BMC Bioinformatics

Research article

No statistical support for correlation between the positions of protein interaction sites and alternatively spliced regions

Marc N Offman[1], Ramil N Nurtdinov[2], Mikhail S Gelfand[3,4] and Dmitrij Frishman*[1]

Address: [1]Department of Genome Oriented Bioinformatics, Technical University of Munich, Wissenschaftszentrum Weihenstephan, 85350 Freising, Germany, [2]Department of Bioengineering and Bioinformatics, Moscow State University, Lab. Bldg. B, Leninskie Gory 1-73, Moscow, 119992, Russia, [3]Institute for Problems of Information Transmission, Russian Academy of Sciences, Bolshoy Karetny per. 19, Moscow, 127994, Russia and [4]State Scientific Center GosNIIGenetika, 1st Dorozhny pr. 1, Moscow 117545, Russia

Email: Marc N Offman - offman@in.tum.de; Ramil N Nurtdinov - n_ramil@mail.ru; Mikhail S Gelfand - gelfand@iitp.ru; Dmitrij Frishman* - d.frishman@wzw.tum.de

* Corresponding author

Published: 19 April 2004

BMC Bioinformatics 2004, **5**:41

Received: 13 January 2004
Accepted: 19 April 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/41

Abstract

Background: Alternative splicing is an efficient mechanism for increasing the variety of functions fulfilled by proteins in a living cell. It has been previously demonstrated that alternatively spliced regions often comprise functionally important and conserved sequence motifs. The objective of this work was to test the hypothesis that alternative splicing is correlated with contact regions of protein-protein interactions.

Results: Protein sequence spans involved in contacts with an interaction partner were delineated from atomic structures of transient interaction complexes and juxtaposed with the location of alternatively spliced regions detected by comparative genome analysis and spliced alignment. The total of 42 alternatively spliced isoforms were identified in 21 amino acid chains involved in biomolecular interactions. Using this limited dataset and a variety of sophisticated counting procedures we were not able to establish a statistically significant correlation between the positions of protein interaction sites and alternatively spliced regions.

Conclusions: This finding contradicts a naïve hypothesis that alternatively spliced regions would correlate with points of contact. One possible explanation for that could be that all alternative splicing events change the spatial structure of the interacting domain to a sufficient degree to preclude interaction. This is indirectly supported by the observed lack of difference in the behaviour of relatively short regions affected by alternative splicing and cases when large portions of proteins are removed. More structural data on complexes of interacting proteins, including structures of alternative isoforms, are needed to test this conjecture.

Background

One of the main surprises resulting from the Human Genome Project was the realization that the enormous complexity of the human metabolism and regulation is

encoded by a relatively small number of genes. While the estimates of this number still vary significantly in the range of 25000 to 35000 [1] it is clear that we have fewer genes than maize [2], and only approximately five times more than the unicellular eukaryote *S. cerevisiae* [3]. Earlier predictions postulated that the human genome should contain between 50.000 and 100.000 genes [4]. These recent findings helped to realize the importance of post-transcriptional gene regulation in generating the proteomic complexity of the human cell.

One of the mechanisms for generating the protein diversity in eukaryotes is alternative splicing (AS). Current estimates of the prevalence of AS in the human genome range from one third of genes [5,6] to about 60% [7]. In fact, analysis of EST data showed that almost all human genes have possible alternatively spliced forms, although many of them seem to be non-functional [8]. AS is also widely observed in genomes of other multicellular eukaryotes [9]. There are two aspects of the AS phenomenon. Firstly, recent results demonstrate that alternative isoforms, especially tissue-specific ones, are often evolutionary young [10-12]. One possible interpretation for that is that AS provides a convenient evolutionary mechanism for generating new proteins without sacrificing existing ones [10]. Secondly, AS leads to generation of proteins identical in some domains and different in others. This is a powerful regulatory mechanism [13,14]. Indeed, it has been shown that AS, compared to a random expectation, tends to avoid disrupting protein domains, to shuffle entire domains, and to target functional sites in proteins [15]. AS is frequent in genes involved in signal transduction and regulatory interactions [16].

To summarize, AS is a powerful mechanism for modulating the protein mode of action. Perhaps the most important type of functional context in which proteins execute their function is constituted by protein-protein interactions (PPI). It is thus tempting to speculate that AS could influence the structure of protein-protein interaction networks by selectively blocking or activating individual interactions dependent on cellular conditions.

Occasional evidence of the interplay between AS and PPI is scattered in the biomedical literature. For example, the domain composition of the Shank postsynaptic density protein is regulated by AS, defining the spectrum of its interaction partners [17]. The interaction of Annexin XI with calcyclin was shown to be isoform specific, owing to partial deletion of the calcyclin-binding site in one of the splice forms [18]. In *Drosophila melanogaster* the regulation of dynein targeting to various cellular organelles via binding to dynactin is modulated by structural variations in the N-terminal portion caused by AS [19]. At the same time, other studies show no interdependence between protein interaction patterns and AS. For example, Liu *et al*. [20] reported that specific protein-protein interactions of cytoplasmic serine hydroxymethyltransferase are not regulated by AS.

To test the hypothesis that protein-protein interactions can be modulated in an alternative splicing dependent manner we have undertaken a study of AS events occurring in proteins known to be involved in transient interactions (see Figure 1). We were able to identify the total of 42 alternatively spliced isoforms in the amino acid chains participating in structurally characterized interaction complexes, as surveyed recently by Thornton and Nooren [21]. In addition to statistics based on the entire dataset we also considered separately AS events changing more than a half of the protein length, where complete domain disruption is a likely result of AS [15], as well AS influencing less than 25% that could be expected to have less drastic effect. In all cases, no statistically significant relationship between AS and PPI could be established.

