Unsolved mysteries

Self-organization of intrinsically disordered proteins with folded N-termini

Philip C. Simister[1], Fred Schaper[2], Nicola O'Reilly[3], Simon McGowan[4], Stephan M. Feller[1]*

* Correspondence

Stephan M. Feller, Cell Signalling Group, Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, University of Oxford, Headley Way, Oxford OX3 9DS, United Kingdom, Tel.: +44-1865-222-438, Fax: +44-1865-222-431, Email: stephan.feller@imm.ox.ac.uk

1 Cell Signalling Group, Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, University of Oxford, Oxford, United Kingdom

2 Department of Systems Biology, Otto-von-Guericke-University Magdeburg, Magdeburg, Germany

3 Peptide Synthesis Laboratory, Cancer Research UK London Research Institute, London, United Kingdom

4 Computational Biology Research Group, Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, University of Oxford, Oxford, United Kingdom

Abstract

Thousands of human proteins lack recognizable tertiary structure in most of their chains. Here we hypothesize that some use their structured N-terminal domains (SNTDs) to organise the remaining protein chain via intramolecular interactions, generating partially structured proteins. This model has several attractive features: as protein chains emerge, SNTDs form spontaneously and serve as nucleation points, creating more compact shapes. This reduces the risk of protein degradation or aggregation. Moreover, an interspersed pattern of SNTD-docked regions and free loops can coordinate assembly of sub-complexes in defined loop-sections and enables novel regulatory mechanisms, for example through posttranslational modifications of docked regions.

Introduction

Proteins were for a long time thought to be mostly well-structured entities made up of one or several domains, which are assembled from spontaneously forming α-helices and/or β-strands, or are folded with the help of chaperones. However, over the last decades this simplistic view has been increasingly questioned.

It is now clear that thousands of human proteins do not fit into simple categories of tertiary structural organisation, that is, in these cases, recognizable domains are completely or mostly missing. Such proteins have been called unstructured or 'intrinsically disordered' (ID) **[1]**.

Approximately one third of human proteins lack recognizable secondary and tertiary structure in most of their protein chain and hence appear to fall into this category **[2]**. Some of them are being collected in the DisProt database **[3]**, but most are excluded from detailed ultrastructural analyses, as they are often considered to be poor targets. Relatively little is known about their shapes, conformations and conformational changes presumably occurring. Nevertheless ID proteins have important functions in multi-protein complex assembly and cell signaling **[4 - 7]** and we need to learn much more about their molecular activities and mechanisms of action.

Their abundance in cells is somewhat puzzling, for example raising questions regarding their escape from proteolytic degradation and lack of aggregate formation. It is well known that misfolded proteins are the cause of several major neurodegenerative disorders and other 'protein deposit' diseases **[8 – 12]**. In fact, almost all proteins contain segments that can in principle form amyloids **[13]**. Poorly folded proteins are also targets for degradation by the proteasome and other proteases **[14 - 16]**. However, structural disorder appears to serve only as a weak signal for intracellular protein degradation and ID-proteins also do not appear to display an overall preference for chaperone binding *in vivo* **[17]**, despite the prominent role that chaperones clearly play in supporting protein folding in general **[18]**.

At least some, if not many, ID proteins may therefore adopt types of order that are not easily recognized by current secondary or tertiary structure prediction programmes. Examples of secondary structure elements that are usually not detected include the poly-proline type II (PPII) **[19 - 20]** and 3_{10} helices **[21 - 22]**. Despite their abundance in human proteins, new examples of these helices often only become apparent through focussed structural analyses of individual proteins.

Beyond these often short, helical regions, it appears likely that several other molecular routes to generating order within unfolded protein chains exist, some of which may still remain to be studied in any detail. The first interesting mechanisms that allow the generation of defined structural states from a disordered conformation have recently emerged. For example, some ID proteins adopt specific, regionally restricted conformations upon binding their partner proteins. This can go as far as adopting multiple distinct conformations depending on which of several interactors is docking **[23]**. Another example are the recently proposed 'disordered domains', stretches longer than 20-30 amino acid residues, which are thought to present functional units for protein interactions **[24]**.

We suggest here a further novel mechanism for rapidly establishing a certain degree of order within long ID protein chains. This N-terminal folding nucleation hypothesis provides an experimentally testable conceptual framework that has many attractive features in terms of explaining how some of the so-called ID proteins can effectively fulfil their functions in cells.

