Intrinsic Structural Disorder Confers Cellular Viability on Oncogenic Fusion Proteins

Hedi Hegyi1, László Buday1,2, Peter Tompa1

1 Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, Budapest, Hungary, 2 Department of Medical Chemistry, Semmelweis University Medical School, Budapest, Hungary

Abstract

Chromosomal translocations, which often generate chimeric proteins by fusing segments of two distinct genes, represent the single major genetic aberration leading to cancer. We suggest that the unifying theme of these events is a high level of intrinsic structural disorder, enabling fusion proteins to evade cellular surveillance mechanisms that eliminate misfolded proteins. Predictions in 406 translocation-related human proteins show that they are significantly enriched in disorder (43.3% vs. 20.7% in all human proteins), they have fewer Pfam domains, and their translocation breakpoints tend to avoid domain splitting. The vicinity of the breakpoint is significantly more disordered than the rest of these already highly disordered fusion proteins. In the unlikely event of domain splitting in fusion it usually spares much of the domain or splits at locations where the newly exposed hydrophobic surface area approximates that of an intact domain. The mechanisms of action of fusion proteins suggest that in most cases their structural disorder is also essential to the acquired oncogenic function, enabling the long-range structural communication of remote binding and/or catalytic elements. In this respect, there are three major mechanisms that contribute to generating an oncogenic signal: (i) a phosphorylation site and a tyrosine-kinase domain are fused, and structural disorder of the intervening region enables intramolecular phosphorylation (e.g., BCR-ABL); (ii) a dimerisation domain fuses with a tyrosine kinase domain and disorder enables the two subunits within the homodimer to engage in permanent intermolecular phosphorylations (e.g., TFG-ALK); (iii) the fusion of a DNA-binding element to a transactivator domain results in an aberrant transcription factor that causes severe misregulation of transcription (e.g., EWS-ATF). Our findings also suggest novel strategies of intervention against the ensuing neoplastic transformations.

Introduction

Chromosomal translocations are the major genetic aberration in cancers, such as leukemias, lymphomas and sarcomas [1–4]. Translocation links two distinct chromosomes, and either fuses one gene to the regulatory region of another gene, or results in a chimera by the fusion of two unrelated genes. The resulting misregulation of the expression of a normal gene or appearance of a unique fusion protein is the cause of neoplastic transformations in many cases. Molecular understanding of the translocation event is of paramount importance in devising strategies against these diseases [3,4]. Translocation has been extensively studied at the genetic level, leading to the recognition that its primary cause is a double-strand break (DSB) of DNA, erroneously repaired by joining two remote chromosomal segments [1]. Fusion events have also been well characterized in terms of the functions of genes/gene products involved. A dominance of DNA-binding and transcription regulatory functions have been observed, whereas at the domain level kinases and DNA-binding motifs occur most frequently [2,5–7].

Much less is known about the structural implications of protein fusion. The proteins involved are often quite long and complex, heterogeneous in sequence and structure, and contain only a few dispersed domains, usually avoided by the translocation breakpoints [3,4,8]. This is particularly true of proteins that appear in chromosomal translocations recurrently, such as MLL [9], CBP [8], or EWS [10]. This has led to the suggestion that the cellular survival of the protein chimera can be explained by its structural disorder, because it enables the cellular viability of a protein generated from segments of two unrelated proteins [8]. The rationale of this notion rests on the prevalence of intrinsically disordered/unstructured proteins (IDPs/IUPs) or protein regions (IDRs), which exist and function without well-defined 3D structures [8,11,12]. Structural disorder reaches high levels in proteins of regulatory and transcriptional functions [13,14], and shows significant evolutionary increase in eukaryotes, compared to prokaryotes [15]. IDPs/IDRs can function as disordered linkers, but most often they carry out their functions by molecular recognition, in which they bind other protein(s), DNA or RNA in a binding-induced folding process [8,11]. Disorder plays a prominent role in cancer-associated proteins [13] and alternative splicing (AS), a process which generates distinct protein products from the same initial transcript [16]. AS may connect disjoint segments of proteins into a viable protein product, in a process...
**Author Summary**

Chromosomal translocations generate chimeric proteins by fusing segments of two distinct genes and are frequently associated with cancer. The proteins involved are large and fairly heterogeneous in sequence and typically have only a few dispersed structural domains connected by long uncharacterized regions. It has never been studied from a structural perspective how these chimeras survive losing significant portions of the original proteins and acquire new oncogenic functions. By analyzing a collection of 406 human translocation proteins we show here that the answer to both questions lies to a large extent in the high level of structural disorder in the fusion partner proteins (on average, they are twice as disordered as all human proteins). The translocation breakpoints usually avoid globular domains. In rare cases when a globular domain is truncated by the fusion, it happens at a location in the domain where the hydrophobicity exposed by the split is favorable (i.e., not too high). Disorder on average is significantly higher in the vicinity of the breakpoint than in the rest of the fusion proteins. Disorder also plays a pivotal role in the acquired oncogenic function by bringing distant/disparate fusion segments together that enables novel intra- and/or intermolecular interactions.