Results

We begin with presenting several examples illustrating the interplay between PPI and AS. In the first case, shown in Figure 2, two alternatively spliced regions were identified in the cyclin-dependent kinase 2 (cdk2). The first (N-terminal) splice event eliminates nearly completely the PSTAIRE region of cdk2 which plays a central role in its interaction with cyclin [22]. The second (C-terminal) sequence span deleted by splicing overlaps partially with the T-loop region of cdk, also involved in cdk2-cyclin interface. In the second example, presented in Figure 3, AS eliminates a large portion of the phosducin N-terminal domain, including one of the regions that contact the βγ-subunits of the retinal G-protein transducin [23]. The corresponding isoform, called phosducin-like orphan protein 1, was reported to be less abundant than phosducin and not to be able to bind transducin, although the functional significance of this and other phosducin orphans remains unclear [24]. Finally, in our third positive example (Figure 4) a significant portion of the interface between the human importin-β and the Ran protein [25] is spliced out, including the entire tandem repeat 7 as well as parts of repeats 6 and 8. This region also includes the so called "stalk region" with unknown function.

In all three examples above there is a clear overlap between the location of alternatively spliced regions and protein interaction sites. In many other cases, though, no such correlation could be established. For example, as seen in Figure 5, AS does not seem to affect the regions of the RhoA protein with which it binds the effector domain of serine-threonine kinases [26]. Based on an extensive manual analysis of our PPI/AS dataset we were not able to

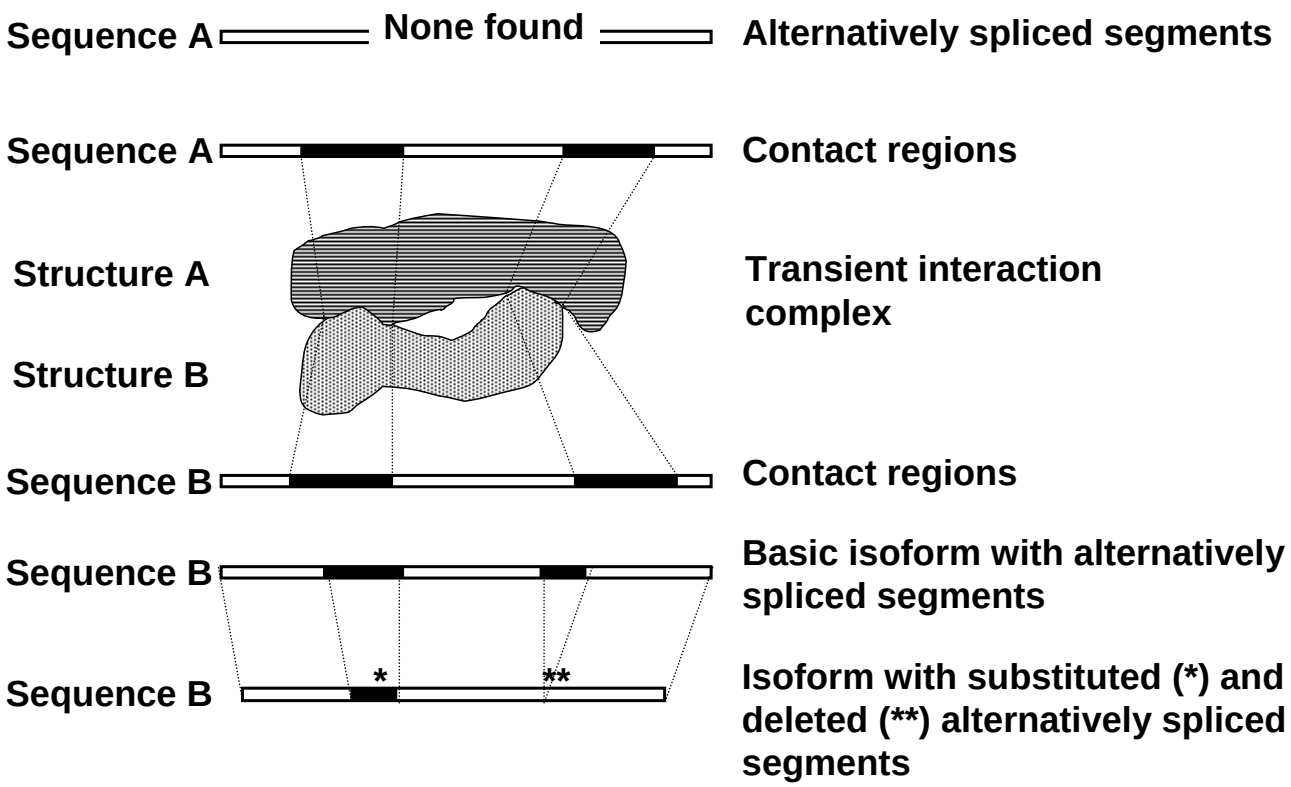


Figure 1
A schematic representation of the study. The positions of sequence regions involved in protein-protein interactions were delineated from structurally characterized protein complexes and juxtaposed with the location of putative alternatively spliced regions.

formulate a definitive opinion as to the degree to which the location of AS regions correlates with PPI interfaces.

We subsequently attempted to investigate the strength of the correlation between the positions of AS and PPI regions using statistical analysis. Tables 1 and 2 present the entire body of structural PPI evidence and AS information used for this purpose. Our analysis resulted in delineation of 30 alternatively spliced isoforms for 16 proteins involved in heterodimer interactions and 12 isoforms for 5 proteins involved in homodimer interactions. These data were used as input to different counting procedures described in the *Methods* section. The overall summary for all types of dimers (with smoothing) is given in Table 3 (for individual positions; [see Additional file 1]) and Table 4 (for entire segments; [see Additional file 2]).

As shown in Table 3, in 15 out of 42 cases, the majority of random windows show less correlation with the contact positions than the real AS segments, in 3 cases the random windows show the same correlation, and in 24 cases the majority of random windows show more correlation with the contact positions than the real AS segment(s). Similarly, for contact segments (Table 4) in 13 cases the majority of windows show less correlation with the contact segments than the real AS segments, in 4 cases the correlation is the same, and in 25 cases most random windows show more correlation. Thus, compared to random control, AS shows a weak tendency to avoid contact regions; however, the chi-squared test shows that in both cases the difference is insignificant. The results of our study are thus negative: the hypothesis that AS might serve as a regulation mechanism for PPI networks could not be confirmed.