The emergence of the N-terminal folding nucleation (NFN) hypothesis

Research in our and many other laboratories has increasingly focussed on the function of large multi-site docking (LMD) protein platforms like the Gab, p130Cas and IRS family proteins **[5 – 6]**, which facilitate the assembly of large signal transduction protein complexes. Such multi-protein complexes are presumably required for the integration and processing of multiple inputs from different upstream signal transducers, with subsequent further propagation of several signals regulating cell survival, proliferation, cytoskeletal structures, migration and/or differentiation. In this context it was noticed that several families of LMD proteins display a similar structural composition, namely a well-structured N-terminal domain (SNTD), for example an SH3, PH or PTB domain, that is followed by a long and, according to secondary structure prediction programs, largely disordered protein chain (**Figure 1**). While this initially just seemed to be a curious feature without any obvious functional explanation, another unexpected and seemingly unrelated finding led to the emergence of a potential rationale for this peculiar LMD protein composition. It was found that in HEK293T cells, which display intrinsically high phosphatidylinositol-3 (PI3) kinase activity, the full-length Gab1 protein is localised in the cytoplasm, instead of at the plasma membrane, as would be expected for cells with constitutively active PI3 kinase **[25]**. This was surprising because the Gab1 PH domain has been shown to preferentially bind PIP3 **[26]**, the membrane-embedded product of PI3 kinase.

When following up this initial observation, it was found that in order to achieve membrane translocation of Gab1 via PIP3 binding of its PH domain, in these cells a further signal is needed. The kinase module Mek - Erk is required to phosphorylate a serine residue (Ser552) localized far away from the PH domain in the disordered tail of the protein **[25]**. This clearly indicates that a distant part of the Gab1 tail region binds, either directly or indirectly, to the Gab1 PH domain, thereby blocking its PIP3 binding pocket. Furthermore, this interaction occurs in a functionally regulated manner. To analyze the interaction between Gab1 PH and the Ser552 epitope further, we employed the peptide array overlay assay **[27 – 29]**. This is a frequently used method for the detection of protein domain interactions with short linear motifs that typically displays little non-specific binding. For this assay, the Gab1 protein is synthesized in the form of a series of overlapping peptides which are immobilized in spots on a cellulose carrier membrane. The affinity-purified Gab1 PH domain, expressed as a GST-fusion protein, is then used as probe to detect *in vitro* short linear Gab1 regions that may bind to the PH directly *in vivo* (**Figure 2**). From this, the Ser552-region emerged as a direct binding motif for the GST-PH probe. In addition, numerous other regions in the Gab1 protein also appeared as putative PH domain binders. None of these regions bound to GST alone. If these potential binding regions are also utilized *in vivo*, the Gab1 PH domain serves not only as a phospholipid-binding module, but also as a nucleation core for the intramolecular binding, and hence compaction, of the supposedly disordered Gab1 tail region.

This N-terminal folding nucleation (NFN) hypothesis has several attractive features:

Firstly, it provides a simple explanation for how disordered proteins can escape protein aggregation or degradation by a co-translational folding mechanism that differs substantially from the classical formation of structured proteins. As the first terminal amino acids emerge from the ‘teflon-like’ ribosomal tunnel, secondary structural elements are spontaneously generated which rapidly assemble into a highly stable SNTD. Once this is built, additional residues emerging from the ribosome can dock onto specific SNTD patches, thereby preventing the unstructured chain from engaging in nonspecific interactions and also preventing those patches on the SNTD from accepting the binding of external protein chains (**Figure 3**).

Moreover, the intramolecular attachment of protein chain segments to the SNTD would seem to generate defined loops which may serve as regions for the assembly of distinct sub-complexes where protein compositions differ from those attached to other loop regions.

In this context, it should be noted that the clustering of specific protein binding sites in LMD proteins of the Gab family was noticed years ago **[30]**. This clustering is presumed to contribute to the spatial organization of complex components. In Gab1, six CRKL SH2 domain candidate binding sites localize to a central region of ca. 170 amino acids, while the other 525 amino acids lack even a single putative binding motif. Combined with the ability of the CRKL adaptor protein to dimerize and possibly tetramerize **[31]**, highly ordered complexes of great stability could be formed in distinct regions of LMD proteins like Gab1.