Conceptually similar to protein fusion. In fact, for two fusion products, EFP [10] and CBP-NOZ [8], the involvement of structural disorder in fusion has been explicitly stated. Nevertheless, its role in either the cellular survival of the fusion product or the ensuing oncogenic function has neither experimentally nor statistically been addressed in these or other works.

Motivated by these inferences, we have tested the association of structural disorder with chromosomal translocations. We collected 406 human fusion proteins (255 with identifiable breakpoints), and analyzed their disorder by the IUPred prediction algorithm [17]. We found that fusion proteins have a very high level of disorder, their translocation breakpoints tend to avoid domains, and disorder appears to play a major role in their oncogenic functions. These findings shed new light on the structural background of how protein products generated by chromosomal translocations are selected for by cellular proliferation and clonal expansion in cancer, and suggest novel strategies of intervention against the ensuing oncogenic transformations.

**Results**

**Length dependence of translocation**

As suggested in the Introduction, translocation is initiated by the DSB of DNA, which may occur either at random or in conjunction with some special feature of DNA sequence/structure. Both scenarios suggest that longer proteins/gene are more likely to undergo DSB and subsequent translocation. This has already been shown in a study on a smaller database of translocation proteins among 291 cancer proteins [18], which suggested that translocated cancer proteins tend to have longer genes, whereas cancer genes with point mutations tend to encode longer proteins, both significantly longer than average human genes/proteins. Because of the noted increase of structural disorder with the length of proteins [19,20], we repeated this analysis on our database of 406 proteins involved in translocations. To this end, the frequency of translocation as a function of the length of the protein (Figure 1A) or the length of the gene (Figure S1) in question was determined. Both functions have an inverse relationship, which suggests that proteins/genes become involved in translocation event(s) roughly in proportion to their length. Nevertheless, the frequency of translocation increases with protein length at a greater power than with gene length (0.82 vs. 0.59), which suggests that additional structural/functional forces of selection besides mere chance of random DNA DSB events operate at the protein level. Figure 1B shows the distribution of the ratio (protein/gene) length for both proteins involved in translocation and all human proteins. Again, a clear difference between the two sets of proteins, where the translocating protein set falls off with a steeper slope, indicate an even greater relative gene length for this set of proteins, which suggests that additional selection forces act at the protein level.

**Disorder in translocation proteins**

Human translocation proteins have an extensive disorder predicted by IUPred [17], with a significant excess of proteins of 70–80% disorder (Figure 2). Comparing the distributions with a chi-square test we found that proteins with an established translocation breakpoint differ very significantly from that of all SwissProt human proteins (p-value<1e-14) whereas those without a known breakpoint differ from all of SwissProt less significantly (p<0.00058). The two sets of translocating proteins also differ from each other (p<0.00067). The mean disorder of all human proteins is 20.7% whereas that of proteins with known translocation breakpoint(s) is 43.3%. Translocation proteins without a known breakpoint have a small local maximum at 60% disorder probably reflecting so far undiscovered breakpoint(s) in some of them, whereas for others the breakpoint is located outside the coding region. The mean disorder for this latter set is 32.1%. The disorder distribution of 500x255 randomly drawn human SwissProt proteins matched in length to translocation proteins with breakpoint is very similar to the total of 18,609 human proteins currently in Swissprot (with a mean disorder 21.9%). The slight difference 1.2% could be attributed to longer proteins having somewhat higher disorder also observed by others [19,20]. This bias, however, does not account for the elevated level of disorder in translocating proteins (cf. Figure 2).

The ratio of the number of all human proteins divided by all translocating proteins with a breakpoint as a function of disorder is shown in Figure S2. The figure shows the increasing frequency of protein translocation with increasing disorder (to be more precise, the chance of such proteins surviving the translocation event increases). As the best fitting trend line is a parable (with an almost perfect fit, R²=0.9587) the relationship appears quadrant, i.e. twice as much disorder entails four times more frequent translocation (with the exception of the disorder range 90–100% where the tendency seems to reverse itself). All the proteins in this set, their disorder, length and breakpoints are listed in Table S1.