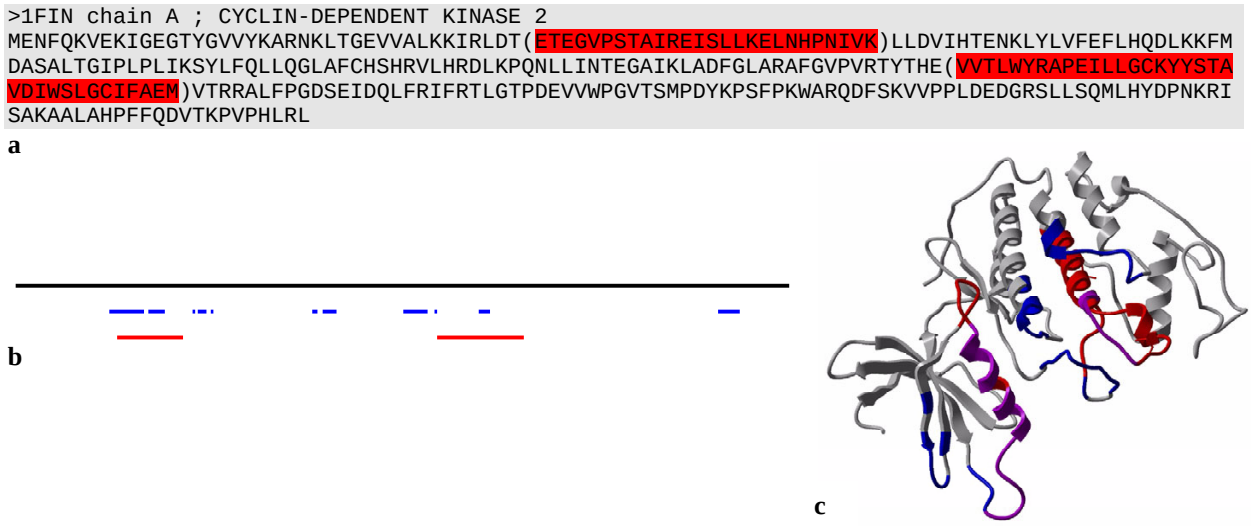


Figure 2
Alternative splicing in the cyclin-dependent kinase 2 (cdk2). a. Amino acid sequence of the protein. Spliced regions are colored red. The first (N-terminal) region was identified based on the homology with the deletion type cdk2 variant in human breast cancer (protein id BAA32794.1). The second (C-terminal) region was identified based on the homology with the human cDNA sequence associated with leiomyosarcoma (GenBank accession number BQ225275). b. Graphical representation of spliced regions (red) and sequence spans involved in protein-protein interactions (blue). The full amino acid sequence is shown as black bar. c. Ribbon representation of the three-dimensional structure of cdk1 (PDB code 1fin, chain a). Regions involved in protein-protein interactions are shown in blue colour, spliced regions are red. Violet colour indicates regions where PPI and AS regions overlap.

Discussion

This study represents the first attempt to investigate the relationship between two complicated biological phenomena – protein-protein interactions and alternative splicing – based on the scarce experimental evidence currently available in the public databases. It was prompted by results of Kriventseva *et al.* [15] that could be interpreted as evidence that AS tends to target functionally important regions in proteins. However, we were not able to obtain convincing evidence of correlation between contact regions in protein-protein interactions and protein segments corresponding to alternatively spliced regions in mRNA. One possible explanation for that could be that our database of AS events was corrupted by aberrant splicing that abounds in EST databases [8]. To guard against this possibility we considered only AS events expected to be reliable by multiple criteria, including presence in multiple ESTs from several clone libraries and conservation in the mouse genome. Another explanation could be that removal by AS of more than a half of the protein sequence would lead to complete disruption of the 3D protein structure, and thus any analysis relying on contact regions in the intact protein becomes meaningless.

However, we obtained essentially the same results on a subset of cases where the influence of AS was relatively modest (less than 25% of the protein sequence removed or substituted by AS; data not shown). Such shorter alternatively spliced regions are still much longer that a typical functionally important sequence motif of the type considered in [15]. On the other hand, as described above, complete removal or substitution of the interaction domain is a well-established mechanism of regulating PPI by AS.

Given more data, it would be interesting to repeat the analysis of the type presented here on the domain level. For instance, one of the remaining open questions is whether AS tends to target (completely remove or partially destroy) interacting domains or other domains of proteins involved in PPI. Partially, this issue was addressed in a recent study [27] where it was shown that domains of types involved in protein-protein interactions frequently are removed or inserted by alternative splicing, although there was no observed increase in the rate of alternative splicing of such domains compared to other types of domains.

