Comparing local search paths with global search paths on protein residue networks: allosteric communication

Susan Khor
slc.khor@gmail.com
(March 8, 2016)

Abstract

Paths constructed on protein residue networks (PRNs) via Euclidean Distance Search (EDS) are compared with paths generated by Breadth First Search (BFS) in terms of their ability to capture key features of allosteric communication in a type III receptor tyrosine kinase (KIT). Findings from this comparison suggest that EDS paths are more plausible discrete models of intra-protein communication pathways than BFS paths, or at least EDS paths can provide different insights not readily available with BFS paths. In general, EDS paths are more stable and have better communication propensity than BFS paths. On the KIT PRN, the EDS paths were better at following the modular contours of KIT’s intra-protein communication network and formed a more organized communication network than the BFS paths. The EDS paths also showed more sensitivity than the BFS paths to subtle but biologically significant structural changes triggered by point mutations to KIT. And finally, the EDS paths were more effective at identifying protein sequence 3-mers containing known KIT mutational hotspots. At a more abstract level, adopting a local view of search on PRNs via EDS paths permitted us to uncover the importance of short-range links, as opposed to long-range links, to the navigability of PRNs. We trace the primary sources of short-range links to secondary structure elements and select solvent-exposed loop regions called independent dynamic segments.

1. Introduction

Mapping out sites, regions and pathways within protein molecules that are functionally critical is an active area of research with important implications for drug delivery and for understanding the mechanics of molecular machines. To date, detailed studies of individual proteins or family of proteins to uncover such maps have been conducted using a combination of computational techniques and time intensive molecular dynamics (MD) simulation. Less computationally demanding but more general studies have also been attempted by applying graph algorithms and complex network concepts to analyze crystallized or NMR derived protein structures represented as a network of interacting residues, e.g. [1] and [2]. Such networks go by several names in the literature. We refer to such a network as a Protein Residue Network (PRN) [3].

In [3] we observed that paths constructed on the PRNs of 166 proteins by the EDS algorithm possess several properties that differ from the shortest paths (BFS). Notably, EDS paths are more varied in length, are less diffusive (have lower search cost) and are less inclined to employ long-range links. Long-range links in a PRN are links between residues which are far apart (> 10) on the protein sequence but close to each other in the tertiary structure. EDS is a greedy Euclidean distance Directed Search algorithm with backtracking similar in principle to Kleinberg’s [4] local search algorithm.

The properties of EDS paths mentioned above are general network properties. In this paper, we conduct a more specific investigation and compare EDS paths with BFS paths in terms of their ability to capture key allosteric communication characteristics of a type III receptor tyrosine kinase (RTK), KIT. Kinase proteins play a major role in signal transduction by turning cellular signaling pathways on and off through phosphorylation of substrate proteins. Some kinases are self-activating. For KIT, this happens when its extra-cellular ligand-binding domains bind with stem cell factor (SCF) [5] which results in phosphorylation of its tyrosine residues: Y568 and Y570 [6]. KIT can also be activated through point mutations, and such abnormal regulation of KIT has been implicated in several cancers.
The key allosteric characteristics of KIT that we will be examining are primarily based on the study by Laine et al. [7]. In that study, allosteric communication in the cytoplasmic region of KIT (PDB: 1T45, chain A, residues 547...694, 753...935) was derived from MD simulations and represented as a network of independent dynamic segments (IDSs) connected by communication pathways (CPs). An IDS is a set of residues close in 3D space, whose motions are highly correlated with each other but are largely decoupled from the rest of the system. A CP is a chain of residues constructed from non-covalently bonded residue pairs which are non-adjacent to each other in the protein sequence, and that are highly stable with low distance variation from each other in a protein’s native dynamics. This network representation was sensitive enough to detect differences between KIT in its wild-type (WT) form which is autoinhibited and inactive, and the active form of KIT induced by an oncogenic point mutation, D816V (aspartate at position 816 is replaced with valine). In WT, at least one CP connects the juxtamembrane region (JMR, residues 547...581) to the spatially distant activation loop region (A-loop, residues 810...835) via the catalytic-loop (residues 790...797). This CP disappears when WT is activated by D816V, which triggers structural reorganization of the JMR and A-loop regions, and decreases communication between the N- and C-lobes of the protein kinase domain (PTK) of KIT. This change in communication pattern reflects the repositioning of the JMR in relation to the PTK, and the dimerization of the PTK upon KIT activation [6]. Communication between the JMR and the A-loop via the catalytic-loop is restored with a second point mutation, D792E (aspartate at position 792 is replaced with glutamate), which reestablishes a bond that Laine et al. [7] proposes is key to the inter-conversion of KIT between its active and inactive states: the H(hydrogen)-bond between residues Y823 and D792.

2. Method and Materials

2.1 Protein Residue Network (PRN)

A PRN is constructed from the coordinates obtained from either PDB (rcsb.org) or snapshots of a MD simulation run. A PRN is a simple undirected connected graph \( G = (V, E) \). Each element in the set of nodes \( V \) represents an amino acid molecule (residue) in a protein sequence. Let the number of nodes \(|V| = N\). Nodes are labeled by the residue id (\( rid \)) given in the coordinates file (so that gaps in node labels correspond in size to gaps in sequence locations).

Our PRN construction is based on the method in [9] which highlights the role side-chain atoms play in identifying well-formed protein structures. Two nodes \( u \) and \( v \) are linked if and only if \( |u - v| > 1 \) (peptide bonds are ignored), and their interaction strength \( I_{uv} \) is above a threshold. \( I_{uv} = \frac{n_{uv} \times 100}{\sqrt{R_u \times R_v}} \) where \( n_{uv} \) is the number of distinct pairs \((i, j)\) such that \( i \) is an atom of residue \( u \), \( j \) is an atom of residue \( v \), and the Euclidean distance between atoms \( i \) and \( j \) is within a cutoff distance. \( R_u \) and \( R_v \) are extracted from a table of normalization values by residue type (Table 1 in [10]). When computing \( I_{uv} \) to demonstrate the role side-chains play in specifying protein structure, only side-chain atoms were used in [9]. In contrast, we use both the side-chain and the protein backbone atoms of an amino acid since the protein backbone plays a significant role in allosteric communication [11-13]. Our cutoff distance is 7.5 Å and \( I_{uv} \geq 5.0 \). Values for these parameters were set through trial and error with the goal of creating PRNs that are singly connected without being unnecessarily dense. Ignoring peptide bonds is appropriate since validation of our model relies on results from [7, 8] where peptide bonds are also ignored in the construction of CPs.

When necessary to distinguish PRNs by source, let \( PRN0 \) be the PRN that is constructed from the protein’s PDB file as opposed to a MD snapshot. The set of PRN edges is partitioned into short-range (\( SE \)) and long-range (\( LE \)) links. A link \((u, v)\) is long-range if and only if \(|u - v| > 10\), and short-range otherwise.
2.2 The Euclidean Directed Search (EDS) algorithm

Pseudo-code for the EDS algorithm is given in Appendix A. At each step of a search, EDS surveys the proximity to target of the current node’s direct neighbors in a PRN, and moves to a node $x$ not yet on the path which is closest (Euclidean distance) amongst all nodes surveyed so far to the target node. It is possible that $x$ is not a direct neighbor of the current node (it is a direct neighbor of some other node already on the path). In this case, EDS retracts its steps (backtrack) until $x$ becomes reachable.

Fig. 1 shows an EDS path (solid lines) connecting 559 and 823 in the WT PRN. A vastly reduced subset of direct neighbors is shown for each node; however if there is a PRN edge between nodes in Fig. 1, it is shown as a dashed line. The number of direct neighbors (degree) in the WT PRN for each node $x$ in Fig. 1 and the Euclidean distance (up to four decimal places, units in Angstrom Å) between $x$’s carbon-alpha ($C\alpha$) atom and that of the target node 823, are listed on the right of Fig. 1. Starting at node 559, EDS moves to 557 because of all 559’s direct neighbors, 557 is closest to 823. From node 557, EDS checks the direct neighborhood of both 559 and 557 to find an as yet unvisited node that is closest to 823, and moves to 792. This penultimate move creates the EDS path $\langle 559, 557, 792, 823 \rangle$.

| Node | Degree | $C\alpha-C\alpha$ Euclidean distance (Å) to 823 |
|------|--------|---------------------------------------------|
| 557  | 22     | 16.9827                                     |
| 559  | 16     | 22.9016                                     |
| 780  | 24     | 21.6446                                     |
| 783  | 20     | 23.2078                                     |
| 787  | 15     | 25.3612                                     |
| 790  | 18     | 15.7188                                     |
| 791  | 16     | 12.8779                                     |
| 792  | 19     | 10.3172                                     |
| 797  | 20     | 13.1320                                     |
| 815  | 23     | 11.0238                                     |
| 816  | 11     | 10.1687                                     |
| 817  | 17     | 6.5876                                      |
| 823  | 20     | 0.0000                                      |

Fig. 1 $\langle 559, 557, 792, 823 \rangle$ is an EDS path in WT that starts in the JMR, visits a node in the catalytic-loop (792) and terminates at a node in the A-loop (823). The length of this EDS path is three. $\langle 559, 787, 783, 780, 790, 792, 823 \rangle$ is the CP that connects the JMR and A-loop regions via the catalytic-loop in inactive KIT [8]. We revisit this CP and the EDS path in section 3.6. The real numbers beside several of the edges give the $C\alpha-C\alpha$ Euclidean distance between the endpoints of the edge.

BFS and EDS are run for all node pairs $(u, v)$ where $u \neq v$. The total number of paths for each is $N(N-1)$. The length of a path is the number of edges in it. Our C/C++ implementation, which is not optimized, took 205 (BFS) and 31 (EDS) seconds on a 32-bit personal computer running Windows OS to generate all EDS paths and all BFS paths for the 1T45 PRN, which has 331 nodes and 2314 edges. These execution times include time to output the paths to file.

Paths can be partitioned into two sets by the sequence distance of their source and target nodes. Short-range paths ($SP$) are paths connecting source and target node-pairs within ($\leq$) 10 residues apart on the protein sequence. Paths that are not short-range are long-range paths ($LP$).

2.3 Communication pathways

Communication pathways are built upon the notion that the communication abilities of a protein’s residues correlate with their equilibrium motions [14]. A communication pathway ($CP$) is composed of a chain of residues non-adjacently located on the protein sequence, such that each link in the chain is non-covalently bonded, stable, and the commute time between any pair of residues in the chain is small [7, 8].
A link is \textit{stable} if it has high occupancy, i.e. is present in a large fraction (above a threshold e.g. ≥ 50%) of the protein’s native ensemble (converged MD simulation of the protein’s native dynamics). The \textit{commute time} between a pair of residues \((i, j)\) is the variance of the Euclidean distance between \((i, j)\) in the protein’s native ensemble [14, 7]. A larger variance increases commute time for a residue pair and decreases its communication propensity. Euclidean distance is calculated between the carbon-alpha (Cα) atoms of a residue pair.

There is a major difference between CPs and both EDS and BFS paths. CPs are akin to a constrained diffusion process from a source node (paths or chains linking residues are extended out from a node until there are no more links with acceptable stability and commute time), than to an unconstrained targeted search for a node. Further, depending on the cutoff values used for required stability and acceptable commute time, there may be no CPs extending out from some residues. Ref. [8] reports that a maximum of 67 CPs extend from a WT residue, and the number of CPs for WT and MU (D816V KIT mutant) is 2404 and 1730 respectively. In contrast, EDS and BFS paths are possible, and algorithmically generated, for all pairs of nodes in a PRN. The PRN edges and/or the generated EDS and BFS paths may be subjected to further screening according to protein specific criteria, but we have opted not to go this route to keep the findings from this study at a general level.

\subsection*{2.4 Path stability and (estimated) path commute time}

Our aim is not to construct equivalent CPs, but to evaluate EDS and BFS paths in terms of their ability to model intra-protein communication. The definition of CPs in section 2.3 gives us two properties that an intra-protein communication path should likely possess: stability and communication propensity. Thus we need a way to quantify stability and commute time for an EDS or BFS path.

Per section 2.3, let \(sb(e)\) be stability of a link \(e\), i.e. fraction of time it is present in a sequence of MD snapshots. Assuming links of a PRN are independent of each other (this is not entirely true because of geometric constraints), we define \textit{stability of a path} \(p\) with \(n\) edges as \(sb(p) = \prod_{i=1}^{n} sb(e_i)\). Paths with larger \(sb(p)\) are more stable since all of its links have high stability.