Clearly, these proposed concepts are for now speculative and in need of further experimental analyses. However, such concepts are urgently required to guide the design of new types of experiments that will help to define the mechanism of how distinct and very large signal transduction protein complexes (a.k.a. stimulus-specific ‘signalosomes’) are rapidly assembled in response to diverse stimuli. The coordinated assembly of well-ordered signaling sub-complexes that can be differentially combined depending on the biological complex is intuitively appealing. It should allow the speedy generation of specific signals in discrete regions of an LMD protein and this must be very desirable for at least some signaling systems. In the case of Gab1/CRKL complexes, which are prominently linked to cell shape change and motility signals through the activation of Rho family GTPases, it is easy to imagine multiple biological contexts where the ability to move swiftly would be advantageous.

Another advantage of discrete regions docking onto the SNTD would be the generation of novel targets for signal regulation, which may, for example, contribute to the undoubtedly important robustness of cell signaling networks **[32]**. This is nicely exemplified by the Gab1 phosphorylation on Ser552. Only upon the generation of two signals, one through PI3 kinase activation and the other one by firing of the mitogenic kinases, will the important LMD protein Gab1 translocate to the membrane, where further phosphorylation leads to the assembly of a potent regulator complex of essential cell behaviours.

Towards a solution

To estimate how commonly NFN is utilized, we initially sought to define how many proteins in the human proteome display an SNTD in combination with a long disordered tail. For this, disordered regions and structural domains were predicted for all human proteins in the UniProt SwissProt database (www.ebi.ac.uk/uniprot/) using DisEMBL (http://dis.embl.de/) and SMART (http://smart.embl-heidelberg.de/), respectively. The two sets of predictions were compared using a custom perl script to identify proteins with a predicted domain or domains in the N-terminus (defined as the first 25% of the protein), no predicted domains in the C terminus (defined here as the remaining 75% of the protein) and predominantly disordered (>80%) in this C-terminus. The initial hits were listed with their corresponding SMART and SwissProt data and then individually inspected to exclude, for example, transmembrane proteins.

This showed that, in addition to the protein families depicted in **Figure 1**, over 50 further proteins display a similar structural organisation (for details see **Supporting Figure 1**). This number is probably an underestimate, since not all domains may yet be recognised by the SMART database and also because proteins with an SNTD and subsequent mostly unstructured tail, but with one or more additional domain(s) more C-terminally, were excluded from the initial search, even though they may use NFN for a part of their amino acid chain. Proteins with an SNTD and a relatively short ID tail were also not scored in this experiment. Of the more than 50 proteins that emerged from this analysis, a major portion is readily known or presumed to be involved in signaling processes.

The NFN candidate proteins now need to be subjected to further biochemical, biological and biophysical analyses. Purification of these proteins from eukaryotic cells and analysis by gel filtration chromatography, analytical ultracentrifugation, mass spectrometry of intact proteins and small angle X-ray scattering (SAXS) should give some information about their shapes. NMR analyses of isolated SNTDs and full-length proteins should identify residues in the SNTDs that contribute to intramolecular contacts with the ID chain. In some cases, even *in vivo* NMR, similar to a study conducted with bacterial FlgM may be possible **[33].**

Mutations of SNTD residues implicated from NMR experiments, and of key residues in the ID tails identified by peptide array overlay blots could then be analysed for functional defects or effects on protein turnover or aggregation in cells. *In vivo* studies with knock-in mutants can subsequently investigate the systemic consequences. It will also be interesting to determine whether some of the proteins utilising NFN are additionally stabilised in their compact shapes by complex formations with other proteins, which should co-purify in stoichiometric amounts.

Clearly nature has found multiple ingenious ways of folding emerging protein chains in highly complex organisms into functional units with great efficacy. As we learn more about these mechanisms, we will also begin to understand better the fundamental principles that govern the assembly and actions of complex signaling networks and their multi-protein hubs.

References

1. Tompa P (2009) Structure and Function of Disordered Proteins. London and New York: CRC Press.

2. Edwards YJ, Lobley AE, Pentony MM, Jones DT (2009) Insights into the regulation of intrinsically disordered proteins in the human proteome by analyzing sequence and gene expression data. Genome Biol 10: R50.

3. Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, et al. (2007) DisProt: the Database of Disordered Proteins. Nucleic Acids Res 35: D786-793.

4. Hegyi H, Schad E, Tompa P (2007) Structural disorder promotes assembly of protein complexes. BMC Struct Biol 7: 65.

5. Mardilovich K, Pankratz SL, Shaw LM (2009) Expression and function of the insulin receptor substrate proteins in cancer. Cell Commun Signal 7: 14.