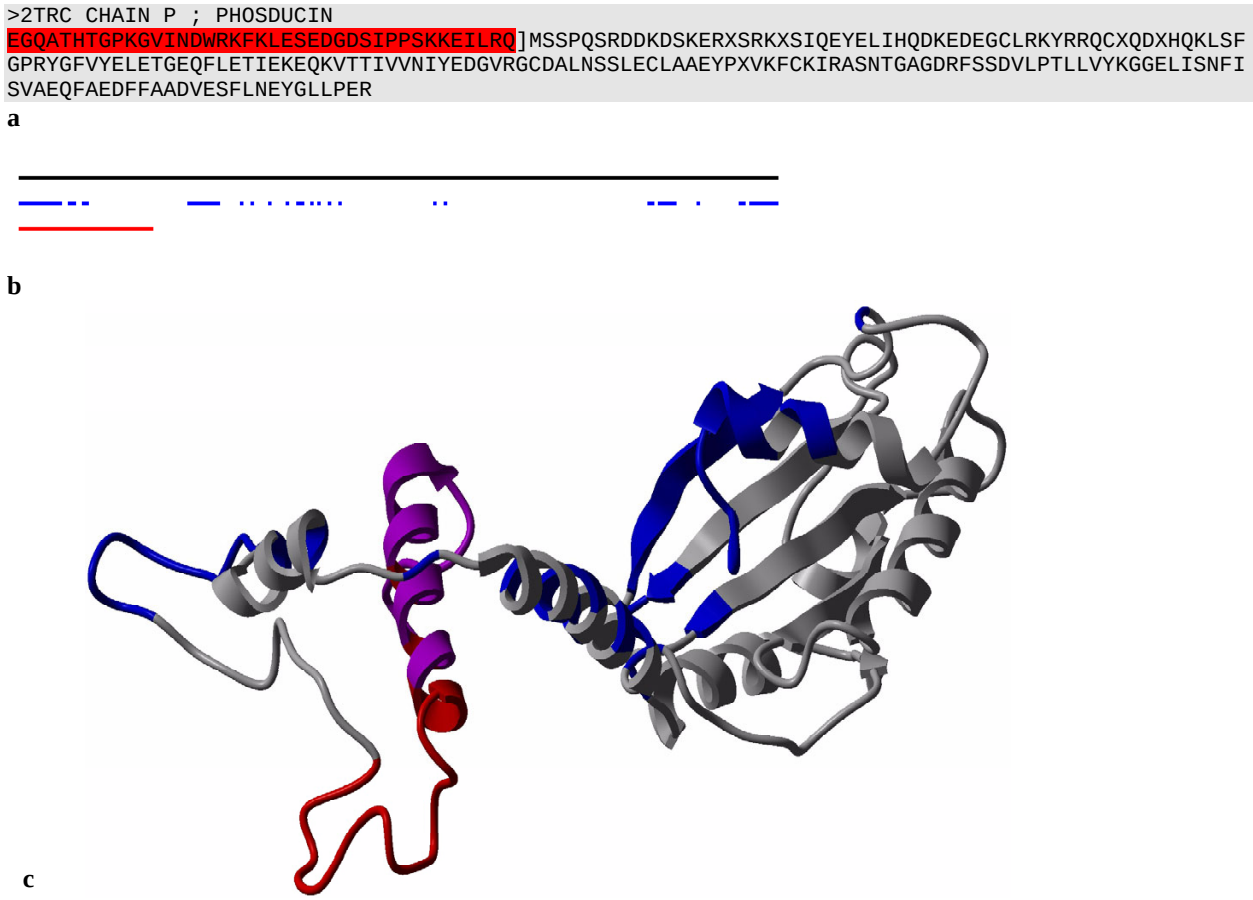


Figure 3
Alternative splicing in phosducin. a. Amino acid sequence of the protein. The spliced region was identified based on the homology with the human phosducin isoform b, also called phosducin-like orphan protein 1 (GenBank accession number NP_072098). b. Graphical representation of the spliced and interacting regions. c. Ribbon representation of the three-dimencional structure (PDB code 2TRC, chain p). Colouring as in Figure 2.

Conclusions

In this study we were not able to establish a statistically significant correlation between the positions of protein interaction sites and alternatively spliced regions. This finding implies that alternative splicing is not a preferred mechanism for controlling protein interaction networks, although many individual protein interactions are known to be isoform-dependent.

Methods

Protein interaction data

A comprehensive survey of structurally characterized transient protein interaction complexes has been recently published [21]. We used tables 3 and 4 from this study which list the PDB [28] codes and chain identifiers of interacting proteins pairs in 24 weak homodimers and 25 heterodimers, respectively. Protein sequence spans involved in physical interactions were identified with the help of the Protein-Protein Interaction Server http://www.biochem.ucl.ac.uk/bsm/PP/server[29]. Contacts between amino acid residues were defined based on the distance threshold of 6 Å between any pair of non-hydrogen atoms. PDB sequences were saved and the positions of residues participating in contacts recorded for each sequence separately.

