Supporting Information

Protein Structure Prediction and Assessment

I-TASSER is a hierarchical algorithm for protein structure prediction (1) which we used to generate the structure models for our splice variants. It first threads the target sequence through a representative PDB structure library by a meta-server threading program LOMETS (2), which includes eight state-of-the-art threading alignment programs from FUGUE (3), HHsearch (4), MUSTER (5), PPA (6), PROSPECT2 (7), SAM-T02 (8), SPARKS (9), and SP3 (9). The continuously aligned structural fragments excised from the top threading templates are then used to assemble the full-length models by replica-exchange Monte Carlo simulations (10), under the guidance of consensus spatial restraints from the LOMETS templates, where the structure of threading unaligned regions is built by *ab initio* modeling. The lowest free-energy models are identified by SPICKER (11), which clusters all structural decoys generated in the low-temperature replicas. In the second round of I-TASSER simulation (12), fragments are excised from the PDB structures that are structurally closest to the cluster centroids, as identified by TM-align (13). Finally, atomic models are constructed from the lowest energy decoys in the second round simulation by REMO through the optimization of the hydrogen-bonding networks (14). A scoring function (C-score) based on the relative clustering structural density and the consensus significance score of multiple threading templates is introduced to estimate the accuracy of the I-TASSER predictions. C-score is strongly correlated with the similarity of the final model to the experimentally determined structures and is typically in the range from −5 to 2, wherein a more positive score reflects a model of better quality. Both false positive and false negative rates are below 0.1 when a C-score cutoff $>$ -1.5 is used for the models of correct topology (15).

For each structure prediction, we used the confidence score, C-score, to assess the quality of the structural models. Based on the benchmark data (15), C-score can be used to reliably predict the TM-score and RMSD of the I-TASSER models to the experimentally determined structures because of the strong correlation of C-score with the actual accuracy of the predictions. Meanwhile, RMSD and TM-score from TM-align were used to characterize and compare the 3-D structural models of the variants (1, 15). In addition, when we observed a structural variation between two models, we manually checked the templates that aligned to this region during I-TASSER modeling. If the structural variation was from random folding and is due to the lack of adequate templates aligning to the sequence, we ignored this difference. Especially in the N- and C-terminal regions, the folding will be flexible if there are not enough templates aligned to the sequence.

Protein Structure Comparison

We use TM-align to compare the structural models of the splice isoforms. TM-align is an algorithm for sequence-independent, automated structure comparison of different proteins (http://zhanglab.ccmb.med.umich.edu/TM-align) (13). For two given structures, it first identifies the optimal alignments based on structural similarity through an iterative Needleman-Wunsch dynamic programming algorithm (16) and structural similarity is quantified using TM-score (17). The value of TM-score lies in the range [0, 1]. Based on the statistical analysis (18), a TM-score >0.5 means the structures share the same fold with a p-value $\leq 5.5\times10^{-7}$, or we need to consider at least 1.8 million random protein pairs to acquire a TM-score of no less than 0.5. A TM-score below 0.3 by TM-align corresponds to the similarity of random structure pairs (19).

Benchmark Analysis for I-TASSER Modeling

To identify splice isoforms of experimentally solved structures for our benchmark testing, alternatively spliced protein sequences from the ASTD database were threaded through the PDB database using the NW-align program (http://zhanglab.ccmb.med.umich.edu/NW-align), an implementation of the standard Needleman-Wunsch dynamic programming algorithm (20). The resulting alignments were filtered using the following criteria to select the variant pairs:

1. $N_{id}/N_{ali} = 100\%$, where N_{id} is the number of the identically aligned residues and N_{ali} is the total number of aligned residues.
2. Alignments with a gap in ASTD sequence were discarded, as the gaps reflected gaps or insertions in the solved PDB structure.
3. Since the same sequence can have multiple hits, as the PDB contains multiple entries for the same protein, the hit with maximum alignment coverage was selected.
4. Finally, the remaining sequences were manually analyzed to check if the proteins were from the same organism and coded by the same gene.

Her2/neu Breast Cancer Dataset

In the mass spectrometric files from LC-MS/MS analyses of normal mouse mammary tissue or mammary tumors derived from doxycycline-inducible, MMTV-rtTA/TetO-NeuNT-mediated Her2/neu transgenic mice, we identified a total of 608 alternative splice variants, of which peptides from 216 proteins were found only in the tumor sample (21). Because the Ensembl database has been updated many times since the previous study, the peptides that were identified from the above 608 proteins were integrated to the latest Ensembl protein IDs (Ensembl version 62) using our custom Michigan Peptide to Protein Integration algorithm (MPPI) (21).

Differential Expression Analysis of Known Alternative Splice Variants

In our paper on the Her2/neu dataset (21), we used only the alternative splice variants that were identified by a unique peptide for differential expression analysis. Hence we had only 53 known splice variants that were differentially expressed. For this study, in order to have a larger sample size in selecting variants for structural comparisons, we analyzed for differential expression of all alternative splice variants identified from both normal and tumor integrated protein lists. A total of 165 distinct known splice variants were differentially expressed with p value <0.001 (Table S4).

Table S1
The Ensembl sequences of the Alternative Splice Variants selected: The alternatively spliced region is in bold

(1) Annexin 6
ENSMUSP00000104511_ Anxa6_001
MAKIAQGAMYRGSVHDFPEFDANQDAEALYTAMKGFGSDKESILELITSRSNKQRQEICQNYKSLYGKDLIEDLKYELTGKFERLIVNLMRPLAYCDAKEIKDAISG
VGTDEKCLIEILASRTNEQMHQLVAAYKDAYERDLESDIIGDTSGHFQKMLVVLLQGTRENDDVVSEDLVQQDVQDLYEAGELKWGTDEAQFIYILGNRSKQHLRLV
FDEYLKTTGKPIEASIRGELSGDFEKLMLAVVKCIRSTPEYFAERLFKAMKGLGTRDNTLIRIMVSRSELDMLDIREIFRTKYEKSLYSMIKNDTSGEYKKALLKLC
GGDDDAAGQFFPEAAQVAYQMWELSAVSRVELKGTVCAANDFNPDADAKALRKAMKGIGTDEATIIDIVTHRSNAQRQQIRQTFKSHFGRDLMADLKSEISGDLARL
ILGLMMPPAHYDAKQLKKAMEGAGTDEKTLIEILATRTNAEIRAINEAYKEDYHKSLEDALSSDTSGHFRRILISLATGNREEGGENRDQAQEDAQ**VAAEIL**EIADT
PSGDKTSLETRFMTVLCTRSYPHLRRVFQEFIKKTNYDIEHVIKKEMSGDVKDAFVAIVQSVKNKPLFFADKLYKSMKGAGTDEKTLTRVMVSRSEIDLLNIRREFI
EKYDKSLHQAIEGDTSGDFMKALLALCGGED