6. Wohrle FU, Daly RJ, Brummer T (2009) Function, regulation and pathological roles of the Gab/DOS docking proteins. Cell Commun Signal 7: 22.

7. Gotoh N (2008) Regulation of growth factor signaling by FRS2 family docking/scaffold adaptor proteins. Cancer Sci 99: 1319-1325.

8. Carrell RW, Lomas DA (1997) Conformational disease. Lancet 350: 134-138.

9. Bucciantini M, Giannoni E, Chiti F, Baroni F, Formigli L, et al. (2002) Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. Nature 416: 507-511.

10. Walsh DM, Klyubin I, Fadeeva JV, Cullen WK, Anwyl R, et al. (2002) Naturally secreted oligomers of amyloid beta protein potently inhibit hippocampal long-term potentiation in vivo. Nature 416: 535-539.

11. Herczenik E, Gebbink MF (2008) Molecular and cellular aspects of protein misfolding and disease. FASEB J 22: 2115-2133.

12. Nakamura T, Lipton SA (2009) Cell death: protein misfolding and neurodegenerative diseases. Apoptosis 14: 455-468.

13. Goldschmidt L, Teng PK, Riek R, Eisenberg D Identifying the amylome, proteins capable of forming amyloid-like fibrils. Proc Natl Acad Sci U S A 107: 3487-3492.

14. Gallastegui N, Groll M The 26S proteasome: assembly and function of a destructive machine. Trends Biochem Sci.

15. Kubota H (2009) Quality control against misfolded proteins in the cytosol: a network for cell survival. J Biochem 146: 609-616.

16. Stolz A, Wolf DH Endoplasmic reticulum associated protein degradation: a chaperone assisted journey to hell. Biochim Biophys Acta 1803: 694-705.

17. Hegyi H, Tompa P (2008) Intrinsically disordered proteins display no preference for chaperone binding in vivo. PLoS Comput Biol 4: e1000017.

18. Hartl FU, Hayer-Hartl M (2009) Converging concepts of protein folding in vitro and in vivo. Nat Struct Mol Biol 16: 574-581.

19. Adzhubei AA, Sternberg MJ (1993) Left-handed polyproline II helices commonly occur in globular proteins. J Mol Biol 229: 472-493.

20. Kay BK, Williamson MP, Sudol M (2000) The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains. FASEB J 14: 231-241.

21. Toniolo C, Benedetti E (1991) The polypeptide 310-helix. Trends Biochem Sci 16: 350-353.

22. Harkiolaki M, Tsirka T, Lewitzky M, Simister PC, Joshi D, Bird LE, Jones YE, O'Reilly N, Feller SM (2009) Distinct binding modes of two epitopes in Gab2 that interact with the SH3C domain of Grb2. Structure 17: 809-822.

23. Oldfield CJ, Meng J, Yang JY, Yang MQ, Uversky VN, et al. (2008) Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. BMC Genomics 9 Suppl 1: S1.

24. Tompa P, Fuxreiter M, Oldfield CJ, Simon I, Dunker AK, et al. (2009) Close encounters of the third kind: disordered domains and the interactions of proteins. Bioessays 31: 328-335.

25. Eulenfeld R, Schaper F (2009) A new mechanism for the regulation of Gab1 recruitment to the plasma membrane. J Cell Sci 122: 55-64.

26. Rodrigues GA, Falasca M, Zhang Z, Ong SH, Schlessinger J (2000) A novel positive feedback loop mediated by the docking protein Gab1 and phosphatidylinositol 3-kinase in epidermal growth factor receptor signaling. Mol Cell Biol 20: 1448-1459.

27. Li SS, Wu C (2009) Using peptide array to identify binding motifs and interaction networks for modular domains. Methods Mol Biol 570: 67-76.

28. Watanabe T, Tsuda M, Makino Y, Konstantinou T, Nishihara H, et al. (2009) Crk adaptor protein-induced phosphorylation of Gab1 on tyrosine 307 via Src is important for organization of focal adhesions and enhanced cell migration. Cell Res 19: 638-650.

29. Pietrek M, Brinkmann MM, Glowacka I, Enlund A, Havemeier A, et al. Role of the Kaposi's Sarcoma-Associated Herpesvirus K15 SH3 Binding Site in Inflammatory Signaling and B-Cell Activation. J Virol 84: 8231-8240.

