SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:

MATHEMATICS AND COMPUTING

COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

Received
20 August 2014

Accepted
3 November 2014

Published
27 February 2015

Correspondence and requests for materials should be addressed to E.C.M. (emwambene@uwc.ac.za)

Self-similarity of human protein interaction networks: a novel strategy of distinguishing proteins

Emad Fadhal[1], Junaid Gamieldien[1] & Eric C. Mwambene[2]

[1]South African Medical Research Council Bioinformatics Unit, South African National Bioinformatics Institute, University of the Western Cape, Bellville, South Africa, [2]Department of Mathematics and Applied Mathematics, University of the Western Cape, P/Bag X17, Bellville, South Africa.

The successful determination of reliable protein interaction networks (PINs) in several species in the post-genomic era has hitherto facilitated the quest to understanding systems and structural properties of such networks. It is envisaged that a clearer understanding of their intrinsic topological properties would elucidate evolutionary and biological topography of organisms. This, in turn, may inform the understanding of diseases' aetiology. By analysing sub-networks that are induced in various layers identified by zones defined as distance from central proteins, we show that zones of human PINs display self-similarity patterns. What is observed at a global level is repeated at lower levels of inducement. Furthermore, it is observed that these levels of strength point to refinement and specialisations in these layers. This may point to the fact that various levels of representations in the self-similarity phenomenon offer a way of measuring and distinguishing the importance of proteins in the network. To consolidate our findings, we have also considered a gene co-expression network and a class of gene regulatory networks in the same framework. In all cases, the phenomenon is significantly evident. In particular, the truly unbiased regulatory networks show finer level of articulation of self-similarity.

Recently, self-repeating phenomena has been observed in remarkably many systems, both natural as well as man made. What piques man's interest in them is often their aesthetic value more than their organising principles. In particular, long-range power-law correlations depicting self-similarities have been discovered in a remarkably wide variety of systems[1]. There have been attempts to identify self-similarity phenomenon in biological complex systems[2,3] through some kind of re-normalisation. For instance, in biology the observation of the self-similarity phenomenon has been observed in surface areas and vesicular distributions of tissues[4,5].

In respect of self-similarity of the general complex systems to which biological networks belong, the work of Song et al[6] is seminal. They analysed a variety of real complex networks and found that these systems consist of self-repeating patterns. This result was achieved by the application of a re-normalisation procedure that coarse-grains the system into boxes containing nodes within a given neighbourhood size. They identified a power-law relation between the number of boxes needed to cover the network and the size of the box, defining a finite self-similar exponent. In the precise terminology of graph theory, they found out that quotients of complex networks defined by covering neighbourhoods of certain distances were also power-law. Others have used variations of the method with some notable improvements[7,8].

However, it is not surprising that coarse-grain self-similarity was weak in PINs. It has been shown that the majority of nodes (over 90% in all cases that have been considered) lie within 3 distances away from the centre[9]. It is therefore not surprising that any coarse-graining beyond 2 distances from the centre would completely destroy the intrinsic power-law behaviour of the system. Coarse-graining requires that the network has a reasonable diameter and nodes are reasonably spread around the centre.

We have, on the other hand, looked at power-law properties of networks from a different perspective; incomparable to the seminal work of Song et al. As has been shown elsewhere, PINs display a certain recognisable structure[9], which for brevity, we call the stingray structure with quills in this sequel. This structure has been both our point of departure and our focus. We contend that PINs are self-repeating from the stingray structural point of view.

There has been an intense and deliberate effort to determine PINs of many organisms with notable successes. The determination of these networks is to help uncover the generic organising principles of functional cellular networks[10–19]. This progress is an important step in our understanding of the evolution and behaviour of such systems.

It is envisaged that an understanding of the organizing principles at systems level of biological networks would elucidate many of the perplexing questions including that of finding therapeutic targets[20–22]. Such effort is under way in many fronts. Whilst this has been the general aim, much of the recent effort has focused on finding functional dependencies amongst the so-called hubs and their topological importance and positions in the network[23,24].

There has been serious undertaking to understand both structural and functional systems level of protein-protein interaction (PPI) networks through graph visualisation and drawing. The most important piece of information that is required in visualisation is spatial distribution of the network. Yet, such information is calculable if networks are treated as metric spaces. Recently it has been shown that, treated as metric spaces, PINs of various organisms are what we have coined, as alluded to, a stingray structure with quills. That is, proteins with high degree coagulate in the centre of the network whilst those in the periphery have low degree and in the fringes we have nodes of single degrees[9].

Further, in that sequel it was shown that the observed stingray structure has significant biological implications. Amongst others, it was observed that proteins involved in sensing pathways tend to be more expressed in central zones and those in the periphery specialise in routine metabolic pathways. Second, it was observed that some zones are uniquely-enriched and represent a far more pronounced specialisation. Third, it was shown that cancer pathways are significantly over represented in zone 2[25].

In this article, we have analysed substructures that are defined by zones from the centre. In other words, we have statistically visualised the human PINs at both global as well as at subsystems level. What has been revealed is as startling as is aesthetic. These substructures display the same phenomenon that is played out on a global scale. The core of human PINs are imposing self-similarity structures. The systems structures and the ensuing organising principles of these human PINs are repeated at macro as well as at lower levels. In other words, if one would appreciate the beauty of the structure and considered it as a flower with many petals; these very petals would also have petals, which would have more petals of the same kind. Moreover, in most cases, central proteins of various levels from human PINs are from same families, playing the same biological role possibly at every level of consideration. This repetition in similarity of centres is observed in gene regulatory networks, albeit with a finer level of articulation[26].

