nature COMMUNICATIONS

ARTICLE

Received 21 Jul 2014 | Accepted 6 May 2015 | Published 9 Jun 2015

DOI: 10.1038/ncomms8412

OPEN

Extreme multifunctional proteins identified from a human protein interaction network

Charles E. Chapple[1,2], Benoit Robisson[1,2], Lionel Spinelli[1,2,3,4,5], Céline Guien[1,2,†], Emmanuelle Becker[1,2,†] & Christine Brun[1,2,6]

Moonlighting proteins are a subclass of multifunctional proteins whose functions are unrelated. Although they may play important roles in cells, there has been no large-scale method to identify them, nor any effort to characterize them as a group. Here, we propose the first method for the identification of 'extreme multifunctional' proteins from an interactome as a first step to characterize moonlighting proteins. By combining network topological information with protein annotations, we identify 430 extreme multifunctional proteins (3% of the human interactome). We show that the candidates form a distinct sub-group of proteins, characterized by specific features, which form a signature of extreme multifunctionality. Overall, extreme multifunctional proteins are enriched in linear motifs and less intrinsically disordered than network hubs. We also provide MoonDB, a database containing information on all the candidates identified in the analysis and a set of manually curated human moonlighting proteins.

[1] Aix-Marseille University, TAGC, Marseille F-13009, France. [2] INSERM UMR_S1090, Marseille F-13009, France. [3] Aix-Marseille University, CIML, Marseille F-13009, France. [4] CNRS, UMR 7280, Marseille F-13009, France. [5] INSERM, U631, Marseille F-13009, France. [6] CNRS, Marseille F-13009, France. † Present address: Aix-Marseille University, GMGF, Marseille, F-13005 France; INSERM UMR_S910, Marseille F-13005, France (C.G.); Rennes I University, IRSET, Rennes F-35042, France; INSERM U1085, Rennes F-35042, France (E.B.). Correspondence and requests for materials should be addressed to C.B. (email: brun@tagc.univ-mrs.fr).

How common are multifunctional proteins? Is protein multifunctionality the exception or the rule? Is there a scale of multifunctionality? These questions are particularly relevant in the context of the C-value paradox, the fact that genome size does not correlate with organismal complexity[1]. Protein multifunctionality may be one of the ways a cell makes more with less.

Like gene/protein function[2], gene/protein multifunctionality can be defined at each of the different organizational levels, ranging from molecular to organismal. At the molecular level, the catalytic promiscuity of a single domain, that is, its ability to catalyse both a primary substrate-specific function and a different, secondary reaction, can account for multifunctionality. So does the presence of multiple different catalytic domains within the same protein[3]. At the cellular level, multifunctionality corresponds to gene products involved in multiple biological processes (BPs), often revealing regulatory roles. Finally, at the organismal level, pleiotropy is detected when changes in a single locus lead to multiple phenotypic effects.

Although protein multifunctionality is widely recognized and certain multifunctional proteins have been extensively studied (for example, TP53), multifunctional proteins as a *group* have not received much attention in the literature. As a result, multifunctional proteins suffer from a lack of characterization, unclear definitions and have only been studied individually, on a case by case basis. This may be partly due to the lingering effects of the one gene, one enzyme and by extension, one function hypothesis[4]. As proteins tend to be studied in terms of their known function, alternate functions—often serendipitously discovered—are arbitrarily relegated to secondary status.

The past few years have seen a growing interest in moonlighting proteins (MPs), a special sub-class of multifunctional proteins[3,5]. They are defined as 'special multifunctional proteins, because they perform multiple autonomous, often unrelated, functions without partitioning these functions into different protein domains'[6]. The human aconitase is an oft-cited example: an enzyme of the tricarboxylic acid cycle that turns into a translational regulator when the iron concentration changes[7]; as is hyaluronan-mediated motility receptor (HMMR), a nuclear microtubule-associated protein that, in certain cancers, is exported to the extracellular matrix where it binds CD44, ultimately promoting metastasis (reviewed in refs 8,9).

MPs may play important roles in normal or pathological contexts and their study could improve our understanding of complex genotype–phenotype relationships. However, owing to the serendipitous nature of their discovery, the number of proteins explicitly described as MPs is still very low, impairing our ability to define common characteristics on which a systematic search can be based. Consequently, there has been no large-scale attempt to identify them. In addition, although sequence analysis tools and computational domain predictions are very useful in establishing a protein's molecular functions, they rarely predict the cellular or physiological functions, and are therefore ill-suited for the identification of MPs. Furthermore, multifunctionality often blurs the possible functional inferences made from sequence similarity searches[10] and current algorithms rarely identify additional functions with a high confidence score[11], hindering conclusive predictions. Finally, that known MPs can switch between functions upon change of (i) subcellular localization, (ii) physicochemical environment, (iii) oligomeric state, (iv) interacting partners or a combination thereof[12] means that predicting alternate functions should go beyond sequence analysis.

Protein–protein interaction (PPI) networks (interactomes) highlight the modularity of cellular processes and allow deciphering protein functions at the cellular level. PPI networks represent the set of all detected interactions between each of their constituent proteins in a time and cell-type independent manner. All possible interactions of a given protein are thus shown simultaneously in the same network, therefore enabling the identification of proteins involved in different processes in different contexts, a prerequisite for the study of MPs. Indeed, MPs are expected to specifically interact with different sets of protein partners, either simultaneously or not, depending on the function performed.

We therefore reasoned that, as a first step towards the large-scale identification of MPs, PPI networks can be used to identify proteins whose multiple functions are very dissimilar to one another. Although these 'extreme multifunctional proteins' (EMFs) will not all adhere to the strict definition of MPs, detecting this form of multifunctionality is interesting in itself and, in addition, we can expect 'classical' MPs to be a subset of EMFs.

By associating network topological information with existing protein annotations, we have identified 430 such EMFs in the human interactome. We show that they form a distinct sub-group of proteins in the human network, characterized by specific features, which set them apart from other multifunctional proteins and which, when combined, form a signature of extreme multifunctionality. We also present MoonDB, a database containing diverse information on our set of 39 manually curated human 'known MPs' and on all the candidates identified in this analysis. Finally, we discuss the relationship between extreme multifunctionality and moonlighting in light of our results.