ENSMUSP00000099788_Anxa6_002
MAKIAQGAMYRGSVHDFPEFDANQDAEALYTAMKGFGSDKESILELITSRSNKQRQEICQNYKSLYGKDLIEDLKYELTGKFERLIVNLMRPLAYCDAKEIKDAISG
VGTDEKCLIEILASRTNEQMHQLVAAYKDAYERDLESDIIGDTSGHFQKMLVVLLQGTRENDDVVSEDLVQQDVQDLYEAGELKWGTDEAQFIYILGNRSKQHLRLV
FDEYLKTTGKPIEASIRGELSGDFEKLMLAVVKCIRSTPEYFAERLFKAMKGLGTRDNTLIRIMVSRSELDMLDIREIFRTKYEKSLYSMIKNDTSGEYKKALLKLC
GGDDDAAGQFFPEAAQVAYQMWELSAVSRVELKGTVCAANDFNPDADAKALRKAMKGIGTDEATIIDIVTHRSNAQRQQIRQTFKSHFGRDLMADLKSEISGDLARL
ILGLMMPPAHYDAKQLKKAMEGAGTDEKTLIEILATRTNAEIRAINEAYKEDYHKSLEDALSSDTSGHFRRILISLATGNREEGGENRDQAQEDAQEIADTPSGDKT
SLETRFMTVLCTRSYPHLRRVFQEFIKKTNYDIEHVIKKEMSGDVKDAFVAIVQSVKNKPLFFADKLYKSMKGAGTDEKTLTRVMVSRSEIDLLNIRREFIEKYDKS
LHQAIEGDTSGDFMKALLALCGGED

(2) Calumenin
ENSMUSP00000031779_Calu_001
MDLRQFLMCLSLCTAFALSKPTEKKDRVHHEPQLSDKVHNDAQNFDYDHDAFLGAEEAKSFDQLTPEESKERLG**KIVSKIDDDKDGFVTVDELKGWIKFAQKRWIHE**
DVERQWKGHDLNEDGLVSWEEYKNATYGYVLDDPDPDDGFNYKQMMVRDERRFKMADKDGDLIATKEEFTAFLHPEEYDYMKDIVVQETMEDIDKNADGFIDLEEYI
GDMYSHDGNADEPEWVKTEREQFVEFRDKNRDGKMDKEETKDWILPSDYDHAEAEARHLVYESDQNKDGKLTKEEIVDKYDLFVGSQATDFGEALVRHDEF
ENSMUSP00000087967_Calu_002
MDLRQFLMCLSLCTAFALSKPTEKKDRVHHEPQLSDKVHNDAQNFDYDHDAFLGAEEAKSFDQLTPEESKERLG**MIVDKIDADKDGFVTEGELKSWIKHAQKKYIYD**
NVENQWQEFDMNQDGLISWDEYRNVTYGTYLDDPDPDDGFNYKQMMVRDERRFKMADKDGDLIATKEEFTAFLHPEEYDYMKDIVVQETMEDIDKNADGFIDLEEYI
GDMYSHDGNADEPEWVKTEREQFVEFRDKNRDGKMDKEETKDWILPSDYDHAEAEARHLVYESDQNKDGKLTKEEIVDKYDLFVGSQATDFGEALVRHDEF

(3) cell division cycle 42 homolog
ENSMUSP00000054634_cdc42_001
MQTIKCVVVGDGAVGKTCLLISYTTNKFPSEYVPTVFDNYAVTVMIGGEPYTLGLFDTAGQEDYDRLRPLSYPQTDVFLVCFSVVSPSSFENVKEKWVPEITHHCPK
TPFLLVGTQIDLRDDPSTIEKLAKNKQKPITPETAEKLARDLKAVKYVECSALTQ**KGLKNVFDEAILAALEPPEPKKSRRCVLL**
ENSMUSP00000030417_cdc42_002
MQTIKCVVVGDGAVGKTCLLISYTTNKFPSEYVPTVFDNYAVTVMIGGEPYTLGLFDTAGQEDYDRLRPLSYPQTDVFLVCFSVVSPSSFENVKEKWVPEITHHCPK
TPFLLVGTQIDLRDDPSTIEKLAKNKQKPITPETAEKLARDLKAVKYVECSALTQRGL**KNVFDEAILAALEPPETQPKRKCCIF**

(4) Polypyrimidine tract binding protein 1

ENSMUSP00000126192_ptbp1_001

MDGIVPDIAVGTKRGSDELFSTCVSNGPFIMSSSASAANGNDSKKFKGDNRSAGVPSRVIHVRKLPSDVTEGEVISLGLPFGKVTNLLMLKGKNQAFIEMNTEEAAN
TMVNYYTSVAPVLRGQPIYIQFSNHKELKTDSSPNQARAQAALQAVNSVQSGNLALAASAAAVDAGMAMAGQSPVLRIIVENLFYPVTLDVLHQIFSKFGTVLKIIT
FTKNNQFQALLQYADPVSAQHAKLSLDGQNIYNACCTLRIDFSKLTSLNVKYNNDKSRDYTRPDLPSGDSQPSLDQTMAAAF**GAPGIMSASPYAGAGFPPTFAIPQA**
AGLSVPNVHGALAPLAIPSAAAAAAASRIAIPGLAGAGNSVLLVSNLNPERVTPQSLFILFGVYGDVQRVKILFNKKENALVQMADGSQAQLAMSHLNGHKLHGKSV
RITLSKHQSVQLPREGQEDQGLTKDYGSSPLHRFKKPGSKNFQNIFPPSATLHLSNIPPSVSEDDLKSLFSSNGGVVKGFKFFQKDRKMALIQMGSVEEAVQALIEL
HNHDLGENHHLRVSFSKSTI

