table S2). In addition,  
267 those surface-critical sites that do not match known binding sites may nevertheless  
268 correspond to latent allosteric regions: even if no known biological function is assigned  
269 to such regions, their occlusion may nevertheless disrupt **hitherto unfound large-scale**  
270 **motions** [[DC2MG(12/11): I actually don't know if I fully agree with this change that  
271 **was introduced: when we talk about latent allosteric sites, the thing this was previously**  
272 **unfound is not the motions themselves, but rather the pockets which were not previously**  
273 **known to disrupt already-known motions. We can discuss during P2 struct]]**

DECLAN CLARKE 12/13/15 2:56 PM

Formatted: Indent: First line: 0.5", Tabs: 3.56", Left

## 275 Dynamical Network Analysis to Identify Interior-Critical Residues

276 The binding leverage framework described above is intended to capture hotspot  
277 regions at the protein surface, but the Monte Carlo search employed is *a priori* excluded  
278 from the protein interior. Allosteric residues often act within the protein interior by  
279 functioning as essential **information flow** ‘bottlenecks’ within the communication  
280 pathways between **distal** regions.

281 To identify **such bottleneck residues**, the protein is first modeled as a network,  
282 wherein residues represent nodes and edges represent contacts between residues (in much  
283 the same way that the protein is modeled as a network in constructing anisotropic  
284 network models, see below). In this regard, the problem of identifying interior-critical  
285 residues is reduced to a problem of identifying nodes that participate in network

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: large-scale motions.

DECLAN CLARKE 12/13/15 2:56 PM

Formatted: Font:Arial, 11 pt, Bold

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: distal

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: An allosteric signal transmitted from one region to another may conceivably take various alternative routes, but many of these routes can share a common set of residues. The removal of such a common set of residues can result in the loss of many or all of the available routes for allosteric signal transmission, thereby making these residues essential information flow bottlenecks.

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: bottlenecks



298 bottlenecks (see Figure 1B and SI Methods section 3.1-b for details). Briefly, the network  
299 | edges are first weighted by the degree of strength in the correlated motions of contacting  
300 | residues: a strong correlation in the motion between contacting residues implies that  
301 | knowing how one residue moves better enables one to predict the motion of the other,  
302 | thereby suggesting a strong information flow between the two residues. The weights are  
303 | used to assign ‘effective distances’ between connecting nodes, with strong correlations  
304 | resulting in shorter effective node-node distances.

305 | Using the motion-weighted network, “communities” of nodes are identified using  
306 | the Girvan-Newman formalism (Girvan *et al.*, 2002). A community is a group of nodes  
307 | such that each node within the community is highly inter-connected, but loosely  
308 | connected to other nodes outside the community. Communities are thus densely inter-  
309 | connected regions within proteins. As tangible examples, the community partitions and  
310 | the resultant critical residues for the canonical set are given in Figures S2.

311 | Finally, the betweenness of each edge is calculated. The betweenness of an edge  
312 | is defined as the number of shortest paths between all pairs of residues that pass through  
313 | that edge, with each path representing the sum of effective node-node distances assigned  
314 | in the weighting scheme above. Those residues that are involved in the highest-  
315 | betweenness edges between pairs of interacting communities are identified as the  
316 | interior-critical residues. These residues are essential for information flow between  
317 | communities, as their removal would result in substantially longer paths between the  
318 | residues of one community to those of another.

319

320 **Software Tool: STRESS (STRucturally-identified ESSential residues)**

DECLAN CLARKE 12/13/15 2:56 PM

Formatted: Font:Italic, Check spelling and grammar

321 We have made the implementations for finding surface- and interior-critical  
322 residues available through a new software tool, STRESS, which may be accessed at  
323 stress.molmovdb.org (Figure 3A). Users may submit a PDB file or a PDB ID  
324 corresponding to a structure to be analyzed, and the output provided constitutes the set of  
325 identified critical residues.

326 Obviating the need for long wait times, the algorithmic implementation of our  
327 software is highly efficient (Figures 3B and 3C). Running times are minimized by using a  
328 scalable server architecture that runs on the Amazon cloud (Figure 3D). Relative to a  
329 naïve global Monte Carlo search implementation, local searches supported with hashing  
330 and additional algorithmic optimizations for computational efficiency also reduce  
331 running times considerably. A typical protein of ~500 residues takes only about 30  
332 minutes on a 2.6GHz CPU.

333 A light front-end server handles incoming user requests, and more powerful back-  
334 end servers, which perform the calculations, are automatically and dynamically scalable,  
335 thereby ensuring that they can handle varying levels of demand both efficiently and  
336 economically.

## 337

## 338 High-Throughput Identification of Alternative

## 339 Conformations

340 We use a generalized approach to systematically identify instances of alternative  
341 conformations throughout the PDB. We first perform multiple structure alignments  
342 (MSAs) across sequence-identical structures that are pre-filtered to ensure structural

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** The

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** both

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** have been made

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** to the scientific community

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** specify a PDB

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** A typical protein of ~500 residues takes only about 30 minutes on a 2.6GHz CPU.

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** also

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** Pronounced conformational change is an essential assumption within our framework for identifying potential allosteric residues.

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** within

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** proteins

357 quality. We then use the resultant pairwise RMSD values to infer distinct conformational  
358 states (Figure S3; see also SI Methods section 3.2).

359 The distributions of the resultant numbers of conformations for domains and  
360 chains are given in Figures S3D and S3E, respectively, and an overview is given in  
361 Figure S3F. We note that the alternative conformations identified arise in an extremely  
362 diverse set of biological contexts, including conformational transitions that accompany  
363 ligand binding, protein-protein or protein-nucleic acid interactions, post-translational  
364 modifications, changes in oxidation or oligomerization states, etc. The dataset of  
365 alternative conformations identified is provided as a resource in File S1 (see also Figure  
366 S3G).