30. Sakkab D, Lewitzky M, Posern G, Schaeper U, Sachs M, et al. (2000) Signaling of hepatocyte growth factor/scatter factor (HGF) to the small GTPase Rap1 via the large docking protein Gab1 and the adapter protein CRKL. J Biol Chem 275: 10772-10778.

31. Harkiolaki M, Gilbert RJ, Jones EY, Feller SM (2006) The C-terminal SH3 domain of CRKL as a dynamic dimerization module transiently exposing a nuclear export signal. Structure 14: 1741-1753.

32. Kitano H (2004) Biological robustness. Nat Rev Genet 5: 826-837.

33. Dedmon MM, Patel CN, Young GB, Pielak GJ (2002) FlgM gains structure in living cells. Proc Natl Acad Sci U S A 99: 12681-12684.

Acknowledgements

The authors would like to thank Toby Gibson (EMBL) for useful discussions. The UK charities Cancer Research UK and Heads Up are acknowledged by SF for their support. FS acknowledges the support by the Deutsche Forschungsgemeinschaft DFG (SFB542) and the German Bundesministerium für Bildung und Forschung BMBF (0313922)

Figure legends

Figure 1. Schematic structure of selected large multi-site docking (LMD) protein families involved in signaling.

The Irs/Dok, Gab, p130Cas and Frs families of LMD proteins are crucial platforms for the assembly of elaborate multi-protein complexes (a.k.a. 'signalosomes') of a wide range of cell membrane receptors involved in regulating cell survival, growth, motility and/or differentiation. They all share a similar composition, that is a well-structured N-terminal domain (SNTD) followed by an apparently largely unstructured amino acid chain. In some cases, short secondary structure motifs like PPII helices, 3_{10} helices etc. have been found or are suspected. A human proteome-wide search subsequently revealed that a large number of proteins exist with a similar structural composition. For further details see **Supporting Figure 1**.

Figure 2. Gab1 peptide array overlay assay identifies potential binding sites for the PH domain.

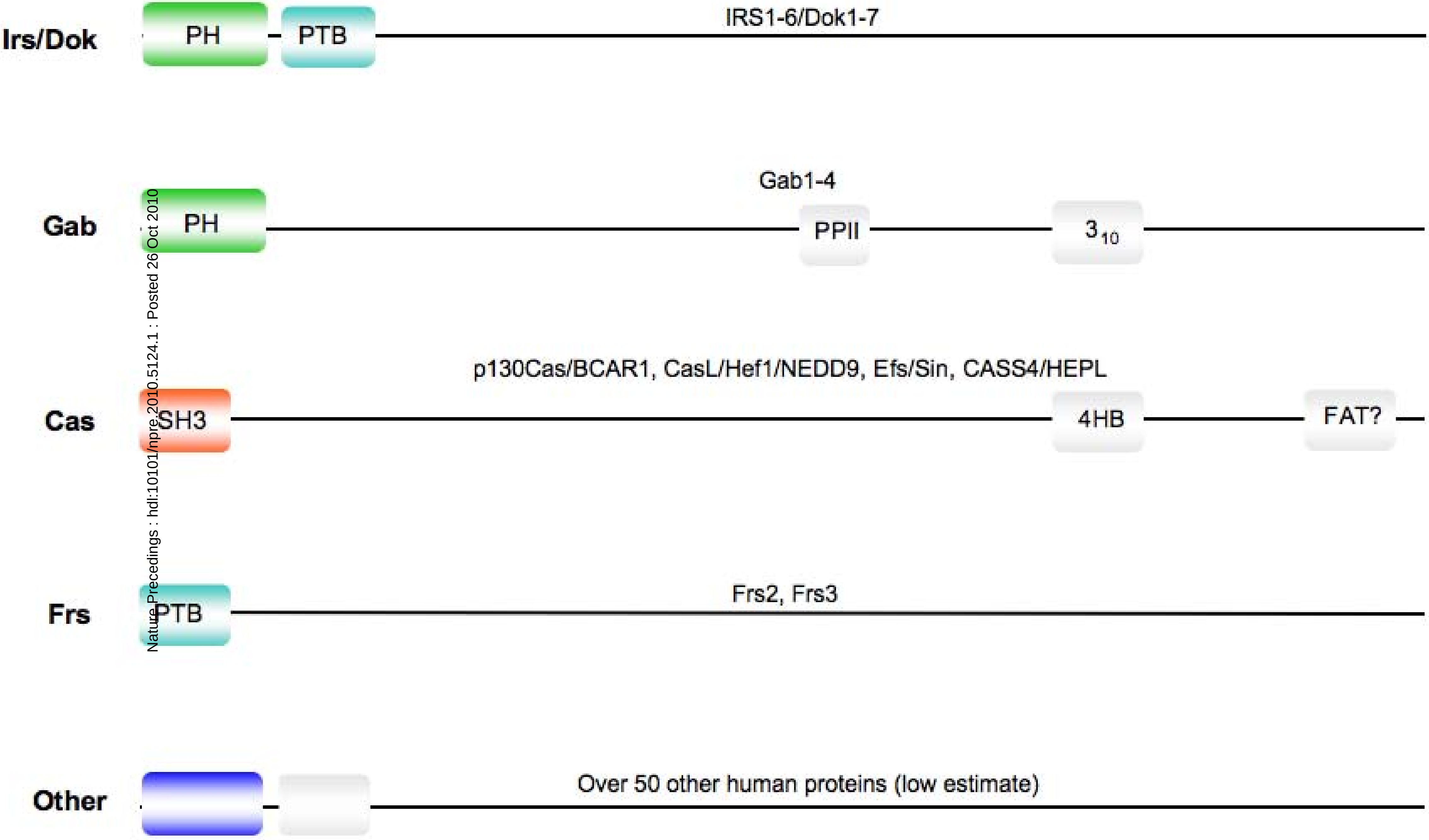
For this assay, the full amino acid sequence of Gab1 from *Mus musculus*, also used in the study of Eulenfeld and Schaper **[25]**, was chemically synthesized as an array of spots of overlapping peptides (Multipep synthesiser [Intavis], peptide lengths 23 amino acids, sliding two residues further with each consecutive peptide), blocked with 5% nonfat dry milk in TrisHCl buffer (pH7.5) with 0.1% Tween 20 added and probed initially with 4 μg/ml GST, followed by incubation with anti-GST, HRP-coupled secondary antibody and ECL detection. No GST binding was detectable to any of the peptides (top panel). The same membrane was then re-probed with 1 μg/ml of affinity-purified GST-PH domain (bottom panel). The red box indicates the Ser552 epitope previously implicated in regulating Gab1 PH domain binding by the work of Eulenfeld & Schaper. Similar results were also obtained when DTT was included in the assay to eliminate potential artefacts from non-specific interactions of Cys residues (data not shown).