**Disorder at the breakpoint**

To address if disorder at the point of fusion is preferred for the survival of the protein chimera we assessed the average level of disorder in translocation proteins. With respect to the actual breakpoint and its vicinity, we predicted disorder around the breakpoint separately for the N- and C-terminal partners within the range of [−50,50] (Figure 3A). In the N-terminal partners the highest values appear in the [−20, 0] region, whereas the values to the right of the breakpoint gradually fall off, in accord with the fact that these parts are eliminated during fusion. In the C-terminal partners it is the values right to the breakpoint that increase in the parts retained in the fusion products. Averaging the averages in the [−50,50] range around the breakpoint we found that the total average for this region is 0.49 (with SD=0.016), which is
significantly higher than in the rest of these proteins (0.43, cf. Figure 2 and Figure 3B; the difference is more than 3 SD-s).

The mean values for the overall disorder of the translocation partners are shown in Figure 3B (with SDs indicated). Apparently they have a very high level of disorder, e.g. more than twice as high as for all human proteins in Swissprot, with a tendency of fusion products to have an even higher one. Comparing them to proteins experimentally determined to be fully disordered (e.g. those in DisProt) and fully ordered (e.g. those in PDB), it can be ascertained that selection following chromosomal translocation strongly favors fusion proteins in which structural disorder dominates.

Translocation and domains in the fusion partner proteins

A clear implication of the above findings is that a protein product is highly disfavored if its site of joining falls in ordered domains, which would most probably lead to the creation of structurally aberrant chimeras. To check this assumption, we have analyzed translocation proteins for the occurrence of Pfam domains. We found that the average coverage of translocation proteins by Pfam domains is 36.3%, whereas this value for a human Swissprot protein is 42.5%. On the other hand, in proteins generated by fusion (where each gene pair is considered only once, with the longest fusion protein for each) Pfam coverage decreases to 30.9%. A chi-square test showed that all three values (and the underlining distributions) differ significantly, each set differs from the total of human Swissprot with p-value < 1e-9, and even the translocation partner proteins and the fusion proteins differ from one another regarding their coverage by Pfam domains with p-value = 0.012.

To check if the breakpoint tends to “avoid” Pfam domains (i.e. such proteins are selected against by cellular proliferation and clonal expansion), we compared the number of actually truncated domains with that of a set of proteins with a randomly generated breakpoint. We found that while the actual number of truncated domains in the 255 translocation partner proteins with an
established breakpoint was 48, the random breakpoints (repeating the random selection process 200 times) resulted in 76.2 domain truncations on average with a standard deviation of 5.6, that is, the difference is highly significant, with a Z-value of 4.87.

In addition, the average disorder of Pfam domains occurring in the fusion products is 14.3% (not significantly different from that in all human proteins, 13.1%), whereas the average disorder of Pfam domains actually truncated is significantly higher, 26.8%. In cases where the missing or retained fragment is more than 10% of the original size of the domain, this disorder value increases to 29.6%.

**Fate of truncated domains in the fusion products**

A closer look at actually truncated Pfam domains located in the fusion proteins (Table 1) provides further evidence for the structural bias of protein products generated by fusion events (for a compendium of all Pfam domain matches where the breakpoint falls in a domain, see Table S2). In several cases the Pfam domain is actually a coiled-coil, which is structurally rather indifferent to the location of truncation. In addition, the remainder of the domain very often has a significant level of disorder, or it is apposed with a rather disordered segment on the fusion partner.

Furthermore, almost all cases of truncation of a globular domain (indicated in Table 1) can be structurally rationalized for the viability of the fusion product (see below), with the exception of the Interleukin 2 domain in IL2-TNFRSF17 where the fusion transcript appears not to be translated into a viable protein [21].

We analyzed in detail the severely truncated protein kinase domain in CDK6-MLL. It has low disorder and the domain retains 40% of its original size, corresponding to 114 amino acids. It appears to apply a clever strategy for cellular survival: for this domain, the accessible non-polar area resulting from sequential truncations calculated by the CHASA program [22] approaches the theoretical value of an intact domain at around 90–94 amino acids only (Figure 4A). Because this length is close to the actual 114 amino-acid truncation, this partial domain almost behaves as intact, and it is most certainly not recognized by the cellular degradation system. Figure 4B shows the missing and retained portions of the kinase fold where it also becomes clear that the fold consists of two sub-domains and only a small part of the second one is retained in the fusion protein, which probably does not fold at all. In contrast, the truncated Interleukin 2 domain does not seem to survive the loss of the last 19 amino acids of its 4-alpha-helix-like fold in the IL2-TNFRSF17 fusion transcript as the exposed hydrophobic surface of the truncated domain is prohibitively high at the site of truncation (Figure S3).