## 367 | **Evaluating Conservation of Critical Residues**

### 369 **Using Various Metrics and Sources of Data**

370 The large dataset of dynamic proteins culled throughout the PDB, coupled with  
371 the high algorithmic efficiency of our critical residue search implementation, provide a  
372 means of evaluating general properties within the large pool of critical residues  
373 identified. In particular, we use a variety of conservation metrics and data sources to  
374 measure the inter- and intra-species conservation of the residues within this pool. As  
375 discussed below, we find that both surface- (Figures 4A-D) and interior-critical residues  
376 (Figures 4E-H) are consistently more conserved than non-critical residues. We emphasize  
377 that the signatures of conservation identified not only provide a means of rationalizing

DECLAN CLARKE 12/13/15 2:56 PM  
Formatted: Indent: First line: 0"

DECLAN CLARKE 12/13/15 2:56 PM  
Deleted: number

DECLAN CLARKE 12/13/15 2:56 PM  
Deleted: of these

DECLAN CLARKE 12/13/15 2:56 PM  
Deleted: on a large scale.

DECLAN CLARKE 12/13/15 2:56 PM  
Deleted: measure their conservation, as evaluated both over long (inter-species) and short (intra-human) evolutionary timescales. Using

DECLAN CLARKE 12/13/15 2:56 PM  
Deleted: data

DECLAN CLARKE 12/13/15 2:56 PM  
Deleted: critical

387 many of the otherwise poorly understood regions of proteins, but they also reinforce the  
388 functional importance of the residues predicted, to be allosteric.

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** believed

### 390 Conservation Across Species

391 When evaluating conservation across species, we find that both surface- and  
392 interior-critical residues tend to be significantly more conserved than non-critical residues

393 with the same degree of burial (Figures 4B and 4F, respectively; note that, negative

394 conservation scores designate stronger conservation – see SI Methods section 3.3-a).

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** ). Surface-critical residue sets have a mean conservation score (i.e., ConSurf score, see SI Methods section 3.3-a) of -0.131, whereas non-critical residue sets with the same degree of burial have a mean score of +0.059 ( $p < 2.2e-16$ ;

395

396 Leveraging Next-Generation Sequencing to Measure Conservation Between

### 397 Humans

398 In addition to measuring inter-species conservation, we have also used fully

399 sequenced human genomes and exomes to investigate conservation among human

400 populations, as many constraints may be species-specific and active in more recent

401 evolutionary history. Commonly, used metrics for quantifying intra-species, conservation

402 include minor allele frequency (MAF) and derived allele frequency (DAF). Low MAF or

403 DAF values are interpreted as signatures of deleteriousness, as purifying selection is

404 prone to reduce the frequencies of harmful variants (see SI Methods section 3.3-b).

405 Non-synonymous single-nucleotide variants (SNVs) from the, 1000 Genomes

406 dataset (McVean *et al.*, 2012) that hit surface-critical residues tend to occur at lower DAF

407 values (Figure 4C). Though this trend is not observed to be significant, the significance

408 improves when examining the shift in DAF distributions, as evaluated with a KS test ( $p=$

409 0.159, Figure S4A), and we point out the limited number of proteins (thirty-two) for,

410 which these 1000 Genomes SNVs coincide with surface-critical sites. Furthermore, the

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** ). Interior-critical residues exhibit a similar trend: the mean conservation score for interior-critical residues and non-critical residues with the same degree of burial is -0.179 and -0.102, respectively ( $p=3.67e-11$

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** Measures of Conservation Amongst Humans from

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** We may

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** use

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** human

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** In this context, commonly

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** evaluating

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** We find that

DECLAN CLARKE 12/13/15 2:56 PM

**Formatted:** Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** single-nucleotide variants (SNVs)

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** in

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** hit these

434 long tail extending to lower DAF values for surface-critical residues may suggest that  
435 only a subset of the residues in our prioritized binding sites is essential. In contrast to  
436 surface-critical residues, however, interior-critical residues are hit by 1000 Genomes  
437 SNVs with significantly lower DAF values than non-critical residues (Figure 4G; see also  
438 Figure S4B).

DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** However 1000 Genomes SNVs tend to hit

439 Given the limited number of proteins to be hit by 1000 Genomes SNVs, we also  
440 analyzed the larger dataset, provided by the Exome Aggregation Consortium (ExAC,  
441 Cambridge MA 2015). ExAC provides sequence data from more than 60,000 individuals,  
442 and samples are sequenced at much higher coverage, thereby ensuring better data quality.  
443 Using MAF as a conservation metric, we performed a similar analysis using this data.

DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** relatively small

DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** data

DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** for many

DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** the ExAC sequencing itself is performed

DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** . Thus, using

444 MAF distributions for surface- and non-critical residues in the same set of proteins are  
445 given in Figure 4D. Although the mean value of the MAF distribution for surface-critical  
446 residues is slightly higher than that of non-critical residues, the median for surface-  
447 critical residues is substantially lower than that for non-critical residues, demonstrating  
448 that the majority of proteins are such that MAF values are lower in surface- than in non-  
449 critical residues. In addition, the overall shifts of these distributions also point to a trend  
450 of lower MAF values in surface-critical residues (Figure S5A, KS test  $p=9.49e-2$ ).

451 Interior-critical residues exhibit significantly lower MAF values than do non-  
452 critical residues in the same set of proteins. MAF distributions for interior- and non-  
453 critical residues are given in Figure 4H (see also Figure S5B).