ENSMUSP00000127783_ptbp1_002

MDGIVPDIAVGTKRGSDELFSTCVSNGPFIMSSSASAANGNDSKKFKGDNRSAGVPSRVIHVRKLPSDVTEGEVISLGLPFGKVTNLLMLKGKNQAFIEMNTEEAAN
TMVNYYTSVAPVLRGQPIYIQFSNHKELKTDSSPNQARAQAALQAVNSVQSGNLALAASAAAVDAGMAMAGQSPVLRIIVENLFYPVTLDVLHQIFSKFGTVLKIIT
FTKNNQFQALLQYADPVSAQHAKLSLDGQNIYNACCTLRIDFSKLTSLNVKYNNDKSRDYTRPDLPSGDSQPSLDQTMAAAFGLSVPNVHGALAPLAIPSAAAAAAA
SRIAIPGLAGAGNSVLLVSNLNPERVTPQSLFILFGVYGDVQRVKILFNKKENALVQMADGSQAQLAMSHLNGHKLHGKSVRITLSKHQSVQLPREGQEDQGLTKDY
GSSPLHRFKKPGSKNFQNIFPPSATLHLSNIPPSVSEDDLKSLFSSNGGVVKGFKFFQKDRKMALIQMGSVEEAVQALIELHNHDLGENHHLRVSFSKSTI

(5) Tax1 (human T-cell leukemia virus type I) binding protein 3

ENSMUSP00000047410_tax1bp3_001

MSYTPGQPVTAVV**QRVEIHKLRQGENLILGFSIGGGIDQDPSQNPFSEDKTDK**GIYVTRVSEGGPAEIAGLQIGDKIMQVNGWDMTMVTHDQARKRLTKRSEEVVRL
LVTRQSLQKAVQQSMLS

ENSMUSP00000104117_tax1bp3_002

MSYTPGQPVTAVVQRVEIHKLRQGENLILGFSIGGGIDQDPSQNPFSEDKTDKVNGWDMTMVTHDQARKRLTKRSEEVVRLLVTRQSLQKAVQQSMLS

Table S2: Benchmark analysis to compare the software (I-TASSER, MODELLER, and ROSETTA) predictions of alternative splice variant structures to experimentally determined structures (Exp) in PDB. RMSD_{ali} is RMSD between aligned residues by TM-align. P1 & P2 are alternative splice protein pairs. RMSD between Exp and Mod (1st model) is calculated using TM-score program. The average RMSD between the experimentally determined and I-TASSER predicted structure was 1.72 Å. These seven pairs are all of the experimental full-length alternatively-spliced pair structures in the Protein Data Bank; all arise from exon swaps. The average RMSD values between the experimentally determined structure and I-TASSER predicted structure was lower when compared to that between predicted structures from MODELLER (2.27 Å) or ROSETTA ( ) to experimentally determined structures.

Gene Name (symbol)	**Variation due to alternative splicing; # of aa in the exon swapped ; percentage sequence identity in spliced region**	**PDB ID (P1)**	**PDB ID (P2)**	**RMSDali Exp-Exp (P1-P2) Å**	**I-TASSER**			**MODELLER**			**ROSETTA**		
					RMSD Mod - Exp(P1) Å	**RMSD Mod - Exp(P2) Å**	**RMSD Mod - Mod (P1-P2) Å**	**RMSD Mod - Exp(P1) Å**	**RMSD Mod - Exp(P2) Å**	**RMSD Mod - Mod (P1-P2) Å**	**RMSD Mod - Exp(P1) Å**	**RMSD Mod - Exp(P2) Å**	**RMSD Mod - Mod (P1-P2) Å**
Acp1 ENSG00000143727	Second exon 19 aa; 47%	5pntA	1xwwA	0.73	0.79	0.75	0.52	2.02	3.49	0.71			
Pfn2 ENSG00000070087	C-terminal exon 32 aa; 79%	1d1jA	2v8fB	0.87	0.94	0.83	0.31	1.17	1.09	0.14			
Nme2 ENSG00000011052	Second exon 34 aa; 82 %	1nskR	3l7uA	0.78	0.98	0.53	0.39	1.19	0.94	0.27			
Gck ENSG00000106633	N-terminal exon 15 aa in 3idhA and 16 aa in 3goiA	3goiA	3idhA	0.7	1.53	1.59	0.42	1.97	1.90	1.02			
Khk ENSG00000138030	Third exon 44 aa; 39 %	2hqqA	3b3lA	2.11	2.61	2.93	2.20	2.31	3.06	0.97			
Mapk8 ENSG00000107643	Sixth exon 24 aa; 69 %	1ukiA	3eljA	2.00	2.15	1.86	1.68	2.36	3.13	1.23			
Pkm2 ENSG00000067225	Ninth exon 55 aa; 60 %	1a49	1t5a	1.96	2.76	3.82	1.97	3.55	3.66	0.86			

Table S3

Mouse and human homologous alternative splice variants

Gene symbol	Selected Splice Variant From Her2/Neu dataset analysis (3)	Corresponding known mouse splice variant from the same gene	Homologous known human splice variant for the selected mouse variant	Homologous known human splice variant for the corresponding mouse variant	Similarity between the variant protein sequences	Similarity between the homologous mouse-human variant protein sequences
calu	ENSMUSP00000087967 (315)	ENSMUSP00000031779 (315)	ENSP00000408838 (315)	ENSP00000249364 (315)	ENSMUSP00000087967 vs ENSMUSP00000031779 93%	ENSMUSP00000087967 vs ENSP00000408838 99%
					ENSP00000408838 vs ENSP00000249364 93%	ENSMUSP00000031779 vs ENSP00000249364 99%
cdc42	ENSMUSP00000054634 (191)	ENSMUSP00000030417 (191)	ENSP00000341072 (191)	ENSP00000314458 (191)	ENSMUSP00000054634 vs ENSMUSP00000030417 96%	ENSMUSP00000054634 vs ENSP00000341072 100%
					ENSP00000341072 vs ENSP00000314458 96%	ENSMUSP00000030417 vs ENSP00000314458 100%
ptbp1	ENSMUSP00000093109 (555)	ENSMUSP00000089978 (529)	ENSP00000349428 (557)	ENSP00000014112 (531)	ENSMUSP00000093109 vs ENSMUSP00000089978 96%	ENSMUSP00000093109 vs ENSP00000349428 97%
					ENSP00000349428 vs ENSP00000014112 96%	ENSMUSP00000089978 vs ENSP00000014112 97%