We found that several other types of domains display similar behavior, such as the von Willebrand factor A, the Fork head domain, the PHD-finger and several others (indicated in **bold italic** in the last column of Table 1). For the DAZAP1-MEF2D fusion protein the truncated RNA recognition motif, 2dgs_A has both its ends disordered within the last 11 and 14 amino acids. The same is true of the N-terminal 18 amino acids of the truncated 1×66_A SAM domain in MN1-ETV6 fusion protein. In the ATIC-ALK fusion protein, the truncated 1p1r_A chain (592 amino acids long) is split into two structural domains by SCOP [23] at residue 200, therefore the truncated segment is in reality only 30 amino acids long, corresponding to a fraction of 0.07 of the second SCOP domain, the latter being 392 residues long.

There are also several occurrences of consecutive short repeats and also one case of an elongated coil of alpha-helical repeats (Clathrin), which also seem to survive because repetitive structures in general are particularly well-suited for survival after truncation. Finally, there are 9 truncated domains (belonging to 7 different Pfam domain families) where there is no representative structure in the PDB, raising the suspicion (especially those with high predicted disorder values) that they could not be crystallized because they are intrinsically disordered.

**Contribution of disorder to the transforming activity of fusion proteins**

The oncogenic function of fusion proteins suggests that disorder not only is involved in the cellular survival of the fusion product,
but also in the novel function generated by the interplay of the two segments that became fused. Thus, clonal selection of cells harboring the fusion protein is promoted by the mechanism enabled by structural disorder. By considering functional information for some of the best characterized fusion proteins (Table 2) and the pattern of their predicted disorder (Figure 5), we argue that there are three basic mechanisms by which disorder contributes to the newly emerging oncogenic function.

(i) The first type is exemplified by BCR-ABL, where a Tyr-kinase phosphorylation motif in BCR gets fused with the Tyr-kinase domain within ABL [5]. Permanent phosphorylation of Tyr177 by ABL Tyr-kinase domain creates a binding site for the adapter protein Grb2 and sends continuous proliferation signals to the nucleus [5]. The phosphorylation site and kinase domain are 465 residues apart in the sequence, with a continuous stretch of disorder of about 257 amino acids (Figure 5A), which permits the chimera to fold back and undergo autocatalytic modification.

(ii) In the second type of mechanism, dimerisation and/or cytoplasmic relocalisation bring about permanent activation of receptor Tyr-kinases. In TFG-ALK, the TRK-fused gene (TFG) segment carries a coiled-coil dimerisation motif at a distance of 175 amino acids from the Tyr-kinase domain of ALK [5], which undergoes
auto-activation due to phosphorylation and presents a novel binding site to SH2 domain containing signaling proteins [24]. The intervening region has 138 of its amino acids disordered, which is probably critical for enabling the Tyr-kinase domains to engage in interdomain interactions.[...]

In the third type of mechanism the fusion of a DNA-binding element to a trans-activator domain results in an aberrant binding motif. For example, the fusion of Ewing sarcoma (EWS) transcription factor, where disorder enables the interplay of remote transcriptional activators that combine the highly repetitive, disordered EWS oncogene with transcription factors ATF1 and Fli1 [10] creates non-trivial cases of fusion proteins are shown where breakpoint falls into a Pfam domain. The abbreviated column identifiers are as follows: Pfam identifier; fplen, truncated domain length; bp, breakpoint; Dlen, domain length; Dbeg, Dend, domain match beginning and end, respectively; N/C, the retained half of the truncated domain; Dfrac, the retained fraction of the truncated domain; IUleft/IUright, the predicted disorder values for the truncated domain and its “mirror” (same number of amino acids as in the truncated domain) on the opposite side of the breakpoint. In the last column possible strategies are shown for the truncated domains to follow to avoid elimination by the proteasomal degradation system [49]. “No PDB” indicates the lack of any PDB structures associated with the protein family in question, which together with high predicted disorder values raises the suspicion of the domain being intrinsically disordered. When a PDB code is shown with a list of numbers (shown in italicized text) they indicate positions in the actual domains that are presumably indifferent to truncation based on the exposed hydrophobic surface of the truncated domain (as shown in detail in Figure 4).