454 In addition to analyzing overall allele frequency distributions, we also evaluate  
455 the *fraction* of rare alleles as a metric for measuring selective pressure. This fraction is  
456 defined as the ratio of the number of rare (i.e., low-DAF or low-MAF) non-synonymous

DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** one may

DECLAN CLARKE 12/13/15 2:56 PM  
**Deleted:** (i.e., rare

467 SNVs to the number of all non-synonymous SNVs in a given protein annotation (such as  
468 all surface-critical residues of the protein, for example; see SI Methods section 3.3-b). A  
469 higher fraction is interpreted as a proxy for greater conservation (Khurana et al., 2013).  
470 Using variable DAF (MAF) cutoffs to define rarity for 1000 Genomes (ExAC) SNVs,  
471 both surface- and interior-critical residues are shown to harbor a higher fraction of rare  
472 alleles than do non-critical residues, further suggesting a greater degree of evolutionary  
473 constraint on critical residues (See Figure 5).

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: in

### 475 Comparisons Between Different Models of Protein Motions

476 The identification of surface- and interior-critical residues entails using sets of  
477 vectors (on each protein residue) to describe conformational change. Notably, our  
478 framework enables one to determine these vectors in multiple ways. Conformational  
479 changes may be modeled using vectors connecting residues in crystal structures from  
480 alternative conformations. We term this approach “ACT”, for “absolute conformational  
481 transitions” (see SI Methods section 3.2-c). The crystal structures of such paired  
482 conformations may be obtained using the framework discussed above. The protein  
483 motions may also be inferred from anisotropic network models (ANMs) (Atilgan *et al.*,  
484 2001). ANMs entail modeling interacting residues as nodes linked by flexible springs, in  
485 a manner similar to elastic network models (Fuglebakk *et al.*, 2015; Tirion, 1996) or  
486 normal modes analysis (Figure 1B). ANMs are not only simple and straightforward to  
487 apply on a database scale, but unlike using alternative crystal structures, the motion  
488 vectors inferred may be generated using a single structure.

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: pairs of corresponding

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: (we

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: ”;

DECLAN CLARKE 12/13/15 2:56 PM

Formatted: Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM

Formatted: Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: , and we thus use ANMs as our primary means of inferring motions

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: Using

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: give

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: results

489 Modeling conformational change using vectors from either ACTs or ANMs gives  
490 the same general trends in terms of the disparities in conservation between critical and

500 non-critical residues. Our framework is thus general with respect to how the motion  
501 vectors are obtained (see Figure 6 and SI Methods section 3.2-c for further details).

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** This method

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** defined

### 503 **Critical Residues in the Context of Human Disease Variants**

504 Directly related to conservation is confidence with which an SNV is believed to  
505 be disease-associated. SIFT (Ng and Henikoff, 2001) and PolyPhen (Adzhubei *et al.*,  
506 2010) are two tools for predicting SNV deleteriousness. ExAC SNVs hitting critical  
507 residues exhibit significantly higher PolyPhen scores relative to non-critical residues,  
508 suggesting the potentially higher disease susceptibility at critical residues (Figure S6).  
509 Significant disparities were not observed in SIFT scores (Figure S7).

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** the concept of SNV deleteriousness: changes in amino acid composition at specific loci may be more or less likely

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** result in

DECLAN CLARKE 12/13/15 2:56 PM

**Formatted:** Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** such effects, and we evaluated these predictions for critical and non-critical residues hit by SNVs in

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** .

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** ), though such significant

DECLAN CLARKE 12/13/15 2:56 PM

**Formatted:** Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM

**Formatted:** Font:Italic, Check spelling and grammar

510 Using HGMD (Stenson *et al.*, 2014) and ClinVar (Landrum *et al.*, 2014), we  
511 identify proteins with critical residues that coincide with disease-associated SNVs (Figure  
512 7A and File S2). Several critical residues coincide with known disease loci for which the  
513 mechanism of pathogenicity is otherwise unclear (File S3). The fibroblast growth factor  
514 receptor (FGFR) is a case-in-point (Figure 7). SNVs in FGFR have been linked to  
515 craniofacial defects. Dotted lines in Figure 7B highlight poorly understood disease SNVs  
516 that coincide with critical residues. In addition, we identify Y328 as a surface-critical  
517 residue, which coincides with a disease-associated SNV from HGMD, despite the lack of  
518 confident predictions of deleteriousness by several widely used tools for predicting  
519 disease-associated SNVs, including PolyPhen (Adzhubei *et al.*, 2010), SIFT (Ng and  
520 Henikoff, 2001), and SNPs&GO (Calabrese *et al.*, 2009). Together, these results suggest  
521 that the incorporation of surface- and interior-critical residues introduces a valuable layer  
522 of annotation to the protein sequence, and may help to explain otherwise poorly  
523 understood disease-associated SNVs.

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** HGDM

DECLAN CLARKE 12/13/15 2:56 PM

**Formatted:** Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM

**Formatted:** Font:Italic, Check spelling and grammar

537

538

## DISCUSSION & CONCLUSIONS

539

540

541

542

543

544

545

546

547

548

549

550

551

552

553

554

555

556

557

558

559

The same principles of energy landscape theory that dictate protein folding are integral to how proteins explore different conformations once they adopt their fully folded states. These landscapes are shaped not only by the protein sequence itself, but also by extrinsic conditions. Such external factors often regulate protein activity by introducing allosteric-induced changes, which ultimately reflect changes in the shapes and population distributions of the energetic landscape. In this regard, allostery provides an ideal platform from which to study protein behavior in the context of their energetic landscapes. To investigate allosteric regulation, and to simultaneously add an extra layer of annotation to conservation patterns, an integrated framework to identify potential allosteric residues is essential. We introduce a framework to select such residues, using knowledge of conformational change.