Table 1 | Metrics of induced subgraphs of HFPIN

PIN	Nodes	Edges	Diameter	Centre	Zones around centre									
					1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
HFPIN	3	181706	13	**MAPK14**	374	4610	3464	578	104	14	2	1	1	Nodes
					86	**32**	**52**	2	2	1	1	2	1	Ave degree
					3	1	1	1	1	1	1	2	1	Min degree
					531	**430**	**393**	14	6	2	2	2	1	Max degree
					0	173	653	307	56	12	1	0	1	# quills
HFPIN1	373	4802	5	**MAPK3**	156	213	3							Nodes
					111	**69**	**14**							Ave degree in the original network
					34	**18**	**6**							Ave degree in the induced network
					1	1	3							Min degree
					144	**80**	**9**							Max degree
					1	4	0							# quills
HFPIN11	155	1587	4	**MAPK1**	103	51								Nodes
					118	**94**								Ave degree in the original network
					22	**14**								Ave degree in the induced network
					2	2								Min degree
					75	**38**								Max degree
					0	0								# quills
HFPIN111	103	866	4	**MAPK11**	64	37	1							Nodes
					106	**143**	**16**							Ave degree in the original network
					15	**18**	**1**							Ave degree in the induced network
					2	1	1							Min degree
					38	**56**	**1**							Max degree
					0	2	1							# quills
HFPIN2	4318	5495	10	**HRAS**	158	1687	2170	262	15	2				Nodes
					72	**46**	**25**	**13**	**4**	**6**				Ave degree in the original network
					53	**35**	**18**	3	1	1				Ave degree in the induced network
					1	1	1	1	1	1				Min degree
					240	**422**	**224**	20	2	1				Max degree
					1	47	493	104	11	2				# quills
HFPIN21	156	1178	7	**NRAS**	85	63	5							Nodes
					85	**60**	**48**							Ave degree in the original network
					19	**9**	**1**							Ave degree in the induced network
					2	1	1							Min degree
					84	**34**	**3**							Max degree
					0	2	3							# quills
HFPIN211	85	575	3	**KRAS**	82	2								Nodes
					85	**51**								Ave degree in the original network
					12	**3**								Ave degree in the induced network
					1	2								Min degree
					52	**4**								Max degree
					9	0								# quills

When pathway and function enrichment analysis are applied to various layers of the induced subgraphs, our results show that there is reinforcement and refinement of these phenomena in various levels of consideration. Moreover, it is clear that there is increased strength in specialisation. Overall, therefore, this self-similarity phenomenon offer a natural way to understanding the biological systems mechanics of the human PINs.

As molecular networks may be biased, we also tested our method and hypothesis on truly unbiased networks such as gene co-expression network and transcriptional regulatory networks. Both cases strongly support the case; and in the case of regulatory network, it is even more pronounced than in PINs.

In other words, we propose that at the core of human PINs, proteins assemble in the same manner of coagulation as systems structures at all levels defined by distance throughout a given network. The key organizing features of the central zones of human PINs are repeated at the level of induced subgraphs defined by distances from the centre. Proteins interact in the same manner, varying only in scale, and refinement of functionality. This recurrence may point to another way of identifying important proteins that may have utility as target drugs.

Results

The general structure of the human PINs. We modeled the human functional protein interaction network (HFPIN)[27], which consists of 9448 nodes and 181706 interactions and the highly curated and currently largest available human signaling network (HSN)[28,29], which consists of 6305 nodes and 62937 interactions. We also looked at the combination of both HFPIN and HSN and produced what we have called the combined human network (CHN), which consists of 10573 nodes and 210689 interactions. Also, a new human protein interaction set based on three-dimensional information with other functional tools has recently been predicted (NHPIS)[30], which consists of 7863 nodes and 23779 interactions. It was equally subjugated to our method. We have also modelled truly unbiased datasets: gene co-expression[31] and regulatory networks[26].

We used a formal method that finds the protein(s) that has the smallest maximal distance to other proteins in the network. The starting point is that in all the networks under consideration, the centres were identified, and nodes were grouped (in zones) according to the distances they are from these central proteins. With this classification, functional enrichment was performed and biological hypotheses were drawn[9,25].

Table 2 | Metrics of induced subgraphs of HSN

PIN	Nodes	Edges	Diameter	Centre	Zones around centre						
					1	**2**	**3**	**4**	**5**	**6**	
HSN	6305	62937	11	**MAPK1**	432	3535	1940	202	38	4	Nodes
					67	**23**	**7**	2	3	3	Ave degree
					1	1	1	1	1	1	Min degree
					451	**377**	**89**	11	9	5	Max degree
					6	401	764	133	20	2	# quills
HSN1	418	4987	5	**MAPK3**	272	142	3				Nodes
					78	**50**	**5**				Ave degree in the original network
					28	**13**	**1**				Ave degree in the induced network
					1	1	1				Min degree
					141	**79**	**2**				Max degree
					16	13	2				# quills
HSN11	254	3020	6	**PIK3CA**	99	145	9				Nodes
					119	**61**	**14**				Ave degree in the original network
					37	**14**	**2**				Ave degree in the induced network
					2	1	1				Min degree
					115	**82**	**7**				Max degree
					0	4	5				# quills
HSN111	99	1362	3	**PIK3R1**	95	3					Nodes
					113	**199**					Ave degree in the original network
					26	**21**					Ave degree in the induced network
					1	1					Min degree
					87	**31**					Max degree
					1	1					# quills
HSN2	2961	27479	9	**AKT1**	198	1558	1082	96	5		Nodes
					51	**38**	**10**	4	3		Ave degree in the original network
					32	**26**	**6**	2	1		Ave degree in the induced network
					1	1	1	1	1		Min degree
					228	**187**	**60**	9	1		Max degree
					8	44	241	63	5		# quills
HSN21	169	728	7	**AKT2**	27	83	46	3	2		Nodes
					81	**61**	**47**	24	27		Ave degree in the original network
					17	**8**	**6**	2	1		Ave degree in the induced network
					4	1	1	1	2		Min degree
					38	**35**	**26**	3	2		Max degree
					0	10	8	2	0		# quills
HSN211	25	84	4	**PDPK1**	15	7	2				Nodes
					101	**48**	**42**				Ave degree in the original network
					7	**4**	**1**				Ave degree in the induced network
					2	1	1				Min degree
					13	**6**	**1**				Max degree
					0	1	2				# quills

Here, we follow the same approach in our consideration of subgraphs of the networks we consider. Before we present self-similarity we are alluding to, let us first summarize the pertinent features of the structure in all the biological networks that were considered. We will argue that the same pattern is evident in induced subgraphs of these networks, determined by distances from central nodes.