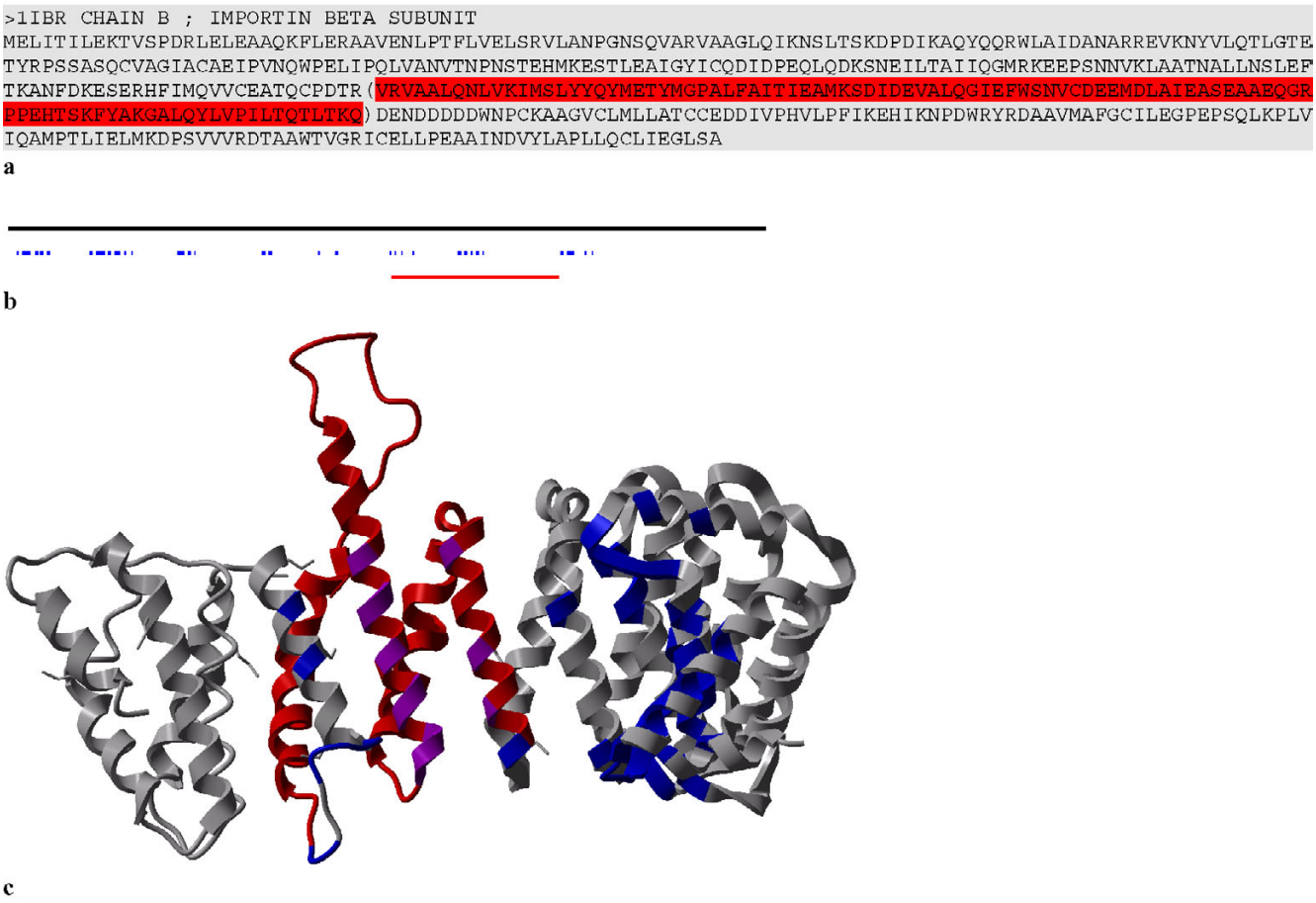


Figure 4
Alternative splicing in importin-β. a. Amino acid sequence of the protein. The spliced region was identified based on the homology with the human cDNA sequence with the GenBank accession number BE744080. b. Graphical representation of the spliced and interacting regions. c. Ribbon representation of the three-dimencional structure (PDB code 1IBR, chain b). Colouring as in Figure 2.

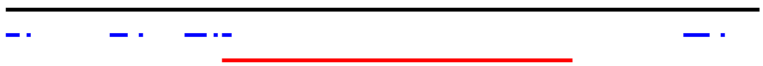
Detection of AS events

Amino acid chains with known positions of interaction regions were further analysed to detect putative AS events. Initial mapping of the protein sequences to the human genome was conducted using TBLASTN [30] searches. Each hit that coincided both with a human gene annotated in LocusLink [31] and the UniGene [31] cluster linked to the latter was used to obtain a complete set of mRNA and EST sequences corresponding to the gene. mRNAs with annotated protein-coding regions were used to derive the corresponding protein sequences. The exon-intron structure of each gene with all known alternatives was established by spliced alignment of ESTs and mRNAs to the genomic DNA sequence using Pro-EST [5]; proteins were spliced-aligned to DNA using Pro-Frame [32]. For further analysis only alternatives changing the coding region covered by the known three-dimensional structure were considered.

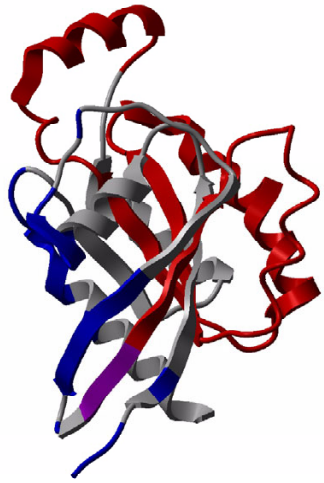
An alternative region (that is, a region present in a fraction of isoforms) was assumed to be reliable if it was supported by a full-length mRNA or a protein, or by several ESTs from different clone libraries. An alternative region was discarded if it was observed in less than three ESTs or ESTs from only one clone library. Further, only conserved AS events, defined by comparison with the mouse genome, were retained. The procedure for establishing the conservation of the alternative splicing events is described in detail in [11]. Briefly, for each human gene the orthologous mouse gene was obtained from HomoloGene [31],

>1CXZ Chain A ; HIS-TAGGED TRANSFORMING PROTEIN RHOA
MAAIRKKLVIVGDVACGKTCLLIVFSKDQFPEVYVPTVFENYVADIEVDGKQ(VELALWDTAGQEDYDRLRPLSYPDTDVILMCFSIDSPDSLENIPEKWTP
EVKHFCPNVPIILVGNKKDLRNDEHTRRELAKMKQ)EPVKPEEGRDMANRIGAFGYMECSAKTKDGVREVFEMATRAALQA

a



b



c

Figure 5
Alternative splicing in the human *rhoa* protein. a. Amino acid sequence of the protein. The splicing region was identified based on the homology with the human cDNA sequence associated with large cell carcinoma (GenBank accession number BQ231766) as well as several other cDNAs. b. Graphical representation of the spliced and interacting regions. c. Ribbon representation of the three-dimensional structure (PDB code 1CXZ, chain a). Colouring as in Figure 2.

Table 1: The dataset of heterodimers.

Interacting protein[a]		Contact Data[b]		Alternative Splicing Data			
Chain	**Length**	**Contact area (Å²)**	**Interacting residues**	**Type[c]**	**Start[d]**	**End[d]**	**Length[e]**
1AM4-A	199	856.04	84–88, 119, 122, 123, 126, 127, 185, 189–194, 196–202, 205, 211–213, 215–218, 220	del., subst.	103	234	28
1AM4-D	174	958.07	511–513, 532–539, 556, 560–564, 566, 567, 570, 571, 586, 588, 592, 596	del., subst.	536	674	6
				del., subst.	663	674	29
				del., subst.	663	674	1
1BKD-R	166	1655.36	5, 12–18, 20, 21, 25, 30–35, 37, 40, 41, 54–71, 73, 95, 98, 99, 102, 103, 105	del., subst.	151	166	16
1C1Y-A	167	695.37	3, 21, 24, 25, 27, 29, 33, 34, 37, 38–42, 52, 54, 56, 63, 71	del., subst.	20	167	15
				del., subst.	62	167	33
1C1Y-B	77	616.01	55, 57, 59, 62, 64–71, 73, 84, 85, 87–91	del.	70	131	0
1CXZ-A	182	892.09	0–3, 5, 25–29, 32, 43–48, 50, 52–54, 163–169, 172	del., subst.	52	181	38
				del.	52	136	0
				del., subst.	139	181	48
				del., subst.	52	181	1
1E96-A	178	579.01	21, 22, 24–33, 36, 40, 41, 159–162	Ins.	75	76	19
				del., subst.	13	178	18
				del., subst.	76	178	40
1FIN-A	289	1609.88	37–50, 52–58, 69, 71–74, 76, 115, 116, 119–124, 150–159, 162, 179–183, 271–279	del.	163	196	0
				del.	40	65	0
					163	196	0
1FIN-B	260	1794.40	173–178, 181, 182, 185, 186, 189, 228, 230, 262, 263, 265–272, 274, 275, 288, 289, 292, 295–300, 302–309, 312–317	del., subst.	266	432	4