When applied to many proteins with distinct conformational changes in the PDB, we investigate the conservation of potential allosteric residues in both inter-species and intra-human genomes contexts, and find that these residues tend to exhibit greater conservation in both cases. In addition, we identify several disease-associated variants for which plausible mechanisms had been unknown, but for which allosteric mechanisms provide a plausible rationale.

Unlike the characterization of many other structural features, such as secondary structure assignment, residue burial, protein-protein interaction interfaces, disorder, and even stability, allostery inherently manifests through dynamic behavior. It is only by considering protein motions and changes in these motions can a fuller understanding of

DECLAN CLARKE 12/13/15 2:56 PM  
Formatted: Line spacing: 1.5 lines

DECLAN CLARKE 12/13/15 2:56 PM  
Deleted: in the context of

DECLAN CLARKE 12/13/15 2:56 PM  
Deleted: previously

DECLAN CLARKE 12/13/15 2:56 PM  
Deleted: unavailable

DECLAN CLARKE 12/13/15 2:56 PM  
Deleted: in the context of



564 allosteric regulation be realized. As such, MD and NMR are some of the most common  
565 means of studying allostery and dynamic behavior (Kornev and Taylor, 2015). However,  
566 these methods have limitations when studying large and diverse protein datasets. MD is  
567 computationally expensive and impractical when studying large numbers of proteins.  
568 NMR structure determination is extremely labor-intensive and better suited to certain  
569 classes of structures or dynamics. In addition, NMR structures constitute a relatively  
570 small fraction of structures currently available.

571 Despite these limitations in MD and NMR, allosteric mechanisms and signaling  
572 pathways may be conserved across many different but related proteins within the same  
573 family, suggesting that such computationally- or labor-intensive approaches for all  
574 proteins may not be entirely essential. Flock *et al.* have carefully demonstrated that the  
575 allosteric mechanisms responsible for regulating G proteins through GPCRs tend to be  
576 conserved (Flock *et al.*, 2015). Investigations into representative families have also been  
577 enlightening in other contexts. In one of the early studies employing network analysis,  
578 del Sol *et al.* conduct a detailed study of several allosteric protein families (including  
579 GPCRs) to demonstrate that residues important for maintaining the integrity of short  
580 paths within residue contact networks are essential to enabling signal transmission  
581 between distant sites (del Sol *et al.*, 2006). Another notable result in the same work is that  
582 these key residues (which match experimental results) may become redistributed when  
583 the protein undergoes conformational change, thereby changing optimal communication  
584 routes as a means of conferring different regulatory properties.

585 There are several notable implications of our dynamics-based analysis across a  
586 database of proteins. Relative to sequence data, allostery and dynamic behavior are far

DECLAN CLARKE 12/13/15 2:56 PM

Formatted: Font:Italic

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** If allosteric mechanisms are similarly shared within other protein families, a detailed analysis with methods such as MD or NMR on one member of a family may help to elucidate the allosteric behavior for other members. Nevertheless, the degree to which these mechanisms are indeed conserved within other groups of proteins is currently unclear, so homology-based predictions of allosteric mechanisms are still not readily available. .

DECLAN CLARKE 12/13/15 2:56 PM

Formatted: Font:Italic, Check spelling and grammar

DECLAN CLARKE 12/13/15 2:56 PM

Formatted: Font:Italic

DECLAN CLARKE 12/13/15 2:56 PM

Formatted: Font:Italic

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** .

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** Notably, many of the key sites identified correspond to residues that had been experimentally determined to be important for allostery.

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** in different conformations

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** -scale analysis.

604 more difficult to evaluate on a large scale. The framework described here enables one to  
605 evaluate dynamic behavior in a systemized and efficient way across many proteins, while  
606 simultaneously capturing residues on both the surface and within the interior. That this  
607 pipeline can be applied in a high-throughput manner enables the investigation of system-  
608 wide phenomena, such as the roles of potential allosteric hotspots in protein-protein  
609 interaction networks.

610 ~~It is only by analyzing a large dataset of proteins can one investigate general~~  
611 ~~trends in predicted allosteric residues. In addition, the implementation detailed here,~~  
612 enables one to match structural features with the high-throughput data generated through  
613 deep sequencing initiatives, which are providing an unprecedented window into  
614 conservation patterns, many of which may be human-specific.

615 We anticipate that, within the next decade, deep sequencing will enable structural  
616 biologists to study evolutionary conservation using sequenced human exomes just as  
617 routinely as cross-species alignments. Furthermore, intra-species metrics for conservation  
618 provide added value in that the confounding factors of cross-species comparisons are  
619 removed: different organisms evolve in different cellular and evolutionary contexts, and  
620 it can be difficult to decouple these different effects from one another. Cross-species  
621 metrics of protein conservation entail comparisons between proteins that may be very  
622 different in structure and function. Sequence-variable regions across species may not be  
623 conserved, but nevertheless impart essential functionality. Intra-species comparisons,  
624 however, can often provide a more direct and sensitive evaluation of constraint.

625 In particular, selective constraints within human populations are particularly  
626 relevant to understanding human disease. Formalisms for analyzing large structural and

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** Knowledge of such sites across many proteins may also be used to identify the best proteins and protein regions for which drugs should be engineered, as well as instances in which specific sequence variants are likely to have the greatest impact.