The essence of the structure is in the following manner. First, the centres consist of single nodes, all heavily involved in signalling pathway[9]. As for the HFPIN, the centre is MAPK14 and that of the HSN the centre is MAPK1. The combined human network has MAPK3 as the centre. Second, nodes in the central positions have higher degrees than those in the periphery. Moreover, degrees distribution is power law. The third feature is that while the diameters are generally large, the majority of proteins are located in the central positions (zone 1 to zone 3). Fourthly, proteins in the periphery are of low degree. They display the quill structure (node with degree 1) in the fringes of the network. To aid in visualising these networks, we have called these imposing structures stingray structures with quills.

The structures of the HFPIN, HSN, CHN and the NHPIS are summarized in Tables 1 to 4.

Central zones of human PINs as induced subgraphs repeat the structure that is observed by the whole network. The key feature of self-similarity is the self-repeating patterns at various levels of consideration. In our case, we reveal that all the networks we dealt with splits into smaller parts that resemble the whole from a structural point of view of graphs. We split the graphs into parts that are defined by the zones from the centre, i.e., we look at the graphs induced by nodes that are zone *i* from the centre, where *i* is 1, 2, and 3. We examine their structure as was done in the global graphs, following closely what was done in our recent work[9]. We show that the structures we observe have similar patterns. What is striking is that centres of these substructures have similar functions and belong to the same families.

When we now examine the repeating substructures of the giant graphs, in all cases, the induced subgraphs of zones 1 and 2, there is a single node for the centres, which are from the same family of the centres of the human PINs. In the first zones, they are from MAPK family, both in the HFPIN and the HSN. In zone 2, the centres of the induced subgraphs of the HFPIN are from the RAS family; those of HSN are from the general kinase family.

The next natural consideration was to look at zones formed from the zones in the first instant to describe the self-similarity phenomenon. We considered a subset of proteins that form a particular zone and their interactions amongst themselves as a separate induced subgraph. Again, the same phenomenon was observed with varying degree of

Table 3 | Metrics of induced subgraphs of CHN

PIN	Nodes	Edges	Diameter	Centre	Zones around centre									
					1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
CHN	10573	210689	13	**MAPK3**	542	6011	3352	367	61	4	1	1	1	Nodes
					95	**34**	**49**	2	2	1	1	1	1	Ave degree
					1	1	1	1	1	1	1	1	1	Min degree
					590	**431**	**394**	12	6	1	1	1	1	Max degree
					1	339	831	212	40	4	1	1	1	# quills
CHN1	530	8978	5	**MAPK1**	370	154	5							Nodes
					109	**68**	**37**							Ave degree in the original network
					38	**20**	**2**							Ave degree in the induced network
					1	1	1							Min degree
					214	**87**	**4**							Max degree
					7	7	2							# quills
CHN11	362	5811	4	**MAPK14**	166	195								Nodes
					160	**68**								Ave degree in the original network
					43	**21**								Ave degree in the induced network
					5	1								Min degree
					137	**96**								Max degree
					0	3								# quills
CHN111	166	2336	3	**MAPK8**	93	69	3							Nodes
					172	**147**	**21**							Ave degree in the original network
					30	**24**	**4**							Ave degree in the induced network
					4	3	3							Min degree
					84	**89**	**5**							Max degree
					0	0	0							# quills
CHN2	5503	72502	8	**PRKACA**	270	2490	2543	173	7					Nodes
					71	**51**	**21**	5	3					Ave degree in the original network
					53	**38**	**13**	2	1					Ave degree in the induced network
					1	1	1	1	1					Min degree
					307	**420**	**135**	36	1					Max degree
					6	42	626	95	7					# quills
CHN21	235	2647	7	**CSNK1E**	80	58	68	14	4					Nodes
					89	**94**	**79**	29	18					Ave degree in the original network
					52	**10**	**6**	4	4					Ave degree in the induced network
					2	1	1	1	3					Min degree
					72	**32**	**29**	9	5					Max degree
					0	2	12	4	0					# quills
CHN211	74	1954	4	**CSNK1D**	67	6								Nodes
					93	**81**								Ave degree in the original network
					57	**2**								Ave degree in the induced network
					5	1								Min degree
					67	**6**								Max degree
					0	2								# quills

Table 4 | Metrics of induced subgraphs of NHPIS

PIN	Nodes	Edges	Diameter	Centre	Zones around centre							
					1	**2**	**3**	**4**	**5**	**6**	**7**	
NHPIS	7863	23779	14	**SNW1**	532	2231	3660	892	159	49	14	Nodes
					14	**11**	**3**	2	2	2	1	Ave degree
					1	1	1	1	1	1	1	Min degree
					477	**458**	**47**	16	9	6	1	Max degree
					25	374	1315	350	109	32	14	# quills
NHPIS1	441	911	7	**CDC5L**	305	102	26	1				Nodes
					3	**5**	**2**	**1**				Ave degree
					1	1	1	1				Min degree
					57	**51**	**3**	1				Max degree
					121	20	19	1				# quills
NHPIS11	142	164	9	**SRRM2**	35	17	32	7	1			Nodes
					2	**4**	**1**	1	1			Ave degree
					1	1	1	1	1			Min degree
					15	**14**	**4**	2	1			Max degree
					17	6	24	4	1			# quills
NHPIS111	12	13	3	**TADA2A**	7	1						Nodes
					1	**1**						Ave degree
					1	1						Min degree
					2	**1**						Max degree
					5	1						# quills
NHPIS2	1713	6052	9	**ESR1**	139	681	774	76	7			Nodes
					14	**10**	**3**	1	1			Ave degree
					1	1	1	1	1			Min degree
					72	**126**	**40**	4	1			Max degree
					4	52	222	51	7			# quills
NHPIS21	90	145	6	**SP1**	10	32	25	8	1			Nodes
					6	**4**	**2**	1	1			Ave degree
					2	1	1	1	1			Min degree
					10	**10**	**11**	2	1			Max degree
					0	10	10	5	1			# quills