Results

Principle of inference for EMFs. Multifunctional proteins are expected to perform their different functions through different interaction partners. We therefore need to identify proteins at the intersection of sets of functionally related proteins. First, overlapping protein sets were identified in the human interactome using overlapping cluster generator (OCG)[13], an algorithm that covers a network with a system of overlapping clusters. These clusters are formed by highly interconnected proteins, which tend to be involved in the same cellular processes and may include protein complexes. We chose OCG because, as we have previously demonstrated[13], it is particularly well suited for the detection of multifunctional proteins and it fares better than other algorithms on sparse graphs such as PPI networks. Second, the cellular process(es) in which the clusters are involved were identified based on the BP Gene Ontology (GO)[14] annotations of their constituent proteins: GO terms annotating at least 50% of a cluster's proteins are assigned to the cluster, which can now be called a 'functional module'. Each individual protein then inherits the annotations of its module(s) in addition to its own. This annotation procedure favours the detection of functionally homogeneous clusters—likelier to represent functional modules—and maximize the number of clusters successfully annotated. Finally, to distinguish extreme from 'classical' multifunctional proteins, proteins found at the intersection of functional modules annotated to *dissimilar functions* were identified.

Function (dis)similarity is given by two metrics of GO term association (made available in the PrOnto database; see Methods and (http://tagc.univ-mrs.fr/pronto)) based on the frequency of co-occurrence of a GO term pair either in a protein's or in a pair of interacting partner's annotations. Using these metrics ensures that the multiple functions in which the candidate protein is found to be involved are very rarely performed (i) by a single protein and (ii) by interacting proteins, two proxies that we consider indicators of *unrelated functions*. Our pipeline (MoonGO) is explained in more detail in the Methods section and is summarized in Fig. 1.

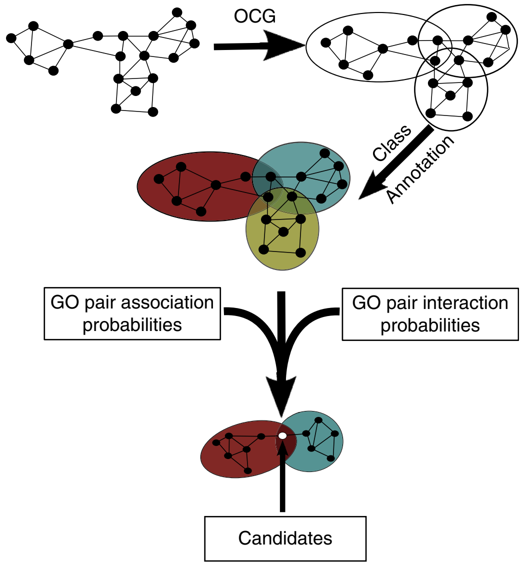
OCG
Class
Annotation
GO pair association probabilities
GO pair interaction probabilities
Candidates

Figure 1 | MoonGO: the EMF identification pipeline. Overlapping clusters are extracted from a PPI network using OCG. Clusters are annotated according to the GO annotations of their constituent proteins. Potential EMFs are then identified at the intersection of clusters involved in unrelated biological processes according to PrOnto GO term association probabilities.

A set of 430 EMF candidates. We have applied our pipeline to a large, high-quality human interactome (74,388 interactions between 12,865 nodes, Supplementary Data 1) built by extracting data from online databases (see Methods). The 855 overlapping clusters returned by OCG contained 33.4 proteins on average. Of these, one or several BPs were assigned to 633 (74%), based on the annotation of their constituent proteins (Supplementary Data 2). All network proteins belonged to at least one annotated cluster. As expected from a previous analysis, around a third of the interactome (3,846 proteins, 29%) belonged to several clusters and can therefore be considered multifunctional[13]. Of these, 430 proteins (10%) are found at the intersection of clusters annotated to dissimilar functions and are considered EMF candidates (Supplementary Data 3).

As candidates are defined with respect to their module's annotations, their identification is dependent on the quality of these annotations. Said quality was assessed by performing three types of randomization tests. First, the annotations of all proteins were shuffled, and candidates were identified using these randomized annotations. Over 100 trials, on average only 104.78 of 855 clusters were annotated compared with 633 for the real data, demonstrating that such high number of functionally homogeneous clusters cannot be found by chance. Consequently, the number of identified candidates markedly decreased in these conditions (7.55 versus 430 for the real data). Second, the network topology was randomized by drawing random edges between the nodes. As expected, since this destroys the modular structure of the network, virtually no candidates were found. The process was repeated ten times and only 0.4 candidates were identified on average, clearly showing that our results cannot be found in a random network. Third, the topology was randomized while keeping the same degree distribution. In other words, the number of nodes with a given degree was the same but which node had which degree was randomized. The process was repeated ten times and our pipeline was applied to these randomized networks. As expected, we found almost no results: only 26.8 EMF candidates were identified on average.

Table 1 | Dissimilar biological processes.

%	Dissimilar biological processes	
53.5	Nitrogen or nucleic acid metabolism	Signalling
7.3	Nucleic acid metabolism	Localization
6.8	Macromolecular metabolic process	Transport
5.9	RNA metabolic process	Signalling
4.7	Nucleic acid metabolism	Transport
4.5	Gene expression	Transport
3.6	Macromolecular metabolic process	Localization
2.9	Nucleic acid metabolism	Physiological processes
2.7	Nucleic acid metabolism	Phosphorus metabolism
1.2	Nucleic acid metabolism	Development

Top 10 dissimilar function pairs by the percentage of candidates identified.

Candidate identification also depends on the annotation probabilities establishing the functional dissimilarity of modules. We therefore recalculated the number of candidates when reshuffling the probabilities of association between GO term pairs. An average of 1.03 candidates were identified over 100 runs, once more indicating that our results cannot be obtained by chance.

The validity of the functional module approach was verified by demonstrating that the GO terms that led to the candidates' discovery are not among their existing annotations but were brought in by the annotation inheritance process depending on their module membership. Indeed, only 71 candidates (17%) were already annotated to both terms used to identify them, 209 (51%) were annotated to one term and inherited the other from their modules' annotations, whereas 128 (31%) were not annotated to either term, therefore inheriting both terms from their modules. This further confirmed the power of interaction network analyses for function prediction.

Finally, the quality of the inferred annotations was assessed using a leave-one-out approach. For each of the clusters identified, we removed one of its proteins' annotations, annotate the cluster and infer the cluster's annotations to the protein. Doing so, we assign at least one of the known annotations correctly in 62.6% of cases, indicating that our approach is capable of rescuing known protein annotations, therefore suggesting that the novel annotations we infer are trustworthy.

As MPs are a subset of EMFs, we compiled a list of 39 known human MPs from the literature and checked whether they were found as candidates. Six of the thirty-nine belonged to dissimilar modules and were found. Although finding 6 out of 39 when identifying 430 candidates out of 12,865 represents an enrichment of 4.6-fold compared with expected, with a significant *P*-value ($1.4e^{-3}$, hypergeometric), 6 of the remaining 33 could not be found by MoonGO since they belonged to clusters that could not be annotated. We cannot thus exclude that the remaining 27 proteins were missed because of ill-annotated clusters or because all their interactions have not been discovered yet. Overall, these different assessments confirmed the specificity of our approach and consolidated our confidence in the identified candidates.