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** We emphasize that it

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** applying this framework over a database of many

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** search for significant disparities in conservation between sites believed to be important in allostery and the rest of the protein. Such

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** may not be apparent when studying a small number or specific classes of proteins. To our knowledge, this is the first study in which the conservation of potential allosteric sites has been measured across a large database of proteins. - ... [2]

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** Full human genomes and exomes are being sequenced at an increasing pace, thereby

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** that can be human-specific or active over short evolutionary timescales. These patterns increasingly serve as detailed signatures of selective constraints which may not only be missing in cross-species comparisons, but are also sometimes difficult to rationalize using static representations of protein structures alone

DECLAN CLARKE 12/13/15 2:56 PM

**Formatted:** Indent: First line: 0.5"

DECLAN CLARKE 12/13/15 2:56 PM

**Deleted:** addition, intra-species

659 sequence datasets will become increasingly important in the context of human health. We  
660 anticipate that the framework and formalisms detailed here, along with the accompanying  
661 web tool we have introduced, will help to further motivate future studies along these  
662 directions.

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: disease. Finally, we

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: our newly developed software

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: will prove to be of great value in enabling investigators to study allostery in diverse contexts. -

## 664 METHODS

DECLAN CLARKE 12/13/15 2:56 PM

Formatted: Line spacing: 1.5 lines

665 An overview of the framework for finding surface- and interior-critical residues is  
666 given in Figure 1. Figure S3 provides a schematic of our pipeline for identifying  
667 alternative conformations throughout the PDB. Cross-species conservation scores were  
668 analyzed in those PDBs for which full ConSurf files are available through the ConSurf  
669 server. 1000 Genomes SNVs were taken from the Phase 3 release, and ExAC SNVs were  
670 downloaded in May 2015. Further details on all protocols are provided in SI Methods.

DECLAN CLARKE 12/13/15 2:56 PM

Formatted: Indent: First line: 0.5"

## 671 672 ACKNOWLEDGMENTS

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: -

673 DC acknowledges the support of the NIH Predoctoral Program in Biophysics (T32  
674 GM008283-24). We thank Simon Mitternacht for sharing the original source code for  
675 binding leverage calculations, as well as Koon-Kiu Yan for helpful discussions and  
676 feedback. The authors would like to thank the Exome Aggregation Consortium and the  
677 groups that provided exome variant data for comparison. A full list of contributing groups  
678 can be found at <http://exac.broadinstitute.org/about>

DECLAN CLARKE 12/13/15 2:56 PM

Formatted: Line spacing: 1.5 lines

679

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: -

# REFERENCES

687

- 688 Adzhubei, I. Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P.,  
689 Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting  
690 damaging missense mutations. *Nat. Methods.* 7, 248–249
- 691 Arnlund, D., Johansson, L.C., Wickstrand, C., Barty, A., Williams, G.J., Malmerberg, E.,  
692 Davidsson, J., Milathianaki, D., DePonte, D.P., Shoeman, R.L., et al. (2014).  
693 Visualizing a protein quake with time-resolved X-ray scattering at a free-electron  
694 laser. *Nat. Methods.* 11, 923–6.
- 695 Atilgan, A.R., Durell, S.R., Jernigan, R.L., Demirel, M.C., Keskin, O., and Bahar, I.  
696 (2001). Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network  
697 Model. *Biophys. J.* 80, 505–515.
- 698 Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H.,  
699 Shindyalov, I.N. and Bourne, P.E. (2000). The Protein Data Bank. *Nucleic Acids*  
700 *Res.* 28, 235–242.
- 701 Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P.L. and Casadio, R. (2009).  
702 Functional annotations improve the predictive score of human disease-related  
703 mutations in proteins. *Hum. Mutat.* 30, 1237–1244.
- 704 Exome Aggregation Consortium (ExAC). (2015) Cambridge, MA.  
705 <http://exac.broadinstitute.org>.
- 706 Capra, J.A., Laskowski, R.A., Thornton, J.M., Singh, M. and Funkhouser, T.A. (2009).  
707 Predicting protein ligand binding sites by combining evolutionary sequence  
708 conservation and 3D structure. *PLoS Comput. Biol.* 5, e1000585.
- 709 Chennubhotla, C. and Bahar, I. (2006). Markov propagation of allosteric effects in

DECLAN CLARKE 12/13/15 2:56 PM

Formatted: Indent: Left: 0", Hanging:  
0.33", Line spacing: 1.5 lines

710 biomolecular systems: application to GroEL–GroES. *Mol. Syst. Biol.* 2.  
711 del Sol, A., Fujihashi, H., Amoros, D., and Nussinov, R. (2006). Residues crucial for  
712 maintaining short paths in network communication mediate signaling in proteins.  
713 *Mol. Syst. Biol.* 2(1).  
714 Csermely, P., Korcsmáros, T., Kiss, H.J.M., London, G., and Nussinov, R. (2013).  
715 Structure and dynamics of molecular networks: A novel paradigm of drug discovery.  
716 *Pharmacol. Ther.* 138, 333–408.  
717 Flock, T., Ravarani, C.N.J., Sun, D., Venkatakrishnan, A.J., Kayikci, M., Tate, C.G.,  
718 Veprintsev, D.B. and Babu, M.M. (2015). Universal allosteric mechanism for Gα  
719 activation by GPCRs. *Nature* 524, 173–179.  
720 Fuglebakk, E., Tiwari, S.P., and Reuter, N. (2015). Comparing the intrinsic dynamics of  
721 multiple protein structures using elastic network models. *Biochim. Biophys. Acta -*  
722 *Gen. Subj.* 1850, 911–922.  
723 Gasper, P.M., Fuglestad, B., Komives, E.A., Markwick, P.R.L., and McCammon, J.A.  
724 (2012). Allosteric networks in thrombin distinguish procoagulant vs. anticoagulant  
725 activities. *Proc. Natl. Acad. Sci. U. S. A.* 109, 21216–22.  
726 Ghosh, A., and Vishveshwara, S. (2008). Variations in Clique and Community Patterns in  
727 Protein Structures during Allosteric Communication: Investigation of Dynamically  
728 Equilibrated Structures of Methionyl tRNA Synthetase Complexes. *Biochemistry.*  
729 47, 11398-11407.  
730 Girvan, M., Girvan, M., Newman, M.E.J., and Newman, M.E.J. (2002). Community  
731 structure in social and biological networks. *Proc. Natl. Acad. Sci. U. S. A.* 99, 7821–  
732 7826.