We define \textit{path commute time} as the average commute time between all pairs of nodes on the path. This definition relies on the existence of edge commute time defined in section 2.3. When MD simulation data is not available, we use \textit{estimated path commute time}, which for a path \(p\) is the average Euclidean distance between all pairs of nodes on \(p\). Estimated path commute time is based on the Cα-Cα Euclidean distance of residue pairs in a static protein structure, whereas path commute time is based on the commute time of residue pairs calculated from MD snapshots.

A path of length \(\lambda\) has \(\lambda(\lambda+1)/2\) node pairs. Some of the node pairs on a backtracking EDS path may not be distinct from each other, and commute time between a residue and itself or a sequence adjacent residue is zero. As an example, the estimated path commute time for \(\langle 783, 780, 790 \rangle\) using the distance values in Fig. 1 is \((4.9172 + 8.3111 + 7.8176)/3 = 7.0153\) Å.

\subsection*{2.5 Independent dynamic segments}

An independent dynamic segment \((IDS)\) is a cluster of residues whose Cα fluctuations are highly correlated with each other but are relatively independent from other IDSs and the rest of the protein [7]. IDSs are grown from seed residues that are identified through Principal Component Analysis and Local Feature Analysis [15, 16]. IDSs correspond to well known functional regions distributed throughout KIT [7]. Residues of the 10 IDS for the wild type (WT), the D816V mutant (MU) and the D816V/D792E double mutant (dbMU) KIT are extracted from Table S1 of [7] and reproduced in Table 1. The WT IDSs are marked on a cartoon of 1T45 in Fig 2 (left).
Table 1 List of residues in each IDS and the region in which each IDS is situated [7]. # is number of residues in an IDS.

| Overlapping region | IDS   | WT     | #  | MU (D816V ) | #  | dbMU (D816V, D792E) | #  |
|--------------------|-------|--------|----|-------------|----|---------------------|----|
| JM-Proximal        | S1    | 547...554 | 8  | 547...553   | 7  | 547...552            | 6  |
| JM-Switch           | S2    | 561...569 | 9  | 561...570   | 10 | 561...569            | 9  |
| JM-Zipper           | S3    | 574...581 | 8  | 571...577   | 7  | 572...581            | 10 |
| Loop-I              | S4    | 588, 609...618 | 11 | 588, 609...618 | 11 | 588, 608...618      | 12 |
| C-loop-I            | S5    | 626...633 | 8  | 598...601, 625...635 | 15 | 626...631            | 6  |
| C-helix (631...647) |       |         |    |             |    |                     |    |
| Loop-II             | S6    | 585...587, 661...666 | 9  | 586, 587, 661...666 | 8  | 586, 587, 661...665 | 7  |
| Pseudo-KID          | S7    | 688...694, 753...762 | 17 | 684, 687, 690...693, 753...763 | 18 | 687...694, 753...762 | 18 |
| A-loop (810...835)  | S8    | 824...831 | 8  | 816...832   | 17 | 814...821, 824...831 | 16 |
| G-helix (865...885) | S9    | 870...882 | 13 | 870...883   | 14 | 874...887            | 14 |
| (binding site)      |       |         |    |             |    |                     |    |
| C-tail (930...935)  | S10   | 926...935 | 10 | 926...935   | 10 | 925...935            | 11 |
| Total number of IDS residues | 101 | 117 | 109 |

Fig. 2 Left: The ten WT IDSs in Table 1 are colored differently on the cartoon of 1T45 and labeled accordingly. S1, S2 and S3 are in the JMR, and S8 is in the A-loop region. Right: A cartoon of 1T45 with the JMR colored orange, the N-lobe (582...684) blue, the C-lobe (762...935) green, and the pseudo-KID (685...694, 753...761) violet. S4, S5 and S6 are located in the N-lobe, while S8, S9 and S10 are in the C-lobe.

2.6 Path compressibility and reduce-ability

The combination of IDSs and CPs capture two complementary ways a perturbation may propagate within a protein: via concerted local atomic fluctuations for short-range communication, and via a network of well-defined pathways for long-range communication. This description suggests a view of intra-protein communication as a network of modules connected in a well-organized manner.

In a modular network, a random walk (RW) on a network with modular structure will spend more time wandering amongst nodes of a module than traversing between modules. This is the common principle exploited by flow-based clustering algorithms [17, 18]. A RW moves to a direct neighbor of the
current node, selected uniformly at random, until the target node is found. It is expected then that if each module in a network were labeled uniquely but nodes of the same module were labeled identically, and a path $p$ is a sequence of node labels in the order the nodes are visited by $p$, paths which follow the modular contours of a network more faithfully would be more *compressible*. A path represented as a string of symbols is compressible if it has a sub-string of length greater than one that is comprised of identical symbols. Let $\text{nodes}(p)$ be the number of nodes in path $p$. The *compression ratio* for a path is $\text{cr}(p) = \frac{\text{nodes}(p) - \text{nodes}(cp)}{\text{nodes}(p)}$ when $\text{nodes}(cp) > 1$, and $\text{cr}(p) = 1$ when $\text{nodes}(cp) = 1$. Larger $\text{cr}$ values imply greater compression. $\text{cr} = 0$ when there is no compression, and a path with maximum compression ($\text{cr} = 1$) stays within a single module.

To discover which type of path can better reflect this modular structure of an intra-protein communication network (presumably embedded in a PRN), we evaluate the compressibility of both EDS paths and BFS paths relative to random walks on the PRNs of KIT with respect to the set of IDSs in Table 1. Each IDS is given a unique identifier such that the set of IDS identifiers are distinct from the set of node labels.

We also evaluate the *reduce-ability* of EDS and BFS path sets to determine which path type creates better organized communication networks. A better organized set of paths will have more well-trodden or re-used sub-paths. A set of paths is more reduce-able if a larger proportion of its paths are sub-paths of other paths in the set.

### 2.7 Hub residues

Ref. [7] defines *hub* residues are residues that lay on many communication pathways. These hub residues are either evolutionarily conserved or have been observed to participate in the regulation of other receptor tyrosine kinases and cytoplasmic kinases. The 71 hub residues identified in [7] for WT are: 649…655 (C-loop-2), 764…785 (E-helix), 790…797 (catalytic-loop), 804…808 ($\beta$-strand B8), 835…843 ($P+1$ loop), 850…865 (F-helix), and 678, 798, 799, 800, 858, and 862 (catalytic spine). These residues are marked in red in Fig. 3 (left).

![Fig. 3](image)

*Fig. 3* Left: The 71 hub residues identified in [7] are in red on a cartoon of 1T45. Right: The seven mutational hotspots referenced in [7] are shown as magenta sticks on a ribbon representation of 1T45.
2.8 Mutational hotspots
Apart from D816, several other gain-of-function point mutations have been documented to induce irregular state change in KIT. These mutational hotspots include V560, V654, T670, D820, N822 and A829 [7]. The sticks in Fig. 3 (right) mark the positions of these seven mutational hotspot residues. The adverse effects of these mutations can be suppressed by kinase inhibitors, e.g. Imatinib is effective against mutations at position 560, and sunitinib is effective against mutations at positions 654 and 670. However, mutations at position 816 have so far proven resistant to both types of kinase inhibitors [7, 8].

2.9 Theoretical models: 1T45_MU and 1T45_dbMU
Since there are no PDB versions of MU or dbMU to date, we rely on theoretical models of MU and dbMU. We use the terms ‘MU’ and ‘dbMU’ when making reference to the structures and findings from Laine et al. [6-8]; and use 1T45_MU and 1T45_dbMU when referencing our theoretical models and findings. A PRN is built in the manner described in section 2.1 for each theoretical model.

The theoretical models were generated with MODELLER 9.15 [19] using 1T45:A, 3G0E:A and 1PKG:A as templates for 1T45_MU, and 1T45:A and 1PKG:B as templates for 1T45_dbMU (HETATOMS were excluded). Using the PyMol align command, the Root-Mean-Squared-Differences (RMSD) between 1T45:A and the respective template proteins are estimated as follows: 0.584 Å for 3G0E:A, 1.678 Å for 1PKG:A and 1.581 Å for 1PKG:B. The respective templates were selected, after some preliminary experimentation, for their ability to produce targets with reasonable resemblance to descriptions of MU and dbMU given in refs. [6-8]. Three evaluation criteria were considered: (i) the overall shape of their energy profiles relative to the profile for 1T45, (ii) whether the theoretical models possess the requisite secondary structures in two key areas: JM-Switch (560…570) and A-loop, and (iii) change in distance between JM-Switch and two residues in the C-lobe, 847 and 912. The target models were optimized by Modeller using the fast.refine option, but no loop modeling was applied.

For each theoretical model, five target models were generated and the one with the lowest DOPE score was selected for further analysis. 1T45_MU and 1T45_dbMU have energy profiles that are similar to the energy profile of 1T45 (Fig. 4). This provides some assurance as to the reasonableness of the target models, and is in harmony with the report in [7] that both MU and dbMU have similar overall folds to WT. The energy profile (DOPE score) is computed by MODELLER with default options. DOPE (Discrete Optimized Protein Energy) is an atomic distance-dependent statistical potential optimized for model assessment [20]. RMSD between 1T45:A (WT) and 1T45_MU, WT and 1T45_dbMU, and 1T45_MU and 1T45_dbMU is 0.266 Å, 0.176 Å and 0.290 Å respectively. The Ramachandran plots for 1T45, 1T45_MU and 1T45_dbMU created with DeepView v4.1 [38] are presented in Appendix G.

Fig. 4 Energy profiles generated by MODELLER for WT (1T45), 1T45_MU and 1T45_dbMU.
However, both MU and dbMU possess two important structural exceptions to WT. MD simulations of MU and dbMU reveal that the small helix (817...819) in the A-loop region becomes partially unfolded in both MU and dbMU. Further, the JM-Switch region in MU exhibits an anti-parallel beta-sheet structure most of the time but this significant fold gain was not observed in the MD simulations of dbMU [7]. In agreement, 1T45_MU has more beta-residues in its JM-Switch region compared to WT, and only turn-residues at positions 817 to 819 (Table 2). 1T45_dbMU also has only turn-residues at positions 817 to 819, but its JM-Switch region is identical to WT's in terms of DSSP secondary structure assignment (Table 2). The D792E mutation does not restore MU to the original WT conformation; instead dbMU is a blend of WT and MU structures [7]. Secondary structure is assigned by DSSP-2.0.4 [21, 22], and we partition the DSSP assignments into three basic groups. Helix-residues are residues assigned either ‘H’, ‘I’ or ‘G’, beta-residues are residues assigned either ‘E’ or ‘B’, and turn-residues are neither helix- nor beta- residues, i.e. encompasses coils, loops, turns and other disordered regions.

| Table 2 Number of residues by secondary structure type assigned by DSSP-2.0.4 [21, 22]. |
|---------------------------------|---------|---------|---------|---------|---------|---------|---------|---------|
| Secondary structure type       | Overall | JM-Switch (560...570) | A-loop (810...835) |
|--------------------------------|---------|-----------------|-----------------|
| Helix (H)                      | WT      | 1T45_MU | 1T45_dbMU      | WT      | 1T45_MU | 1T45_dbMU      | WT      | 1T45_MU | 1T45_dbMU      |
|                                | 132     | 123     | 129             | 0       | 0       | 0             | 9       | 6       | 6             |
| Beta (B)                       | 61      | 66      | 59              | 2       | 7       | 2             | 4       | 6       | 4             |
| Turn (T)                       | 138     | 142     | 143             | 9       | 4       | 9             | 13      | 14      | 16            |

One of the parameters monitored by [8] in their MD simulations of MU to evidence the detachment of the JMR from the PTK as part of KIT activation is $D_{\text{MS}}$, which is the distance between the centre-of-mass of JM-Switch and the centre-of-mass of two closest C-lobe residues (847 and 912). An increase in mean $D_{\text{MS}}$ was observed in MU relative to WT. To obtain support for similar phenomenon in 1T45_MU, the average distance (Cα-Cα Euclidean distance) between JM-Switch residues and residue 847 (JMS_847), and the average distance between JM-Switch residues and residue 912 (JMS_912), were measured. Both of these distances are significantly longer in 1T45_MU relative to WT (one-sided paired t.test p-values are 1.216E-06 and 1.15E-06 respectively). In WT, JMS_847 is 20.4231 Å with a std. dev. of 6.1511, and JMS_912 is 29.8747 Å with a std. dev. of 5.8974. In 1T45_MU, JMS_847 is 20.7342 Å with a std. dev. of 6.2100, and JMS_912 is 30.2246 Å with a std. dev. of 5.9935.