The candidates were found linking 141 different pairwise function combinations between 55 different GO terms. These annotation pairs are functionally dissimilar by both annotation and interaction probabilities (Supplementary Data 4). Most of the candidates (>90%) were found annotated to dissimilar functions involving on the one hand nucleic acid-linked metabolic processes and on the other (i) signalling activity, (ii) localization or (iii) transport (see Table 1 and Supplementary Data 4).

Candidate characterization. To investigate whether EMF candidates (Cands) form a distinct group of proteins with respect to the other proteins of the network, the candidates were analysed to identify common trends and features that characterize them (Table 2). For each characteristic studied, they were compared with several categories: (i) the entire network's proteins, (ii) proteins that belong to multiple clusters but are not candidates as those clusters are not annotated to significantly different GO terms (Multi-non-candidate (NC)), (iii) all multi-clustered proteins (Multi), (iv) proteins belonging to a single cluster (Mono), (v) all NC proteins and (vi) network hubs, defined here as those nodes whose degree is at least twice the network average (≥25). Note that these categories are not all mutually exclusive. Since the candidates are defined as those proteins found at the intersection of modules annotated to dissimilar terms, they are all, by definition, multiclustered. Therefore, to identify characteristics common to EMFs that distinguish them from other multifunctional proteins, features shared by candidates but not by the Multi-NCs need to be found.

Table 2 | Candidate features.

Candidates compared with:	Multi-NC	Hubs	Network
Annotations (BP)	⇑	⇑	⇑
Betweeness	⇑	⇑	⇑
Clusters	⇑	⇑	⇑
Conservation	⇑		⇑
Degree	⇑	⇑	⇑
Disorder		⇓	
ELMs	⇑		⇑
Domains	⇑		⇑
Expression	⇑		⇑
Isoforms			⇑
Length			⇑
OMIM	⇑	⇑	⇑
Kinases			⇑
Shortest Paths	⇓	⇑	⇓

BP, Biological Process; ELM, eukaryotic linear motif; OMIM, online mendelian inheritance in man; NC, non-candidate.

Comparison of the mean values of different features of candidate proteins as compared with Multi-NC, hubs and the network as a whole. ⇓ Indicate that candidates had significantly lower mean values than the compared group and ⇑ significantly lower. Cells are left blank when there was no significant difference in means. The Wilcoxon test with a significance threshold of ≤0.05 was used to determine significance. 'Kinases' refers to the number of different kinases predicted to phosphorylate the protein. For details, see Methods.

Candidate network topological features. Candidates have a significantly higher degree than the Multi-NCs. On average, they interact with 74.6 proteins compared with 21.9 for the Multi-NCs (Wilcoxon test *P*-value = 1.27e-96; Fig. 2a). They consequently belong to more network clusters (mean Cands (15.2) versus Multi-NCs (3.8), Wilcoxon test *P*-value = 1.15e-150; Supplementary Fig. 1) and are significantly more central to the network than Multi-NCs (mean Cands (235,005.6) versus Multi-NCs (46,177.2), Wilcoxon test *P*-value = 1.44e-112) as shown by their betweeness scores (Supplementary Fig. 2), a measure of how central a given node is in the network, calculated by quantifying the number of times a node acts as a bridge along the shortest

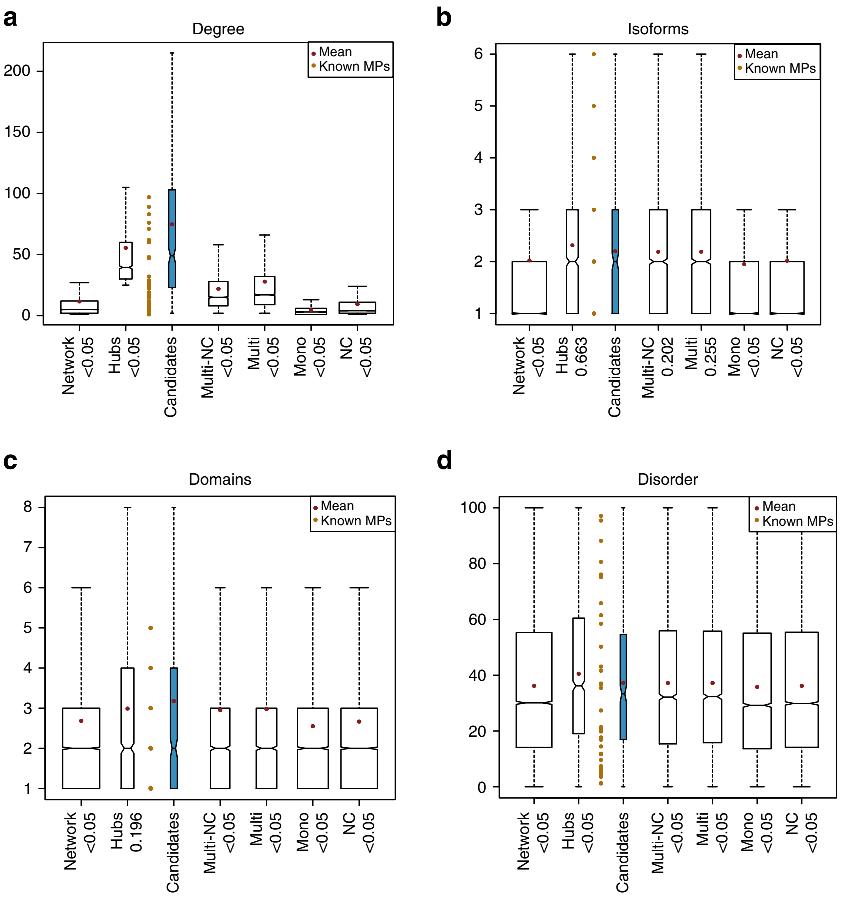

Figure 2 | Protein features. (**a**) Protein degree. (**b**) Protein isoforms. (**c**) The number of Pfam domains (including PfamB) predicted on each protein. (**d**) Protein disorder as calculated by disopred. The numbers shown are the percentage of a protein's residues that are disordered. Outliers are not shown. Red dots indicate mean values and the camel dots between Candidates and Hubs are the values of the known moonlighting proteins.

path between two other nodes. Not surprisingly, candidates are also more connected to each other according to a shortest path analysis than the Multi-NCs (Supplementary Fig. 3).

Note that although candidates tend to be more connected than hubs (mean Cands (74.6) versus Hubs (55.5)) not all candidates have a high degree (see Fig. 2a) and only 20% of network hubs were found as candidates, demonstrating that a high degree is neither necessary nor sufficient for consideration as a candidate.