733 Halabi, N., Rivoire, O., Leibler, S., and Ranganathan, R. (2009). Protein Sectors:  
734 Evolutionary Units of Three-Dimensional Structure. *Cell* 138, 774–786.

735 Huang, B., and Schroeder, M. (2006). LIGSITEcsc: predicting ligand binding sites using  
736 the Connolly surface and degree of conservation. *BMC Struct. Biol.* 6, 19.

737 Huang, W., Lu, S., Huang, Z., Liu, X., Mou, L., Luo, Y., Zhao, Y., Liu, Y., Chen, Z.,  
738 Hou, T., et al. (2013). AlloSite: A method for predicting allosteric sites.  
739 *Bioinformatics* 29, 2357–2359.

740 Khurana, E., Fu, Y., Colonna, V., Mu, X.J., Kang, H.M., Lappalainen, T., Sboner, A.,  
741 Lochovsky, L., Chen, J., Harmanci, A., et al. (2013). Integrative Annotation of  
742 Variants from 1092 Humans: Application to Cancer Genomics. *Science*. 342,  
743 1235587–1235587.

744 Kornev, A.P. and Taylor, S.S. (2015). Dynamics-Driven Allostery in Protein Kinases.  
745 *Trends Biochem. Sci.* xx, 1–20.

746 Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and  
747 Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence  
748 variation and human phenotype. *Nucleic Acids Res.* 42, D980–5.

749 Lee, J., Natarajan, M., Nashine, V.C., Socolich, M., Vo, T., Russ, W.P., Benkovic, S.J.,  
750 and Ranganathan, R. (2008). Surface Sites for Engineering Allosteric Control in  
751 Proteins. *Science* 322, 438–442.

752 Lockless, S.W., Ranganathan, R., Kukic, P., Mirabello, C., Tradigo, G., Walsh, I., Veltri,  
753 P., Pollastri, G., Socolich, M., Lockless, S.W., et al. (1999). Evolutionarily  
754 conserved pathways of energetic connectivity in protein families. *BMC*  
755 *Bioinformatics* 15, 295–299.

756 McVean, G.A., Altshuler (Co-Chair), D.M., Durbin (Co-Chair), R.M., Abecasis, G.R.,  
757 Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P.,  
758 et al. (2012). An integrated map of genetic variation from 1,092 human genomes.  
759 *Nature* *491*, 56–65.

760 Ming, D. and Wall, M.E. (2005). Quantifying allosteric effects in proteins. *Proteins* *59*,  
761 697–707.

762 Mitternacht, S. and Berezovsky, I.N. (2011). Binding leverage as a molecular basis for  
763 allosteric regulation. *PLoS Comput. Biol.* *7*, e1002148.

764 Miyashita, O., Onuchic, J.N., and Wolynes, P.G. (2003). Nonlinear elasticity, protein  
765 quakes, and the energy landscapes of functional transitions in proteins. *Proc. Natl.*  
766 *Acad. Sci.* *100*, 12570–12575.

767 Ng, P.C. and Henikoff, S. (2001). Predicting Deleterious Amino Acid Substitutions.  
768 *Genome Res.* *11*, 863–874.

769 Panjkovich, A. and Daura, X. (2012). Exploiting protein flexibility to predict the location  
770 of allosteric sites. *BMC Bioinformatics* *13*, 273.

771 Panjkovich, A. and Daura, X. (2010). Assessing the structural conservation of protein  
772 pockets to study functional and allosteric sites: implications for drug discovery.  
773 *BMC Struct. Biol.* *10*, 9.

774 Reynolds, K.A., McLaughlin, R.N., and Ranganathan, R. (2011). Hot Spots for Allosteric  
775 Regulation on Protein Surfaces. *Cell* *147*, 1564–1575.

776 Rivalta, I., Sultan, M.M., Lee, N.-S., Manley, G. a., Loria, J.P., and Batista, V.S. (2012).  
777 PNAS Plus: Allosteric pathways in imidazole glycerol phosphate synthase. *Proc.*  
778 *Natl. Acad. Sci.* *109*, E1428–E1436.

779 Rodgers, T.L., Townsend, P.D., Burnell, D., Jones, M.L., Richards, S.A., McLeish,  
780 T.C.B., Pohl, E., Wilson, M.R., and Cann, M.J. (2013). Modulation of Global Low-  
781 Frequency Motions Underlies Allosteric Regulation: Demonstration in CRP/FNR  
782 Family Transcription Factors. *PLoS Biol.* *11*, e1001651.

783 Rousseau, F. and Schymkowitz, J. (2005). A systems biology perspective on protein  
784 structural dynamics and signal transduction. *Curr. Opin. Struct. Biol.* *15*, 23–30.

785 Sethi, A., Eargle, J., Black, A.A., and Luthey-Schulten, Z. (2009). Dynamical networks  
786 in tRNA:protein complexes. *Proc. Natl. Acad. Sci. U. S. A.* *106*, 6620–5.

787 Shulman, A.I., Larson, C., Mangelsdorf, D.J., and Ranganathan, R. (2004). Structural  
788 determinants of allosteric ligand activation in RXR heterodimers. *Cell* *116*, 417–  
789 429.