There are limitations to the use of theoretical models without MD simulation data. Proteins, even when in their “native state”, are dynamic spatial networks. So both the crystallized structure of 1T45 and the theoretical models of MU and dbMU are at best merely theoretical snapshots of the respective protein’s behavior. We do not claim that the theoretical models or their respective PRNs capture all the salient details of WT, MU or dbMU. For instance, the unfolding of the small helix (817...819) in MU is further evidenced by an increase in Ca-Ca Euclidean distance between 816 and 823 relative to WT [8]. However, this distance decreased in 1T45_MU relative to WT. The Ca-Ca Euclidean distance between 816 and 823 is 10.1687 Å in WT, 9.90216 Å in 1T45_MU and 10.089 Å 1T45_dbMU. To reiterate, the main objective of this work is to compare EDS with BFS paths on PRNs; not to reproduce results from [6-8] in all of its detail or to gain new biological insights into the functioning of KIT, although this would be nice. The KIT allostery studies cited here provide a biological context for this comparison.

3. Results and Discussion

3.1 Path stability and path commute time

We evaluated path stability and path commute time of EDS and BFS paths on the PRN0 of 2EZN. 2EZN (chain A) is a mainly beta-sheet protein with 101 residues. All 51,000 snapshots of the native dynamics (298K) MD simulation of 2EZN were retrieved from the Dynamcomics database [23, 24] to compute link stability and link commute time, which were then used to calculate path stability and path commute time (section 2.3). We examine paths of length greater than one to exclude residue pairs directly connected to
each other by a PRN edge. Significance of difference hypothesis is tested with R’s Wilcoxon one-sided test, paired when possible, and a p-value of < 0.01 is taken to confirm a hypothesis.

We find that over all paths of length greater than one, EDS paths are significantly more stable, and have significantly smaller commute times (i.e. better communication propensity) than BFS paths. This conclusion also holds when the analysis is broken down by path range, i.e. short or long (section 2.2). Both short- and long-range EDS paths are significantly more stable and have significantly better communication propensity than BFS paths of the same range. Since having high link stability and smaller commute times are two foundational characteristics of communication pathways (CPs), these findings support the notion that EDS paths are more plausible intra-protein communication pathways than BFS paths. Incidentally, short-range paths exhibit significantly better stability and significantly higher communication propensity than long-range paths for both EDS and BFS.

The above findings follow from differences in the link usage pattern previously observed in [3], coupled with differences in stability and commute times of links by different ranges. Compared with BFS paths, EDS paths have a significantly weaker propensity to use long-range links than short-range links [3]. Short-range links are also significantly more stable and have significantly smaller commute times than long-range links. A pair of residues with a small commute time means their Cα-Cα Euclidean distance has not varied much over the MD simulation. It stands to reason that if a link exists between such a pair, the link is expected to be highly stable. We repeated this analysis of path and edge stability and commute times on the native dynamics of 11 other proteins and observed the same results (Appendix B).

For PRNs without available MD simulation data to compute commute time of edges, we proposed in section 2.4 the use of estimated path commute time. To check that this is a reasonable thing to do, the following correlations, which are all significant (p-value < 0.01), were calculated with the 2EZN data: (i) Cα-Cα Euclidean distance and edge commute time of PRN0 edges is positively (0.2403457) correlated; (ii) path commute time is positively (0.4241025 for EDS, 0.3975642 for BFS) correlated with estimated path commute time; (iii) path stability is negatively (-0.3500514 for EDS, -0.3670511 for BFS) correlated with path commute time; and (iv) path stability is negatively (-0.5074292 for EDS, -0.5451615 for BFS) correlated with estimated path commute time.

3.2 The 1T45 (WT), 1T45_MU and 1T45_dbMU PRNs

The observations in this and the next sections serve to reassure that the PRNs are reasonable and meaningful in the main, with respect to WT, MU and dbMU. In this section, we take a closer look at the PRNs to evaluate how well they manifest the KIT mutants by comparing the Cα-Cα Euclidean distance of several PRN edges that correspond to the H-bonds mentioned in [7] and [8].

In its autoinhibited inactive form (WT), the JMR is a mainly solvent-exposed coil, attached to the C-lobe of the PTK by four stable H-bonds: V560-T787, K558-I789, Y568-F848 and Y570-Y846 [8]. Post-phosphorylation of its tyrosine residues, the JMR is likely to be less attached to the PTK [25], and in MU (D816V), the JMR becomes more ordered with its JMTSwitch region taking on a well-formed anti-parallel beta-sheet structure [6]. In contrast, the JM-Switch region of dbMU (D816V/D792E) is less ordered and resembles the JM-Switch of WT [7]. We confirm that these four residue pairs are linked in all three PRNs. Three of these four links are longer in 1T45_MU than in WT, but are shorter in 1T45_dbMU than in 1T45_MU (Table 3). These distance changes suggest the JMR has displaced itself relative to the PTK in the expected direction. The formation of the beta-sheet in the JM-Switch region is evidenced via two intra-JMR H-bonds: V559-V570 and E561-V569, which are rarely observed in the MD simulation of WT, but are highly present (stable) in the MD simulation of KIT mutants [8]. These two residue pairs are also linked in all three PRNs, and changes in their distances concur with observations from the MD simulations. Both links are shorter in 1T45_MU relative to WT, and longer in 1T45_dbMU relative to 1T45_MU (Table 3).
Table 3 Cα-Cα Euclidean distance (units in Angstrom Å) of PRN links associated with H-bonds that are related to changes in the JMR [8].

| PRN link | WT | 1T45_MU | 1T45_MU – WT | 1T45_dbMU | 1T45_dbMU – 1T45_MU |
|----------|----|---------|-------------|----------|---------------------|
| Distance increases evidence the JMR is less attached to the PTK in 1T45_MU. | 560-787 | 5.63535 | 5.56056 | -0.07479 | 5.76068 | 0.20012 |
| 558-789 | 5.29895 | 5.35059 | 0.05164 | 5.12070 | -0.22852 |
| 568-848 | 9.11620 | 9.69018 | 0.57398 | 9.29650 | -0.39368 |
| 570-846 | 9.84692 | 10.4424 | 0.59548 | 9.88363 | -0.55877 |
| Distance decreases evidence the beta-sheet formation in the JM-Switch region of 1T45_MU. | 559-570 | 5.77561 | 5.71981 | -0.0558 | 5.74989 | 0.03008 |
| 561-569 | 5.64005 | 5.40700 | -0.23305 | 5.66234 | 0.25534 |

Three very stable H-bonds were identified in [7] as playing a key role in the allosteric communication between the JMR and A-loop regions of inactive KIT: Y823-D792, D792-H790, and H790-N797. The stability (occupancy rates) for these three H-bonds in the MD simulation of WT were 95%, 95%, and 93% respectively. The D816V mutation affects the stability of these three H-bonds to different extents. In the MD simulation of MU, the stability of Y823-D792 deteriorated to 45%, while N797 showed an increased preference for interacting with D792 (85%) over H790. The stability of Y823-D792 is restored in dbMU, but N797 spends almost equal time bonding with H790 and D792. Hence, dbMU’s H-bond network between Y823 and the catalytic-loop residues (H790, D792 and N797) is a hybrid of the same H-bond network in WT and in MU [7]. H790-D792 is a very stable bond in all conformational ensembles of the three KIT variants. We confirm that all four residue pairs mentioned here are linked in all three PRNs, and present evidence for these local effects of the point mutations in terms of changes in Euclidean distance (Table 4). The Euclidean distance between 823 and 792 increased in 1T45_MU (relative to WT) but decreased in 1T45_dbMU (relative to 1T45_MU). The Euclidean distance between 790 and 797 increased while the Euclidean distance between 792 and 797 decreased in 1T45_MU. The Euclidean distance of both 792-797 and 790-797 decreased in 1T45_dbMU.

Table 4 Increases (decreases) in Cα-Cα and Cβ –Cβ Euclidean distances (units in Angstrom Å) evidence the deterioration (improvement) in stability of select H-bonds reported in [7].

| PRN link | WT | 1T45_MU | 1T45_dbMU |
|----------|----|---------|----------|
| 792-823 | 10.3172 | 9.1561 | 10.4924 | 9.7760 |
| 790-792 | 5.5726 | 6.1617 | 5.4349 | 5.7161 |
| 790-797 | 9.7923 | 7.5358 | 9.8715 | 7.4993 |
| 792-797 | 7.4072 | 5.7035 | 7.3048 | 5.2225 |

| PRN link | 1T45_MU – WT | 1T45_dbMU – 1T45_MU |
|----------|-------------|---------------------|
| 823-792 | 0.1793 | 1.0959 | -1.2752 | -0.0041 | -0.4760 | -0.4801 |
| 792-790 | -0.1071 | -0.4352 | -0.5423 | -0.0306 | -0.0104 | -0.0410 |
| 790-797 | 0.1578 | 0.0075 | 0.1653 | -0.0786 | -0.0440 | -0.1226 |
| 792-797 | -0.1024 | -0.4810 | -0.5834 | -0.0696 | -0.0724 | -0.1420 |

To support our use of Cα-Cα and Cβ –Cβ Euclidean distance as proxy for interaction stability in an MD simulation, we use 2EZN data from section 3.1 to obtain the following correlations: the Cα-Cα and Cβ –Cβ Euclidean distance of PRN0 edges are significantly and negatively (-0.5118428 and -0.4924738 respectively) correlated with link stability. The average Cα-Cα Euclidean distance of PRN0 edges with stability > 0.5 is 7.770 (std. dev. 1.949029), while for links with stability ≤ 0.5, it is 10.150
The average $C_\beta-C_\beta$ Euclidean distance of PRN0 edges with stability $> 0.5$ is 7.969 (std. dev. 1.902419), while for links with stability $\leq 0.5$, it is 10.220 (std. dev. 1.43093). Hence shorter PRN0 edges tend to be more stable and vice versa. The edges of a PRN tend to be stable in general. 

Several of the links in Tables 3 and 4 have rather large $Ca-Ca$ distance values. Using larger distance cut-offs can be problematic for a pure $Ca-Ca$ residue interaction network (RIN) as important structural variations may be smeared away. In a RIN, a pair of amino acids (nodes) is connected by an edge if the Euclidean distance between its endpoint $Ca$ atoms is within a user specified threshold range. To be compatible with PRNs, we also exclude peptide bonds from RINs. We illustrate this problem in Appendix C on 1T45 with the WT PRN and two WT RINs: one with a 10.3172Å distance cutoff (just long enough for 792-823 to be an edge), and another with a cutoff distance of 15.5677Å (which is the maximum $Ca-Ca$ Euclidean distance for the edges in WT PRN). The figure in Appendix C shows that the RINs have many more edges that cross large cavities on the molecular surface. PRNs are less susceptible to this problem as they are weaker expander graphs (and therefore more modular) than RINs [3]. We revisit this point in section 3.3 with further evidence that at the same maximum edge distance cutoff, PRNs exhibit more structural definition than RINs.

### 3.3 Complex network characteristics of IDS and Hub nodes

In this section, we examine the 1T45 (WT) PRN in terms of node degree and node clustering coefficient. We discuss how these two network structural measures can be related to IDS and hub nodes, and to their respective secondary structure characteristics. The degree of a node is the number of edges incident on the node in a network, viz., a PRN which is a simple, undirected and unweighted graph. The clustering coefficient of a node gives the link density of the node’s direct neighborhood, e.g. the clustering coefficient for a node with degree $k$ is $2e/(k(k-1))$ where $e$ is the number of links between the $k$ nodes. Both the node degrees and the node clustering coefficients of all three KIT PRNs are positively and strongly correlated with each other (p-values from R’s one-sided t.test are reported in parentheses where applicable.