The number in parentheses is the protein length of the variant

Functional Motifs found in the regions of structural differences between the human calu, cdc42, and ptbp1 variant pairs

Gene Name	Region where Structural difference is observed	RMSD Å	Functional motif or residue
Calu	35-37 44-50	5.3 (in the region of structural difference)	Ser-35, Ser-44, and, Tyr-47 can be potentially phosphorylated
Cdc42	No change	1.34	-
Ptbp1	Multiple locations	4.09	The RNA Recognition Motifs

Table S4: Alternative Splice Variants that are differentially expressed between tumor and normal samples by spectral counting method with p value < 0.001

Protein	symbol	change	Protein	symbol	change
ENSMUSP00000038755	Abhd14b	up	ENSMUSP00000099634	Acadvl	down
ENSMUSP00000100038	Aco1	up	ENSMUSP00000007131	Acly	down
ENSMUSP00000130611	Actb	up	ENSMUSP00000087736	Actc1	down
ENSMUSP00000066068	Actn4	up	ENSMUSP00000087043	Actg1	down
ENSMUSP00000035829	Akap12	up	ENSMUSP00000016105	Adss	down
ENSMUSP00000017534	Aldoc	up	ENSMUSP00000068479	Ak1	down
ENSMUSP00000025561	Anxa1	up	ENSMUSP00000030583	Ak2	down
ENSMUSP00000109305	Anxa4	up	ENSMUSP00000100045	Akr1b3	down
ENSMUSP00000104511	Anxa6	up	ENSMUSP00000030090	Alad	down
ENSMUSP00000098405	Anxa7	up	ENSMUSP00000032934	Aldoa	down
ENSMUSP00000101803	Arhgdia	up	ENSMUSP00000118417	Aldoa	down
ENSMUSP00000111942	Arl1	up	ENSMUSP00000099394	Aoc3	down
ENSMUSP00000104485	Atox1	up	ENSMUSP00000032974	Atp2a1	down
ENSMUSP00000099803	Btf3l4	up	ENSMUSP00000104124	Atp2a3	down
ENSMUSP00000110673	Cald1	up	ENSMUSP00000026495	Atp5a1	down
ENSMUSP00000031779	Calu	up	ENSMUSP00000101006	Atp5d	down
ENSMUSP00000101862	Cap1	up	ENSMUSP00000028610	Cat	down
ENSMUSP00000063389	Capg	up	ENSMUSP00000026148	Cbr2	down
ENSMUSP00000078640	Cbx1	up	ENSMUSP00000030345	Cpt2	down
ENSMUSP00000054634	Cdc42	up	ENSMUSP00000034562	Cryab	down
ENSMUSP00000077349	Ckmt1	up	ENSMUSP00000106481	Dld	down
ENSMUSP00000103477	Clta	up	ENSMUSP00000072620	Eno3	down
ENSMUSP00000109249	Cnbp	up	ENSMUSP00000069209	Ephx2	down
ENSMUSP00000095169	Csrp1	up	ENSMUSP00000075945	Fcgbp	down
ENSMUSP00000121203	Ctsd	up	ENSMUSP00000023854	Fhl1	down
ENSMUSP00000034539	Dcps	up	ENSMUSP00000116725	Fhl1	down
ENSMUSP00000032992	Eif3c	up	ENSMUSP00000040150	Fhl3	down
ENSMUSP00000099649	Eif4a1	up	ENSMUSP00000114019	Gyg	down
ENSMUSP00000090876	Eif4a2	up	ENSMUSP00000015800	Hspa8	down
ENSMUSP00000104250	Eif5a	up	ENSMUSP00000039172	Hspb6	down
ENSMUSP00000079045	Eno1	up	ENSMUSP00000095316	Idh1	down
ENSMUSP00000063734	Ezr	up	ENSMUSP00000103007	Idh2	down
ENSMUSP00000130145	Fubp1	up	ENSMUSP00000087494	Ldb3	down
ENSMUSP00000018727	G3bp1	up	ENSMUSP00000103267	Ldha	down

