Chapter 20

Annotation of Alternatively Spliced Proteins and Transcripts with Protein-Folding Algorithms and Isoform-Level Functional Networks

Hongdong Li, Yang Zhang, Yuanfang Guan, Rajasree Menon, and Gilbert S. Omenn

Abstract

Tens of thousands of splice isoforms of proteins have been catalogued as predicted sequences from transcripts in humans and other species. Relatively few have been characterized biochemically or structurally. With the extensive development of protein bioinformatics, the characterization and modeling of isoform features, isoform functions, and isoform-level networks have advanced notably. Here we present applications of the I-TASSER family of algorithms for folding and functional predictions and the IsoFunc, MIsoMine, and Hisonet data resources for isoform-level analyses of network and pathway-based functional predictions and protein-protein interactions. Hopefully, predictions and insights from protein bioinformatics will stimulate many experimental validation studies.

Key words Functional prediction, Isoform network, Protein folding, Splice isoforms

1 Introduction

One of the most remarkable developments in biological evolution is the emergence in multicellular organisms of gene structures with exons and introns. An elaborate splicing machinery in cells processes heterogeneous nuclear RNAs and generates several different mRNA transcripts from individual genes that can be translated into protein products. Just describing gene or protein expression as "upregulated" or "downregulated" ignores the fact that these transcripts and proteins are mixtures. These splice variants can and often do have dramatically different functions; when the proteins fold similarly and compete for target sites, they may, in fact, have opposing actions, such as proapoptotic and antiapoptotic activities [1].

Alternative splicing generates protein diversity without increasing genome size. This phenomenon seems to explain how humans can "get by" with only 20,000 protein-coding genes, whereas there

Cathy H. Wu et al. (eds.), *Protein Bioinformatics: From Protein Modifications and Networks to Proteomics*, Methods in Molecular Biology, vol. 1558, DOI 10.1007/978-1-4939-6783-4_20, © Springer Science+Business Media LLC 2017

were predictions of 50,000 to 100,000 or more protein-coding genes when the Human Genome Project was launched. The splice variants cannot be identified in genome sequences, but the splicing can be mapped to the gene exon/intron structures. There are multiple kinds of splicing events, including alternative 5 ′ or 3′ start sites, mutually exclusive exons (exon swaps), intron retention, alternative promoters, and alternative polyadenylation. There are examples of every kind of splicing in cancers, for example, and complex combinations of splice isoforms are well described in the nervous system. Ensembl, UniProt, neXtProt, RefSeq, and ECgene are databases with extensive information about protein splice variants.

2 Materials

Depending on the biological or clinical question being investigated, either primary experimental data or publicly available datasets for protein and transcript isoforms from appropriate specimens may be utilized for annotation and characterization of splice isoforms. For example, we have characterized splice isoforms of Her2/ERBB2+ breast cancers from humans and in mouse models [2, 3].

1. Use the current version of UniProt (e.g., release 2015/03) at http://www.uniprot.org to obtain a reliable, high-quality set of protein isoforms which are consistently annotated both in Ensembl (version 75) at http://wwwensemblorgHomo sapiens/Info/Index and in NCBI RefSeq (e.g., release 70) at http://www.ncbi.nlm.nih.gov/refseq/.

   Note that the annotation on splice isoforms varies from database to database and version to version. In Ensembl, information on protein-coding transcripts for a gene is updated or changed whenever the database version is changed. RefSeq is generally less inclusive than Ensembl.

   The identification of a "canonical isoform" often defaults to the longest product (protein sequence). UniProt curators choose a canonical variant from among several protein isoforms encoded by one gene using some mixture of the following criteria: highest expressed (varies by tissue and conditions); most conserved across species; the amino acid sequence that allows the clearest description of domains, isoforms, polymorphisms, and post-translational modifications; or, finally, the longest sequence. The other sequences are called "noncanonical" isoforms: see http://www.uniprot.org/help/canonical*and*isoforms. As described below, we propose a method to identify the "most highly connected isoform" and consider the highest connected isoform (HCI) the canonical form.

   Generally, the functional annotation of genes is based on their widely studied canonical protein or, more crudely, on the mixture of unrecognized isoforms. There is only very limited experimental evidence or computational annotation of functions of noncanonical protein isoforms. Gene ontology (GO) and Kyoto encyclopedia of genes and genomes (KEGG) are widely used; we recommend also using GeneCards for gene-level annotations (http://www.genecards.org/).