On average, compared with all nodes in the WT PRN, the WT IDS nodes have a significantly smaller node degree (2.19e-15), and a significantly larger fraction of their links is short-range (6.105e-09) (Table 5). On average, compared with all nodes in the WT PRN, the hub nodes (section 2.7) have a significantly larger node degree (2.014e-10), and a significantly larger fraction of their links is long-range (6.855e-05) (Table 5). These significant differences in node degree are in accord with the different roles IDS nodes and hub nodes play in intra-KIT communication. The IDS nodes facilitate short-range while the hub nodes facilitate long-range intra-protein communication.

#### Table 5 Connectivity (mean ± std. dev.) of nodes in the 1T45 PRN. Hubs are most connected, while IDS nodes are least. Compared to all nodes, a significantly larger proportion of links incident on IDS nodes are short-range (SE), while a significantly larger proportion of links incident on hub nodes are long-range (LE).

| Node type | Number of nodes | Node degree | Fraction of node links that are SE | Fraction of node links that are LE |
|-----------|-----------------|-------------|-----------------------------------|-----------------------------------|
| All       | 331             | 13.9800 ± 5.2766 | 0.5191 ± 0.2214 | 0.4809 ± 0.2214 |
| IDS       | 101             | 9.7430 ± 4.1028  | 0.6709 ± 0.2229  | 0.3291 ± 0.2229  |
| Hubs      | 71              | 18.0100 ± 4.3408 | 0.4350 ± 0.1490  | 0.5650 ± 0.1490  |

There is a solvent-exposed loop component to each IDS which gives them the flexibility to have their own independent dynamics (Fig. 2-left). On 1T45, the WT IDSs comprise 24 helix-residues, 9 beta-residues and 68 turn-residues. On 1T45_MU, the MU IDSs are composed of 26 helix-residues, 19 beta-residues and 72 turn-residues. On 1T45_dbMU, the dbMU IDSs comprise 29 helix-residues, 8 beta-residues and 72 turn-residues. In contrast, the hub residues tend to be situated in the more rigid parts of a protein (Fig. 3-left). The (WT) hub residues are made up of 46 helix-residues, 8 beta-residues and 17 turn-
residues. This difference in rigidity profiles between IDS residues and hub residues was also observed in four other proteins [26].

The term ‘independent dynamic segment’ appeared previously in [27] within the context of protein binding sites and their ability to propagate signals within a protein a great distance away from the original point of perturbation. This ability was associated with the notion of ‘discrete breathers’ (DB) which are sites that can be found in a spatially discrete and topologically inhomogeneous network of oscillating nodes. Within the context of proteins, namely the nonlinear network model, DB sites were observed to harvest energy from the background of the system and other network sites (energy is redistributed in favour of DB sites) [28], and to ‘jump’ to other sites distantly located in the network with good energy yield and specificity under suitable conditions [29]. DB sites tend to form in the stiffest parts of a protein [28], and energy is transferred most efficiently along rigid sites [29]. Ref. [30] characterized the rigidity of a site in terms of its node degree and node clustering coefficient, and found that the energy gap (excitation threshold) of a DB is negatively correlated with node degree, and positively correlated with node clustering coefficient. (An alternative way to quantify the level of clustering in a network is to count the number of triangles or cycles of length three in the network, and this is the method used in [30].) DBs spontaneously emerge more easily at sites with lower excitation thresholds. In short, DB sites form more readily in the stiffer parts of a protein which tend to be richly connected globally but poorly connected locally, and less readily in the more flexible parts of a protein which tend to be poorly connected globally but richly connected locally.

The aforementioned relationship between rigidity, as indicated by node degree and node clustering coefficient, and residue secondary structure type is observed in the WT PRN, and it coincides with the respective roles IDS and hub nodes play in WT’s intra-protein communication. On average, the helix-residues have a significantly larger node degree (4.304e-07) and a significantly smaller node clustering coefficient (3.431e-05) than the turn-residues (Table 6, Fig. 5 top-right). On average, the hub nodes, which are helix dominant, have a significantly larger node degree (< 2.2e-16), and a significantly smaller node clustering coefficient (6.868e-08) than the IDS nodes, which are enriched with turn-residues (Table 7, Fig. 5 top-left). The stronger node clustering coefficient of IDS nodes also supports the characterisation of IDSs as modules (albeit modularity and clustering are related but distinct concepts).

**Table 6** Node degree and node clustering coefficient (mean ± std. dev.) for WT residues by secondary structure.

| Node type | Number of nodes | WT PRN | WT RIN 10.3172Å distance cutoff |
|-----------|----------------|--------|---------------------------------|
| Helix     | 132            | 15.35 ± 5.5327 0.3701 ± 0.0794 | 18.88 ± 6.4524 0.4109 ± 0.0801 |
| Beta      | 61             | 15.36 ± 3.3468 0.3606 ± 0.0503 | 20.93 ± 5.0096 0.4167 ± 0.0757 |
| Turn      | 138            | 12.07 ± 5.1520 0.4220 ± 0.1268 | 14.80 ± 6.1914 0.4295 ± 0.1233 |

**Table 7** Node degree and node clustering coefficient (mean ± std. dev.) for WT residues.

| Node type | Number of nodes | WT PRN | WT RIN 10.3172Å distance cutoff |
|-----------|----------------|--------|---------------------------------|
| All       | 331            | 13.98 ± 5.2766 0.3900 ± 0.1019 | 17.56 ± 6.5560 0.4197 ± 0.0999 |
| Hubs      | 71             | 18.01 ± 4.3408 0.3568 ± 0.0483 | 23.0 ± 5.8089 0.3828 ± 0.0679 |
| IDS       | 101            | 9.743 ± 4.1028 0.4402 ± 0.1388 | 11.57 ± 4.4505 0.4598 ± 0.1329 |

When this analysis is repeated with WT’s RIN (10.3172Å cutoff distance), the same conclusions can be made except that there is less distinction between the secondary structures of RIN nodes (Fig. 5 bottom-right). At α = 0.05, both the helix- and the beta- residues do not have significantly smaller clustering coefficients than the turn-residues (0.07039, 0.1861). In contrast, both the helix- and the beta-residues in WT’s PRN, have significantly smaller clustering coefficients on average than the turn-residues (3.431e-05, 1.081e-06). This reinforces the point made in section 3.2 about structural blurring in RINs. The WT RIN (2906) has more edges than the WT PRN (2314), but clustering and modularity also depends on the placement of edges in the network.
3.4 Path compressibility and reduction

All three path types (RW, EDS and BFS) exhibit some level of path compressibility on all three KIT PRNs (Table 8). The RW paths are significantly more compressible than both the EDS and BFS paths (largest p-value < 2.2e-16 for WT, for 1T45_MU and for 1T45_dbMU). This observation supports the presence of modules in the PRNs in the form of IDSs. Further, the compression ratio is affected by IDS size. Paths on the 1T45_MU PRN are significantly more compressible than paths on the other two PRNs because it has more IDS nodes (Table 1) (largest p-value < 2.2e-16 for RW, for EDS and for BFS).

The EDS paths are significantly more compressible than the BFS paths (p-value < 2.2e-16 for WT, for 1T45_MU and for 1T45_dbMU). When the number of highly (cr ≥ 0.5) compressed paths are considered, EDS outperforms BFS by a factor of at least two (Table 9). We attribute this to the nature of EDS search which is less diffusive than BFS [3]. These two findings imply that the EDS paths follow the modular contours of the PRNs more faithfully than the BFS paths. EDS also produces about 10 times more highly compressed paths than RW (Table 9). When comparing the significance of compression ratios, the one-sided Wilcoxon test, paired where applicable, is used since the distribution of cr values is not normal and highly skewed (Appendix D).

So far we have examined how the modular parts of the PRNs in terms of IDSs are better expressed by EDS paths than by BFS paths. In a well-structured communication network, there should also be well-defined inter-module routes. We find that the set of EDS paths is about 1.5 times more reduce-able than
the set of BFS paths on all three PRNs (Table 10), and therefore the EDS paths carve more well-structured communication networks on the PRNs than the BFS paths.

**Table 8** Mean ± std. dev. compression ratio $cr$ for all paths.

| PRN       | RW       | EDS       | BFS       |
|-----------|----------|-----------|-----------|
| 1T45 (WT) | 0.1025 ± 0.0643 | 0.0721 ± 0.1569 | 0.0501 ± 0.1246 |
| 1T45_MU   | 0.1431 ± 0.0777 | 0.1004 ± 0.1837 | 0.0780 ± 0.1540 |
| 1T45_dbMU | 0.1222 ± 0.0723 | 0.0822 ± 0.1666 | 0.0616 ± 0.1390 |

**Table 9** Number of paths with $cr \geq 0.5$

| PRN       | RW | EDS | BFS | EDS/BFS | EDS/RW |
|-----------|----|-----|-----|---------|--------|
| 1T45 (WT) | 317 | 3792 | 1294 | 2.93    | 11.96  |
| 1T45_MU   | 620 | 6257 | 2667 | 2.34    | 10.09  |
| 1T45_dbMU | 473 | 4412 | 1757 | 2.51    | 9.33   |

**Table 10** Number of unique paths after the set of all paths (which comprises 109230 paths) is reduced, i.e. after the removal of every path that is a sub-path of some other path in the set.

| PRN       | EDS       | BFS       | BFS/EDS |
|-----------|-----------|-----------|---------|
| 1T45 (WT) | 39,311    | 62,337    | 1.59    |
| 1T45_MU   | 39,486    | 62,688    | 1.59    |
| 1T45_dbMU | 39,694    | 62,771    | 1.58    |

### 3.5 Inter-IDS communication

To obtain an overall grasp of the communication network within proteins, we constructed a weighted complete graph comprised of all pairs of IDSs for each KIT variant. An IDS pair is linked in this graph if a path runs through them. A path may connect one or more IDS pairs, or none at all. It is easier to work with the compressed paths (section 2.6). A compressed path $cp$ connects an IDS pair $(y, x)$ if both $x$ and $y$ appear in $cp$. The weight of the link between $x$ and $y$ is the number of such $cp$ compressed paths, normalized so that weights of a graph sum to 1.0. The weight of a link in an IDS interaction network indicates the communication strength between an IDS node pair, with heavier weights indicating stronger communication.

Fig. 6 (top) compares the weights for all links in three inter-IDS communication networks built for WT. Weights in the RW network is fairly uniform, as expected from an unbiased random walk. However, both the EDS and the BFS networks show some strong biases, with heavier weights for links between IDS-pairs S1-S5, S1-S6, S4-S6, S5-S6, and S8-S9. The more intense communication between the last three IDS pairs may be the result of their co-location within the same lobe, i.e. S4, S5, and S6 are located within the N-lobe, while IDSs S8 and S9 reside in the C-lobe (Fig. 2-right). More importantly, the weight assigned to IDS-pair S1-S8 by EDS is 2.6 times stronger than that assigned by BFS, whose weight is almost the same as RW’s, and no other IDS-pair displays such a large weight difference (Fig. 6-bottom). This finding implies that the JMR and A-loop regions of inactive KIT are better connected by EDS paths than by BFS paths.

Fig. 7 shows the inter-IDS communication profiles generated with EDS paths (top) and with BFS paths (bottom) for the three PRNs: WT, 1T45_MU and 1T45_dbMU. Changes in communication strength are partly due to change in IDS size. Communication strength is positively influenced by IDS size. Except for S5 and S8, both IDS residue range and IDS size do not vary much across the three KIT variants (Table 1). The size and number of IDSs can and do vary more considerably in other KIT mutants [8]. MU’s S5 is almost twice the size of WT’s S5, while S8 in both MU and dbMU are at least twice the size of S8 in WT. Consequently IDS-pairs involving S5 generally increase (or at least do not decrease) in communication strength in 1T45_MU relative to WT. The same effect is seen for IDS-pairs involving S8 in 1T45_MU and in 1T45_dbMU, but with one notable exception.
EDS communication strength between S1-S8 actually decreased in 1T45_MU relative to WT. In contrast the communication strength between S1-S8 increased in 1T45_MU relative to WT when BFS paths are used to construct the inter-IDS network. S1-S8 communication strength improved in 1T45_dbMU with both EDS and BFS paths. The direction of these changes in communication strengths are not observed with RW (Fig. 8). Hence, the reported deterioration in allosteric communication between the JMR and A-loop regions as a result of D816V in MU, and subsequent restoration of said allosteric communication as a result of D792E in dbMU are more clearly manifested by EDS paths than by BFS paths on the PRNs of our theoretical models. The JMR contains IDSs S1, S2 and S3, while the A-loop region contains S8. Both EDS and BFS show only increase in communication strength in 1T45_MU relative to WT for the other two IDS-pairs: S2-S8 and S3-S8 (Fig. 7).