Candidate sequence and structural features. Protein interaction and functional annotation databases tend to refer to genes rather than gene products and do not differentiate between different protein isoforms[15]. It is therefore possible that the multiple functions of our candidates are actually carried out by different isoforms of the same gene. However, candidates do not have significantly more isoforms than the Multi-NCs (mean Cands (2.2) versus Multi-NCs (2.1), Wilcoxon test *P*-value = 0.2; Fig. 2b), suggesting that the number of isoforms is not a defining characteristic. As expected for highly multifunctional proteins, candidates have more domains than the Multi-NCs (mean Candidates (3.2) versus Multi-NCs (3.0), Wilcoxon test *P*-value = 0.0001, Fig. 2c). They are not, however, significantly longer (mean Cands (664.7) versus Multi-NCs (614.4), Wilcoxon test *P*-value = 0.09, Supplementary Fig. 4).

The candidates were also, like the hubs, more conserved than the Multi-NCs (mean Cands (43.7) versus Multi-NCs (42.6), Wilcoxon test *P*-value = 0.007; see Methods and Supplementary Fig. 5).

As structural disorder can allow conformational changes, we used DISOPRED[16] to predict disordered residues and analysed both the percentage of disordered residues per protein (Fig. 2d) and the number of stretches of consecutive disordered residues of different lengths (Supplementary Fig. 6). Interestingly, although candidates are not distinguishable from either Multi-NCs or the network average, they are significantly less disordered than hubs (mean Cands (37.3) versus Hubs (40.5), Wilcoxon test *P*-value = 0.028, see Fig. 2d). This trend was confirmed by the results of ten other disorder predictors whose results were retrieved from the D2P2 database[17] (see Supplementary Figs. 7–16). These results suggest that, despite their high average degree, EMFs, unlike regular hubs, are under stronger selective pressure to maintain a stable secondary structure.

Eukaryotic linear motifs (ELMs) are short stretches of amino acids often located within intrinsically disordered regions, which have been shown to help the targeting of proteins to specific subcellular localizations, determining the modification state of proteins, or regulating protein activity in a context-dependent manner[18]. We checked their numbers in proteins of each group, particularly in their disordered regions. Interestingly, candidates are enriched in ELMs per residue compared with the Multi-NCs (mean Cands (0.0051) versus Multi-NCs (0.0035), Wilcoxon test *P*-value = 3e-4, Fig. 3), a trend which is more pronounced when considering only ELMS that fall within disordered regions (mean Cands (0.009) versus Multi-NCs (0.006), Wilcoxon test *P*-value = 8e-4, Supplementary Fig. 17). Notably, ELMs are more enriched among disordered residues in candidates (1.8-fold, mean per disordered residue (0.009) versus per residue overall (0.005), Wilcoxon test *P*-value = 0.03) than in Multi-NCs (1.5-fold, mean per disordered residue (0.006) versus per residue overall (0.004), Wilcoxon test *P*-value = 0.29). These results indicate that candidates contain more ELMs, particularly in disordered regions. With respect to hubs, the same trend was observed, although not statistically significant. It therefore appears that candidates differ from hubs in their disorder content and from Multi-NCs in the number of linear motifs per residue.

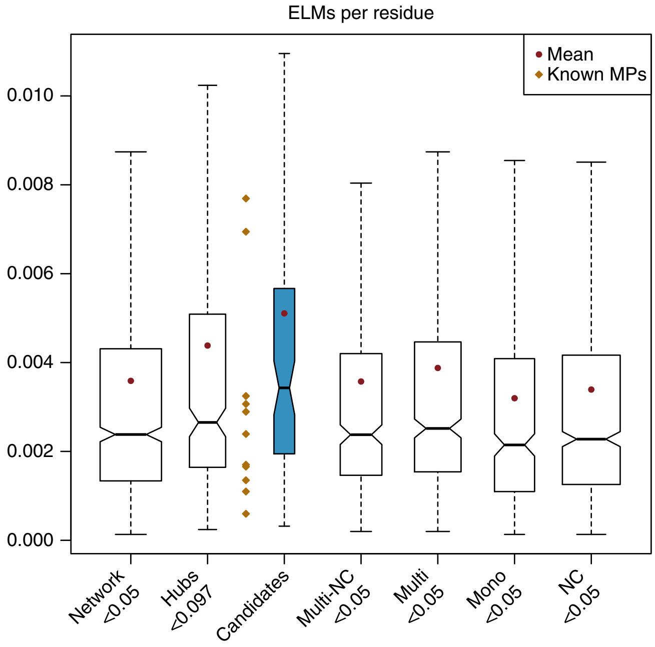

Figure 3 | ELMs per residue. The plots show the number of ELMs divided by the length of each protein.

Finally, the different groups were also checked for the presence of proteins identified as containing ELMs involved in functional switches, the status of which ultimately affects the function of the ELM-containing protein (collected in the switchELM database[19]). Candidates show a 6.2-fold enrichment in such proteins (hypergeometric *P*-value = 1.2e-27), compared with a 1.8-fold among Multi-NCs (*P*-value = 2.44e-14), 4.68-fold for hubs (*P*-value = 1.07e-62) and a depletion in Mono (2.7-fold less, *P*-value = 2e-49). When different types of functional switches are considered, the candidates are enriched in proteins containing binary switch motifs (with an ON/OFF state) modulated by allosteric effects (1.6-fold, *P*-value = 4.85e-2), compared with Multi-NCs and hubs, which show no enrichment. As numbers are low (27 such proteins in the interactome, 9 of which are EMF candidates, *P*-value = 4.85e-2), we cannot reasonably extrapolate this observation to the complete EMF data set. However, this particular finding combined with the higher occurrence of ELMs in the candidates reinforces their functional significance as extreme multifunctional and potential MPs.

Candidate annotations and expression. The candidates have significantly more BP annotations (mean = 16.8, Supplementary Fig. 18) than both hubs (mean = 13.0, Wilcoxon test *P*-value = 0.00014) and Multi-NCs (mean = 9.3, Wilcoxon test *P*-value = 1.56e-21), as expected for proteins involved in multiple functions. This is not introducing a bias in the analysis as only 17% of the candidates were already annotated to the dissimilar GO pairs used to identify them as candidates (see ‘A set of 430 EMF candidates’). Finally, the candidates are more ubiquitously expressed at the mRNA level (mean = 24.8 tissues, Supplementary Fig. 19) than Multi-NCs (mean = 19.8 tissues, Wilcoxon test *P*-value = 5.53e-05), raising the possibility that their different functions could be performed in different tissues.

Candidate involvement in disease. Multifunctional proteins in general and EMFs in particular are expected to be involved in disease since impairing their function can affect multiple cellular processes. We therefore used online mendelian inheritance in man (OMIM)[20] annotations to test our candidate's involvement in human diseases and found a total of 113 out of 430 candidates associated with 229 different diseases.