2. RNA-Seq provides far more precise measurement of levels of expression of transcripts and their isoforms than microarray analyses do. Available datasets can be downloaded from the NCBI Short Read Archive (http://www.ncbi.nomlnih.gov/sra). Often there are biological replicates available.

   Analyze the RNA-Seq data using Sailfish [4], an alignment-free algorithm (unlike Tophat/Cufflinks) for estimation of the isoform abundances. Sailfish builds a unique index of all k-mers (short and consecutive sequences containing k nucleic acids), counts the occurrences of the k-mers in the RNA-Seq fragments, and quantitates the transcripts by the number of occurrences of the k-mers through an EM algorithm. The index file for the Sailfish quantitation process can be generated from Ensembl cDNA file (GRCh38 version).

   Use the R package edgeR to quantitate differential expression of transcript isoforms between two tissues or tumor types being compared. This method uses an over-dispersed Poisson model to account for biological and technical variability. Bonferroni correction of p values for multiple hypothesis testing can be performed with the p.adjust function in stats, R package (http://stat.ethz.ch/R-manual/R-patched/library/stats/html/p.adjust.html). Use the Sailfish estimates of read counts for each transcript isoform in calculations of differential expression.

3 Methods

3.1 Inferring Structure and Function of Protein Isoforms Using Structural Bioinformatics Tools

We developed an integrated computational approach with tools for 3D protein structure prediction and structure-based protein function annotation, which can be used for inferring potential folding and functions of the noncanonical splice proteins compared with those of the canonical proteins. As examples, we first applied this approach to examining the function and folding stability of pyruvate kinase M2 isoforms, whose 3D structures were known, and three cancer-related isoform pairs, Bcl-x, caspase 3, and odd-skipped related 2, which were reported to have opposite functions, but lacked experimentally derived structures [5].

3.1.1 Protein 3D Structure Prediction

We developed and recommend two methods, I-TASSER [6–11] and QUARK [12–14] for template-based and ab initio protein structure predictions, respectively (Fig. 1).