An IDS pair whose large change in communication strength may not be explained by size is S2-S3. S2 is contained in the JM-Switch region which undergoes significant structural changes in MU. Consequently, the substantially stronger communication between S2 and S3 seen in 1T45_MU, but not in WT or 1T45_dbMU may be due to the effects of increased beta-residues in the JM-Switch region (Table 2).
Fig. 7 Inter-IDS communication profiles for EDS (top) and for BFS (bottom) on the three KIT PRNs: 1T45 (WT), 1T45_MU and 1T45_dbMU.

Fig. 8 Change in communication strength between IDS S1 and S8 as a result of the point mutations. Only EDS paths show a decrease in communication strength between S1 and S8.

3.6 Communication between JMR and A-loop through the catalytic-loop

In the inactive form of KIT, the JMR and the A-loop regions are allosterically connected which permits the effects of a point mutation such as D816V in the A-loop region to be felt in the distantly located JMR. Ref. [7] reports the presence of a CP in WT which gets disrupted by D816V. The seven residues involved
in this CP are V559, C787, L783, M780, H790, D792 and Y823 [8] (Fig. 9). We refer to this CP as CP_WT. It is possible to trace the CP_WT path along the edges of a PRN in each of the three PRNs (Fig. 1). However, neither EDS nor BFS generates a path that traverses the CP_WT route. Nonetheless, the estimated path commute time for \(\langle 559, 787, 783, 780, 790, 792, 823 \rangle\) did increase in 1T45_MU relative to WT, and did decrease in 1T45_dbMU relative to 1T45_MU. The respective estimated path commute times are 12.5562Å, 12.5745Å and 12.5061Å for WT, 1T45_MU and 1T45_dbMU.

![Fig. 9 Residues involved in CP_WT and the JCA path that connects 559 with 823 are shown as sticks in a ribbon view of 1T45 (WT). The JMR, catalytic-loop, A-loop and E-helix are colored yellow, green, red and blue respectively.](image)

There are EDS and BFS paths that connect the JMR and the A-loop through the catalytic-loop. We refer to such paths as JCA/ACJ paths. The JCA/ACJ paths are restricted to the nodes of the JMR, catalytic-loop and A-loop regions only. More formally, let \(j_x\) be a residue in the JMR, \(c_x\) be a catalytic-loop residue, \(a_x\) be a residue in the A-loop region and \(1 \leq p < q < k\), then a JCA path has the form \(\langle j_1, \ldots, j_p, c_{p+1}, \ldots, c_q, a_{q+1}, \ldots, a_k \rangle\) while a ACJ path takes the form \(\langle a_1, \ldots, a_p, c_{p+1}, \ldots, c_q, j_{q+1}, \ldots, j_k \rangle\). In all three KIT PRNs, almost all JCA/ACJ paths visited only one catalytic-loop node, and the three most visited catalytic-loop nodes are 791, 792 and 797. Nodes 792 and 797 are part of the H-bond network in Table 4. Over half (52.38%) of WT’s EDS JCA/ACJ paths but only a third (29.17%) of 1T45_MU’s EDS JCA/ACJ paths pass through node 792. Node 791 is most popular with the EDS JCA/ACJ paths of both 1T45_MU (66.67%) and 1T45_dbMU (76.89%). Since CP_WT also involves the E-helix (764…785), we could have defined JECA/ACEJ paths, where E stands for any E-helix residue, in a similar manner to JCA/ACJ paths. However, the set of JECA/ACEJ paths is empty for both EDS and BFS.

In all three PRNs, node 559 is connected to node 823 by the same EDS JCA/ACJ path: \(\langle 559, 557, 792, 823 \rangle\). In agreement with expected behavior, the estimated path commute time of this JCA path is longer in the 1T45_MU PRN (13.2921Å) than in the WT (13.1855Å) or the 1T45_dbMU (13.1459Å) PRNs. In all three PRNs, the BFS JCA/ACJ paths between nodes 559 and 823 follow either \(\langle 559, 557, 797, 823 \rangle\) or \(\langle 823, 792, 557, 559 \rangle\) routes. The average estimated path commute times for these BFS JAC/ACJ paths in WT, 1T45_MU and 1T45_dbMU are 14.04235Å, 14.2011Å and 14.0328Å respectively. Thus, the BFS JAC/ACJ paths between nodes 559 and 823 also exhibit an increase in estimated path commute time in 1T45_MU relative to WT, and a decrease in 1T45_dbMU relative to 1T45_MU.

Another JAC/ACJ path that shows signs of being disturbed in 1T45_MU and restored in 1T45_dbMU is the one connecting node 559 with 817. Node 817 is remarkable because it is adjacent to the point mutation site (D816V) and is part of the small-helix (817…819) that unfolds as a result of the point mutation. In both WT and 1T45_dbMU, the connecting EDS JCA path is \(\langle 559, 557, 792, 817 \rangle\) with
estimated path commute times of 13.4932Å and 13.4274Å respectively. In 1T45_MU, this path is \(559, 557, 792, 815, 817\) with an estimated path commute time of 13.8322Å. The addition of 815 is a consequence of there being no edges in the 1T45_MU PRN between nodes of the unfolded small-helix and any of the catalytic-loop nodes.

The above are two select examples of EDS JCA/ACJ paths that undergo changes in the different KIT PRNs which echo the effects of the point mutations in MU and dbMU. The first example is inspired by CP WT, and the second is newly observed by us. These communication effects are also detectable statistically at an aggregate level (Tables 11 & 12). In WT, EDS produced 1.4 times more JCA/ACJ paths than BFS, and the EDS JCA/ACJ paths have significantly (0.002151) smaller estimated path commute time on average than the BFS JCA/ACJ paths. This indicates that communication between JMR and A-loop via the catalytic-loop in WT is stronger with EDS than with BFS paths. This ratio drops to 1.167 for both 1T45_MU and 1T45_dbMU. In all three PRNs, the EDS JCA/ACJ paths are also more collapsible than the BFS JCA/ACJ paths. This implies that the EDS JCA/ACJ paths form a more well-organized communication network between the JMR and A-loop regions via the catalytic-loop than the BFS JCA/ACJ paths.

The number of JCA/ACJ paths increased on 1T45_MU for both EDS and BFS. However, this increase is much larger for BFS, and the difference in estimated path commute time between BFS JCA/ACJ paths in WT and in 1T45_MU is not significant. In contrast, the EDS JCA/ACJ paths in 1T45_MU have significantly (0.01104) larger estimated path commute times on average than those in WT. This signals a decline in communication propensity between the JMR and A-loop region through he catalytic-loop in 1T45_MU.

The number of JCA/ACJ paths increased in 1T45_dbMU for both EDS and BFS by about 16%. This increase indicates the availability of more diverse communication channels between the JMR and the A-loop region through the catalytic-loop in 1T45_dbMU. The greater diversity of EDS JCA/ACJ paths on 1T45_dbMU is also accompanied with a significant (0.001595) improvement in communication propensity, relative to the EDS JCA/ACJ paths on 1T45_MU. This improvement is large enough that there is no significant difference between the estimated path commute times of EDS JCA/ACJ paths in 1T45_dbMU and in WT. In other words, EDS provides evidence of decline in 1T45_MU and subsequent restoration in 1T45_dbMU, of the communication propensity between the JMR and A-loop region through the catalytic-loop. This is not the case with the BFS JCA/ACJ paths however. For BFS JCA/ACJ paths, there is no significant difference in terms of estimated path commute time between WT, 1T45_MU or 1T45_dbMU.

### Table 11 JCA/ACJ path statistics

| PRN          | EDS          | BFS          | EDS/BFS          |
|--------------|--------------|--------------|------------------|
|              | Number of paths (A) | Number of reduced paths (B) | Ratio (A/B) | Number of paths (C) | Number of reduced paths (D) | Ratio (C/D) | Ratio (A/C) |
| 1T45 (WT)    | 210          | 86           | 2.44            | 150            | 101           | 1.49        | 1.40        |
| 1T45_MU      | 216          | 88           | 2.45            | 185            | 122           | 1.52        | 1.17        |
| 1T45_dbMU    | 251          | 115          | 2.18            | 215            | 169           | 1.27        | 1.17        |

### Table 12 The mean ± std. dev. estimated path commute time and SE fraction of EDS and BFS JCA/ACJ paths. SE fraction of a path is the fraction of edges on the path which is short-range.

| PRN          | EDS          | BFS          |
|--------------|--------------|--------------|
|              | Est. path commute time | SE fraction of path | Est. path commute time | SE fraction of path |
| 1T45 (WT)    | 13.30 ± 1.9748 | 0.4546 ± 0.2156 | 13.93 ± 2.0980 | 0.3001 ± 0.1710 |
| 1T45_MU      | 13.74 ± 1.9696 | 0.4059 ± 0.2192 | 13.79 ± 2.1636 | 0.2850 ± 0.1975 |
| 1T45_dbMU    | 13.20 ± 1.9995 | 0.3745 ± 0.2171 | 13.87 ± 2.0841 | 0.2580 ± 0.1893 |

For all three KIT PRNs, compared with the BFS JCA/ACJ paths, the EDS JCA/ACJ paths have significantly shorter (or at least not significantly longer) estimated path commute time, 0.3978 for
Having shorter estimated path commute time and utilizing more SE bonds well for both path stability and path commute time, as SE tend to be more stable than LE, and path stability is negatively correlated with both path commute time and estimated path commute time (section 3.1).

3.7 Identifying hub residues

The task here is to identify the 71 hub residues (section 2.7) using information generated by EDS and BFS on the 1T45 PRN. A number of complex network approaches have been proposed to identify key residues for different purposes in a network of interacting protein residues. These approaches typically employ some notion of network centrality either by node degree, number of paths (betweenness) or graph distance (closeness) [1, 2]. We tried several of these centrality measures and found closeness based on estimated path commute time (section 2.4) gave the best outcome. That closeness based on Euclidean distance works best to identify the hub residues is not unexpected since hub residues display fast commute times [7].

Closeness for a node $x$ is the total estimated path commute time from $x$ to all other nodes in the network, and from all other nodes in the network to $x$. $Closeness(x) = \sum_{y \neq x}^{N} [epct(x, y) + epct(y, x)]$. Since a BFS path from node $v$ to node $w$ may differ from a BFS path from node $w$ to node $v$, and similarly, the EDS path from node $i$ to node $j$ may differ from the EDS path from node $j$ to node $i$, the estimated path commute time of a path from $x$ to $y$ may differ from the estimated path commute time of a path from $y$ to $x$, i.e. it is not necessarily that $epct(x, y) = epct(y, x)$.

Nodes with small closeness values are nearer on average to all other nodes in the network than nodes with large closeness values. Closeness was computed with both EDS and BFS paths. To check the reasonableness of this measure, we computed closeness for residues within the two lobes of WT’s PTK and found as expected, that residues residing in the same lobe are closer to each other than residues residing in different lobes (Appendix E).

To recover the hub residues, the nodes were sorted in non-decreasing order of their closeness values. Nodes with smaller closeness values are enriched with hub residues. Compared with all nodes, hub nodes have significantly smaller closeness values (one sided t.test pvalue < 2.2e-16 with both EDS and BFS), and (WT) IDS nodes have significantly larger closeness values (one sided t.test pvalue < 2.2e-16 with both EDS and BFS) (Table in Fig. 10). It is possible to recover at least 60% of the 71 hubs residues at the cost of 26 false positives (10% of 260) with both EDS and BFS closeness values (Fig. 10). Full recovery is slightly quicker with EDS closeness (61.15% FPR) than with BFS closeness (66.92%). The catalytic-loop plays an important role in channeling communication between the JMR and A-loop regions [7] and the eight hub residues in the catalytic-loop (790…797) were the easiest to recover. They are amongst the top 13 residues with the smallest EDS closeness value, and are amongst the top 18 with the smallest BFS closeness value.