There was a 7.6-fold overrepresentation of disease-associated proteins among candidates (hypergeometric P-value = 2.1e-07), but only a 6.2-fold and 5.8-fold in all Multis (hypergeometric P-value = 9.3e-19) and hubs (hypergeometric P-value = 0.0003), respectively. As proteins that are involved in multiple processes are more likely to cause disease when perturbed, these results support our claim that multi-clustered proteins are likely multi-functional and reinforce the differences between candidates and hubs.

We also used the list of 435 cancer genes from ref. 21 to check for overrepresentation of cancer-associated genes among candidates. Once more, we found that those are clearly overrepresented (3.8-fold, hypergeometric P-value = 6.8e-2) among the candidates, whereas their overrepresentation among all Multis (twofold, hypergeometric P-value = 2.7e-38) and hubs (threefold, hypergeometric P-value = 2.8e-43) is less important. This suggests that although all multi-clustered protein groups are enriched in cancer genes, candidates are once more different from other multi-clustered proteins and hubs.

A signature of EMFs. The signature was built by combining all tested features that displayed a significant statistical difference between the candidates and the Multi-NCs on the one hand and the hubs on the other.

Our analyses therefore describe a first set of characteristics (summarized in Fig. 4) of EMFs that differentiates them from other multi-functional proteins. They tend to have more interactors, to belong to more clusters, to be more central and more connected to each other in the network; they also have more annotations, more domains are more conserved and contain more linear motifs. They have a greater tendency to be involved in disease and tend to be expressed more ubiquitously.

Another set of characteristics was defined with respect to hubs. Candidates tend to have more interactors, to be more central to the network although less connected to each other and to belong to a greater number of network clusters. They are more likely to be involved in disease and have more BP annotations. Very interestingly, they tend to be less disordered than hubs, with the same average disorder as the network.

To ensure that these signatures are not influenced by highly studied proteins (such proteins often have an artificially high degree in PPI networks because their interactions have been exhaustively characterized), we repeated the analysis on a smaller human PPI network built exclusively from large-scale yeast two-hybrid data (CCSB network[22]). Despite the much smaller size of this network (15,617 interactions between 4,494 proteins), we could still observe the same global trends in the 43 candidates found by our pipeline for most of the features of the signatures (Supplementary Figs 20–31). Although the low number of candidates kept most of these observations below the significance threshold, it is interesting to note that the 43 candidates were still significantly less disordered than the hubs (mean Cands (36.0) versus hubs (47.5), Wilcoxon test P-value = 0.01). As these results were obtained in a bias-free network, they reinforce the robustness of our findings on the large interactome.

Example candidates. Although a discussion of each of our candidates is clearly beyond the scope of a single paper, we highlight a few particularly interesting cases here. Note that the cluster's Cellular Component (CC) annotations shown below are only indicative and were not used in the prediction of candidates.

Receptor tyrosine-protein kinase erbB-2 (ERBB2) is a member of the epidermal growth factor receptor family and an essential component of the neuregulin-receptor complex, which regulates outgrowth and stabilization of peripheral microtubules[23]. Besides its signalling role, ERBB2 is also a transcription factor involved in the transcription of rRNA genes by RNA Pol I (ref. 24). We find ERBB2 at the intersection of two clusters, one annotated to 'cellular nitrogen compound metabolic process' (BP, GO:0034641), a parent term of 'transcription, DNA-templated', and to 'cytosol' and 'nucleus' (CC), the other annotated to 'signal transduction' (BP, GO:0007165) and 'plasma membrane' (CC). Our method, therefore, correctly identified ERRB2 as an EMF and assigned it to its different and unrelated real functions.

Protein RPP40, a component of the nuclear ribonuclease P, known to cleave the 5' end of tRNA molecules during their processing, was found at the intersection of a cluster annotated to 'cellular nitrogen compound metabolic process' (BP, GO:0034641)—a parent term of its *bona fide* annotation 'tRNA processing'—and 'nucleus' (CC), and another annotated to 'signal transduction' (BP, GO:0007165) and 'plasma membrane', and

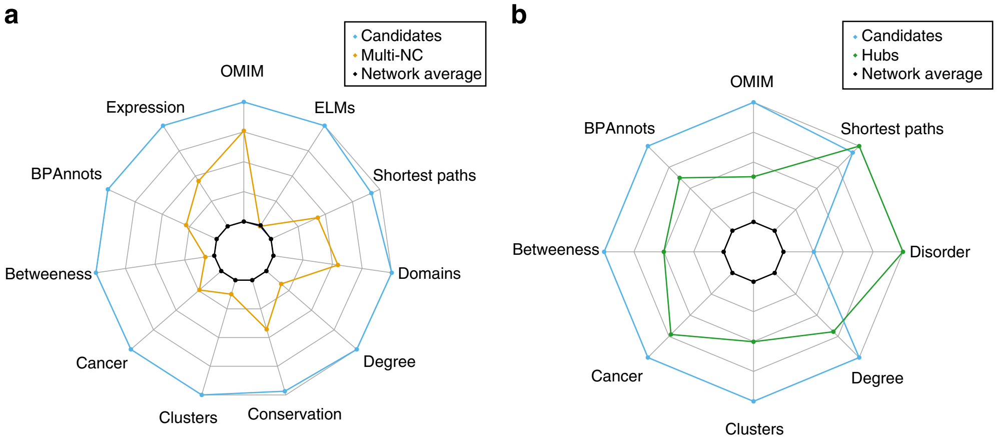

Figure 4 | Radar plots. Radar plots showing the characteristics that were significantly different in candidates with respect to Multi-NCs (**a**) and hubs (**b**). Mean values are plotted for all features except association with cancer, where fold overrepresentation is shown. For Shortest Paths, the outermost data point is the *most* connected, that is, has the *smallest* Shortest Path value. For all others, the outermost data point is the one with the greatest value. These plots are purely descriptive and were built once the candidates were found. The features they describe were not used to identify the candidates.

'cytosol' (CC), suggesting a possible signaling role for this nuclear protein. Interestingly, the mitochondrial counterpart of RNase P assumes the same function, although formed by three proteins not related at the sequence level to the nuclear form. It has been proposed that this mtRNase P complex 'was not built simply from components of a preexisting nucleolytic pathway but by combining components from different, essentially unrelated biochemical pathways'[25], which would strongly suggest possible moonlighting functions for the proteins dedicated to this cellular process.

WBP4 is a spliceosome-associated protein, which promotes pre-mRNA splicing. It is found at the intersection of a cluster annotated to 'RNA splicing, via transesterification reactions with bulged adenosine as nucleophile' (BP, GO:0000377) and 'nucleoplasm' (CC), which would correspond to its known function, and another annotated to 'response to endogenous stimulus' (BP, GO:0009719) and 'cytosol','nucleoplasm part' and 'intracellular non-membrane-bounded organelle' (CC), suggesting a novel role for this protein. Although no involvement in a signalling pathway has been shown to date, the protein contains two WW domains able to interact with proline- or phosphoserine-phosphothreonine-containing motifs and known to mediate regulatory interactions in various signalling pathways such as Hippo[26].

