

**Supplementary Table 1 - Summary of network statistics: Correlation between organic laid-out network distances and the mathematically ideal BLAST E-value distances**

	Sequences <sup>a</sup>	Edges <sup>b</sup>	E-value Threshold <sup>c</sup>	Correlation
Amine-binding GPCRs (Fig. 2B)	42 / 42	324 / 324	1×10 <sup>-33</sup> 30% ID / 280 amino acids	R: 0.906 ± 0.034 Z: 11.87 P: 8.04 × 10 <sup>-33</sup>
STE and WNK kinases (Supplementary Fig. 3B)	51 / 51	821 / 821	1×10 <sup>-27</sup> 30% ID / 270 amino acids	R: 0.846 ± 0.026 Z: 9.76 P: 8.18 × 10 <sup>-26</sup>
Enoyl-CoA hydratase family (Fig. 6)	329 / 410	15,723 / 16,054	1×10 <sup>-50</sup> 40% ID / 260 amino acids	R: 0.873 ± 0.004 Z: 49.0 P: 0.0
Kinase superfamily (Fig. 3)	429 / 513	17,213 / 17,355	1×10 <sup>-25</sup> 29% ID / 260 amino acids	R: 0.936 ± 0.003 Z: 40.9 P: 0.0
Crotonase superfamily: domain-only sequences (Fig. 5C)	825 <sup>d</sup> / 1170	<40,014 <sup>d</sup> / 40,946	1×10 <sup>-29</sup> 38% ID / 180 amino acids	R: 0.838 ± 0.002 Z: 31.5 P: 9.81 × 10 <sup>-219</sup>
Crotonase superfamily: full-length sequences (Fig. 5A)	825 <sup>d</sup> / 1170	<64,168 <sup>d</sup> / 74,470	1×10 <sup>-30</sup> 33% ID / 250 amino acids	R: 0.867 ± 0.002 Z: 35.4 P: 9.68 × 10 <sup>-275</sup>
Class A Rhodopsin-like GPCRs (Fig. 4A)	603 / 605	75,820 / 75,820	1×10 <sup>-11</sup> 24% ID / 210 amino acids	R: 0.921 ± 0.002 Z: 54.4 P: 0.0
GPCR suprafamily (Fig. 4B)	766 / 766	140,544 / 140,544	1×10 <sup>-02</sup> 22% ID / 120 amino acids	R: 0.924 ± 0.002 Z: 31.9 P: 1.99 × 10 <sup>-223</sup>

<sup>a</sup>In the Sequences column, the first number reflects the number of sequences in the largest connected cluster that was considered for the correlation analysis. The second number reflects the total number of sequences in the dataset.

<sup>b</sup>The first number in the Edges column is the number of edges in the largest connected cluster that was considered in the calculations here. The second number reflects the total number of edges in the dataset.

<sup>c</sup>Listed with the network E-value threshold is the median percent identity and median alignment length for edges corresponding to the threshold E-value. These are the “worst” edges included in the analysis.

<sup>d</sup>The statistics on the crotonase superfamily networks are based on the distances between the 825 sequences in common in the large connected cluster between the full-length and domain-only networks. There are 974 nodes connected by 40,014 edges in the large connected cluster in the domain-only network, and 931 nodes connected by 64,168 edges in the full-length network.

**Supplementary Table 2 - Comparison of mathematically ideal and displayed pairwise network distances between 51 human STE and WNK kinases**

A. BLAST E-values: (from pairwise alignments)	A		
B. Organic layout	R: 0.846 ± 0.026 Z: 9.76 P: 8.18 × 10 <sup>-23</sup>	B	
C. Neighbor Joining tree	R: 0.854 ± 0.026 Z: 10.63 P: 1.08 × 10 <sup>-26</sup>	R: 0.714 ± 0.026 Z: 8.23 P: 5.52 × 10 <sup>-19</sup>	C
D. Distances from multiple sequence alignment	R: 0.851 ± 0.026 Z: 10.77 P: 2.50 × 10 <sup>-27</sup>	R: 0.713 ± 0.026 Z: 8.96 P: 1.59 × 10 <sup>-19</sup>	R: 0.974 ± 0.026 Z: 12.98 P: 7.71 × 10 <sup>-39</sup>

Pearson's correlations (R) and associated Z-scores (Z) and P-values (P) describing the similarity between the relative pairwise distances between 51 STE and WNK kinase domain sequences as assessed by all shortest paths between  $-\log_{10}(\text{BLAST E-values})$ , the shortest paths between sequences as displayed by a two-dimensional graph layout algorithm, the shortest paths between sequences in a Neighbor-Joining tree, and the relative pairwise distances calculated from a multiple sequence alignment. The pairwise BLAST E-values and the graph layout algorithm correspond to a network thresholded at an E-value of  $1 \times 10^{-27}$ .

**Supplementary Table 3 - Comparison of mathematically ideal and displayed pairwise distances between networks of the crotonase superfamily, using either full-length sequences or just the crotonase domain**

A. BLAST E-values: full-length sequences	A		
B. Organic layout: full-length sequences	R: 0.867 ± 0.002 Z: 35.4 P: 9.68 × 10 <sup>-275</sup>	B	
C. Organic layout: domain only	R: 0.838 ± 0.002 Z: 31.5 P: 9.81 × 10 <sup>-219</sup>	R: 0.868 ± 0.002 Z: 33.2 P: 1.12 × 10 <sup>-241</sup>	C
D. BLAST E-values: domain only	R: 0.872 ± 0.002 Z: 33.0 P: 1.16 × 10 <sup>-238</sup>	R: 0.826 ± 0.002 Z: 30.7 P: 1.04 × 10 <sup>-207</sup>	R: 0.918 ± 0.002 Z: 33.4 P: 4.42 × 10 <sup>-245</sup>

Pearson's correlations (R) and associated Z-scores (Z) and P-values (P) describing the similarity between the relative pairwise distances between 825 crotonase superfamily sequences found in common in the large connected clusters in Fig. 6A and 6B, as assessed by all shortest paths between  $-\log_{10}(\text{BLAST E-values})$  using the full-length or domain-only sequence, the shortest paths between full-length and domain-only sequences as displayed by a two-dimensional graph layout algorithm. Additionally, pairwise BLAST E-values and the graph layout algorithms were used to analyze sequence similarity networks thresholded at a BLAST E-value of  $1 \times 10^{-30}$  (full length) or  $1 \times 10^{-29}$  (domain only).  $1 \times 10^{-30}$  corresponds to a median of 33% identity over alignments of 250 amino acids, and  $1 \times 10^{-29}$  corresponds to a median of 38% identity over alignments of 180 amino acids. Note that the layouts (B and C) are visual representations of the underlying distances in A and D, respectively. A and D cannot be visualized exactly in fewer than N-1 dimensions.