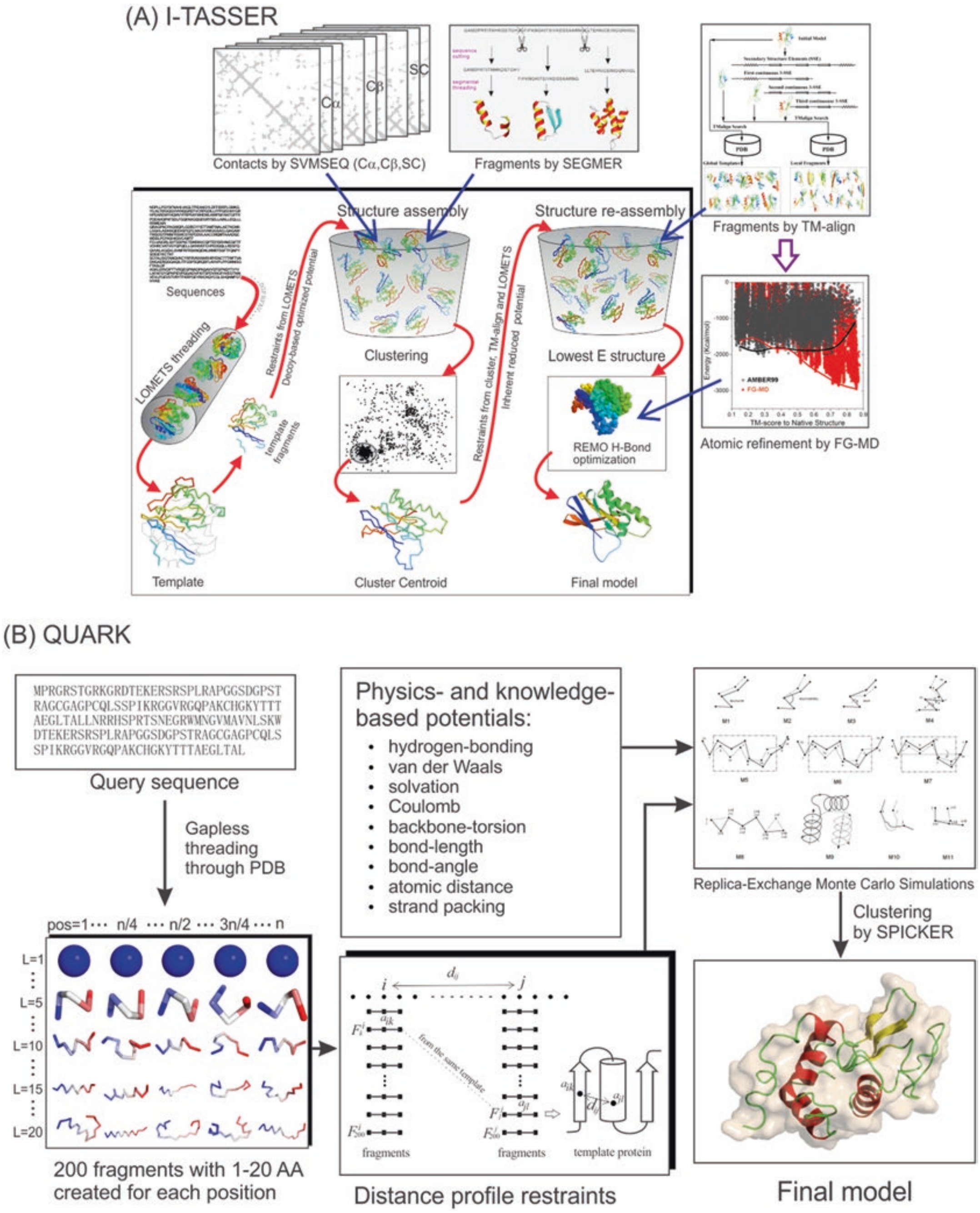


Fig. 1 Flowcharts of template-based structure prediction by I-TASSER (**a**) and ab initio structural folding by QUARK (**b**)

For additional comments about the I-TASSER family of tools and alternative methods, *see* Subheading 4.

I-TASSER pipeline for template-based structure prediction. The pipeline of I-TASSER is presented in Fig. 1a. The classic I-TASSER is depicted in the left-bottom panel of Fig. 1a, which consists of three steps:

1. Thread the query sequence through the PDB library by a locally installed meta-threading-server (LOMETS) method [15] to identify structure templates. Because LOMETS combines multiple threading programs based on different but complementary alignment algorithms, the templates often have improved structural coverage and alignment accuracy compared to individual threading algorithms.
2. Construct full-length model by reassembling the continuous segments excised from the templates in the threading-aligned regions using iterative replica-exchange Monte Carlo (REMC) simulations, where the structure in the threading-unaligned regions is built by an on-lattice ab initio modeling procedure. The REMC simulations are guided by a highly optimized knowledge-based force field that consists of generic backbone contact and correlation interactions, hydrogen bonding, and threading template-based restraints [11, 16].
3. Select the lowest free energy models by the clustering of all structure decoys as generated by the REMC simulations [17, 18]; the atomic models are then refined by REMO via the optimization of the hydrogen-bonding networks [19].

In recent developments, we proposed three methods to improve the accuracy of I-TASSER for distant homology modeling. These new developments are depicted in the upper-right panel of Fig. 1a, which was proved efficient in improving the modeling accuracy for proteins that lack close homologous templates [20]. These include:

1. Extended SVMSEQ [21] to generate multiple-scale residue contacts to guide the long-range I-TASSER structure assembly [22]
2. Developed SEGMER [23] to detect super-secondary structural motifs by segmental threading which improve spatial restraints of medium-range I-TASSER simulations
3. Applied FG-MD [24] to refine I-TASSER models at the atomic level based on fragment-guided molecular dynamic simulations, where the fragments from the PDB are shown to be able to improve the funnel of landscape of the physics-based force field.

QUARK pipeline for ab initio protein structure folding. Because the accuracy of I-TASSER predictions often relies on the existence of PDB templates and cannot be used to model proteins with new folds, we developed QUARK for ab initio protein structure prediction (Fig. 1b), [12, 25], which consists of three steps:

1. Select 200 short fragments at each position of the target sequence, each with 1–20 residues, from unrelated proteins by gapless matches.
2. Use the selected fragments to assemble full-length models by REMC simulations, under a composite physics- and knowledge-based potential that consists of H-bond, van der Waals, solvation, Coulomb, backbone torsion, bond length and angle, and statistical strand and helix packing interactions [12]. Meanwhile, non-covalent contact and distance profiles are derived from the short-range fragments by consistency analysis and used to accommodate long-range packing simulations [25].
3. Select final models by SPICKER clustering program, and then refine the atomic models by ModRefiner [26] through a two-step atomic-level minimization to improve H-bond networks and physical realism.