Closeness based on EDS paths performed as well as closeness based on BFS paths. This result is not unexpected as a strong positive correlation between EDS and BFS betweenness centrality was reported in [3]. Both EDS and BFS closeness outperformed RW closeness convincingly. Using closeness based only on short-range paths to recover hub residues is not a good strategy as evidenced by the poor performance of both $eds\_sp$ and $bfs\_sp$. This is expected since hub residues lay on the intersection of many communication pathways, and CPs serve as channels for long-range communication. In contrast, IDS residues originate few or no CPs at all [7].
3.8 Identifying mutational hotspots

D816 is a mutational hotspot but not one of the 71 hub residues. Nor is it a hub in the conventional sense, i.e. its node degree is only 11 and 7/11 of its links are SE. In the WT PRN, node 816 ranks 98 by EDS closeness, and 115 by BFS closeness. (A node with a smaller closeness value has a smaller/higher closeness rank.) Of the seven mutational hotspots listed in section 2.8, only one (V654) is a hub residue. Contrary to complex network thinking where removal or damage to hub nodes is expected to cause major disruption to network connectivity and functionality, oncogenic mutations in protein kinases tend to appear in the more flexible or locally frustrated regions [31]. Five of the mutational hotspots are turn-residues and the remaining two are beta-residues. This suggests that a strategy different from both closeness based on estimated path commute time and node degree is needed for recovering or predicting mutational hotspots.

We noticed from our attempts to recover hub residues in section 3.7 that residues located next to a mutational hotspot in the KIT protein sequence tend to occupy highly central positions, i.e. have large node betweenness. Nodes that are traversed by many EDS or BFS paths have larger betweenness values. Therefore instead of trying to recover the seven mutational hotspots directly, we extended the target set to seven triples. Each triple comprises the hotspot residue and its immediate neighbors on the protein sequence. Our extended target set of residues is \{(559, 560, 561), (653, 654, 655), (669, 670, 671), (815, 816, 817), (819, 820, 821), (821, 822, 823), (828, 829, 830)\}. We need only recover one member from each triple to achieve 100% recovery.

Using this strategy of an extended target set and sorting the nodes in non-ascending order of their betweenness values, full recovery was possible with a False Positive Rate (FPR) of 26.54% (incurring the cost of 86 false positives). This best performance is yielded by EDS betweenness calculated using short-range paths only (\textit{eds} \_sp in Fig. 11). However, full recovery may not be necessary and the complete set of mutational hotspots may not be known \textit{a priori}. At smaller FPRs, \textit{eds} \_sp also yielded the best performance. Small FPRs are preferably \textit{per se}, but more so for this problem since there are potentially three tests associated with each candidate residue: the residue itself and its two sequence neighbors. This is where heuristics based on amino-acid chemistry and domain expertise can help further \textit{in silico} winnowing of the candidates.
There are also qualitative differences in the order in which the mutational hotspot residues are recovered. The target residues and their respective ranks are listed in Table 13. Residues that neighbor 816 to its immediate left and right on the protein sequence (815 and 817) are easier to recover when all paths are used to compute node betweenness, i.e. \textit{eds} and \textit{bfs} columns in Table 13. We also note that 791, the left sequence neighbor of residue 792, which is mutated to restore communication interrupted by D816V, is also ranked highly by both \textit{eds} (rank 5) and \textit{bfs} (rank 12). In contrast, residue 560, which borders IDS S2 in the JM-Switch region, and whose mutation is studied extensively in [8], is recovered most easily with \textit{eds sp}. This difference in effectiveness of short- and long-range paths may be a reflection of the difference in impact range of the mutational hotspots. While mutations at 816 produced long-range structural effects [7, 8], mutation at 560 did not appear to have a structural effect on the distant A-loop [8].

For both EDS and BFS, node betweenness computed using short-range paths only, \textit{eds sp} and \textit{bfs sp} columns in Table 13, produced better results than node betweenness based on all paths (Fig. 11-top). We attribute this difference to the proximity of the mutational hotspots to IDSs. It may be that using only short-range paths to compute betweenness is helping to clarify the centrality of a node to the short-range communications around or within IDSs. Ref. [31] proposes that oncogenic mutation sites in protein kinases harness the local frustration present in the more flexible regions of proteins where the oncogenic mutations are commonly found, to produce global effects.

**Fig. 11** Top: Recovery of the seven mutational hotspots using node betweenness and an extended target set. Bottom: \textit{eds sp} outperforms random ordering of nodes. \textit{rs} and \textit{rs17} are the two best results from 20 random node permutations.
In contrast to the previous problem of recovering hub residues, the results produced with EDS betweenness for this task differs from the results produced with BFS betweenness. We attribute this divergence to increased problem difficulty. Compared to the hub recovery problem, the mutational hotspot recovery problem has fewer targets (7 vs. 71) in the same search space and therefore any difference in rank becomes more noticeable. With fewer needles in the haystack, we were concerned about the competitiveness of a random strategy, i.e. ranking the residues in random order. 20 permutations were generated and we report the best two in Fig. 11-bottom. Permutation rs2 performed well at low FPRs, but took a long time to fully recover all targets. On the other hand, permutation rs17 was slow to start but finished as quickly as eds_sp. It appears then that eds_sp gives the best of both worlds and is thus a better strategy than random.

Table 13: Extended set of target mutational hotspot residues and their respective centrality rank. A more central residue has a larger betweenness value and a smaller/higher rank. Each bolded residue number increases the True Positive count by one to a maximum of seven in Fig. 11.

| Rank | Residue | Rank | Residue | Rank | Residue | Rank | Residue |
|------|---------|------|---------|------|---------|------|---------|
| 8    | 817     | 4    | 560     | 1    | 815     | 6    | 669     |
| 9    | 815     | 6    | 669     | 3    | 817     | 7    | 670     |
| 16   | 669     | 16   | 653     | 28   | 823     | 11   | 559     |
| 18   | 653     | 21   | 815     | 33   | 671     | 16   | 655     |
| 21   | 823     | 22   | 670     | 37   | 670     | 21   | 823     |
| 25   | 671     | 26   | 817     | 38   | 655     | 22   | 830     |
| 35   | 654     | 27   | 823     | 40   | 654     | 36   | 817     |
| 37   | 830     | 29   | 830     | 49   | 653     | 40   | 560     |
| 59   | 670     | 43   | 655     | 53   | 669     | 97   | 561     |
| 64   | 655     | 93   | 820     | 64   | 822     | 101  | 815     |
| 77   | 822     | 125  | 559     | 119  | 560     | 134  | 653     |
| 104  | 560     | 130  | 816     | 127  | 559     | 135  | 671     |
| 117  | 559     | 156  | 671     | 181  | 830     | 152  | 654     |
| 181  | 828     | 230  | 829     | 230  | 821     | 208  | 816     |
| 203  | 829     | 249  | 561     | 235  | 816     | 243  | 820     |
| 216  | 816     | 253  | 828     | 264  | 820     | 288  | 821     |
| 220  | 820     | 295  | 822     | 283  | 561     | 303  | 829     |
| 261  | 821     | 308  | 821     | 294  | 829     | 311  | 822     |
| 296  | 561     | 315  | 819     | 317  | 819     | 327  | 819     |
| 321  | 819     | 322  | 654     | 329  | 828     | 328  | 828     |

3.9 Contribution of short-range links to navigability of PRNs

The discussion in section 3.3 points to the role the more rigid parts of a protein, in particular helix structures, could play in long-range energy transport in proteins and thus act as communication channels for long-range allosteric effects in proteins. The energy transport role of helix structures in proteins (not necessarily through the DB mechanism) has been observed directly in experimental studies, e.g. [11, 12]. Ref [14] observed from the five enzymes examined that helix-residues tend to have better communication efficiency than turn-residues. In the WT PRN, the helix-residues have significantly smaller (one-sided t.test pvalue: 2.235e-05 with EDS and 1.409e-05 with BFS) closeness values than the turn-residues, and consequently hub nodes, which are primarily helix-residues, have significantly smaller closeness values than IDS nodes, which are primarily turn-residues (section 3.7). But despite their lower communication efficiencies and shorter communication range, turn-residues are also important to allosteric communication, particularly when they are part of an IDS. Further, some turn-residues are central to allosteric communication, e.g. the catalytic-loop residues in KIT (despite the name, 795…797 are helix-residues). On average, the catalytic-loop residues (790…797) have a significantly larger node degree (one-sided t.test pvalue 0.0003074) and a significantly smaller node clustering coefficient (one-sided t.test
p-value 0.002525) than all nodes in the 1T45 (WT) PRN, which places the catalytic-loop residues on the stiffer end of the rigidity scale (section 3.3).

Since we are proposing that EDS paths are more representative of intra-protein communication than BFS paths, and the discussion so far supports the notion that secondary structures mediate allostery communication in complementary ways, it stands to reason that secondary structures should also influence inter-nodal distances in PRNs. More generally, we want to understand how the presence or absence of secondary structures influences navigability of PRNs, i.e. how average path length scales with number of nodes in a network. In a navigable network, the average length of local search paths scales logarithmically with network size [4].

To examine the contribution of secondary structures to EDS and BFS path lengths, the endpoints of short-range edges (SE) in the 1T45 (WT) PRN were rewired randomly (such that the resulting network remains a simple graph and no edge connects residues consecutively located on the protein sequence) to create a randSE network. In the 1T45 (WT) PRN, SE make up 46% (1075/2314) of all edges, but covers 89.4% (211/236) of all intra-IDS edges, and 98% (617/630) of intra-SSE edges (Table 14). An intra-IDS edge is a network edge whose endpoints are residues of the same IDS. 88% (236/268) of edges whose endpoints are both IDS nodes are intra-IDS edges. An intra-SSE edge is a network edge whose endpoints are within the same secondary structure element (SSE). SSEs locate alpha-helices, beta-strands and linear stretches of turn-residues in a protein sequence. We identify SSEs in a manner similar to [26] as follows: a SSE is a set of at least three residues, consecutively located on the protein sequence, which have the same (DSSP) secondary structure assignment. Intra-SSE edges cover only 64.4% (152/236) of intra-IDS edges.

For comparison, we also created a randLE network by rewiring the long-range edges (LE) of the 1T45 (WT) PRN in the aforementioned manner. In the 1T45 (WT) PRN, LE make up 54% of all edges but encompasses about 73% (1226/1684) of inter-SSE edges. While over half (57%) of SE are intra-SSE edges, almost 99% (1226/1239) of LE are inter-SSE edges. Therefore, a randLE is more disruptive to inter-SSE edges, while a randSE network is more disruptive to both intra-SSE edges and intra-IDS edges. Fig. F1 shows the contact maps for the 1T45 PRN, a randSE network and a randLE network.

Table 14 Breakdown of 1T45 (WT) PRN edges by intra-SSE, inter-SSE, SE, LE and intra-IDS. Intra-SSE and intra-IDS edges generate 63.72% ((617+68)/1075) of SE.

| DSSP pair | SE  | LE  | Total (SE+LE) | Intra-IDS | Intra-IDS and SE |
|-----------|-----|-----|---------------|-----------|-----------------|
| Intra-SSE |     |     |               |           |                 |
| HH        | 344 | 0   | 344           |       18  | 18              |
| BB        | 37  | 0   | 37            |       0   | 0               |
| TT        | 236 | 13  | 249           |     134  | 125             |
| Sub-total | 617 | 13  | 630           |     152  | 143             |
| Inter-SSE |     |     |               |           |                 |
| HH        | 6   | 265 | 271           |       0   | 0               |
| HB        | 33  | 115 | 148           |       2   | 1               |
| HT        | 288 | 429 | 717           |     64   | 49              |
| BB        | 43  | 127 | 170           |       2   | 2               |
| BT        | 80  | 144 | 224           |     16   | 16              |
| TT        | 8   | 146 | 154           |       0   | 0               |
| Sub-total | 458 | 1226| 1684          |     84   | 68              |
| Total (X+Z)| 1075  | 1239 | 2314          |   236    | 211             |

We find that the EDS paths are significantly affected in a negative way by the random shuffling of short-range edges, i.e. disruption to intra-SSE and intra-IDS edges. Significance tests of hypotheses involving path lengths and estimated path commute times are performed with R’s paired one-sided Wilcoxon test. On average, the randSE network has significantly (p-value < 2.2e-16) longer EDS paths than both the 1T45 (WT) PRN and the randLE network (Table 15). In contrast, randSE has a significantly (p-value < 2.2e-16) shorter average BFS path length than both 1T45 (WT) PRN and randLE.