Table 5 | Summary of increases in percentage of pathways as one moves into deeper levels of HFPIN1

Enriched pathways	Zone 1 of HFPIN	Zone 1 of HFPIN1	Zone 1 of HFPIN11	Zone 1 HFPIN111
Signal transduction	38.1%	52%	52%	42.1%
Immune system	31.3%	48%	55.3%	54.6%
MAPK signalling pathway	26.6%	35.8%	48.5%	54.6%
Pathways in cancer	22%	26.2%	31%	18.7%
TRAF6 Mediated Induction of proinflammatory cytokines	10.4%	20.5%	26.2%	28.1%

connectivity and expressed level of manifestation of this organizing principle, depending of the distance of the zone from the centre.

In all the induced subgraphs, we observed the same organizing principles. Nodes with high degree coagulate in central positions and those with low degree are in the periphery of the graphs. Of particular importance, the degree distribution of proteins in these induced subgraphs follow similar patterns (see supplementary figures S1 to S7). The centre of the whole graph is MAPK14 for the HFPIN and MAPK1 for the HSN. As for the HFPIN, at the centre of the induced subgraph of nodes in the first zone is MAPK3. When one considers the zone 1 nodes at MAPK3, the centre is MAPK1 of which its zone 1 subgraph has centre MAPK11 (table 1). In which case, we repeatedly look at induced subgraphs of induced subgraphs. While the level of expression may weaken as we consider the induced subgraphs of these subgraphs, the centres at zones 1 all belong to the MAPK family, a critical family of proteins in signalling. The same is observed for the HSN (table 2).

It is not particularly surprising that, considering that the combined human network has more data, the features of the self-similarity is more pronounced (table 3).

This repeatedness is also observed in zones 2 of the human PINs. Centres are from KRAS family for HFPIN and AKT for the HSN respectively (tables 1 and 2). Both of these families are heavily implicated in cancer pathways[32,33].

Table 6 | Summary of increases percentage of pathways as one moves into deeper levels of HFPIN2

Enriched pathways	Zone 2 of HFPIN	Zone 1 of HFPIN2	Zone 1 of HFPIN21	Zone 1 of HFPIN211
Signal transduction	51.2%	52%	53%	52.5%
Immune system	32.6%	48.1%	55.3%	45%
MAPK signalling pathway	14.1%	35.9%	48.5%	54.7%
Pathways in cancer	28.2%	26.3%	31.1%	45%

Table 7 | Cancer pathways' zonal distribution in HFPIN

Type of cancer	# of proteins	Zone 1 (374)	Zone 2 (4610)	Zone 3 (3464)	Zone 4 (578)	Zone 5 (104)
Breast	330	11 (3.3%)	189 (57.2%)	121 (36.6%)	9 (2.7%)	-
Cervical	711	26 (3.6%)	425 (59.7%)	230 (32.3%)	23 (3.2%)	7 (0.9%)
Endometrial	1515	57 (3.7%)	839 (55.3%)	514 (33.9%)	83 (5.4%)	20 (1.3%)
Fallopian	1292	49 (3.7%)	715 (55.3%)	446 (34.5%)	67 (5.1%)	14 (1%)
Glioblastoma	1046	38 (3.6%)	589 (56.3%)	368 (35.1%)	44 (4.2%)	6 (0.5%)
Glioma	1180	40 (3.3%)	621 (57.7%)	440 (37.2%)	63 (5.3%)	13 (1.1%)
Kidney	561	14 (2.4%)	331 (59%)	193 (34.4%)	23 (4%)	-
Liver	715	29 (4%)	402 (56.2%)	247 (34.5%)	33 (4.6%)	4 (0.5%)
Lung	532	19 (3.5%)	314 (59%)	175 (32.8%)	22 (4.1%)	2 (0.3%)
Ovarian	775	26 (3.3%)	432 (55.7%)	279 (36%)	32 (4.1%)	6 (0.7%)
Pancreatic	717	30 (4.1%)	411 (57.3%)	244 (34%)	28 (3.9%)	4 (0.5%)
Pituitary	1126	37 (3.2%)	591 (52.4%)	421 (37.3%)	61 (5.4%)	15 (1.3%)
Rectal	1597	69 (4.3%)	861 (53.9%)	552 (34.5%)	90 (5.6%)	23 (1.4%)
Average		**3.5%**	**56.5%**	**34.8%**	**4.4%**	**0.7%**

Table 8 | Cancer pathways' zonal distribution in HSN

Type of cancer	# of proteins	Zone 1 (432)	Zone 2 (3535)	Zone 3 (1940)	Zone 4 (202)	Zone 5 (38)
Breast	236	12 (5%)	151 (63.9%)	70 (29.6%)	2 (0.8%)	1 (0.4%)
Cervical	533	42 (7.8%)	323 (60.6%)	157 (29.4%)	9 (1.6%)	1 (0.1%)
Endometrial	1092	89 (8.1%)	647 (59.2%)	336 (30.7%)	17 (1.5%)	2 (0.1%)
Fallopian	941	72 (7.6%)	563 (59.8%)	287 (30.4%)	16 (1.7%)	2 (0.2%)
Glioblastoma	767	64 (8.3%)	471 (61.4%)	216 (28.1%)	13 (1.6%)	2 (0.2%)
Glioma	824	35 (8%)	278 (64%)	114 (62.2%)	5 (1.1%)	1 (0.2%)
Kidney	434	14 (2.4%)	331 (59%)	193 (34.4%)	23 (4%)	-
Liver	537	45 (8.3%)	328 (61%)	155 (28.8%)	7 (1.3%)	1 (0.1%)
Lung	422	31 (7.3%)	260 (61.6%)	121 (28.6%)	8 (1.9%)	2 (0.4%)
Ovarian	557	39 (7%)	334 (59.9%)	174 (31.2%)	8 (1.4%)	1 (0.1%)
Pancreatic	536	46 (8.5%)	332 (61.9%)	148 (27.6%)	8 (1.4%)	1 (0.1%)
Pituitary	789	56 (7%)	458 (58%)	253 (32%)	19 (2.4%)	2 (0.2%)
Rectal	1162	95 (8.1%)	677 (58.2%)	365 (31.4%)	21 (1.8%)	3 (0.2%)
Average		**7.5%**	**60.6%**	**29.6%**	**1.5%**	**0.2%**