ENSMUSP00000062996	Gdi2	up	ENSMUSP00000003207	Lipe	down
ENSMUSP00000107593	Hnrnpa3	up	ENSMUSP00000022148	Mccc2	down
ENSMUSP00000072533	Hnrnpd	up	ENSMUSP00000030742	Mecr	down
ENSMUSP00000126817	Hnrnpf	up	ENSMUSP00000018632	Myh4	down
ENSMUSP00000074483	Hnrnph2	up	ENSMUSP00000027151	Myl1	down
ENSMUSP00000039269	Hnrnpk	up	ENSMUSP00000112861	Myl1	down
ENSMUSP00000049407	Hnrnpl	up	ENSMUSP00000004673	Ndrg2	down
ENSMUSP00000037268	Hnrnpul1	up	ENSMUSP00000030805	Park7	down
ENSMUSP00000091921	Hsp90aa1	up	ENSMUSP00000063825	Pcx	down
ENSMUSP00000118189	Hsp90aa1	up	ENSMUSP00000061227	Pgm2	down
ENSMUSP00000024739	Hsp90ab1	up	ENSMUSP00000128770	Pkm2	down
ENSMUSP00000028222	Hspa5	up	ENSMUSP00000035220	Prkar2a	down
ENSMUSP00000113722	Hspa8	up	ENSMUSP00000039797	Prkar2b	down
ENSMUSP00000034426	Kars	up	ENSMUSP00000058321	Ptrf	down
ENSMUSP00000007814	Khsrp	up	ENSMUSP00000005860	Pvalb	down
ENSMUSP00000079053	Krt5	up	ENSMUSP00000071231	Pygl	down
ENSMUSP00000053962	Lcn2	up	ENSMUSP00000047564	Pygm	down
ENSMUSP00000121201	Lcp1	up	ENSMUSP00000072652	Serpina1a	down
ENSMUSP00000051619	Mapk3	up	ENSMUSP00000044033	Serpina6	down
ENSMUSP00000113071	Msn	up	ENSMUSP00000023161	Srl	down
ENSMUSP00000044827	Mybbp1a	up	ENSMUSP00000101563	Tnnt3	down
ENSMUSP00000016771	Myh9	up	ENSMUSP00000103546	Tpm2	down
ENSMUSP00000089680	Naca	up	ENSMUSP00000101855	Trim72	down
ENSMUSP00000075067	Npm1	up	ENSMUSP00000095656	Tufm	down
ENSMUSP00000086542	Nsfl1c	up			
ENSMUSP00000021082	Nt5c	up			
ENSMUSP00000077794	Pabpc4	up			
ENSMUSP00000021646	Papln	up			
ENSMUSP00000029941	Pdlim5	up			
ENSMUSP00000021282	Pfas	up			
ENSMUSP00000072773	Postn	up			
ENSMUSP00000039109	Ppp1ca	up			
ENSMUSP00000114159	Prdx1	up			
ENSMUSP00000105356	Prdx2	up			
ENSMUSP00000071636	Prdx6	up			
ENSMUSP00000004316	Psap	up			
ENSMUSP00000030769	Psmc2	up			
ENSMUSP00000126192	Ptbp1	up			

ENSMUSP00000078745	Rab1	up			
ENSMUSP00000111309	Ranbp1	up			
ENSMUSP00000038964	Rbm3	up			
ENSMUSP00000017548	Rpl19	up			
ENSMUSP00000110094	Rpl24	up			
ENSMUSP00000081474	Rplp2	up			
ENSMUSP00000069004	Rps15	up			
ENSMUSP00000103940	Rps16	up			
ENSMUSP00000032998	Rps3	up			
ENSMUSP00000004554	Rps5	up			
ENSMUSP00000016072	Rrbp1	up			
ENSMUSP00000099907	Rtn4	up			
ENSMUSP00000058237	S100a1	up			
ENSMUSP00000092697	Spna2	up			
ENSMUSP00000090059	Srsf2	up			
ENSMUSP00000020501	Sumo3	up			
ENSMUSP00000106861	Tagln2	up			
ENSMUSP00000060538	Tardbp	up			
ENSMUSP00000081142	Tardbp	up			
ENSMUSP00000047410	Tax1bp3	up			
ENSMUSP00000030187	Tln1	up			
ENSMUSP00000030056	Tnc	up			
ENSMUSP00000113219	Tpm3	up			
ENSMUSP00000106519	Tpt1	up			
ENSMUSP00000086626	Uba1	up			
ENSMUSP00000099807	Uba2	up			
ENSMUSP00000111363	Ube2l3	up			
ENSMUSP00000075782	Ubqln1	up			
ENSMUSP00000041299	Usp5	up			
ENSMUSP00000022369	Vcl	up			
ENSMUSP00000024866	Xdh	up			
ENSMUSP00000070993	Ywhae	up			
ENSMUSP00000100067	Ywhaq	up			
ENSMUSP00000022894	Ywhaz	up			

Figure S1: The experimentally determined and predicted structures of the variants of Ketohexokinase (khk) are shown here. (a & b) show superimposed PDB and predicted khk variant structures. The alternatively spliced region (shown in blue and cyan) shows a spatial shift in the beta chain in both PDB and predicted structures (the arrow points to this shifted region).