doi:10.1371/journal.pcbi.1000552.t001
Discussion

Chromosomal translocations generating novel oncogenic fusion proteins represent the leading cause of neoplastic transformations in leukemias, lymphomas and sarcomas [1–4], but the structural characterization of proteins involved in fusion and/or the ensuing fusion proteins is largely incomplete thus far. Often, long regions of the proteins involved in the fusion event lack recognizable similarity to any other known protein, which is an indication of their likely structural disorder, as suggested in the case of CBP [9] and EEF [10]. Here we report that chromosomal translocations are highly correlated with structural disorder, and disorder also contributes to the oncogenic function elicited by the fusion event. Whereas translocating proteins tend to be longer than average human proteins, and longer proteins tend to be more disordered [19,20], the elevated disorder of translocation proteins cannot be accounted for by their increased length. Rather, there is strong and specific selection for proteins with elevated structural disorder following translocation/fusion, which has many interesting implications.

The signs of selection at the protein level have already been suggested in a previous study on the non-random position of breakpoints in translocation genes [30]. The major findings, i.e. that the breakpoint is almost invariably located in-frame and there is only a limited and highly biased set of domain combinations in translocating proteins, point to selection forces that act at the level of proteins. Clearly, a protein that can be translated into a viable product which has functional advantages in cellular proliferation and clonal expansion, is significantly more likely to be observed in cancer. Our results on structural disorder of the protein chimera extend these findings and provide a structural rationale at the whole-protein level.

Our key conclusion is that structural disorder makes the fusion product of two unrelated proteins look like a natural protein that does not activate cellular degradation pathways. Joining two truncated globular domains at random would generate a folding-incompetent protein, unable to pass the quality control of the cell; thus it would be rapidly degraded without any chance to confer a proliferative advantage on the cell that harbors it. Because IDPs are depleted of hydrophobic amino acids [11,12], the fusion of two IDRs does not expose more hydrophobic regions and does not cause major structural disturbances. Our findings are in accord with the suggestion that the cost the cell has to pay for the functional advantages conferred by structural disorder is that it is structurally permissive to joining unrelated proteins [8]. This is also underlined by the observation that breakpoints tend to avoid domains, the domains involved have an increased level of disorder and/or a close-to-normal ratio of hydrophobic residues exposed upon fusion. Overall, these preferences suggest that a high level and proper spacing of disordered regions increase solubility of truncated folding-incompetent segments probably in accord with an intramolecular chaperone effect [31].

In addition to survival in the cell, disorder may also be involved in the novel function(s) gained by fusion proteins. In several cases, the conformational freedom provided by long disordered segments enable the pathological interplay of remote functional elements brought together by the fusion event. There are several considerations in support of the role of disorder in these functions. For example, the transforming function of fusion products is rather insensitive to the actual position of the breakpoint, as shown by both the frequent occurrence of multiple breakpoints within the same protein, and a range of mutation studies where deletion of large regions of the fusion proteins leave their transforming activity intact [9,10]. These observations are highly reminiscent of the insensitivity of the function of disordered proteins to deleting/scrambling their residues, which have led to the notion of fuzziness [32]. In addition, disorder is critical in the physiological functions of proteins analogous to the oncogenic functions discussed here. Autophosphorylation of remote regulatory elements enabled by disorder is a recurring theme in the activation of signaling Tyr-
Acquiring and/or reconstructing the fusion proteins

Methods

Acquiring and/or reconstructing the fusion proteins

Human chromosomal translocations and the relevant genes/proteins were collected by sifting through Swissprot, NCBI's GenBank and TICdb [37–39], searching the annotations for key expressions such as “chromosomal translocation”, “chromosome translocation” or “fusion protein”. Breakpoints and gene names were recorded when this information was available in the databank entries. We focused on protein-coding entries: whenever there was a partial peptide sequence in the annotation part of a GenBank fusion entry, we used NCBI’s Tblastn [40] to compare the nucleotide sequence of the GenBank entry to the non-redundant set of all human proteins. The two most closely matching proteins (with percentage identity >95%), matching either end of the GenBank entry, were picked. This procedure resulted in 739 highly redundant protein identifiers. After comparing these proteins to one another with Blastp and replacing 42 gene names with their synonyms, the 739 proteins were found to belong to 305 different genes. We culled 101 more genes from the latest version of TicDB [39], altogether resulting in 406 translocation-related genes (Table S1, Supplementary material).