Biological ramifications of the self-similarity structure in the HFPIN and similar networks. It has recently been observed that there is some level of specialization by proteins in various zones of the HFPIN[25]. Also, while some pathways cut across zones, of importance is that sensing pathways are far more pronounced in central zones than in periphery. Zones in the periphery tend to be involved in gene expression and metabolic pathways more than those in the centre. In addition, it was also observed that zone 2 bear the significant burnt of pathways involved in cancers.

It is therefore natural that we understand how this phenomenon is played out from the point of view of the self-repeating topology we have alluded to in this article in biological terms. What is made clear is that there seem to be some level of strengthening in terms of pathways.

Four issues are worthy noting. First, the fact that some zones have uniquely-enriched pathways is a clear indication that in those zones, there is a strong representation of proteins that are associated with such pathways. Consider for instance the TRAF6 Mediated Induction of Proin-flammatory cytokines pathway, which is uniquely-enriched in zone 1 in the entirety of the network in the HFPIN. In zone 1 of the induced subgraph of zone 1, as a percentage of proteins involved in this pathway, there is an increase to 20.5% from 10.4%. In the second layer, (zone 1 of zone 1 of zone 1), the

Table 9 | Cancer pathways' zonal distribution in CHN

Type of cancer	# of proteins	Zone 1 (542)	Zone 2 (6011)	Zone 3 (3352)	Zone 4 (367)	Zone 5 (61)
Breast	350	24 (6.8%)	224 (64%)	95 (27.1%)	7 (2%)	-
Cervical	760	43 (5.6%)	496 (65.2%)	203 (26.7%)	16 (2.1%)	2 (0.2%)
Endometrial	1644	91 (5.5%)	1007 (61.2%)	474 (28.8%)	61 (3.7%)	11 (0.6%)
Fallopian	1408	71 (5%)	869 (61.7%)	409 (29%)	51 (3.6%)	8 (0.5%)
Glioblastoma	1128	63 (5.5%)	719 (63.7%)	311 (27.5%)	30 (2.6%)	5 (0.4%)
Glioma	1270	67 (5.2%)	765 (60.2%)	380 (29.9%)	48 (3.7%)	10 (0.7%)
Kidney	593	44 (7.4%)	389 (65.5%)	150 (25.2%)	10 (1.6%)	-
Liver	769	51 (6.6%)	475 (61.7%)	221 (28.9%)	21 (2.7%)	1 (0.1%)
Lung	571	39 (6.8%)	369 (64.6%)	153 (26.7%)	9 (1.5%)	1 (0.1%)
Ovarian	823	37 (4.4%)	524 (63.6%)	236 (28.6%)	23 (2.7%)	3 (0.3%)
Pancreatic	771	44 (5.7%)	483 (62.6%)	223 (28.9%)	21 (2.7%)	-
Pituitary	1228	60 (4.8%)	738 (60%)	373 (30.3%)	47 (3.8%)	10 (0.8%)
Rectal	1753	96 (5.4%)	1061 (60.5%)	515 (29.3%)	70 (3.9%)	11 (0.6%)
Average		**5.7%**	**62.7%**	**28.2%**	**2.8%**	**0.3%**

Table 10 | Cancer pathway distribution in induced zone 2 of HFPIN in self-similarity terms

Type of cancer	# of proteins	Zone 1 (158)	Zone 2 (1687)	Zone 3 (2170)	Zone 4 (262)	Zone 5 (15)
Breast	182	2 (1%)	71 (39%)	98 (53.8%)	11 (6%)	-
Cervix	407	13 (3.1%)	147 (36.1%)	224 (55%)	20 (4.9%)	3 (0.7%)
Endometrium	790	28 (3.5%)	304 (38.4%)	407 (51.5%)	45 (5.8%)	5 (0.6%)
Fallopian	673	17 (2.5%)	241 (35.8%)	374 (55.5%)	37 (5.4%)	4 (0.5%)
Glioblastoma	563	19 (3.3%)	220 (39%)	290 (51.5%)	32 (5.6%)	2 (0.3%)
Glioma	587	22 (3.7%)	217 (36.9%)	316 (53.8%)	30 (5.1%)	2 (0.3%)
Kidney	314	10 (3.1%)	130 (41.4%)	156 (49.6%)	14 (4.4%)	4 (1.2%)
Liver	381	13 (3.4%)	141 (37%)	207 (54.3%)	17 (4.4%)	3 (0.7%)
Lung	299	5 (1.6%)	126 (42.1%)	148 (49.4%)	18 (6%)	2 (0.6%)
Ovarian	411	9 (2.1%)	151 (36.7%)	230 (55.9%)	19 (4.6%)	2 (0.4%)
Pancreas	300	9 (3%)	143 (47.6%)	126 (42%)	19 (6.3%)	3 (1%)
Pituitary	569	19 (3.3%)	220 (38.6%)	301 (52.8%)	27 (4.7%)	2 (0.3%)
Rectal	811	26 (3.2%)	305 (37.6%)	431 (53.1%)	46 (5.6%)	3 (0.3%)
Average		**2.8%**	**38.9%**	**52.1%**	**5.2%**	**0.5%**

percentage incereases to 26.2%. In the next level, it increases to 28.1%. This points to the fact that as one moves into deeper levels, one sees that there is a coagulation of proteins that are highly specialised in specific pathways (table 5).