MoonDB. We have collected our results in MoonDB, a database that includes the 39 human MPs used here as well as the full list of EMF candidates. For all proteins, MoonDB provides easy access to diverse information (sequence, domain organization, functional annotations, involvement in disease and so on). In addition, for the candidates identified here, information about the functional modules from the graph and the functional dissimilarity of the GO pairs, which allowed their identification, are provided. MoonDB is available at http://tagc.univ-mrs.fr/MoonDB.

Discussion

What are MPs except proteins whose different functions have been serendipitously discovered and whose current definition has been molded to fit them? Do MPs really form a protein class unto themselves, with their own as yet undiscovered characteristics? Are they proteins that play truly different roles in the cell, or do we consider these roles different because we have not discovered their functional links yet?

In this context, we have chosen to avoid the term *moonlighting* here, largely because the current definition is too stringent. According to the primary definition[5], a multifunctional protein must not partition its functions into different domains to be considered moonlighting. However, it is very likely that when no second domain is identified *in silico*, it is simply because the domain in question is either unknown or below detection thresholds. To establish that a protein's two functions are performed by the same domain would therefore require experimental analyses. Human protein XRRC5, for example, was known to be involved in DNA repair yet was found to interact with metalloproteinases of the extracellular matrix[27]. It was only after this discovery that the protein was shown to contain a vWF domain whose sequence had diverged to the point that it was below the detection thresholds. We have, therefore, coined the term EMFs to describe proteins whose multiple functions are very different to one another. We feel that such proteins are of interest irrespective of whether they split their functions across multiple domains or whether their functions are independent. In addition, some multifunctional proteins are not considered moonlighting simply because their alternate functions are very well known. TP53, for example, fits the definition of MPs yet has never been considered as one. This is just one of many such 'sunlighting' proteins, which, were they to be discovered today, would be classed as moonlighting. EMFs include both classes of protein.

We have previously shown that using algorithms like OCG that is able to cluster proteins into multiple graph modules allows the identification of multifunctional proteins[13]. These modules correspond to the functional units of the network, are composed of groups of highly connected proteins involved in the same cellular function[28] and permit function prediction when containing uncharacterized proteins[29,30]. In the present work, we have used the functional module approach to identify those multifunctional proteins that exhibit extreme multifunctionality. This was made possible by the tailored development of the MoonGO pipeline, which uses an original function dissimilarity measurement, the PrOnto probabilities. The latter are based on the frequency of co-occurrence of GO term pairs in protein annotations or among the annotations of interacting protein pairs and are specific to the species being studied. These probabilities reflect biases towards infrequent association implying functional dissimilarity or, conversely, frequent association indicating functional similarity. Finally, although there have long been calls for PPI network analyses for the identification of highly multifunctional and MPs[31], such an extensive and global approach has never been undertaken. We show here that EMF candidates can be identified without *a priori* by the combination of interaction data analysis and the processing of current knowledge on protein function such as GO annotations.

Our method identified 3.3% of the human interactome (430 of 12,865 proteins) as candidate EMFs. Although an estimation of the expected proportion of EMFs is rather difficult, it is of note that our candidates from the bias-free CCSB network represent a similar proportion (1%), another argument confirming the significance of our results.

We have used the different characteristics shared by this group of proteins to define a signature of extreme multifunctionality, which distinguishes them from other network proteins. Importantly, this signature is reinforced by the fact that the same trends were observable in the candidates identified in the much smaller CCSB network, which is free of the biases associated with highly studied proteins.

Although EMFs are in many ways similar to hubs (degree of at least twice the network average, ≥ 25) and have many of the same characteristics, not all EMFs are hubs. It is particularly interesting that EMFs tend to be less disordered than hubs given that intrinsic protein disorder can enable proteins to adopt different conformations that can assist EMFs in their multiple functions (as already suggested for MPs[32]). Furthermore, although intrinsic disorder has been shown to be important for hub's multiple interactions (for example, see refs 33,34), candidates are significantly less disordered than hubs (see Fig. 2d) despite 361 of them (88%) being hubs. Disorder therefore appears to be important only for a specific subset of hubs, those which are not EMFs. That EMFs are just as disordered as the network average extends the observation made by Hernandez *et al.*[35] on a small number of known MPs to the broader class of EMFs.

So, what makes EMFs special compared with other multifunctional proteins? In terms of network topology, a typical EMF is likely to have a higher degree, to belong to more network modules and to be more central to the network. It is more likely to be involved in multiple diseases and to be expressed more ubiquitously, suggesting that it can perform different functions in different tissues. It will also have more domains, be more conserved than a classical multifunctional protein, and contain more short linear motifs (ELMs). These ELMs are short conserved sequences mostly located in disordered regions.

They form low-affinity interaction interfaces, are involved in transient interactions and, importantly, mediate molecular decision-making in cell regulation[19,36,37]. That ELMs (i) can bind competitively or sequentially to different interaction partners in a context-dependant manner, (ii) provide a large panoply of conditional regulatory types through interactions[19] and (iii) are more numerous in EMFs, provides a possible molecular explanation of the functional versatility of these proteins. This clearly calls for further studies.

Overall, the signatures we defined clearly show that EMFs form a distinct subgroup of multifunctional proteins exhibiting characteristics that distinguish them from hubs, classical multifunctional proteins and the network in general and can pave the way towards a better understanding of protein moonlighting.

Methods

Networks. Interaction data were retrieved using the PSIQUIC[38] interfaces of the APID[39], BioGrid[40], IntAct[41], DIP[42], MINT[43], MatrixDB[44], Reactome[45], InnateDB[46], MolCon, Spike[47] and TopFind[48] databases. They were filtered according to their identification methods and only binary interactions between proteins were kept. Protein names were mapped to UniProt IDs, and sequences clustered using CD-HIT[49]. TrEMBL/SwissProt protein pairs sharing $\geq$95% similarity were considered to be the same protein: interactions of the TrEMBL protein were then inherited by the Swiss-Prot protein. Self interactions were discarded. The final result was high-quality interactomes consisting entirely of experimentally verified, direct, binary interaction pairs.

The CCSB network was downloaded from the CCSB Human Interactome database[22].

Cluster Identification and annotation. Clusters were generated using OCG[13] and default options. The clusters were annotated according to the BP GO annotations of its constituent proteins. A cluster will be annotated to a GO term *iff* $\geq$50% of annotated proteins in that cluster share that GO term and all member proteins will inherit the annotation(s) of the cluster. Both direct GO annotations and all parent terms are taken into account. Note that those clusters that can only be annotated to the root of the ontology (that being the only term shared by $\geq$50% of their constituent proteins) are given the annotation 'BP unknown'. Because the quality of computationally inferred GO annotations has now been shown to rival that of curated non-experimental annotations[50], we use all BP GO annotations in this study, irrespective of their evidence codes.