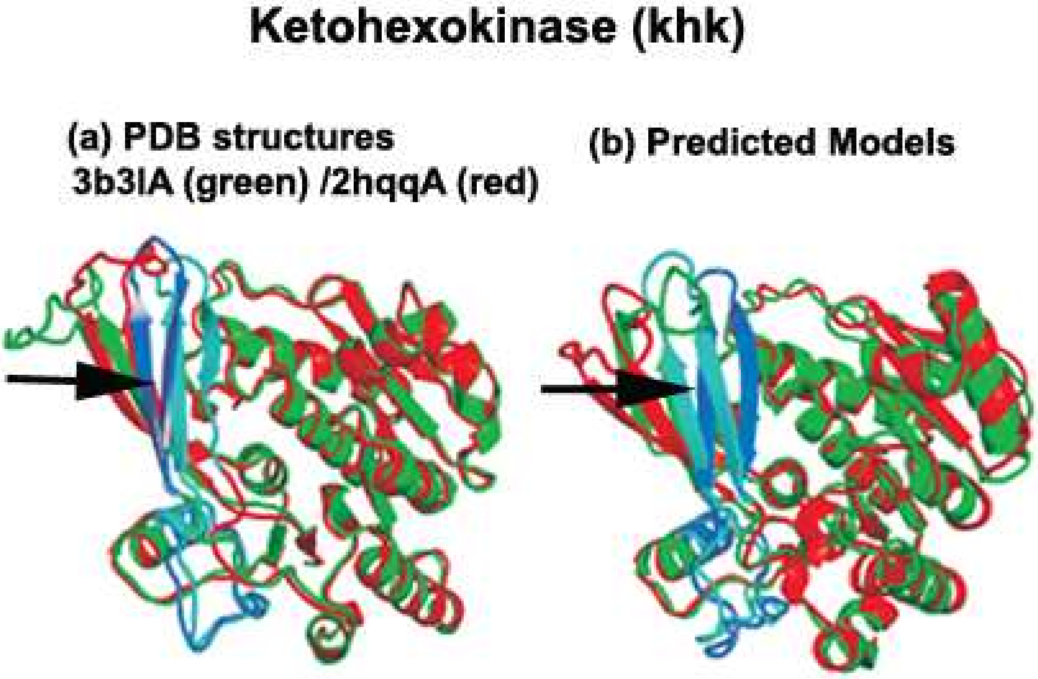


Figure S2: The experimentally determined and predicted structures of the variants of Acid phosphatase 1(acp1) (a and b) and Mitogen-activated protein kinase 8 (mapk8) (c and d) are shown here. (a) Superimposed PDB structures of acp1 variants; PDB IDs 5pntA (green), 1xwwA (red). (b) Superimposed 3-D models predicted by I-TASSER for the acp1 variants. (c) Superimposed PDB structures of mapk8 variants; PDB IDs 1ukiA (green), 3eljA (red). (d) Superimposed 3-D models predicted by I-TASSER for the mapk8 variants. The blue and the purple colors show the loop region where the splicing occurs. Alternative splicing does not seem to change the structure of the variants. The predicted models are very similar to the experimentally determined structure with an average RMSD between the PDB and the predicted structures of acp1 and mapk8 variants being 0.75 Å and 2 Å.

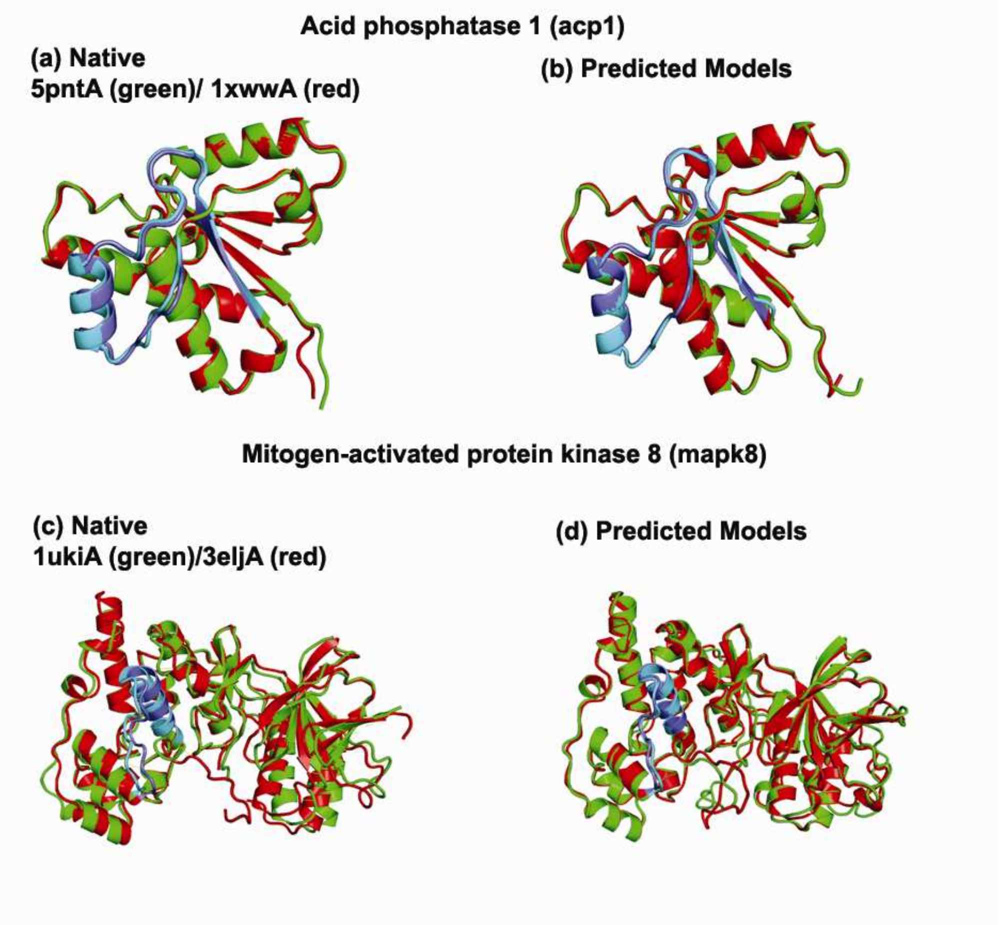

Acid phosphatase 1 (acp1)
(a) Native
5pntA (green)/ 1xwwA (red)
(b) Predicted Models
Mitogen-activated protein kinase 8 (mapk8)
(c) Native
1ukiA (green)/3eljA (red)
(d) Predicted Models

Figure S3: The figure below shows the backbone structure and side chain positions of the residues in the alternatively spliced regions of the predicted 3D models of acp1 variants; I-TASSER modeling shows that the back bone structures are aligned but the side chain rotomers are different and do not superimpose on one another.

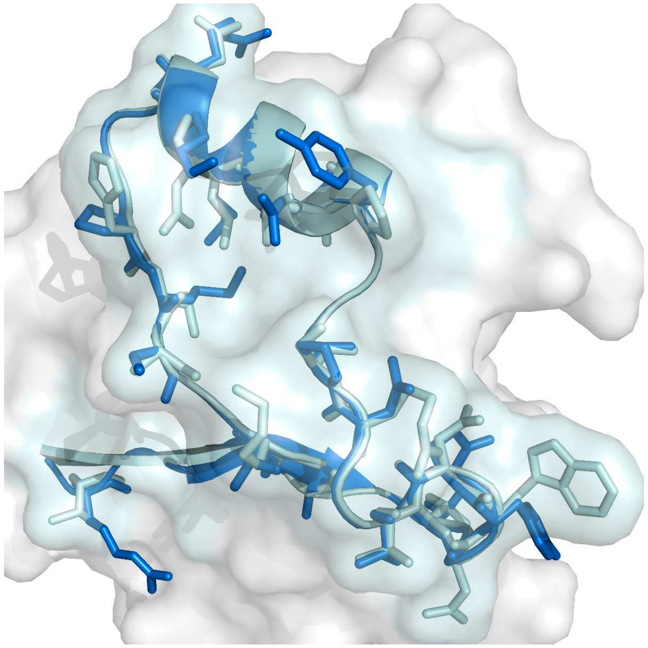


Figure S4: The figure below shows the backbone structure and side chain positions of the residues in the alternatively spliced regions of the predicted 3D models of mapk8 variants; I-TASSER modeling shows that the back bone structure are aligned but the side chains rotomers are different and do not superimpose on one another.