Second, this phenomenon of strengthening is not restricted to uniquely-enriched pathways. Consider the top 4 pathways in zone 1: signal transduction (38.1%), immune system (31.3%), MAPK (26.6%), pathways in cancer (22%). In the third level of consideration (zone 1 of zone 1 of zone 1), the order changes: immune system (55.3%), signal transduction (52%), MAPK (48.5%), pathway in cancer (31%). By the time the next level is considered, the MAPK signalling pathway dominates, with 54.6% (table 5).

Third, some pathways are more highly represented in the periphery of central zones. For instance, it is interesting to note that signal transduction has an ebbing effect as one moves deeper into central zones of central zones; it still leads in zone 2 of induced subgraph of zone 1. In zone 2 of zone 1 of the induced subgraph, the percentage of proteins involved in signal transduction is highest with 52.5% of proteins involved in this pathway (table 6).

Finally, while it was noted that zones in periphery have a tendency to diversify in metabolic functions, it is important to note that such pathways are ubiquitous. However, there are more enriched in periphery of zones of central zones. Consider for instance, gene expression, metabolism and membrane trafficking. In the induced subgraph of zone 1, the gene expression pathway is uniquely-enriched in zone 2, whilst in the induced subgraph of zone 2, it is significant in zones 2. In the induced subgraph of zone 3, it is the main theme of central zones.

These observations are equally evident in the HSN, CHN and NHPIS (see supplementary tables S1 to S6).

In summary, therefore, we see that the self-repeating structure is played out even from the biological point of view. Signalling pathways continue to be significant in central zones; routine metabolic pathways are significant in the periphery of the network, at all levels of consideration. However, the consideration of the self-repeating structure renders specialisation even more prominent: there are cases where pathways are highly distinguished or uniquely-enriched. Using the self-similarity structure, it is possible to group proteins in some order of importance, a theme we discuss below.

Cancer pathways' zonal distribution in self-similarity terms. In our recent work when we considered the distribution of proteins that consistently expressed in 13 types of cancer[25], it was shown that most of these proteins are prominent in zone 2 of the HFPIN, HSN and CHN (tables 7 to 9). Here, the same methods were applied as we analysed each of the subgraphs from each zone. While on the whole network, cancer proteins are in zone 2, the critical compartment is zone 3 of zone 2 for the HFPIN (table 10) and zone 2 of zone 2 for both HSN and CHN referred to in Tables 11 and 12.

Distinguishing proteins using the self-similarity edifice. It is generally accepted that the degree of the node is a strong indicator of the importance and/or essentiality of the protein in the network[23,24]. As one looks at various layers of zones, central zones of central zones tend to have higher degree in the entirety of the network than the other zones. For instance, proteins from zone 1 of zone 1 in HFPIN have an average degree of 118 and that of zone 1 of zone 2 is 85 (table 1).

Table 11 | Cancer pathway distribution in induced zone 2 of HSN in self-similarity terms

Type of cancer	# of proteins	Zone 1 (198)	Zone 2 (1558)	Zone 3 (1082)	Zone 4 (96)	Zone 5 (5)
Breast	136	7 (5.1%)	89 (65.4%)	33 (24.2%)	7 (5.1%)	-
Cervical	285	17 (5.9%)	189 (66.3%)	72 (25.2%)	7 (2.4%)	-
Endometrial	562	44 (7.8%)	325 (57.8%)	176 (31.3%)	17 (3%)	-
Fallopian	489	37 (7.5%)	295 (60.3%)	143 (29.2%)	14 (2.8%)	-
Glioblastoma	409	32 (7.8%)	252 (61.6%)	113 (27.6%)	12 (2.9%)	-
Glioma	422	33 (7.8%)	259 (61.3%)	120 (28.4%)	10 (2.3%)	-
Kidney	248	19 (7.6%)	156 (62.9%)	61 (24.5%)	12 (4.8%)	-
Liver	281	21 (7.4%)	175 (62.2%)	76 (27%)	9 (3.2%)	-
Lung	229	15 (5%)	143 (62.4%)	63 (27.5%)	8 (3.4%)	-
Ovarian	290	16 (5.1%)	190 (65.5%)	74 (25.5%)	10 (4.3%)	-
Pancreatic	285	19 (6.6%)	189 (66.3%)	71 (24.9%)	6 (2.1%)	-
Pituitary	393	34 (8.6%)	242 (61.5%)	107 (27.2%)	10 (2.5%)	-
Rectal	581	47 (8%)	340 (58.5%)	177 (30.4%)	17 (2.9%)	-
Average		**6.3%**	**62.4%**	**27.1%**	**3.2%**	**-**

Table 12 | Cancer pathway distribution in induced zone 2 of CHN in self-similarity terms

Type of cancer	# of proteins	Zone 1 (270)	Zone 2 (2490)	Zone 3 (2543)	Zone 4 (173)	Zone 5 (7)
Breast	212	7 (3.1%)	97 (45.7%)	104 (49%)	4 (1.8%)	-
Cervical	470	18 (3.8%)	257 (54.6%)	184 (39.1%)	11 (2.3%)	-
Endometrial	939	35 (3.7%)	516 (54.9%)	362 (38.5%)	26 (2.7%)	-
Fallopian	810	33 (4%)	448 (55.3%)	306 (37.7%)	23 (2.8%)	-
Glioblastoma	676	30 (4.4%)	361 (53.4%)	267 (39.4%)	18 (2.6%)	-
Glioma	721	35 (2.8%)	383 (53.1%)	280 (38.8%)	23 (3.1%)	-
Kidney	371	15 (4%)	172 (46.3%)	173 (46.6%)	11 (2.9%)	-
Liver	451	17 (3.7%)	237 (52.5%)	184 (40.7%)	13 (2.8%)	-
Lung	351	15 (4.2%)	169 (48.1%)	158 (45%)	9 (2.5%)	-
Ovarian	495	19 (3.8%)	276 (55.7%)	183 (36.9%)	17 (3.4%)	-
Pancreatic	453	21 (4.6%)	240 (52.9%)	180 (39.7%)	12 (2.6%)	-
Pituitary	693	27 (3.8%)	372 (53.6%)	274 (39.5%)	20 (2.8%)	-
Rectal	991	43 (4.3%)	526 (53%)	391 (39.4%)	30 (3%)	1 (0.1%)
Average		**4%**	**52.2%**	**40.7%**	**2.7%**	**0.007%**

It has also been shown that, in general, both sensing pathways and proteins implicated in diseases tend to be pronounced in central positions[25]. While there is some disagreements about what is more important between sensing pathways and metabolic ones, we contend that sensing pathways are more important as they are likely to elicit a metabolic response to facilitate homeostasis.