Function association probabilities. We have developed two measures of GO term functional similarity, one (the annotation probabilities) measures whether two GO terms are found annotating the same protein more often than expected by chance. The second (interaction probabilities) measures whether there are more interactions between proteins annotated to GO term X and proteins annotated to GO term Y than would be expected by chance.

For both metrics, we have calculated the probability of association between two GO terms GO_1 and GO_2 using the hypergeometric distribution $\mathscr{H}$ (N, K, n, k), where, for the annotation probabilities, N is the number of proteins with at least two different direct annotations, K is the number of proteins directly annotated to GO_1, n is the number of proteins annotated to GO_2 and k is the number of proteins annotated to both terms.

For the interaction probabilities, N is the number of interactions in our network between proteins with at least two different annotations, K is the number of interactions involving proteins annotated to GO_1, n the number of interactions involving proteins annotated to GO_2 and k the the number of interactions between a protein annotated to GO_1 and one annotated to GO_2. To be considered dissimilar by our pipeline, two GO terms must have a probability of $\leq$0.05 for both metrics.

We have made these probabilities available in the PrOnto database, which is freely available at http://tagc.univ-mrs.fr/pronto/. Note that although PrOnto offers cross-ontology probabilities, only the BP probabilities were used in the present work to identify candidates.

A more detailed explanation of our choice of developing PrOnto rather than using existing semantic similarity measures and a comparision of PrOnto and Semantic Similarity measures are provided as a Supplementary Note 1.

MoonGO. MoonGO, our EMF prediction tool, uses the annotated clusters and PrOnto probabilities to search the network for proteins found connecting clusters annotated to dissimilar BP GOs. It then looks for nodes that are members of both clusters (see Fig. 1)

As the number of GO term pairs analysed is very large (106618), MoonGO will correct for multiple testing by multiplying the P-value of association between two terms by the number of tests performed to obtain a corrected e-value. MoonGO uses both the annotation and the interaction probabilities to identify its candidates. A candidate is kept if both the annotation and interaction e-values of the BP GO term pairs associated with it are significant. For the work presented here, we have used an e-value significance threshold of $\leq$0.05.

Analyses. The betweeness and shortest path analyses were done using the R igraph library[51]. Domain predictions were made using pfam_scan.pl[52], the results shown in Fig. 2c include both PfamA and PfamB domain predictions. Protein disorder values were obtained using disopred[16] with default settings. Figure 2d was generated by plotting the sum of disordered residues of each protein. Protein isoform information and OMIM annotations were downloaded from UniProt and protein annotations from the EBI's QuickGO server (https://www.ebi.ac.uk/QuickGO/). Expression data were taken from ref. 53. Protein phosphorylation predictions were made using GPS 2.1 (ref. 54).

Conservation. The protein sequences of all network proteins and their annotated homologues in yeast, mouse, fly and worm were aligned against each other using t_coffee[55]. The homologous sequences were taken from EnsEMBL[56]. To obtain an indicative conservation value, the bit score of the alignments was divided against the length of the human homologue and the resulting values were compared.

ELMs. The coordinates of all annotated ELMs in our network's protein were obtained from the ELM database[57]. These were mapped to the disordered regions predicted by DISOPRED. Only ELMs that fell entirely within or which overlapped perfectly with a disordered region were counted. ELMs that overlapped but extended beyond the disordered region were ignored. To avoid misrepresentations caused by differing protein or disordered region lengths, the number of ELMs in disordered regions was divided by the total number of disordered residues of each protein. Similarly, the number of ELMs per protein was divided by the total number of residues of each protein.

Known MPs and MoonDB database. The MoonDB database was written using a combination of HTML 4.01, PHP 5.5 and Javascript, the data are stored in a MySQL 5.5 database. We have compiled a manually curated list of *bona fide* human MPs, which we have made available, along with our own predictions, at the MoonDB database.

References

1. Doolittle, W. F. Is junk dna bunk? a critique of encode. *Proc. Natl Acad. Sci. USA* **110,** 5294–5300 (2013).
2. Jacq, B. Protein function from the perspective of molecular interactions and genetic networks. *Brief. Bioinform.* **2,** 38–50 (2001).
3. Copley, S. D. Moonlighting is mainstream: paradigm adjustment required. *Bioessays* **34,** 578–588 (2012).
4. Tatum, E. L. & Beadle, G. W. Genetic control of biochemical reactions in neurospora: an 'aminobenzoicless' mutant. *Proc. Natl Acad. Sci. USA* 234–243 (1942).
5. Jeffery, C. J. Moonlighting proteins. *Trends Biochem. Sci.* **24,** 8–11 (1999).
6. Huberts, D. H. E. W., Venselaar, H., Vriend, G., Veenhuis, M. & van der Klei, I. J. The moonlighting function of pyruvate carboxylase resides in the non-catalytic end of the tim barrel. *Biochim. Biophys. Acta* **1803,** 1038–1042 (2010).
7. Volz, K. The functional duality of iron regulatory protein 1. *Curr. Opin. Struct. Biol.* **18,** 106–111 (2008).
8. Maxwell, C. A., McCarthy, J. & Turley, E. Cell-surface and mitotic-spindle rhamm: moonlighting or dual oncogenic functions? *J. Cell Sci.* **121,** 925–932 (2008).
9. Jiang, J. *et al.* Multifunctional proteins bridge mitosis with motility and cancer with inflammation and arthritis. *Sci. World J.* **10,** 1244–1257 (2010).
10. Gómez, A., Domedel, N., Cedano, J., Piñol, J. & Querol, E. Do current sequence analysis algorithms disclose multifunctional (moonlighting) proteins? *Bioinformatics* **19,** 895–896 (2003).
11. Khan, I., Chitale, M., Rayon, C. & Kihara, D. Evaluation of function predictions by pfp, esg,and psi-blast for moonlighting proteins. *BMC Proc.* **6,** S5 (2012).
12. Jeffery, C. J. Proteins with neomorphic moonlighting functions in disease. *IUBMB Life* **63,** 489–494 (2011).
13. Becker, E., Robisson, B., Chapple, C. E., Guénoche, A. & Brun, C. Multifunctional proteins revealed by overlapping clustering in protein interaction network. *Bioinformatics* **28,** 84–90 (2012).
14. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genet.* **25,** 25–29 (2000).
15. Talavera, D., Robertson, D. L. & Lovell, S. C. Alternative splicing and protein interaction data sets. *Nature Biotechnol.* **31,** 292–293 (2013).
16. Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F. & Jones, D. T. The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20,** 2138–2139 (2004).
17. Oates, M. *et al.* D2P2: database of disordered protein predictions. *Nucleic Acids Res.* **41,** D508–D516 (2013).