Table 1: The dataset of heterodimers. *(Continued)*

1I2M-A	165	1438.50	17–21, 23, 67–77, 91–103, 106–108, 110, 133, 134, 137, 138, 140	del., subst.	84	173	6
				Ins.	83	84	17
				del.	9	12	0
				del., ins.	9	12	0
					83	84	17
1I2M-B	388	1341.05	42, 44, 45, 55, 56, 75–77, 93–96, 106, 109, 128, 129, 147–152, 181, 200, 201, 249, 250, 266, 268–271, 278, 279, 303, 304, 320, 322, 323, 325, 334, 354, 355, 371, 373, 374, 382, 384, 407–409	del., subst.	182	411	5
				Ins.	43	44	17
				Ins.	43	44	31
				ins., del., subst.	43	44	17
					182	411	5
1IBR-B	458	1646.59	7, 10–15, 18, 19, 21, 22, 25, 26, 51, 52, 55–60, 62, 63, 66–69, 72, 76, 104–108, 110, 111, 114, 155–157, 159, 160, 189, 199, 200, 232, 235, 239, 246, 273–275, 277, 278, 281, 284, 285, 288, 335, 336, 339–343, 350, 354	del.	233	334	101
				del.	263	334	71
1LFD-A	87	583.07	18, 20, 27–35, 51–54, 56	del., subst.	37	100	17
1WQ1-G	320	1391.63	745, 746, 749, 750, 782–793, 795, 796, 799, 802, 803, 831, 833, 894–898, 901–904, 906, 907, 910, 911, 914, 927, 928, 931, 934, 935, 938, 939, 942, 944, 947–952	del., subst.	897	1037	15
2TRC-P	212	2251.56	14–26, 28–30, 32, 33, 62–71, 77, 80, 85, 90, 93–95, 97, 99, 102, 105, 132, 135, 193, 194, 196–201, 207, 219, 220, 222–230	del	14	52	0
2TRC-B	340	2175.25	8, 12, 42, 44–48, 55–57, 59, 75–77, 96–101, 116–119, 143–148, 161–164, 184–186, 188, 203–206, 226–230, 246, 266, 268, 270–274, 288–292, 295, 304, 306–317, 332, 333, 335, 337, 339	del	20	33	0

[a] Nooren & Thornton, 2003 [b] http://www.biochem.ucl.ac.uk/bsm/PP/server [c] del. = deletion of residues, subst. = substitution of residues. [d] residue numbers according to PDB file numbering [e] Length of the substituted part.

Table 2: The dataset of homodimers.

Interacting protein[a]		**Contact data[b]**		**Alternative splicing data**			
Chain	**Length**	**Contact area (Å²)**	**Interacting residues**	**Type[c]**	**Start[d]**	**End[d]**	**Length[e]**
1IKN-A	285	698.19	195, 197–201, 211, 213–218, 242, 243, 245, 246, 248–251	subst.	19	24	7
				del.	85	112	0
				del. subst.	85	290	18
				del.	113	290	0
				del. subst.	186	291	1
				del.	186	219	0
1A15-A	67	752.65	20–31, 35, 36, 61, 62, 64–67	del.	16	39	0
				del. subst.	63	67	10
				del.	40	49	0
1DOM-A	76	849.23	3–18, 20, 29, 31, 33–38, 40–43, 49–53	del. subst.	42	76	1
1CNT-1	177	954.74	11, 19, 81, 84, 91, 92, 95, 96, 100–103, 106–111, 113–115, 117, 118, 120–122, 124, 125, 128, 129, 133–137, 140	del. subst.	49	187	24
1TRZ-A	21	74.10	21	1	15		

[a] Nooren & Thornton, 2003 [b] http://www.biochem.ucl.ac.uk/bsm/PP/server [c] del. = deletion of residues, subst. = substitution of residues. [d] residue numbers according to PDB file numbering [e] Length of the substituted part

and human protein isoforms were spliced-aligned to the mouse gene DNA using Pro-Frame. An alternative region was considered as conserved if it could be aligned with the same similarity level as the rest of the protein, and the corresponding mouse exons were bounded by canonical splicing sites.

Counting coincidences of AS and PPI regions

Positions of contact regions in protein sequences were compared with positions of AS regions to determine whether there was a significant correlation between them. For each protein sequence considered, two bit strings were generated representing the involvement of each amino acid position in AS and PPI, as depicted in Figure 6. One technical difficulty is that many contact regions are very short (often comprised by just one amino acid residue) and/or may be interrupted by individual residues or groups of residues that do not take part in interactions. We thus used two smoothing parameters – N and M – to group interacting residues in a given amino acid chain into contiguous contact segments. N was defined as the minimal allowed length of a spacer between two residues involved in interactions, and M represents the minimal allowed length of a contact segment. Contact segments separated by spacers shorter than N were merged, and then the segments shorter than M were deleted. We tested different combinations of N and M, and found that the obtained results were robust as regards these parameters (data not shown). The results for N = 5 and M = 1 are reported below.