In view of the foregoing, we propose that proteins in zone 1 have a higher weighting than those in zone 2 and so on. So, for instance, nodes in zone 3 of zone 1 would have more weight than those in zone 1 of zone 2.

Self-similarity in other biological networks. Both gene co-expression and regulatory networks show stingray structures. When gene co-expression network is subjugated to sub-structure analysis, the majority of the induced subgraphs have single centres. However, as we delve further, we do not obtain single centres. Also,

Table 13 | Metrics of induced subgraphs of Co-expression network

Network	Nodes	Edges	Diameter	Centre	Zones around centre									
					1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
Co-exp	7171	254260	16	**TFRC**	575	2061	2621	789	205	45	6	8	6	Nodes
					367	**105**	**27**	6	6	4	5	6	9	Ave degree
					2	2	1	2	2	2	2	2	2	Min degree
					1228	**654**	**324**	46	32	20	8	12	14	Max degree
					0	0	1	0	0	0	0	0	0	# quills
Co-exp1	573	92575	4	**GP1BB**	527	41	4							Nodes
					345	**50**	**10**							Ave degree
					6	2	4							Min degree
					948	**222**	**14**							Max degree
					0	0	0							# quills
Co-exp11	527	89504	3	**SCNN1A**	464	62								Nodes
					367	**126**								Ave degree
					26	4								Min degree
					922	**354**								Max degree
					0	0								# quills
Co-exp111	464	80869	3	**GNAS, HAB1**	454	8								Nodes
					348	**203**								Ave degree
					24	96								Min degree
					847	**390**								Max degree
					0	0								# quills
Co-exp2	1790	79494	10	**PRR11**	314	634	550	137	27	2				Nodes
					271	**79**	**34**	18	3	2				Ave degree
					28	2	2	2	2	2				Min degree
					470	**372**	**270**	98	10	2				Max degree
					0	0	0	0	0	0				# quills
Co-exp21	314	37268	5	**FEN1**	162	117	34							Nodes
					294	**204**	**82**							Ave degree
					120	32	24							Min degree
					333	**292**	**156**							Max degree
					0	0	0							# quills
Co-exp211	162	23123	3	**32 GENES**	129	1								Nodes
					279	**2**								Ave degree
					118	2								Min degree
					319	**2**								Max degree
					0	0								# quills

that centres are from the same family cannot fully be established (table 13).

However, the gene regulatory networks we looked at, despite that the networks have small orders, show a much more pronounced articulation of the phenomenon (see supplementary tables S7 to S10).

Methods

Evaluation of biological networks as metric spaces. We considered human PINs (HFPIN, HSN, NHPIS) and gene co-expression and regulatory networks as metric spaces by defining the usual graph theoretic distance between nodes of a graph. Using a python wrapper around the C++ BOOST graph library (http://www.boost.org/), we used the Dijkstra algorithm to compute the shortest distances between *all pairs* of nodes and then identifyied the node or *all* nodes whose greatest distance to other nodes is/are smallest. This is the network center(s). From here, nodes were classified according to their distances from the centre and divided into zones based on distance from the topological centre(s). From each distance class, we calculated their degree distributions and also considered their connectivity of the graphs induced for each zone.

Pathway and function enrichment analysis. In order to determine whether zones of the human PINs we considered have biological significance, we divided proteins into subsets based on their distance from the true topological centre. Protein sets representing each zone were then subjected to a pathway over-representation analysis in order to determine whether the zones were specialised for specific functions. The Comparative Toxigenomics Databases Gene Set Enricher web service (http://ctdbase.org/tools/enricher.go and Gene Ontology enrichment (http://geneontology.org/page/go-enrichment-analysis) was used to perform the enrichment analysis and a corrected P-value of 0.01 was chosen as a statistical significance cutoff. Lastly, when such enrichment was observed, we calculated the proportion of proteins involved in each enriched pathway as a way to assess whether any zone displayed functional specialization.

Cancer gene expression data sources. We considered gene expression absence/presence calls from the following cancers types: breast, lung, kidney, pancreas, liver, cervix, ovary, glioblastoma, pituitary, glioma, fallopian, endometrium and rectum, which was downloaded from Gene Expression Barcode database (http://barcode.luhs.org/index.php?page=genesexp). Genes expressed in at least 99% of samples of a cancer of interest based on the Human HGU133 platform were downloaded. Gene expression was used as a proxy for protein expression and was mapped onto the PINs of interest in order to identify the zones in which gene product is located in.

Testing the difference between proportions. We performed a z-test for the difference between two population proportions p_1 and p_2. We identified the null and alternative hypotheses and we specified the level of significance to be $P < 0.01$. After that we determined the critical value(s) from the statistic table. Finally we found the standardized test statistic as shown below.

Statistical significance of the proportional analysis of pathway representation of zones. To test differences between proportions among zones, we need a statistical comparison of observed differences. A two-sample z-test for the differences between proportions for the top statistically enriched REACTOME pathways among zones was conducted. We defined the null hypothesis H_0 to be: classification proportions of zones in the periphery in human PINs have as high proportion significance as zones closest to the centre, i.e the accuracy of the sensing functions in zones closest to the centre and the accuracy of metabolic functions in zones in the periphery. If the $P < 0.01$, we rejected H_0 and concluded that the proportions support our claim that zones closest to the centre have high proportion significance than the zones in the periphery. In the other words, we have enough evidence at the 1% level to conclude that zones closest to the centre have high proportion significance than the zones in the periphery.