The average path length is dominated by lengths of long-range paths (LP), which make up 94% of all paths. While both short- and long-range EDS paths are significantly (p-value < 2.2e-16) longer in randSE.
on average, short-range (SP) BFS paths deviate from long-range BFS paths and are significantly (p-value < 2.2e-16) longer on average in randSE, compared to PRN. This deviation is in accordance with [32], which views a small-world network as a structure charged with both local and global communication efficiencies. Local communication efficiency is a function of the average (BFS) path length of subgraphs centered on every node in a network. High levels of network clustering improves local communication efficiency since short (BFS) path lengths amongst the direct neighbors of a node \( x \) correlates with a large clustering coefficient for \( x \).

randLE has a significantly (p-value < 2.2e-16) shorter average EDS path length than 1T45 (WT) PRN, but a significantly (p-value < 2.2e-16) longer average estimated path commute time than 1T45 (WT) PRN, which may help explain why the PRN and not the randLE configuration is the native one. Incidentally, 1T45 (WT) also has a significantly (p-value < 2.2e-16) shorter average BFS estimated path commute time than both randSE and randLE.

The significant effects on BFS and EDS average path length reported here for the KIT protein are also observed in 166 other (possibly multimeric) proteins (top left and right panels of Fig. F2 in Appendix F). Importantly, the randSE networks are not navigable, i.e. their average EDS path length no longer scales logarithmically with network size, and therefore the presence of SSEs do contribute to the navigability of PRNs by EDS. Next, we discuss the nature of this contribution.

### Table 15: Path length and estimated path commute time statistics for EDS and BFS.

|         | EDS path length | EDS estimated path commute time | BFS path length | BFS estimated path commute time |
|---------|----------------|--------------------------------|----------------|---------------------------------|
|         | Mean Std. dev. | Mean Std. dev. | Mean Std. dev. | Mean Std. dev. |
| 1T45 (WT) | 4.5578 2.0760 | 13.9008 4.0553 | 3.7077 1.3716 | 14.7060 4.0071 |
| PRN     | 12.9480 63.8351 | 14.9411 5.8459 | 2.5811 0.6207 | 20.5246 7.2604 |
| randSE  | 4.5774 2.6192 | 15.0843 5.7357 | 2.6228 0.6814 | 20.0722 6.6611 |
| randLE  | 4.5774 2.6192 | 15.0843 5.7357 | 2.6228 0.6814 | 20.0722 6.6611 |

Navigability in a small-world network rests on the assumption that there are local (short-range) and global (long-range) links, and that local links connect more similar nodes while global links connect less similar ones. In social networks, similarity may be assessed in terms of geography, language, occupation, ethnicity, education level, interests and so forth. In many theoretical small-world network models, locality is measured as distance, e.g. Manhattan distance in Kleinberg’s grid model [4]. And the same distance norm is used to direct local searches in the network. In PRNs, we use edge sequence distance as a measure of locality, but we use Euclidean distance to direct EDS. Hence some positive correlation between sequence distance and Euclidean distance of edges is assumed. The average Spearman correlation between edge Euclidean distance and edge sequence distance for the 166 PRNs in [3] and their randSE and randLE networks are 0.4075 (std. dev. 0.1163), 0.5129 (std. dev. 0.1143) and 0.8011 (std. dev. 0.0570) respectively. All correlation values are significant. In contrast, there is almost zero correlation between edge Euclidean distance and edge sequence distance in random geometric networks.

The edge shuffling process to produce randSE and randLE networks preserves node degree, but affects the ratio of number of long-range edges to number of short-range edges (|LE|/|SE|), edge sequence distance distribution and node clustering coefficient distribution. For KIT, |LE|/|SE| increases to 8.8 in randSE and decreases slightly to 1.1 in randLE relative to PRN (Table 16). |LE|/|SE| is a significant metric as it influences the distribution of edge sequence distance. For the 166 PRNs and their randLE networks, |LE|/|SE| is fairly constant with network size, but for the randSE networks, |LE|/|SE| increases with network size (center left panel in Fig. F2). Numerous studies have observed or subscribed to the notion that a right skewed link length distribution, i.e. the probability of a link is inversely related to its length, is necessary for navigability of small-world networks [4, 33]. We use the coefficient of variation COV = std. dev. / mean of edge sequence distances as a proxy for skew-ness of the edge sequence distance.
distribution. Since median edge sequence distance is less than mean edge sequence distance for the 166 PRNs, a larger COV implies a skew to the right, or a longer right-tail. The edge sequence distance COVs for the 166 PRNs and their randLE networks range between 1.0 and 2.0, but they are below 1.0 for the randSE networks (center right panel in Fig. F2). For KIT, randSE has a smaller edge sequence distance COV than both PRN and randLE (Table 16), and consequently, a less right skewed edge sequence distance distribution (Fig. 12 top-left).

Table 16 Network structural characteristics of the WT PRN, an instance of its randSE network and an instance of its randLE network. COV = std. dev./mean is the coefficient of variance.

| 1T45 (WT) | Edge sequence distance | Node clustering |
|-----------|------------------------|-----------------|
|           | Median | Mean | Std. dev. | COV | Mean | Std. dev. | Max |
| PRN       | 1.1525 (1239/1075) | 14.00 | 46.42 | 63.3847 | 1.3655 | 0.3900 | 0.1019 | 0.8333 |
| randSE    | 8.8051 (2078/236)  | 72.00 | 102.40 | 90.8341 | 0.8871 | 0.1109 | 0.0624 | 0.3333 |
| randLE    | 1.0998 (1212/1102) | 17.00 | 78.70 | 98.6237 | 1.2532 | 0.1676 | 0.1254 | 0.7000 |

![Graphs showing edge sequence distance and node clustering distributions](image)

Fig. 12 Edge sequence distance distribution (top-left) and node clustering coefficient distribution (top-right) for the 1T45 (WT) PRN and its randSE and randLE networks. Bottom: Rigidity profiles for 1T45 (WT) PRN, randSE and randLE networks measured by node degree and node clustering coefficient.

For the 166 PRNs, average node clustering (i.e. the network clustering coefficient) is significantly reduced in both their randSE and randLE networks (bottom-left panel in Fig. F2), but the randLE networks have significantly larger maximum node clustering coefficients (bottom-right panel in Fig. F2).
A wider node clustering coefficient distribution contributes to diversity in topological rigidity which is a contributing factor for successful allosteric communication in protein kinases [31]. For KIT, average node clustering is slightly but significantly (one-sided t-test paired p-value = 3.117e-10) smaller in randLE than in randSE (Table 16). More pertinent, the randLE network has a larger maximum node clustering coefficient than randSE (Table 16), and a wider node clustering coefficient distribution (Fig. 12 top-right). As a result, randLE (randSE) has a broader (narrower) rigidity profile than randSE (randLE) (Fig. 12 bottom), and this difference in rigidity profile seems to contribute to randLE’s navigability (randSE’s non-navigability). In contrast, observations with BFS average path length would lead one to conclude that a narrower rigidity profile is preferable for allosteric communication. A study of PRNs as a hybrid of SE and LE, or rigid and flexible, sub-networks may prove interesting.

3.10 Suggestions for future local search strategies on PRNs

To reiterate, the goal of this paper is not to replicate the findings in [6-8] with PRNs and EDS paths. Nonetheless, conducting this research has intimated to us additional search objectives that may improve the efficiency and/or the accuracy (e.g. paths that more closely resemble CPs) of local search on PRNs. Currently, the only objective of EDS is to minimize (in a greedy way) the Euclidean distance to target node, and backtracking is used as an escape from local minima.

The discussion on the communication adeptness of helix residues and their network structural characteristics (sections 3.3 & 3.9) suggests that a local search strategy on PRNs could benefit from node degree and/or secondary structure type information. The explicit use of node degree to guide local search has been demonstrated in other complex networks [33]. The influence of node degree on both EDS and BFS paths already happens implicitly, as evidenced by the presence of hierarchical EDS and BFS paths. A path with monotonically increasing, or monotonically decreasing, or monotonically increasing then monotonically decreasing node degrees is hierarchical [34]. For the 166 PRNs in [3], on average, 37.5% (std. dev. 12.0%) of all EDS and 50.7% (std. dev. 15.1%) of all BFS paths are hierarchical. These proportions may seem on the low side but keep in mind that proteins are biological and dynamic systems, so a less strict application of the definition of hierarchical paths could be more appropriate. A large (60%) proportion of the hierarchical EDS and BFS paths have a zoom-in zoom-out navigational pattern, i.e. their node degrees monotonically increase then monotonically decrease, which is a hallmark of successful local navigation in other real-world networks [35].

In terms of graph distance, EDS paths tend to be longer than BFS paths on average. But because EDS is less inclined to use long-range edges than BFS, EDS paths are more stable and have smaller estimated path commute time than BFS paths in general (section 3.1). From the perspective of wiring cost in a spatial network, long-range edges are seen as more expensive than short-range edges. This implies to us that the addition of a wiring cost minimization objective could further reduce the estimated path commute time of EDS paths. The room for reducing estimated path commute time of EDS paths exists and this gap may need to be narrowed to produce realistic CPs. For instance, CP_WT, although longer than the EDS JCA path that connects 559 and 823 in terms of graph distance, has a smaller estimated path commute time than the EDS JCA path (section 3.6). Finding local search paths with minimum estimated path commute time efficiently requires a clever algorithm.

Finally, EDS backtracks so there are no non-edges present in paths. However, in light of the possibility of ‘jumps’ between DB sites (section 3.3), jumps between nodes not connected by a PRN edge could be allowed for a local search to escape local minima. On the other hand, allowing some local search paths to be trapped in local minima may serve a purpose [1, 36].

4. Conclusion

We have shown that EDS paths on PRNs can give different and biologically meaningful outcomes from BFS paths on PRNs. These outcomes suggest that EDS paths make better models of intra-protein communication than BFS paths, or can at least provide useful information not readily available from BFS paths. EDS paths are significantly more stable and possess significantly faster commute times than BFS
paths. EDS paths conform to the modular contours of KIT’s intra-protein communication network better than BFS paths, and produce a more organized set of paths. EDS paths are better able than BFS paths to capture at a general level, the characteristics of allosteric communication in inactive KIT, and the changes in allosteric communication effected by the point mutations even though the point mutations did not induce large fold changes. We attribute this to EDS path construction which is more sensitive to changes in the position of residues. At a detailed level however, and using our theoretical models of the KIT mutants, EDS was unable to reproduce results derived with CPs in [6T8]. Whilst constructing CPs on PRNs with EDS is not the goal of this research, possible modifications to EDS are suggested. Information generated by EDS paths show a clear advantage over BFS paths in the recovery of mutational hotspots known to deregulate KIT. Evidence is also presented to support our use of PRNs over RINs. Finally, we examine the network characteristics of IDSs, CPs, secondary structure elements and rigidity of protein regions and discuss how their network characteristics contribute to navigability of PRNs by EDS. We hope that the findings and observations reported here will inspire others to consider local search or other forms of search in addition to BFS when studying protein structure networks to design new approaches for analyzing proteins on a large scale in a high-throughput manner.

Acknowledgments
This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET:www.sharcnet.ca) and Compute/Calcul Canada. Thanks to the Dynaomemics group for providing access to the MD data and to E. Laine for helpful discussions.