We next recorded the breakpoints in the annotation of the proteins or assigned them whenever translocation proteins from chimeric nucleic acid sequences in GenBank could be reconstructed; at least one breakpoint within the coding region could be identified in 146 genes. Using the transcript information in TicDB and the corresponding proteins in the Ensembl [41] database, the final number of translocation partner proteins increased to 255. Fusion proteins were next reconstructed from the chimeric nucleic acid sequences in GenBank by running Blastx to query NCBI’s non-redundant protein database, and also from the coordinates provided by TicDB, which resulted in the reconstruction of 187 fusion proteins. These correspond to 171 non-redundant fusion sequences at a sequence identity threshold of 90%.

| Fusion protein (breakpoint 1/breakpoint 2) | GenBank/Swissprot id (length) | Elements of oncogenic function in the fusion proteins | Distance/disorder between oncogenic elements | Reference |
|------------------------------------------|-------------------------------|-----------------------------------------------------|-------------------------------------------|-----------|
| BCR-ABL (426; 26)                        | AB60388 (1271)                | Oligomerization domain (BCR, 1-79)                  | 562/355                                   | [5,50]    |
| NPM-ALK (117; 1057)                      | NPM_HUMAN (294)               | Oligomerisation domain (NPM, 1-117)                | 60/2                                      | [26]      |
| EML4-ALK (496; 1057)                     | EMAL4_HUMAN (981)             | Basic dimerisation domain (EML4, 31-140)          | 415/89                                    | [27]      |
| EML4-ALK (496; 1057)                     | ALK_HUMAN (1620)              | Dimerisation domain (ALK, 555-822)                | 228/12                                    | [51]      |
| TPR-MET (218; 1009)                      | NP_003283.2 (2363)            | Leu-zipper (TPR, 75-99, 117-141)                  | 147/26                                    | [25]      |
| TFG-ALK (240; 1057)                      | NP_006061 (400)               | Coiled-coil dimerisation domain (TFG, 93-124)     | 175/138                                   | [24]      |
| TEL-JAK2 (337; 504)                      | NP_001978.1 (452)             | Dimerisation domain (TEL, 38-124)                 | 558/203                                   | [6]       |
| EWS-ATF1 (264; 109)                      | NP_005234.1 (656)             | EAD (EWS, 1-86)                                   | 280/265                                   | [28]      |
| MLL-CBP (1362; 266)                      | HRX_HUMAN (3969)              | AT-hooks (MLL, 169-180, 217-227, 301-309)         | 2403/1620                                 | [29]      |
| MLL-ENL (1362; 92)                       | HRX_HUMAN (3969)              | AT-hooks (MLL, 169-180, 217-227, 301-309)         | 1436/1270                                 | [7]       |

We collected functional information for 9 fusion proteins, which suggests that disorder of the region intervening newly joined functional regions contributes to oncogenic function. The table shows the distance and length of disorder separating the two elements required for the oncogenic function.

doi:10.1371/journal.pcbi.1000552.t002
Fusion proteins were analyzed for the occurrence of Pfam domains running Blastp against the entire set of Pfam-A domain sequences [43] and also the human subset of Pfam domains derived from Swissprot proteins. We set the thresholds for a domain match at an e-value<1e-06 and sequence similarity >60%. We found that at this similarity level there was less than 1% difference between the two sets of domain matches, so analyzed all 18,609 human Swissprot proteins for the occurrence of Pfam domains using only the Swissprot-derived human subset of Pfam. (Ideally, when looking for non-overlapping domain matches, choosing the best ones we would find only identity matches. We found that out of 29,946 non-overlapping Pfam domains in 14,541 human Swissprot proteins with at least one domain match there were only 1275 domain matches with less than 100% sequence similarity.) We further analyzed the translocation proteins with in-house Perl scripts, namely, (i) statistical significance of the difference between any two distributions was evaluated with the chi-square test; (ii) p-values corresponding to the calculated chi-square values and degrees of freedom were calculated by a computer program courtesy of Zsuzsa Dosztanyi; (iii) percentage values of disorder, length, etc. distributions were also calculated by own Perl scripts.

Calculation of exposed hydrophobic surface in the truncated Pfam domains

Actual values for the accessible nonpolar surface area (Anp) of truncated domains were determined as follows: The C-terminus of the protein structure was gradually truncated and the actual values of Anp for the truncated fragments were determined with the CHASA program [22] as suggested by [44]. The theoretical values were calculated using the formula by Chothia [45] and Janin [46].

The truncated domain structures were drawn and annotated using NCBI’s Cn3D program [47]. The high-resolution images in Figures 4 and S3 were created using the Polyview-3D server [48].