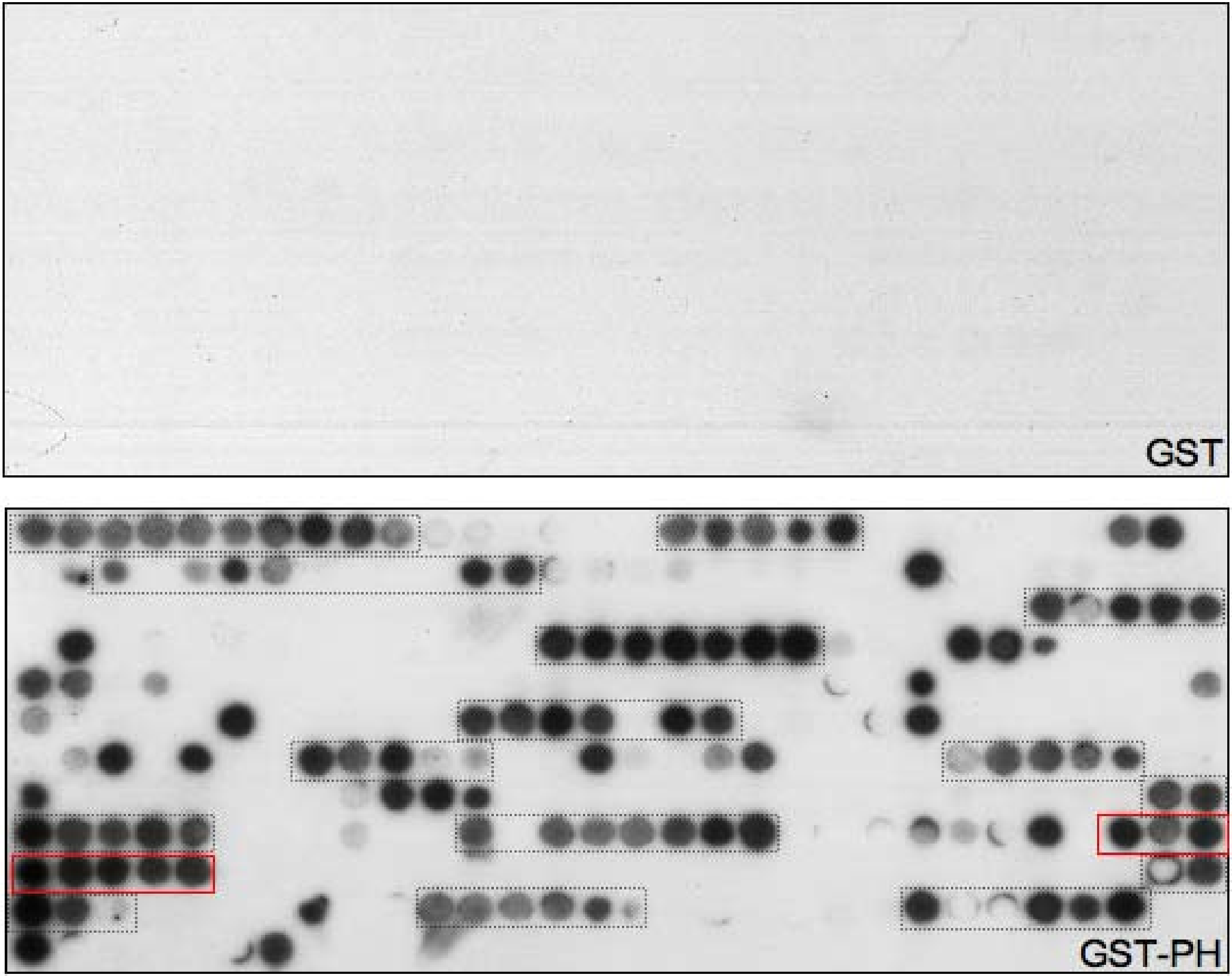
Figure 3. Graphical illustration of the N-terminal folding nucleation (NFN) hypothesis.