790 Stenson, P.D., Mort, M., Ball, E. V., Shaw, K., Phillips, A.D., and Cooper, D.N. (2014).  
791 The Human Gene Mutation Database: building a comprehensive mutation repository  
792 for clinical and molecular genetics, diagnostic testing and personalized genomic  
793 medicine. *Hum. Genet.* *133*, 1–9.

794 Süel, G.M., Lockless, S.W., Wall, M.A., and Ranganathan, R. (2003). Evolutionarily  
795 conserved networks of residues mediate allosteric communication in proteins. *Nat.*  
796 *Struct. Biol.* *10*, 59–69.

797 Tirion, M.M. (1996). Large Amplitude Elastic Motions in Proteins from a Single-  
798 Parameter, Atomic Analysis. *Phys. Rev. Lett.* *77*, 1905–1908.

799 Tsai, C., Ma, B. and Nussinov, R. (1999). Folding and binding cascades: Shifts in energy  
800 landscapes. *Proc. Natl. Acad. Sci. U. S. A.* *96*, 9970–9972.

801 Vanwart, A.T., Eargle, J., Luthey-Schulten, Z., and Amaro, R.E. (2012). Exploring



802 residue component contributions to dynamical network models of allostery. J.  
803 Chem. Theory Comput. 8, 2949–2961.  
804 Yang, L.W. and Bahar, I. (2005). Coupling between catalytic site and collective  
805 dynamics: A requirement for mechanochemical activity of enzymes. Structure 13,  
806 893–904.  
807  
808

## 809 CAPTIONS

810 **Figure 1. Schematic overviews of methods for finding surface- and interior-critical**  
811 **residues.** (A) A simulated ligand probes the protein surface in a series of Monte Carlo  
812 simulations (top-left). The cavities identified may be such that occlusion by the ligand  
813 strongly interferes with conformational change (top-right; such a site is likely to be  
814 identified as surface-critical, in red), or they may have little effect on conformational  
815 change, as in the case of shallow pockets (bottom-left) or pockets in which large-scale  
816 motions do not drastically affect pocket volume (bottom-right). (B) Interior-critical  
817 residues are identified by weighting residue-residue contacts (edges) on the basis of  
818 correlated motions, and then identifying communities within the weighted network.  
819 Residues involved in the highest-betweenness interactions between communities (in red)  
820 are selected as interior-critical residues.

821  
822 **Figure 2. Summary statistics for surface-critical sites.** The distributions of the  
823 numbers of surface-critical sites per domain and per complex are given in (A) and (B),  
824 respectively. Panel (C) gives the distributions of the number of surface-critical sites per

DECLAN CLARKE 12/13/15 2:56 PM  
Formatted: Line spacing: 1.5 lines

825 complex without thresholding. Complexes are taken from the the PDB biological  
826 assembly files. Without applying thresholds to the list of ranked surface-critical sites, the  
827 protein is often covered with an excess of identified critical sites.

828

829 **Figure 3. STRESS web server front page, running times, and server architecture.**

830 (A) The server enables users to either provide PDB IDs or to upload their own PDB files  
831 for proteins of interest. Users may opt to identify surface-critical residues, interior-critical  
832 residues, or both. (B) Running times are shown for systems of various sizes. Shown in  
833 red are the running times without optimizing for speed, and green shows running times  
834 with algorithmic optimization. (C) The same data is represented as a log-log plot. The  
835 slopes of these two approaches demonstrate that our algorithm reduces the computational  
836 complexity by an order of magnitude. Our speed-optimized algorithm scales at  $O(n^{1.3})$ ,  
837 where n is the number of residues. (D) A thin front-end server handles incoming user  
838 requests, and more powerful back-end servers perform the heavier algorithmic  
839 calculations. The back-end servers are dynamically scalable, making them capable of  
840 handling wide fluctuations in user demand. Amazon's Simple Queue Service is used to  
841 coordinate between user requests at the front end and the back-end compute nodes: when  
842 the front-end server receives a request, it adds the job to the queue, and back-end servers  
843 pull that job from the queue when ready. Source code is available through Github  
844 ([github.com/gersteinlab/STRESS](https://github.com/gersteinlab/STRESS)).

845

846 **Figure 4. Multiple metrics and datasets reveal that critical residues tend to be**

847 **conserved.** Surface- and interior-critical residues (red) in phosphofructokinase (PDB

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: :

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: Performing local searching supported with hashing

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: implementing additional algorithmic optimizations for computational efficiency reduce

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: considerably (in green), relative to a more naive approach without

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: (in red).

DECLAN CLARKE 12/13/15 2:56 PM

Deleted:

858 3PFK) are given in (A) and (E), respectively. Distributions of cross-species conservation  
859 scores, 1000 Genomes SNV DAF averages, and ExAC SNV MAF averages for surface-  
860 and non-critical residue sets are given in (B), (C), and (D), respectively. The same  
861 distributions corresponding to interior- and non-critical residue sets are given in (F), (G),  
862 and (H), respectively. In (B), mean inter-species conservation scores for surface-critical  
863 sets are -0.131, whereas non-critical residue sets with the same degree of burial have a  
864 mean score of +0.059 ( $p < 2.2e-16$ ). In (F), mean inter-species conservation scores for  
865 interior-critical sets are -0.179, whereas non-critical residue sets with the same degree of  
866 burial have a mean score of -0.102 ( $p=3.67e-11$ ). In (C), means for surface- and non-  
867 critical sets are  $9.10e-4$  and  $8.34e-4$ , respectively ( $p=0.309$ ); corresponding means in (D)  
868 are  $4.09e-04$  and  $2.26e-04$ , respectively ( $p=1.49e-3$ ). In (G), means for interior- and non-  
869 critical sets are  $2.82e-4$  and  $3.12e-3$ , respectively ( $p=1.80e-05$ ); corresponding means in  
870 (H) are  $3.08e-05$  and  $3.27e-04$ , respectively ( $p=7.98e-09$ ).  $N = 421, 32, 84, 517, 31,$  and  
871 90 structures for panels B, C, D, F, G, and H, respectively. P-values are based on  
872 Wilcoxon-rank sum tests. See SI Methods for further details.