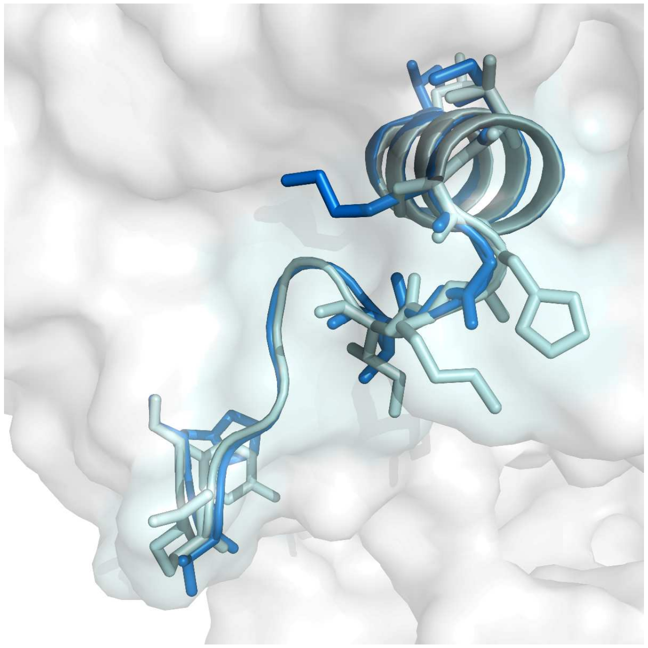


Figure S5: The experimentally solved and I-TASSER predicted 3D structures of the first domains of the mouse ryanodine receptor 2 (ryr2) splice variants. The domains differ by 35 amino acids; the shorter variant (RYR2_2) does not contain these amino acids (exon 3 is missing) compared to the longer variant (RYR2_1). Since the solved structure of the longer domain contained most of the alternatively spliced region resolved (residues VPPDLSICTFVLEQSLSVRALQEMLANTV), we used this pair to illustrate structural difference due to deletion. It is important to note here that part of the alternatively spliced region (residues KSEG) and part of the sequence found in both domains (residues QVDVEKWKFMMKTAQGGG) are missing in the resolved structure of the first domain of the longer variant (3IM5). However, distinct structural difference is observed between the experimentally solved structures (3IM5 and 3QR5) due to alternative splicing.

(a) Alignment of the first domain sequences; the residues in blue are in the alternatively spliced region

```
RYR2_1 1    MADAGEGEDEIQFLRTDDEVVLQCTATIHKEQQKLCLAAEGFGNRLCFLESTSNSKNVPP
            ::::::::::::::::::::::::::::::::::::::::::::::::::::::::
RYR2_2 4    MADAGEGEDEIQFLRTDDEVVLQCTATIHKEQQKLCLAAEGFGNRLCFLESTSNSK----

RYR2_1 61   DLSICTFVLEQSLSVRALQEMLANTVEKSEGQVDVEKWKFMMKTAQGGGHRTLLYGHAIL
                                           :::::::::::::::::::::::::::::
RYR2_2 60   -------------------------------QVDVEKWKFMMKTAQGGGHRTLLYGHAIL

RYR2_1 121  LRHSYSGMYLCCLSTSRSSTDKLAFDVGLQEDTTGEACWWTIHPASKQRSEGEKVRVGDD
            ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
RYR2_2 89   LRHSYSGMYLCCLSTSRSSTDKLAFDVGLQEDTTGEACWWTIHPASKQRSEGEKVRVGDD

RYR2_1 181  LILVSVSSERYLHLSYGNSSWHVDAAFQQTLWSVAPI   217
            :::::::::::::::::::::::::::::::::::::
RYR2_2 149  LILVSVSSERYLHLSYGNSSWHVDAAFQQTLWSVAPI   185
```

(b) Superimposed 3D structures of the domain 1 of ryr2 variants; the region colored in blue is the alternatively spliced region. The experimentally solved structures and the predicted models of the domain looked very similar with RMSD between them being 1.05 Å (between 3IM5 and predicted model for the longer rry2 variant) and 1.12 Å (between 3QR5 and predicted model for shorter rry2 variant) Residues KSEGQVDVEKWKFMMKTAQGGG is missing in 3IM5(shown in magenta in the predicted model for longer rry2 variant). Similar structural differences were observed in both experimentally solved and predicted structures due to deletion. The RMSD between 3IM5 and 3QR5 is 1.09 Å and between the predicted models is 1.88 Å.