Test of structure prediction methods in blind CASP experiments. I-TASSER and QUARK have been tested in both benchmark [6, 7, 12, 13, 27] and blind experiments [8, 9, 10, 28–30]. For the blind test, community-wide CASP experiments have been organized every two years since 1994 to examine the state of the art of structure prediction methods (http://predictioncenter.org) [31, 32]. Structure predictions of a set of >100 protein targets are made *before* the experimental structures are released, and the modeling results by predictors are assessed by independent scientists. The CASP experiments have attracted hundreds of predictors from the community in the last two decades. I-TASSER was tested (as "Zhang-server") in the seventh–eleventh CASP competitions in 2006–2014, and QUARK participated in the ninth and tenth CASPs. I-TASSER and QUARK have consistently ranked in the top two positions in the automated server section for generating the most accurate protein structure predictions [29, 30, 33–35]. These results demonstrate the advantage of these pipelines over other state-of-the-art methods for high-resolution protein structure predictions.

3.1.2 Structure-Based Function Annotations

To annotate the biological functions of proteins, we first developed a high-quality protein function database, BioLiP [36], semi-manually curated from databases and PubMed literature. Two complementary approaches, COFACTOR [37, 38] and COACH [39], were then proposed to predict the protein functions by structurally matching the prediction structure models with the known proteins in BioLiP.

Development of protein functional databases. Many structure-based protein function analyses and prediction studies use known proteins solved in the PDB [40] as templates to infer biological functions of unknown proteins. However, numerous proteins in the PDB contain redundant entries and/or misordered residues and functions. In particular, many proteins were solved using artificial molecules as additives to facilitate the structural determination experiments. These ligands do not necessarily represent biologically relevant binding. Therefore, it is essential to develop cleaned protein libraries with biological functions carefully validated. We proposed a hierarchical procedure, which consists of three steps of computational filtering and manual literature validation for assessing the biological relevance of the annotated protein functions.

1. Download 3D structure for each entry in the PDB, with the modified residues (i.e., residues modified post-translationally, enzymatically, or by design) translated to their precursor standard residues based on the MODRES record. To exclude crystallization neighbors, the biological unit files rather than the asymmetric unit files are used for evaluating the ligand-protein contacts.
2. Extract ligands, which are defined as small molecules, from the PDB file. Three types of ligand molecules are collected in the BioLiP database, the molecules from the HETATM records (excluding water and modified residues), small DNA/RNA, and peptides with less than 30 residues. If the closest interatomic distance between two HET group ligands is smaller than 2 Å, the two ligands are merged as a single ligand and are regarded as a *k*-mer ligand.
3. Submit each ligand molecule to a composite automated and manual procedure to decide its biological relevance. If the ligand molecule is evaluated as biologically relevant, its interaction with the receptor (i.e., binding site residues in the receptor) is deposited into the BioLiP database. Additionally, the ligand-binding affinity, catalytic site residues, EC numbers, GO terms, and cross-links to the PDB, UniProt, PDBsum, PDBe, and PubMed databases are also collected and deposited into BioLiP.

BioLiP is updated weekly and is freely available for the community at http://zhanglab.ccmb.med.umich.edu/BioLiP. The current release of BioLiP contains 344,990 entries constructed from 72,005 unique PDB proteins, in which 40,078 entries are for DNA/RNA-protein interactions, 15,648 for peptide-protein interactions, 94,907 for metal ion-protein interactions, and 184,357 for regular small molecule-protein interactions. There are in total 23,492 entries with binding affinity data collected from Binding MOAD [41], PDBbind-CN [42], and BindingDB [43] databases and from a manual survey of the literature. It also contains proteins of known enzyme commission (EC) [44] and gene ontology (GO) [45]. Currently, the EC domain of BioLiP involves 7674 protein chains with 203 unique first three-digit and 1900 unique four-digit enzyme commission numbers. The GO domain contains 26,004 chains/domains associated with 11,686 unique gene ontology terms. These data provide important resources for function annotation studies.

COFACTOR method for EC, GO, and ligand-binding predictions. COFACTOR is a template-based approach developed for structure-based function annotation [37, 38].

1. Identify functional templates by global and local structural matches of the target proteins with the known proteins in the BioLiP [36]. The target structures can be either from experimental determination or computational structure prediction.
2. Infer functional insights, including ligand-binding sites, enzyme commission number, and gene ontology terms, from the highest scoring function templates.