1. Havlin, S. *et al.* Fractals in biology and medicine. *Chaos, Solitons, and Fractals* **6**, 171–201 (1995).
2. Tao, S. & Zhang, Y. Self-similarity formed of complex networks. In *Circuits, Communications and Systems, 2009. PACCS'09. Pacific-Asia Conference on*, 155–158 (IEEE, 2009).
3. Serrano, M. A., Krioukov, D. & Boguná, M. Self-similarity of complex networks and hidden metric spaces. *Phy. Rev. Lett.* **100**, 078701 (2008).
4. Avnir, D., Farin, D. & Pfeifer, P. Molecular fractal surfaces. *Nature* **308**, 261–263 (1984).
5. Huet, S. *et al.* Relevance and limitations of crowding, fractal, and polymer models to describe nuclear architecture. *International Review of Cell and Molecular Bio* **307**, 443–79 (2014).
6. Song, C., Havlin, S. & Makse, H. A. Self-similarity of complex networks. *Nature* **433**, 392–395 (2005).
7. Zhou, W.-X., Jiang, Z.-Q. & Sornette, D. Exploring self-similarity of complex cellular networks: The edge-covering method with simulated annealing and log-periodic sampling. *Physica A: Statistical Mechanics and its Applications* **375**, 741–752 (2007).
8. Gallos, L. K., Song, C. & Makse, H. A. A review of fractality and self-similarity in complex networks. *Physica A: Statistical Mechanics and its Applications* **386**, 686–691 (2007).
9. Fadhal, E., Gamieldien, J. & Mwambene, E. C. Protein interaction networks as metric spaces: a novel perspective on distribution of hubs. *BMC Sys Bio* **8**, 6 (2014).
10. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47–C52 (1999).
11. Ouzounis, C. A. & Karp, P. D. Global properties of the metabolic map of escherichia coli. *Genome research* **10**, 568–576 (2000).
12. McAdams, H. H. & Arkin, A. Gene regulation: Towards a circuit engineering discipline. *Current Bio* **10**, R318–R320 (2000).
13. Elowitz, M. B. & Leibler, S. A synthetic oscillatory network of transcriptional regulators. *Nature* **403**, 335–338 (2000).
14. Savageau, M. A. Design principles for elementary gene circuits: Elements, methods, and examples. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **11**, 142–159 (2001).
15. Bolouri, H. & Davidson, E. H. Modeling transcriptional regulatory networks. *BioEssays* **24**, 1118–1129 (2002).
16. Hasty, J., McMillen, D., Isaacs, F. & Collins, J. J. Computational studies of gene regulatory networks: in numero molecular biology. *Nature Reviews Genetics* **2**, 268–279 (2001).
17. Guet, C. C., Elowitz, M. B., Hsing, W. & Leibler, S. Combinatorial synthesis of genetic networks. *Science* **296**, 1466–1470 (2002).
18. Newman, M. E. The structure and function of complex networks. *SIAM review* **45**, 167–256 (2003).
19. Spirin, V. & Mirny, L. A. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences* **100**, 12123–12128 (2003).
20. Fry, D. C. Protein–protein interactions as targets for small molecule drug discovery. *Peptide Science* **84**, 535–552 (2006).
21. White, A. W., Westwell, A. D. & Brahemi, G. Protein–protein interactions as targets for small-molecule therapeutics in cancer. *Expert reviews in molecular medicine* **10**, e8 (2008).
22. Strosberg, A. D. Protein–protein interactions as targets for novel therapeutics. *Drug Discov* (2007).
23. He, X. & Zhang, J. Why do hubs tend to be essential in protein networks? *PLoS Genetics* **2**, e88 (2006).
24. Jeong, H., Mason, S., Barabasi, A. & Oltvai, Z. Lethality and centrality in protein networks. *Arxiv preprint cond-mat/0105306* (2001).
25. Fadhal, E., Mwambene, E. C. & Gamieldien, J. Modeling human protein interaction networks as metric spaces has potential in disease research and drug target discovery. *BMC Sys Bio* **8**, 68 (2014).
26. Neph, S. *et al.* Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150**, 1274–1286 (2012).
27. Wu, G., Feng, X. & Stein, L. A human functional protein interaction network and its application to cancer data analysis. *Genome Bio* **11**, R53 (2010).
28. Li, L. *et al.* The human phosphotyrosine signaling network: evolution and hotspots of hijacking in cancer. *Genome research* **22**, 1222–1230 (2012).
29. Awan, A. *et al.* Regulatory network motifs and hotspots of cancer genes in a mammalian cellular signalling network. *IET Sys Bio* **1**, 292–297 (2007).
30. Zhang, Q. C. *et al.* Structure-based prediction of protein-protein interactions on a genomewide scale. *Nature* **490**, 556–560 (2012).
31. Obayashi, T. *et al.* Atted-ii: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in arabidopsis. *Nucleic acids research* **35**, D863–D869 (2007).
32. Staal, S. P. Molecular cloning of the akt oncogene and its human homologues akt1 and akt2: amplification of akt1 in a primary human gastric adenocarcinoma. *Proceedings of the National Academy of Sciences* **84**, 5034–5037 (1987).
33. Amado, R. G. *et al.* Wild-type kras is required for panitumumab efficacy in patients with metastatic colorectal cancer. *J. of Clinical Oncology* **26**, 1626–1634 (2008).

Acknowledgments

This work was supported in part by a grant from South Africa National Research Foundation (NRF).

Author contributions

E.F. implemented the algorithms, performed the analyses and drafted the original manuscript. E.C.M. proposed the concept of analyzing PINs as a self-similarity structure and oversaw the topological and statistical analyses. J.G. designed and oversaw and assisted in the functional evaluation tests and the biological interpretation of the results. E.C.M. and J.G. supervised the study and edited the manuscript. All authors have read and approved the final manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Fadhal, E., Gamieldien, J. & Mwambene, E.C. Self-similarity of human protein interaction networks: a novel strategy of distinguishing proteins. *Sci. Rep.* **5**, 7628; DOI:10.1038/srep07628 (2015).