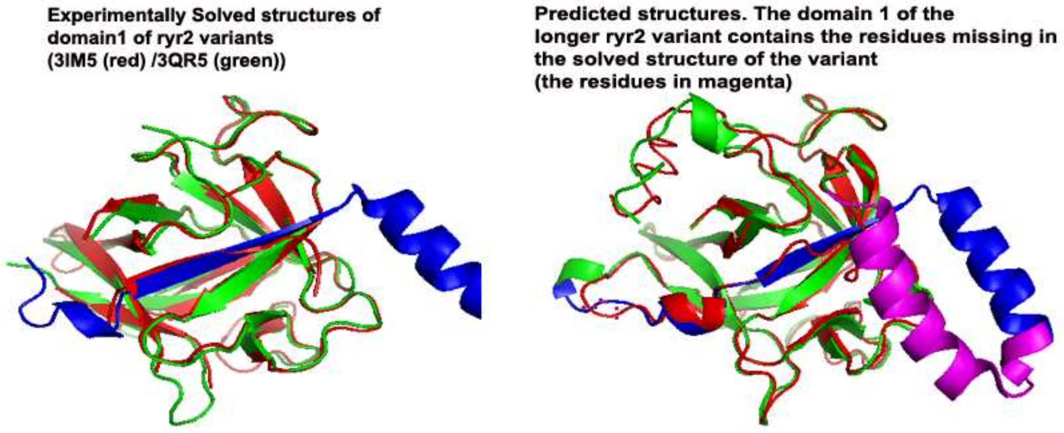
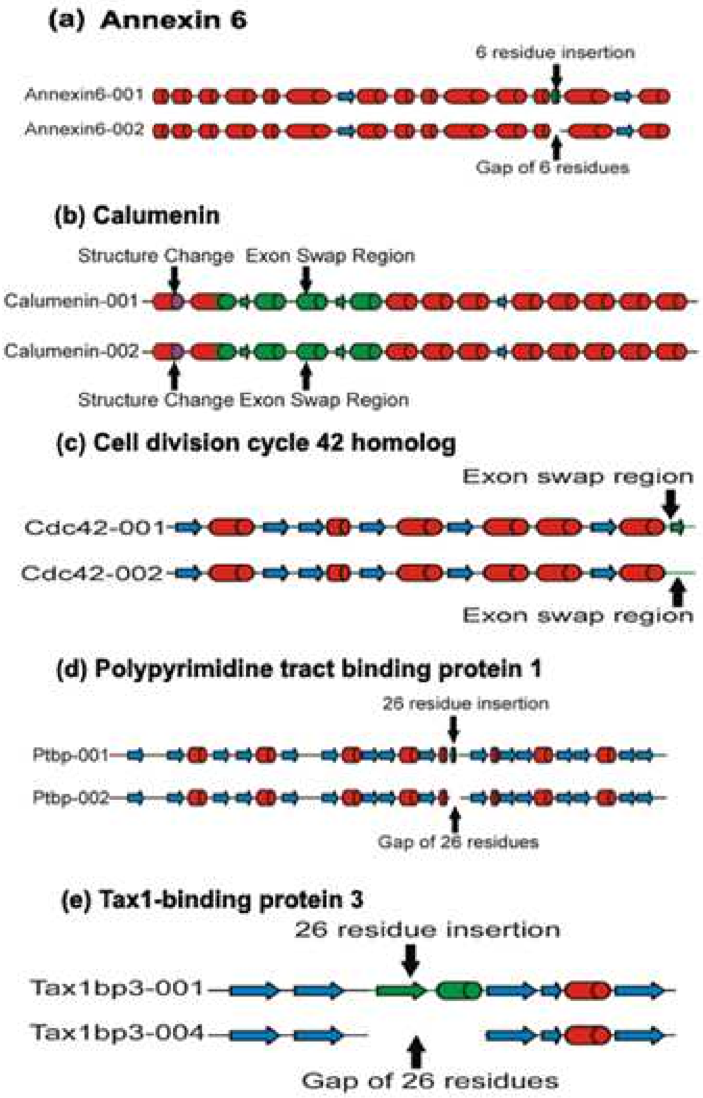

Experimentally Solved structures of domain1 of ryr2 variants (3IM5 (red) /3QR5 (green))
Predicted structures. The domain 1 of the longer ryr2 variant contains the residues missing in the solved structure of the variant (the residues in magenta)

Figure S6: Schematic representation of the five breast cancer-related splice variants. The helices are shown as red cylinders; the beta sheets as blue arrows; the alternatively spliced regions in green.



References

1. Roy, A., Kucukural, A., & Zhang, Y. (2010) I-TASSER: a unified platform for automated protein structure and function prediction *Nat. Protocols* 5: 725-738.

2. Wu, S. T. & Zhang, Y. (2007) LOMETS: A local meta-threading-server for protein structure prediction *Nucl. Acids. Res.* 35: 3375-3382.

3. Shi, J., Blundell, T. L., & Mizuguchi, K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties *J Mol Biol* 310: 243-257.

4. Soding, J. (2005) Protein homology detection by HMM-HMM comparison *Bioinformatics* 21: 951-960.

5. Wu, S. & Zhang, Y. (2008) MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information *Proteins* 72: 547-556.

6. Zhang, Y. (2007) Template-based modeling and free modeling by I-TASSER in CASP7 *Proteins* 69: 108-117.

7. Xu, Y., Xu, D., Crawford, O. H., Einstein, Larimer, F., Uberbacher, E., Unseren, M. A., & Zhang, G. (1999) Protein threading by PROSPECT: a prediction experiment in CASP3 *Protein Eng* 12: 899-907.

8. Karplus, K., Barrett, C., & Hughey, R. (1998) Hidden Markov models for detecting remote protein homologies *Bioinformatics* 14: 846-856.

9. Zhou, H. & Zhou, Y. (2005) Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments *Proteins* 58: 321-328.

10. Zhang, Y., Kihara, D., & Skolnick, J. (2002) Local energy landscape flattening: Parallel hyperbolic Monte Carlo sampling of protein folding *Proteins* 48: 192-201.

11. Zhang, Y. & Skolnick, J. (2004) SPICKER: a clustering approach to identify near-native protein folds *J Comput Chem* 25: 865-871.

12. Wu, S., Skolnick, J., & Zhang, Y. (2007) Ab initio modeling of small proteins by iterative TASSER simulations *BMC Biol* 5: 17.

13. Zhang, Y. & Skolnick, J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score *Nucleic. Acids Res.* 33: 2302-2309.

14. Li, Y. & Zhang, Y. (2009) REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks *Proteins* 76: 665-676.

15. Zhang, Y. (2008) I-TASSER server for protein 3D structure prediction *BMC Bioinformatics* 9: 40.

16. Needleman, S. B. & Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins *J Mol Biol* 48: 443-453.

17. Zhang, Y. & Skolnick, J. (2004) Scoring function for automated assessment of protein structure template quality *Proteins* 57: 702-710.

18. Xu, J. & Zhang, Y. (2010) How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* 26: 889-895.

19. Zhang, Y., Hubner, I., Arakaki, A., Shakhnovich, E., & Skolnick, J. (2006) On the origin and completeness of highly likely single domain protein structures *Proc. Natl. Acad. Sci. USA* 103: 2605-2610.

20. Needleman, S. B. & Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins *Journal of Molecular Biology* 48: 443 - 453.

21. Menon, R. & Omenn, G. S. (2010) Proteomic characterization of novel aternative splice variant proteins in Human epidermal growth factor receptor 2/neu induced breast cancers *Cancer Research* 70: 3440-3449.