References
1 Atilgan AR, Akan P and Baysal C (2004) Small-world communication of residues and significance for protein dynamics. Biophysical Journal 86:85-91.
2 Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, Venger I and Pietrokovski S. (2004) Network analysis of protein structures identifies functional residues. Journal of Molecular Biology 344 1135-1146.
3 Laine E, Auclair C and Tchertanov L (2012) Allosteric communication across the native and mutated KIT Receptor Tyrosine Kinase. PLoS Comput. Biol 8(8):e1002661.
4 Laine E, Chauvet de Beauchene I et al (2014) Hotspot mutations in KIT receptor differentially modulate its allosterically coupled conformational dynamics: Impact on activation and drug sensitivity. PLoS Comput. Biol. 10(7):e1003749
15 Prenev PS and Atick JJ (1996) Local feature analysis: a general statistical theory for object representation. Network: Computation in Neural systems 7:477-500. LFA
16 Zhang Z and Wriggers W (2006). Local feature analysis: a statistical theory for reproducible essential dynamics of large macromolecules. Proteins 64(2):391-403.
17 Rosvall M and Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. PNAS 105(4):1118-1123.
18 S. van Dongen (2000) Graph clustering by flow simulation. PhD thesis, University of Utrecht.
19 Šali A and Blundell TL (1993) Comparative protein modeling by satisfaction of spatial restraints. J. Mol. Biol. 234:779-815. http://salilab.org/modeller/
20 Shen M-Y and Šali A (2006) Statistical potential for assessment and prediction of protein structures. Protein Sci. 15:2507-2524.
21 Kabsch W and Sander C (1983) Dictionary of Protein Secondary Structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22(12):2577-2637. http://swift.cmbi.ru.nl/gv/dssp/
22 Touw WG, et al. (2015) A series of PDB related databases for everyday needs. Nucleic Acids Research January; 43(Database issue): D364-D368.
23 Beck DAC, et al. (2008) Dynameomics: Mass Annotation of Protein Dynamics by All-Atom Molecular Dynamics Simulations. Protein Engineering Design & Selection 21: 353-368.
24 Van der Kamp MW et al. (2010) Dynameomics: A comprehensive database of protein dynamics. Structure, 18: 423-435.
25 Zou J, Wang YD, Ma FX, Xiang ML, Shi B, et al. (2008) Detailed conformational dynamics of juxtamembrane region and activation loop in e-Kit kinase activation process. Proteins 72: 323–332.
26 Karami Y, Laine E and Carbone A (2016) Dissecting protein architecture with communication blocks and communication segment pairs. BMC Bioinformatics 17(Suppl 2):13.
27 Csermely P, Palotai R and Nussinov R (2010) Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. Trends Biochem Sci. October 35(10):539-546.
28 Juancio B, Sanejouand Y-H, Piazza F and De Los Rios P (2007) Discrete breathers in nonlinear network models of proteins. Phys. Rev. Lett. 99, 238104.
29 Piazza F and Sanejouand Y-H (2009) Long-range energy transfer in proteins. Physical Biology 6, 046014.
30 Piazza F and Sanejouand Y-H (2008) Discrete breathers in protein structures. Physical Biology 5, 026001.
31 Dixit A and Verkhivker GM (2011) The energy landscape analysis of cancer mutations in protein kinases. PLoS ONE 6(1): e26071.
32 Latora V and Marchiori M. (2001) Efficient behavior of small-world networks. Phys. Rev. Lett. 87, 198701.
33 Huang W, Chen S and Wang W (2014) Navigation in spatial networks: A survey. Physica A 393:132-154.
34 Gao L (2001). On inferring autonomous system relationships in the Internet. IEEE/ACM Trans. On Networking 9: 733-745.
35 Boguna M, Krioukov D and Claffy KC (2008). Navigability of complex networks. Nature Physics 5:74-80.
36 Leitner DM (2008). Energy flow in proteins. Annu. Rev. Phys. Chem. 59:233-259.
37 Humphrey W, Dalke A and Schulten K (1996) VMD - Visual Molecular Dynamics. J. Molec. Graphics, 14: 33-38. http://www.ks.uiuc.edu/Research/vmd/
38 Guex N and Peitsch MC (1997) SWISS-MODEL and the Swiss-PDBViewer: An environment for comparative protein modeling. Electrophoresis 18:2714-2723. http://www.expasy.org/spdbv/
Appendix A Pseudo-code for EDS

**EDS**\((s, t)\)

Input: source node \(s\), target node \(t\), a graph \(G\)

Output: EDS path \(p\) //A sequence of nodes in order of EDS visit with \(s\) as the leftmost node and \(t\) as the rightmost node when EDS is successful.

Main variable: \(\text{inspected}\) //The set of direct neighbors of nodes currently in \(p\), excluding the nodes in \(p\), sorted in ascending order by their Euclidean distance to \(t\). The leftmost \(\text{node of inspected}\) is a node currently closest to \(t\) and not already in \(p\).

```plaintext
1: append \(s\) to \(p\) //\(p = \langle s \rangle\)
2: \(\text{inspected}\) is empty initially
3: do
4: \(x :=\) the rightmost node of \(p\)
5: for each node \(i\) in the set of direct neighbors of \(x\) in \(G\) do
6: if \(i = t\) then
7: append \(i\) to \(p\)
8: stop //path \(p\) from \(s\) to \(t\) is found
9: end if
10: if \(i\) is not in \(p\) then
11: calculate distance \(d\) between \(i\) and \(t\)
12: add \(i\) to \(\text{inspected}\) //nodes in \(\text{inspected}\) are sorted by their distance to \(t\)
13: end if
14: end for
15: let \(y\) be the leftmost node in \(\text{inspected}\) //closest to \(t\) amongst all nodes currently in \(\text{inspected}\)
16: if \(x\) and \(y\) are not linked in \(G\) then
17: inspect \(p\) from right to left starting at \(x\) for a node \(z\) that is a direct neighbor of \(y\) //backtrack
18: append to \(p\) the sub-path of \(p\) starting from \(x\) to \(z\) //\(p = \langle s \ldots z \ldots x \ldots z \rangle\)
19: end if
20: append \(y\) to \(p\) //\(p = \langle s \ldots z \ldots x \ldots z y \rangle\)
21: remove \(y\) from \(\text{inspected}\)
22: while \(\text{inspected}\) is not empty
23: //path \(p\) from \(s\) to \(t\) is not found
```
Appendix B Path stability and path commute time

Snapshots of the native dynamics (298K) of 11 randomly selected proteins and 2EZN were downloaded from the Dynameomics database [23, 24]. A PRN0 was constructed for the chain of each protein within the residue range simulated in Dynameomics (Table B1). Except for 2EZN where the entire MD simulation was used, stability and commute times of links were computed using the first $x$ of the $y$ available MD native dynamics snapshots (this is due to data download constraints). With these link stability and link commute time information, we evaluated the path stability and path commute time for the set of all EDS and BFS paths of each protein’s PRN0.

Table B1 PDB code and basic statistics for the 12 proteins.

| PDB code (residue range used) | MD snapshots used/total | Number of PRN0 nodes | Number of PRN0 links | Number of paths with > 1 edge in PRN0 |
|-------------------------------|-------------------------|----------------------|----------------------|--------------------------------------|
|                               |                         |                      | Short-range (SE)     | Long-range (LE)                       | Short-range (SP) | Long-range (LP) |
| 1CUK-A (156-203)              | 20,000/51,163           | 48                   | 178                  | 65                                   | 494             | 1,276           |
| 1EZG-A (2-83)                 | 20,000/52,490           | 82                   | 326                  | 324                                  | 878             | 4,464           |
| 1ELP-A (1-83)                 | 20,000/52,318           | 83                   | 215                  | 328                                  | 1,120           | 4,600           |
| 2EZN-A (1-101)                | 51,000/51,000           | 101                  | 346                  | 491                                  | 1,218           | 7,208           |
| 3GRS-A (366-478)              | 20,000/53,650           | 113                  | 330                  | 329                                  | 1,490           | 9,848           |
| 1EBD-A (155-271)              | 20,000/53,224           | 117                  | 335                  | 354                                  | 1,560           | 10,634          |
| 1D0N-A (27-159)               | 20,000/51,311           | 133                  | 386                  | 431                                  | 1,778           | 14,144          |
| 1HFB-A (5-160)                | 20,000/51,867           | 156                  | 587                  | 489                                  | 1,836           | 20,192          |
| 1BFD-A (2-181)                | 20,000/52,997           | 180                  | 561                  | 712                                  | 2,368           | 27,306          |
| 1ESJ-A (1-272)                | 20,000/52,274           | 272                  | 939                  | 1,065                                | 3,452           | 66,252          |
| 1BS2-A (136-482)              | 12,000/51,989           | 347                  | 1,208                | 1,164                                | 4,414           | 110,904         |
| 1EHE-A (5-404)                | 12,000/51,560           | 399                  | 1,430                | 1,391                                | 5,010           | 148,150         |
Protein structure

Path stability by path type.

- **SP EDS >= 0.5** denotes the fraction of EDS short-range paths with > 1 edge with path stability of at least 0.5.
- **LP BFS = 1.0** denotes the fraction of BFS long-range paths with > 1 edge with path stability of 1.0.

Fig. B1 Top: Regardless of search type (BFS or EDS), short-range paths are more stable than long-range paths. Regardless of path range (short or long), EDS paths are more stable than BFS paths. **Bottom:** Regardless of search type (EDS or BFS), long-range paths have longer commute times than short-range paths. Regardless of path range (short or long), BFS paths have longer commute times than EDS paths.

Table B2 p-values generated with R’s Wilcoxon one-sided test, paired when possible (path comparisons). For all the 11 PRN0s, LP EDS paths are significantly (p-value < 0.05) more stable than LP BFS paths, and LP EDS paths have significantly smaller path commute time than LP BFS paths. Except for 1EZG, short-range links (SE) in PRN0s are significantly more stable and have significantly smaller commute time than long-range links (LE) in PRN0s.

| PRN0  | LP path stability | LP path commute time | Edge stability | Edge commute time |
|-------|-------------------|----------------------|----------------|------------------|
| BFS < EDS | BFS > EDS | SE > LE | SE < LE |
| 1CUK-A | 2.07E-02 | 5.20E-19 | 2.68E-09 | 1.32E-25 |
| 1EZG-A | 9.02E-29 | 1.45E-02 | 3.68E-01 | 2.57E-07 |
| 1ELP-A | 2.13E-09 | 4.21E-25 | 1.83E-03 | 1.39E-09 |
| 3GRS-A | 4.42E-51 | 1.90E-106 | 4.94E-14 | 1.87E-24 |
| 1EBD-A | 5.35E-60 | 3.57E-115 | 3.31E-10 | 9.06E-31 |
| 1DON-A | 5.25E-46 | 7.34E-86 | 3.05E-19 | 9.84E-29 |
| 1HBB-A | 1.18E-59 | 2.11E-140 | 1.46E-73 | 2.40E-103 |
| 1BFDA | 1.30E-25 | 4.61E-236 | 2.68E-43 | 5.09E-63 |
| 1ESJ-A | 1.57E-165 | 2.22E-267 | 8.06E-67 | 9.46E-147 |
| 1BS2-A | 2.33E-281 | 0.00E+00 | 1.01E-113 | 5.28E-187 |
| 1EHE-A | 0.00E+00 | 0.00E+00 | 4.92E-143 | 8.34E-247 |

LP BFS avg denotes the average path commute time of long-range BFS paths.

SP EDS mid denotes the median path commute time of short-range EDS paths.
Appendix C

Fig. C1 Six views of 1T45 showing cavity crossings by PRN and RIN edges. The surface of the 1T45 molecule was drawn with VMD 1.9.2 [37] SURF at a probe radius of 1.4 Å. The dotted lines in black are PRN edges. The orange solid lines are RIN edges. The 1T45 RIN edges in the first five panels have a maximum Euclidean distance (measured Ca to Ca) of 10.3172 Å (just long enough for 792-823 to be an edge). This RIN has 2906 edges. The 1T45 PRN edges have a maximum Euclidean distance (measured Ca to Ca) of 15.5677 Å. The bottom right panel is a view of 1T45 with RIN edges at a cutoff of 15.5677 Å. This RIN is more dense; it has 9139 edges and show many more cavity crossing edges.
Appendix D

![Graph showing fraction of paths on WT PRN and number of paths on the WT PRN (log scale) for different compressibility ratios. The graphs compare between different search algorithms: eds, bfs, and rw.](image-url)
Appendix E

Closeness \( (x') = \sum_{y \neq x}^{N} \left[ \text{epct}(x', y') + \text{epct}(y', x') \right] \) gives the closeness of node \( x \) in region \( i \) to all other nodes in region \( j \). In Fig. E1, \( x \) and \( y \) are restricted to two regions: the N-lobe (582…684) and the C-lobe (762…935).

![Fig. E1 Closeness of residues within and between the two lobes of WT's PTK.](image)

Appendix F

![Fig. F1 From left to right: Contact map (adjacency matrix) of the 1T45 (WT) PRN, a randSE network and a randLE network. A non-white cell \((x, y)\) denotes the presence of an edge \((x, y)\). Short-range edges are marked in green.](image)
Fig. F2 Network statistics for the 166 PRNs in [3] and their randSE and randLE networks.
Appendix G Ramachandran Plots for 1T45, 1T45_MU and 1T45_dbMU.