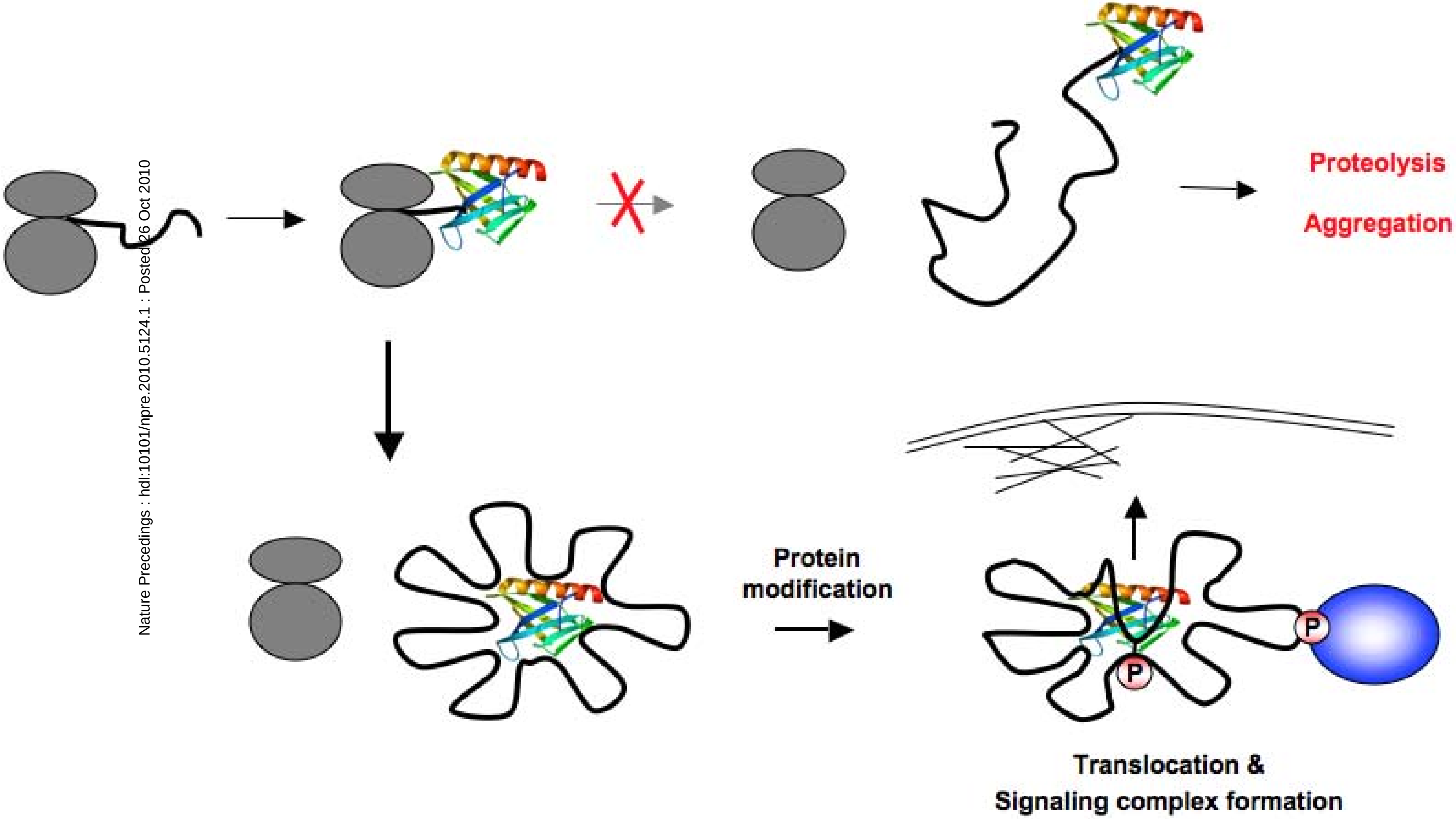
The NFN hypothesis proposes that, as the nascent amino acid chain of an LMD protein emerges from the ribosome (depicted in grey), the SNTD is rapidly and spontaneously generated and this SNTD then serves as a nucleation core for additional and specific intramolecular protein chain contacts, which generate a more compact protein shape. The compaction avoids the presence of a long disordered amino acid chain, which may lead to protein proteolysis or aggregation. Instead, the interspersed pattern of docked regions and loops generates defined subsections in the protein that may serve as functional subunits. Protein modifications like phosphorylation in some of these defined regions may lead, for example, to the liberation of docked regions, allowing the SNTD to engage in novel types of interactions which could allow the anchorage of the LMD protein in specific subcellular locations. Other modifications may generate docking points for interaction domains of signaling partner proteins, resulting in the rapid assembly of defined sub-complexes on specific loops. Taken together, the features of LMD proteins emerging through NFN are expected to increase the ability of cells to respond rapidly and selectively to a diverse set of incoming stimuli.