Supporting Information

Figure S1 Translocation frequency as a function of gene length. Found at: doi:10.1371/journal.pcbi.1000552.s001 (1.45 MB TIF)
Figure S2 Translocation frequency as a function of protein disorder. Ratio of the number of all human proteins and translocation proteins with breakpoint as a function of disorder (%IU). The best fitting trendline is a parable. Found at: doi:10.1371/journal.pcbi.1000552.s002 (0.13 MB EPS)
Figure S3 Effect of truncation on the Interleukin 2 domain. (A) On the top the theoretical and actual values for the accessible nonpolar surface area (Anp) of truncated domains are shown, calculated as described for Figure 4; (B) Structure of the IL2 domain (pdb code: 1irl). The C-terminal portion missing due to the translocation is colored grey. Found at: doi:10.1371/journal.pcbi.1000552.s003 (1.29 MB TIF)
Table S1 List of proteins observed in chromosomal translocations. 151 proteins without (Table S1A) and 255 proteins with (Table S1B) breakpoint (latter also indicated). The percentage disorder is also shown. Found at: doi:10.1371/journal.pcbi.1000552.s004 (0.07 MB XLS)
Table S2 List of all truncated Pfam domains and their mapping to the fusion proteins using Blastp. Found at: doi:10.1371/journal.pcbi.1000552.s005 (0.01 MB TXT)

Author Contributions

Conceived and designed the experiments: PT. Analyzed the data: HH LB. Wrote the paper: HH PT.
References