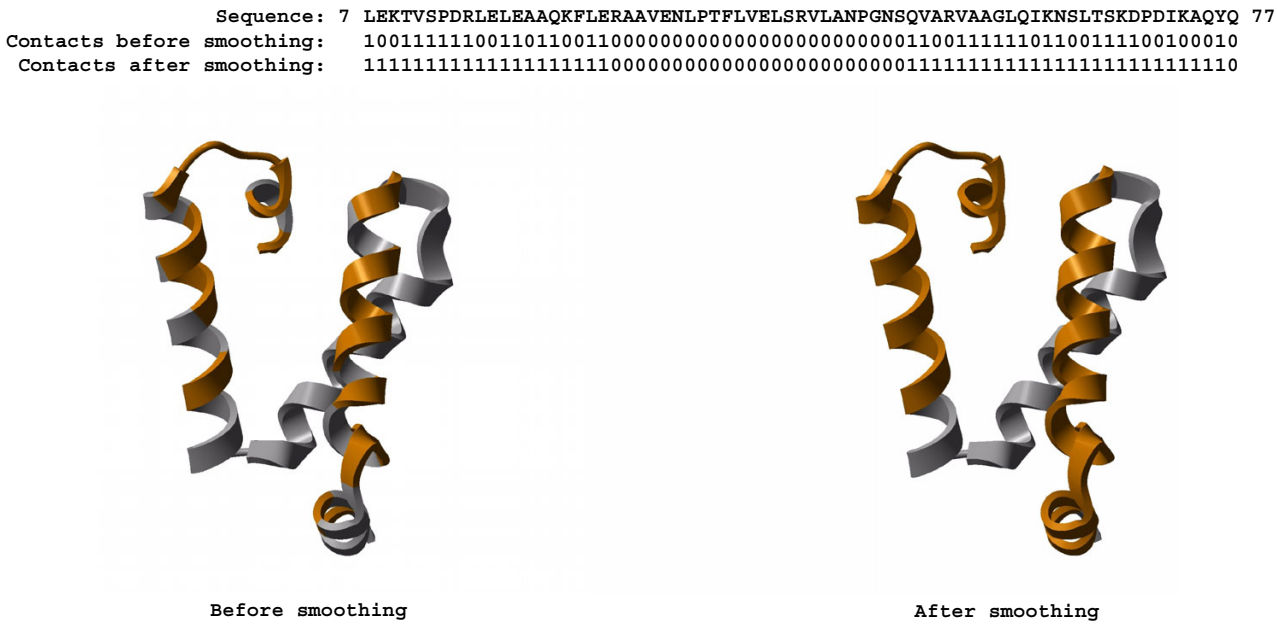


Figure 6

Smoothing positions of PPI regions. Amino acid sequence is represented as a bit string in which 1 indicates those residue positions that are involved in PPI, and 0 – those that are not. Subsequently smoothing is conducted such that contact segments separated by spacers shorter than N residues are merged, and then the segments shorter than M are deleted. In this example a sequence fragment of the human importin β-chain (PDB code 1ibr; residue positions from 7 to 77) is shown which interacts with the GTP-binding nuclear protein RAN. The interaction involves residues situated on one side of two α-helices. After smoothing with N = 5 and M = 1 entire helices are considered interacting regions.

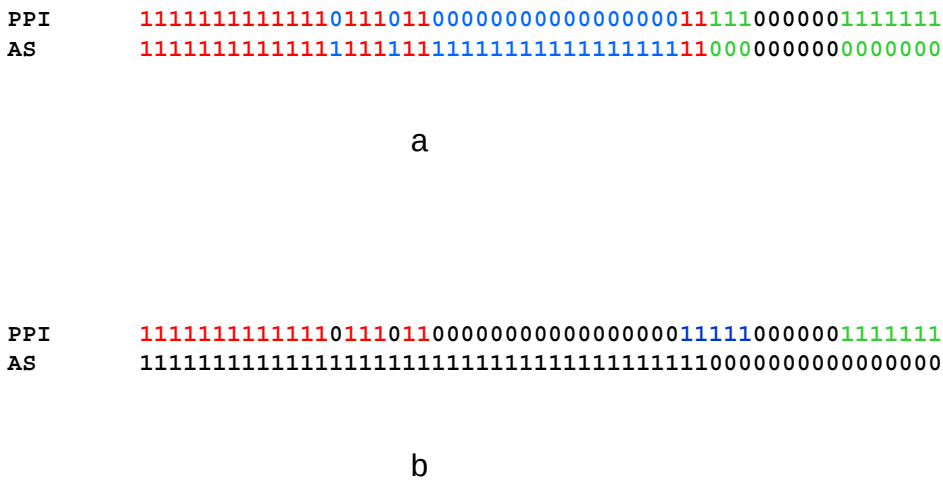


Figure 7
Counting coinciding PPI and AS regions. For clarity, no smoothing was made. a. By individual residues. In this example, there are 20 amino acid positions involved both in PPI and AS (PPI/AS, red colour), 19 non-PPI/AS positions (blue colour), 10 PPI/non-AS positions (green colour), and 6 non-PPI/non-AS positions (black colour). b. By entire sequence segments. In this example there are three PPI regions entirely covered by AS regions (red), one PPI region partially overlapping with an AS region (blue), and one PPI region not overlapping with AS.

Several different approaches to calculate the correlation between AS and PPI bit strings with and without smoothing were tested. The first, residue-by-residue approach involved considering each amino acid position separately and counting the number of positions in which the values of AS and PPI bit strings are equal or differ (Figure 7a). Each position can thus be classified as belonging to one of the four classes: AS/PPI, AS/non-PPI, non-AS/PPI, or non-AS/non-PPI. In a second, area-based approach (Figure 7b), we first grouped the adjacent positions of bit strings possessing identical values and then calculated the correlation between occurrences of such groups as a whole. Entire PPI segments were classified into three classes: entirely overlapping with AS, partially overlapping with AS, and non-overlapping with AS.

To estimate the degree of correlation, two types of statistical analysis were employed. Initially we applied the standard χ^2 test to the contingency tables formed by two parameters describing each position (PPI/non-PPI, AS/non-AS). However, as the results of this analysis were inconclusive (data not shown), we compared the observed correlation with that assuming fixed contacts and random placement of AS segments (*cf*. the procedure in [15]). Formally, for each AS isoform we considered a window of the same length as the length of the alternatively spliced segment. For each position of the window the correlation between the contact positions (or entire segments) and the current position of the window was computed as described in the previous paragraph. Then for the current position of the window the computed correlation and the correlation of the real AS segment were compared. The current window was subsequently classified into one of three classes: higher correlation, same correlation, and lower correlation. The same procedure was generalized for the case of two (or more) AS segments in an isoform. In this case, all non-overlapping pairs of windows were considered and the same procedure was applied. Again, the window sizes were set to the sizes of the real AS segments.

Authors' contributions

MO conducted protein structure analysis and implemented statistical techniques used in this work. He also

did a major part of data analysis work, created statistical tables and figures presented in this paper. RNN identified the positions of alternatively spliced regions based on comparative genome analysis and spliced alignment. MSG proposed the statistical methods used in this work and supervised their development in collaboration with MO. DF and MSG, together, conceived the original hypothesis of this study. DF proposed the specific approach adopted in this study and supervised the work of MO in the part related to protein structure analysis. The manuscript was written by DF and MSG. All authors read and approved the manuscript.