18. Van Roey, K. *et al.* Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chem. Rev.* **114,** 6733–6778 (2014).
19. Van Roey, K., Dinkel, H., Weatheritt, R. J., Gibson, T. J. & Davey, N. E. The switches.elm resource: a compendium of conditional regulatory interaction interfaces. *Sci. Signal. 6, rs7* (2013).
20. McKusick-Nathans Institute of Genetic Medicine, J. H. U. Online Mendelian Inheritance in Man, omim (2013). URL http://www.omim.orgAccessed on May 2013]..
21. Futreal, P. A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4,** 177–183 (2004).
22. Rolland, T. *et al.* A proteome-scale map of the human interactome network. *Cell* **159,** 1212–1226 (2014).
23. Zaoui, K., Benseddik, K., Daou, P., Salaün, D. & Badache, A. Erbb2 receptor controls microtubule capture by recruiting acf7 to the plasma membrane of migrating cells. *Proc. Natl Acad. Sci. USA* **107,** 18517–18522 (2010).
24. Li, L.-Y. *et al.* Nuclear erbb2 enhances translation and cell growth by activating transcription of ribosomal rna genes. *Cancer Res.* **71,** 4269–4279 (2011).
25. Holzmann, J. *et al.* Rnase p without rna: identification and functional reconstitution of the human mitochondrial trna processing enzyme. *Cell* **135,** 462–474 (2008).
26. Sudol, M. & Harvey, K. F. Modularity in the hippo signaling pathway. *Trends Biochem. Sci.* **35,** 627–633 (2010).
27. Monferran, S., Muller, C., Mourey, L., Frit, P. & Salles, B. The membrane-associated form of the dna repair protein ku is involved in cell adhesion to fibronectin. *J. Mol. Biol.* **337,** 503–511 (2004).
28. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402,** C47–C52 (1999).
29. Brun, C. *et al.* Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biol.* **5,** R6 (2003).
30. Sharan, R., Ulitsky, I. & Shamir, R. Network-based prediction of protein function. *Mol. Syst. Biol.* **3,** 88 (2007).
31. Gómez, A. *et al.* Do protein-protein interaction databases identify moonlighting proteins? *Mol. Biosyst.* **7,** 2379–2382 (2011).
32. Tompa, P., Szàsz, C. & Buday, L. Structural disorder throws new light on moonlighting. *Trends Biochem. Sci.* **30,** 484–489 (2005).
33. Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M. & Uversky, V. N. Flexible nets. the roles of intrinsic disorder in protein interaction networks. *FEBS J.* **272,** 5129–5148 (2005).
34. Patil, A., Kinoshita, K. & Nakamura, H. Domain distribution and intrinsic disorder in hubs in the human protein-protein interaction network. *Protein Sci.* **19,** 1461–1468 (2010).
35. Hernández, S. *et al.* Do moonlighting proteins belong to the intrinsically disordered protein class? *Proteomics Bioinformatics* **5,** 262–264 (2012).
36. Davey, N. E. *et al.* Attributes of short linear motifs. *Mol. Biosyst.* **8,** 268–281 (2012).
37. Fuxreiter, M., Tompa, P. & Simon, I. Local structural disorder imparts plasticity on linear motifs. *Bioinformatics* **23,** 950–956 (2007).
38. Aranda, B. *et al.* Psicquic and psiscore: accessing and scoring molecular interactions. *Nat. Methods* **8,** 528–529 (2011).
39. Prieto, C. & Rivas, J. D. L. Apid: Agile protein interaction dataanalyzer. *Nucleic Acids Res.* **34,** W298–W302 (2006).
40. Chatr-Aryamontri, A. *et al.* The biogrid interaction database: 2013 update. *Nucleic Acids Res.* **41,** D816–D823 (2013).
41. Kerrien, S. *et al.* The intact molecular interaction database in 2012. *Nucleic Acids Res.* **40,** D841–D846 (2012).
42. Salwinski, L. *et al.* The database of interacting proteins: 2004 update. *Nucleic Acids Res.* **32,** D449–D451 (2004).
43. Ceol, A. *et al.* Mint, the molecular interaction database: 2009 update. *Nucleic Acids Res.* **38,** D532–D539 (2010).
44. Chautard, E., Ballut, L., Thierry-Mieg, N. & Ricard-Blum, S. Matrixdb, a database focused on extracellular protein-protein and protein-carbohydrate interactions. *Bioinformatics* **25,** 690–691 (2009).
45. Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* **39,** D691–D697 (2011).
46. Lynn, D. J. *et al.* Innatedb: facilitating systems-level analyses of the mammalian innate immune response. *Mol. Syst. Biol.* **4,** 218 (2008).
47. Elkon, R. *et al.* Spike-a database, visualization and analysis tool of cellular signaling pathways. *BMC Bioinformatics* **9,** 110 (2008).
48. Lange, P. F. & Overall, C. M. Topfind, a knowledgebase linking protein termini with function. *Nat. Methods* **8,** 703–704 (2011).
49. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28,** 3150–3152 (2012).
50. Skunca, N., Altenhoff, A. & Dessimoz, C. Quality of computationally inferred gene ontology annotations. *PLoS Comp.* **8,** e1002533 (2012).
51. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal,* **Complex Systems,** 1695 (2006).
52. Mistry, J., Bateman, A. & Finn, R. D. Predicting active site residue annotations in the pfam database. *BMC Bioinformatics* **8,** 298–312 (2007).
53. Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA* **101,** 6062–6067 (2004).
54. Xue, Y. *et al.* Gps 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Mol. Cell Proteomics* **7,** 1598–1608 (2008).
55. Notredame, C., Higgins, D. G. & Heringa, J. T-coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302,** 205–217 (2000).
56. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res.* **42,** D749–D755 (2014).
57. Puntervoll, P. *et al.* Elm server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.* **31,** 3625–3630 (2003).

Acknowledgements

We thank Carl Herrmann and Anaïs Baudot for helpful discussions and advice and to Matt Oates for his kind help with the D2P2 database. The work was funded by the ANR (Agence Nationale pour la Recherche) as a PIRIbio grant 09-PIRI-0028, Moonlight project to C.B. and supported by an A*MIDEX project (ANR-11-IDEX-0001-02). C.E.C. is a postdoctoral fellow of the Fondation pour la Recherche Médicale and B.R. is the recipient of a PhD fellowship from the AXA Research Fund.

Author contributions

C.E.C. conceived and developed the MoonGO pipeline and made the analyses. B.R. participated to the analyses. L.S. and C.G. build the MoonDB database. E.B. built the set of known MPs. C.B. conceived and managed the work. The manuscript was written by C.E.C. and C.B.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

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

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Chapple, C. E. *et al.* Extreme multifunctional proteins identified from a human protein interaction network. *Nat. Commun.* 6:7412 doi: 10.1038/ncomms8412 (2015).