1. Aplan PD (2006) Causes of oncogenic chromosomal translocations. Trends Genet 22(1): 46–55.
2. Fattal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR (2004) A census of human cancer genes. Nat Rev Cancer 4(3): 177–83.
3. Rowley JD (2001) Chromosome translocations: dangerous liaisons revisited. Nat Rev Cancer 1(3): 245–50.
4. Goga A, McLaughlin J, Arai DE, Saffran DC, Witte ON (1995) Alternative signals to RAS for hematopoietic transformation by the BCR-ABL oncogene. Cell 82(6): 981–10.
5. Lacromièque V, Bourieux A, Vallée VD, Poirel H, Quang CT, Mauchauffe M, Berthou C, Lesnard M, Berger K, Ghyselen J, Bernard OA (1997) A TEL-JAK2 fusion protein with constitutive kinase activity in human leukemia. Science 278(5341): 1309–12.
6. Slany RK, Lavau C, Cleary ML (1998) The oncogenic capacity of HRX-ENL requires the transcriptional transactivation activity of ENL and the DNA binding motif of HRX. Mol Cell Biol 18(1): 122–9.
7. Pickard SJ. page 10. In: Molecular biology of the cell, 4th edn. Garland Science, New York, p 10.
8. Nk KP, Nk, S, Ennin RN, Denny CT, Uversky VN, Laee (2007) Multiple aromatic side chains within a disordered structure are critical for transcription and transforming activity of EWS family oncoproteins. Proc Natl Acad Sci U S A 104(2): 479–84.
9. Tompa P (2002) Intrinsically unstructured proteins. Trends Biochem Sci 27(10): 527–33.
10. Tsozoomin E, Oldfield CJ, Dunker AK (2003) showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. J Mol Recognit 15(1): 343–84.
11. Iakoucheva L, Brown G, Lawson J, Obradovic Z, Dunker AK (2002) Intrinsic Disorder in Cell-signaling and Cancer-associated Proteins. J Mol Biol 323(8): 573–584.
12. Minezaki Y, Honna K, Kinjo AR, Nishikawa K (2006) Human transcription factors contain a high fraction of intrinsically disordered regions essential for transcriptional regulation. J Mol Biol 361(4): 1137–49.
13. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 343(1): 635–45.
14. Romero PR, Zaidi S, Fang YY, Uversky VN, Radiovic P, Oldfield CJ, Cortese MS, Sickmeier M, LeGall T, Obradovic Z, Dunker AK (2006) Aligned multiple consensus in protein intrinsic disorder enables increased functional diversity in multicellular organisms. Proc Natl Acad Sci U S A 103(22): 8390–5.
15. Dozsmány Z, Czúmski V, Tompa P, Simon I (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21(16): 3433–4.
16. Farrow SJ, Higgins DG, Ouzounis CA, Lopez-Bigas N (2006) Structural and functional properties of genes involved in human cancer. BMC Genomics 7: 5.
17. Dozsmány Z, Chen J, Dunker AK, Simon I, Tompa P (2006) Disorder and sequence repeats in hub proteins and their implications for network evolution. J Proteome Res 5(11): 2985–95.
18. Ekman D, Light S, Björndal AK, Eklöf A (2006) What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae? Genome Biol 7(8): R45.
19. Lahti Y, Grass MP, Carbonnel F, Boullet JC, Berger R, Larsen CJ, Taipas A (1992) A new gene, BCM, on chromosome 16 is fixed to the interleukin 2 gene by a 4.16×0.26p11 translocation in a malignant T cell lymphoma. Embryo J 11(11): 3097–904.
20. Fleming PJ, Finzke NG, Mezari M, Srivivasan R, Rose GD (2005) A novel method reveals that solvent water favors polyproline II over beta-strand accessible surface area (CHASA). Protein Sci 14(4): 111–8.
21. Andreoca A, Howorth D, Chadomnia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2008) Data growth and its impact on the SCOP database: new accessible surface area (CHASA). Protein Sci 14(1): 111–8.
22. Mak HH, Peuschel P, Lin T, Nanjokas MA, Zuo D, Park M (2007) Oncogenic activation of the Met receptor tyrosine kinase fusion protein. Tp-Met, involves exclusion from the endocytic degradative pathway. Oncogene 26(51): 7211–21.
23. Bischof D, Pulkos K, Caden D, De Sousa J, Kim J, Hanler S, Zseby P, Smith IS, Broome S, Nishikawa K, Sugiyama Y, Ishikawa K, Ueno T, Hamada T, Haruta H, Kurashina K, Hatanaka H, Bando M, Ohno S, Ishikawa Y, Aburatani H, Niki T, Sahara Y, Sugiyama Y, Mano H (2007) Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature 448(7153): 561–6.
24. Pan S, Ming KY, Dunn TA, Li KK, Lee KA (1998) The EWS/AF11 fusion protein contains a dispersed activation domain that functions directly. Oncogene 16(12): 1625–31.
25. Shiryev SA, Papadopoulos JS, Schaller AF, Agarwala R (2007) Improved BLAST searches using longer words for protein seeding. Bioinformatics 23(21): 2949–51.
26. Flick P, Aken BL, Beal K, Ballestar E, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cottle T, Down T, Dyer SC, Eyer T, Fitzgerald S, Fernandez-Banet J, Grañ S, Haidar S, Hammond M, Holland R, How KI, Howe K, Johnson N, Junkins K, Kariha K, Kelle D, Kokoris F, Kulesha E, Lawson D, Longend L, Mogi K, Meindl P, Owenham B, Parker A, Pritchard B, Pril A, Rice S, Rieos D, Schuster M, Sealy I, Slater G, Smalley D, Spudich G, Tervahan S, Valla A, Vogel J, White S, Wood M, Barney E, Cox T, Curwen V, Durbin R, Fernandez-Saiz XM, Herrero J, Hubbard TJ, Kasprzyk A, Proctor G, Smith J, Ureña-Vidal A, Searle S (2008) Ensembl 2008. Nucleic Acids Res 36(Database issue): D707–14.
27. Martin MJ, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Sickmeier M, Hamilton JA, LeGall T, Obradovic Z, Dunker AK (2006) Improved mapped translocation breakpoints in cancer. BMC Genomics 8: 33.
28. Pan S, Ming KY, Dunn TA, Li KK, Lee KA (1998) The EWS/AF11 fusion protein contains a dispersed activation domain that functions directly. Oncogene 16(12): 1625–31.
29. Shiryev SA, Papadopoulos JS, Schaller AF, Agarwala R (2007) Improved BLAST searches using longer words for protein seeding. Bioinformatics 23(21): 2949–51.
30. Flick P, Aken BL, Beal K, Ballestar E, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cottle T, Down T, Dyer SC, Eyer T, Fitzgerald S, Fernandez-Banet J, Grañ S, Haidar S, Hammond M, Holland R, How KI, Howe K, Johnson N, Junkins K, Kariha K, Kelle D, Kokoris F, Kulesha E, Lawson D, Longend L, Mogi K, Meindl P, Owenham B, Parker A, Pritchard B, Pril A, Rice S, Rieos D, Schuster M, Sealy I, Slater G, Smalley D, Spudich G, Tervahan S, Valla A, Vogel J, White S, Wood M, Barney E, Cox T, Curwen V, Durbin R, Fernandez-Saiz XM, Herrero J, Hubbard TJ, Kasprzyk A, Proctor G, Smith J, Ureña-Vidal A, Searle S (2008) Ensembl 2008. Nucleic Acids Res 36(Database issue): D707–14.
31. Tompa P, Csermely P (2004) The role of structural disorder in the function of genes involved in human cancer. BMC Genomics 7: 4.