Acknowledgements

This study was partially supported by grants from the Ludwig Institute of Cancer Research (CRDF RB0-1268), the Howard Hughes Medical Institute (55000309), Russian Fund for Basic Research (04-04-49440), the Fund for Support of the Russian Science, and the Program in Molecular and Cellular Biology of the Russian Academy of Sciences.

References

1. **Ensembl Genome Browser** [http://www.ensembl.org/Homo_sapiens/]
2. **maizegenome.org – Why Maize?** [http://www.maizegenome.org/why_maize.html]
3. **Comprehensive Yeast Genome Database** [http://mips.gsf.de/genre/proj/yeast/index.jsp]
4. **Gene Sweep 2003–2004** [http://www.ensembl.org/Genesweep]
5. Mironov AA, Fickett JW, Gelfand MS: **Frequent alternative splicing of human genes.** *Genome Res* 1999, **9:**1288-1293.
6. Brett D, Hanke J, Lehmann G, Haase S, Delbruck S, Krueger S, Reich J, Bork P: **EST comparison indicates 38% of human mRNAs contain possible alternative splice forms.** *FEBS Lett* 2000, **474:**83-86.
7. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, Szustakowki J, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
8. Kan Z, States D, Gish W: **Selecting for functional alternative splices in ESTs.** *Genome Res* 2002, **12:**1837-1845.
9. Brett D, Pospisil H, Valcarcel J, Reich J, Bork P: **Alternative splicing and genome complexity.** *Nat Genet* 2002, **30:**29-30.
10. Modrek B, Lee CJ: **Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss.** *Nat Genet* 2003, **34:**177-180.
11. Nurtdinov RN, Artamonova II, Mironov AA, Gelfand MS: **Low conservation of alternative splicing patterns in the human and mouse genomes.** *Hum Mol Genet* 2003, **12:**1313-1320.
12. Thanaraj TA, Clark F, Muilu J: **Conservation of human alternative splice events in mouse.** *Nucleic Acids Res* 2003, **31:**2544-2552.
13. Boise LH, Gonzalez-Garcia M, Postema CE, Ding L, Lindsten T, Turka LA, Mao X, Nunez G, Thompson CB: **bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death.** *Cell* 1993, **74:**597-608.
14. Peneff C, Ferrari P, Charrier V, Taburet Y, Monnier C, Zamboni V, Winter J, Harnois M, Fassy F, Bourne Y: **Crystal structures of two human pyrophosphorylase isoforms in complexes with UDPGlc(Gal)NAc: role of the alternatively spliced insert in the enzyme oligomeric assembly and active site architecture.** *EMBO J* 2001, **20:**6191-6202.
15. Kriventseva EV, Koch I, Apweiler R, Vingron M, Bork P, Gelfand MS, Sunyaev S: **Increase of functional diversity by alternative splicing.** *Trends Genet* 2003, **19:**124-128.
16. Modrek B, Resch A, Grasso C, Lee C: **Genome-wide detection of alternative splicing in expressed sequences of human genes.** *Nucleic Acids Res* 2001, **29:**2850-2859.
17. Lim S, Naisbitt S, Yoon J, Hwang JI, Suh PG, Sheng M, Kim E: **Characterization of the Shank family of synaptic proteins. Multiple genes, alternative splicing, and differential expression in brain and development.** *J Biol Chem* 1999, **274:**29510-29518.
18. Sudo T, Hidaka H: **Regulation of calcyclin (S100A6) binding by alternative splicing in the N-terminal regulatory domain of annexin XI isoforms.** *J Biol Chem* 1998, **273:**6351-6357.
19. Nurminsky DI, Nurminskaya MV, Benevolenskaya EV, Shevelyov YY, Hartl DL, Gvozdev VA: **Cytoplasmic dynein intermediate-chain isoforms with different targeting properties created by tissue-specific alternative splicing.** *Mol Cell Biol* 1998, **18:**6816-6825.
20. Liu X, Szebenyi DM, Anguera MC, Thiel DJ, Stover PJ: **Lack of catalytic activity of a murine mRNA cytoplasmic serine hydroxymethyltransferase splice variant: evidence against alternative splicing as a regulatory mechanism.** *Biochemistry* 2001, **40:**4932-4939.
21. Nooren IM, Thornton JM: **Structural characterisation and functional significance of transient protein-protein interactions.** *J Mol Biol* 2003, **325:**991-1018.
22. Jeffrey PD, Russo AA, Polyak K, Gibbs E, Hurwitz J, Massague J, Pavletich NP: **Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex.** *Nature* 1995, **376:**313-320.
23. Gaudet R, Bohm A, Sigler PB: **Crystal structure at 2.4 angstroms resolution of the complex of transducin betagamma and its regulator, phosducin.** *Cell* 1996, **87:**577-588.
24. Craft CM, Xu J, Slepak VZ, Zhan-Poe X, Zhu X, Brown B, Lolley RN: **PhLPs and PhLOPs in the phosducin family of G beta gamma binding proteins.** *Biochemistry* 1998, **37:**15758-15772.
25. Vetter IR, Arndt A, Kutay U, Gorlich D, Wittinghofer A: **Structural view of the Ran-Importin beta interaction at 2.3 A resolution.** *Cell* 1999, **97:**635-646.
26. Maesaki R, Ihara K, Shimizu T, Kuroda S, Kaibuchi K, Hakoshima T: **The structural basis of Rho effector recognition revealed by the crystal structure of human RhoA complexed with the effector domain of PKN/PRK1.** *Mol Cell* 1999, **4:**793-803.

27. Resch A, Xing Y, Modrek B, Gorlick M, Riley R, Lee C: **Assessing the impact of alternative splicing on domain interactions in the human proteome.** *J Proteome Res* 2004, **3:**76-83.
28. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C: **The Protein Data Bank.** *Acta Crystallogr D Biol Crystallogr* 2002, **58:**899-907.
29. **Protein-Protein Interaction Server** [http://www.biochem.ucl.ac.uk/bsm/PP/server]
30. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
31. Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, Wagner L: **Database resources of the National Center for Biotechnology.** *Nucleic Acids Res* 2003, **31:**28-33.
32. Mironov AA, Novichkov PS, Gelfand MS: **Pro-Frame: similarity-based gene recognition in eukaryotic DNA sequences with errors.** *Bioinformatics* 2001, **17:**13-15.