COFACTOR was tested in the function prediction section of the ninth CASP experiment, where COFACTOR (as "I-TASSER_FUNCTION" in the server section and "ZHANG" in the human section) was ranked in the first two positions [46].

COACH protocol for ligand-binding prediction. Because individual methods can only generate predictions for specific protein targets, we proposed a new protocol, COACH, which aims to extend the coverage of function predictions by combining results from multiple methods (Fig. 2) [39]. The pipeline consists of two steps:

1. Use two complementary algorithms TM-SITE and S-SITE to derive functional templates from BioLiP library. TM-SITE was designed to recognize functional templates by binding-specific substructure comparison combined with biochemical feature alignment of binding pockets. S-SITE uses protein sequence profile collected by PSI-BLAST that is then searched through BioLiP.

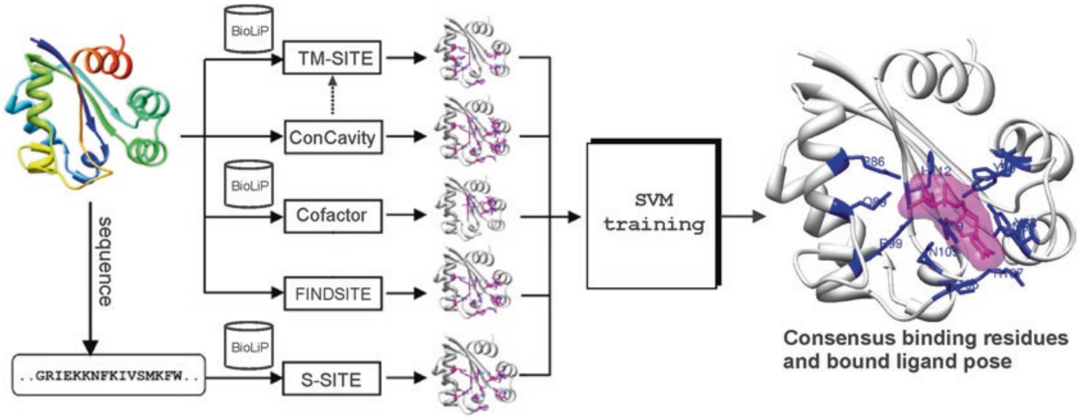


Fig. 2 Protocol of COACH for structure-based protein function annotation. Prediction results from five different programs are combined using support vector machine to increase accuracy and coverage of function annotations

2. Combine multiple predictions by TM-SITE and S-SITE, together with three other predictors (COFACTOR [38], FINDSITE [47], and Concavity [48]), to derive consensus function annotations using support vector machine (Fig. 2).

The COACH protocol was examined in the recent community-wide COMEO experiment designed to test ligand-binding predictions using prereleased PDB sequences as a continuous base [49]. COACH was consistently ranked as the best method in the last 22 individual datasets with its average AUC score, the area under the curve of the true positive rate versus false positive rate plot, 22.5 % higher than the second best method [49].

3.1.3 Structure-Based Function Annotations of Alternative Splice Proteins Expressed in Her2/Neu-Induced Breast Cancers

Identification of splice variant peptides in tumor tissue of mice with her2/neu-amplified breast cancer. We analyzed tumor and normal mammary tissue LC-MS/MS datasets from the Chodosh mouse model of Her2/neu-driven breast cancer, which accounts for 15–20 % of breast cancers in humans [50]. A total of 608 distinct alternative splice variants, 540 known and 68 novel, were identified [3]. There were 216 more from the tumor lysate than from the normal sample (505 vs. 289), probably reflecting greater cellularity and higher expression per cell. We chose 32 of the 45 novel proteins expressed only in tumor specimens for confirmation with qRT-PCR; all were confirmed except for one primer which did not work, and 29 of 31 showed increased mRNA expression. Of the 15 biomarker candidates that Whiteaker et al. [50] confirmed as overexpressed in tumor lysates with MRM-MS, we found that 10 had splice variants in our analysis, although we had no information on the functional activities of the different isoforms of these or any other proteins from proteomic analyses.