873

874 **Figure 5. Critical residues are shown to be more conserved, as measured by the**  
875 **fraction of rare alleles.** Protein regions with high fractions of *rare* variants are believed  
876 to be more sensitive to sequence variants than other regions, thereby explaining why such  
877 variants occur infrequently in the population. Panels (A) and (C) show distributions for  
878 rare (low DAF) non-synonymous SNVs (taken from the 1000 Genomes dataset) in which  
879 the critical residues are defined to be the surface-critical (A) and interior-critical (C)  
880 residues. Panels (B) and (D) show distributions for rare (low MAF) non-synonymous

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: Statistics for panels (B) and (F) are given in the main text.

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: :

884 SNVs (taken from the ExAC dataset) in which the critical residues are defined to be the  
885 surface-critical (*B*) and interior-critical (*D*) residues. For varying thresholds to define  
886 rarity, there are more structures in which the fraction of rare variants is higher in critical  
887 residues than in non-critical residues. Cases in which the fraction is equal in both  
888 categories are not shown. We consider all structures such that at least one critical and at  
889 least one non-critical residue are hit by a non-synonymous SNV. Panels (*A*), (*B*), (*C*), and  
890 (*D*) represent data from 31, 90, 32, and 84 structures, respectively.

891

892 | **Figure 6. Modeling protein conformational change through a direct use of crystal**  
893 **structures from alternative conformations using absolute conformational transitions**

894 (ACT). (*A*) Distributions (155 structures) of the mean conservation scores on surface-  
895 critical (red) and non-critical residues with the same degree of burial (blue). (*B*)  
896 Distributions (159 structures) of the mean conservation scores for interior-critical (red)  
897 and non-critical residues with the same degree of burial (blue). Mean values are given in  
898 parentheses. Results for single-chain proteins are shown, and p-values were calculated  
899 using a Wilcoxon rank sum test.

900

901 **Figure 7. Potential allosteric residues add a layer of annotation to structures in the**  
902 **context of disease-associated SNVs.** The structure shown (*A*) is that of the fibroblast  
903 growth-factor receptor (FGFR) in VMD Surf rendering, with HGMD SNVs shown in  
904 orange, bound to FGF2, in ribbon rendering (PDB 1HIL). (*B*) A linear representation of  
905 structural annotation for FGFR. Dotted lines highlight loci which correspond to HGMD  
906 sites that coincide with critical residues, but for which other annotations fail to coincide.

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: :

908 Deeply-buried residues are defined to be those that exhibit a relative solvent-exposed  
909 surface area of 5% or less, and binding site residues are defined as those for which at  
910 least one heavy atom falls within 4.5 Angstroms of any heavy atom in the binding partner  
911 (heparin-binding growth factor 2). The loci of PTM sites were taken from UniProt  
912 (accession P21802).

913

914 | **Table 1. Statistics on the surfaces of *apo* structures within the canonical set of**  
915 **proteins**

916 For each *apo* structure within the canonical set of proteins, statistics relating surface-  
917 critical sites to known ligand-binding sites are reported. The surface of a given structure  
918 is defined to be the set of all residues that have a relative solvent accessibility of at least  
919 50%, where relative solvent accessibility is evaluated using all heavy atoms in both the  
920 main-chain and side-chain of a given residue. Mean values are given in the bottom row.  
921 NACCESS is used to calculate relative solvent accessibility (Hubbard and Thornton,  
922 1993) . *Column 1*: PDB IDs for each structure; *Column 2*: among these surface residues,  
923 the fraction that constitute surface-critical residues; *Column 3*: among surface residues,  
924 the fraction that constitute known ligand-binding residues (known ligand-binding  
925 residues are taken to be those within 4.5 Angstroms of the ligand in the *holo* structure;  
926 Table S1); *Column 4*: the Jaccard similarity between the sets of residues represented in  
927 columns 2 and 3 (i.e., surface-critical and known-ligand binding residues), where values  
928 given in parentheses represent the expected Jaccard similarity, given a null model in  
929 which surface-critical and ligand-binding residues are randomly distributed throughout  
930 the surface (for each structure, 10,000 simulations are performed to produce random

DECLAN CLARKE 12/13/15 2:56 PM

Deleted: :

932 distributions, and the expected values reported here constitute the mean Jaccard similarity  
933 among the 10,000 simulations for each structure); *Column 5*: the number of distinct  
934 surface-critical sites identified in each structure; *Column 6*: the number of known ligand-  
935 binding sites in each structure; *Column 7*: the number of known ligand-binding sites  
936 which are positively identified within the set of surface-critical sites, where a positive  
937 match occurs if a majority of the residues in a surface-critical site coincide with the  
938 known ligand-binding site; *Column 8*: The fraction of ligand-binding sites captured is  
939 simply the ratio of the values in column 7 to those in column 6.  
940

Predicting Allosteric Hotspots Using Dynamics-Based Formalisms with Sequence Analyses Across Diverse Evolutionary Timescales

may not be apparent when studying a small number or specific classes of proteins.

To our knowledge, this is the first study in which the conservation of potential allosteric sites has been measured across a large database of proteins.

The ability to leverage our framework in a high-throughput manner also better