Supporting Figure 1. Human proteins identified as NFN candidates by bioinformatics analysis.

Schematic representation of proteins identified by the prediction of disordered regions and structural domains for all human proteins in the UniProt SwissProt database (www.ebi.ac.uk/uniprot/) using DisEMBL (http://dis.embl.de/) and SMART (http://smart.embl-heidelberg.de/), respectively. The two sets of predictions were compared using a custom perl script to identify proteins with a predicted domain or domains in the N-terminus (defined as the first 25% of the protein), no predicted domains in the C terminus (defined here as the remaining 75% of the protein) and predominantly disordered (>80%) in this C-terminus. Initial hits were listed with their corresponding SMART and SwissProt data and then individually inspected to exclude, for example, transmembrane proteins. Proteins shown here clearly represent an underestimate of actual candidates in the human proteome, since, for example, proteins with additional domains in the amino acid chain following the folded N-terminus were excluded, even if several hundred disordered amino acids follow the N-terminal domain. If multiple splice variants occur, only a single representative is shown for each protein. Proteins are alphabetically listed according to the gene names following the HGNC nomenclature (7-2010; http://www.genenames.org/), identifiers below the names and SNTD designations are according to the SMART database. Protein domains and chain lengths are not drawn to scale. Values on the right side indicate the number of amino acids in each protein. Many of the NFN candidates depicted here are known or suspected to act in cell signaling. Please note that LMD proteins already shown in Figure 1 and again found in the bioinformatic analysis (FRS2, FRS3, GAB1, GAB2, GAB3, IRS1, IRS2,) are not shown again in this figure.


Irs/Dok
PH
PTB
IRS1-6/Dok1-7
Gab
PH
Gab1-4
PPII
3 10
Cas
SH3
p130Cas/BCAR1, CasL/Hef1/NEDD9, Efs/Sin, CASS4/HEPL
4HB
FAT?
Frs
PTB
Frs2, Frs3
Other
Over 50 other human proteins (low estimate)


GST
GST-PH


Proteolysis
Aggregation
Protein modification
P
P
Translocation & Signaling complex formation