Among the 68 novel proteins, we demonstrated variants resulting from new translation start sites, new splice sites, extension or shortening of exons, deletion or swap of exons, retention of introns, and translation in an alternative reading frame. Our annotations revealed multiple variants with potential significant functional motifs, including two relating to BRCA1 through binding to its BRCT domain. The peptide sequence "FSRAEAEGPGQACPPRPFPC" is in the second intronic region of leucine zipper-containing LF (Rogdi) gene. Using Splice Site Prediction by Neural Network from the Berkeley Drosophila Genome Project (http://w.fruitfly.org/seq_tools/splice.html), we found a predicted donor splice site "gactgaggtgaggtg" where the novel peptide was identified as coding sequence with a splice site prediction score of 0.93. Functional motifs identified in expressed intronic sequences include LIG_BRCT_BRCA1_1, a phosphopeptide motif which interacts directly with the carboxy-terminal domain of BRCA1. The peptide "GSGLVPTLGRGAETPVSGAGATRGLSR" aligned to the first intronic region of transcription factor *sox7*; the very same LIG-BRCT_BRCA1_1 motif was found in this intronic region.

Structure modeling-based analysis on alternative splice proteins in Her2/neu-induced breast cancers. Experimentally determined structures of protein splice isoforms remain rare; in fact, there were only seven full-length pairs of such isoforms in the enormous Protein Data Bank (PDB) and the Alternative Splicing and Transcript Diversity (ASTD) database as of 2011 [5]. Homology modeling methods are poor at predicting atomic-level structural differences because of the high sequence identity between the isoforms. We exploited the bioinformatics tools [6, 11, 38, 39] described above (Subheading 3.1) to analyze the folding, structural conformation, and likely functional consequences of alternative splicing of proteins identified in the Her2/neu-induced breast cancer model [5]. The procedure of the application consists of three steps:

1. Based on the I-TASSER modeling, we have demonstrated that attributes in ab initio structural assembly and template refinement can partially differentiate atomic details of splice protein variant pairs [5]. The structure modeling approach was benchmarked on the seven pairs of protein splice isoforms with solved structures in PDB, which resulted in structural models with an average RMSD = 1.72 Å to the native, after excluding all homologous templates to the targets. Some of the structural variations in the isoform pairs were due to exon swapping. Even alternative splice variants whose structures are very similar may have functional differences due to the absence of a functionally critical residue or altered post-translational modifications of residues in the swapped exon. For example, in the case of acid phosphatase (acp1) variants, the Mg^{2+}-binding site is missing in the 1xwwA variant.
2. We used the strategy to model three cancer-related variant pairs reported to have opposite functions, but lacking experimentally derived structures: Bcl-x, caspase 3, and odd-skipped related 2. In each isoform pair, we observed structural differences in regions where the presence or absence of a motif can directly influence the distinctive functions of the variants. For example, an additional 63 amino acids (AA 129–191) create an extra domain in the core structure of bclx-L (233 AA) compared with bclx-S (170 AA); the shorter variant is missing the two Bcl-2 family motifs BH1 and BH2, while the longer variant contains all four Bcl-2 homology motifs (BH1–4). This difference results in a completely different topology and function; bclx-L is antiapoptotic, while bclx-S is proapoptotic.
3. We applied I-TASSER to five splice-variant pairs overexpressed in the mouse Her2/neu mammary tumor: annexin 6, calumenin, cell division cycle 42 (cdc42), polypyrimidine tract-binding protein 1 (ptbp1), and tax1-binding protein 3 (tax1bp3). These pairs were chosen based on the following five criteria: differential expression, annotated as a known protein in Ensembl,

at least 75 % sequence identity with the canonical protein, known homologous variants of the protein pair in *Homo sapiens*, and an I-TASSER confidence score (C-score) for both variants >−1.5 to ensure the quality of structure prediction.

Despite the high sequence identity between the variant pairs (99, 92, 95, 95, and 79 %, respectively), structural differences were revealed in biologically important regions of these protein pairs. For example, the only difference between anxa6-001 and anxa6-002 at the sequence level is the presence of six residues in anxa6-001 (VAAEIL, AA 525–530) that are missing in anxa6-002. The global topology of the I-TASSER models of the two isoforms is almost identical, with RMSD = 0.38 Å and TM-score = 0.99. However, there is an obvious local structural change in the region due to the absence of "VAAEIL" residues (AA 525−530 in anxa6-001), as identified by the structural alignment algorithm, TM-align [51]. As reported, these six residues are in the end of a helical region (blue-colored in the original figure) which is followed by a loop. Because of the absence of the six residues, the loop is smaller in the shorter variant. The nearby proline-directed kinase phosphorylation ([ST]P) site followed by a serine phosphorylation site moves from 535–537 to 529–531, inside the helix region in anxa6-002, where phosphorylation is less probable than for anxa6-001.

The I-TASSER models in Fig. 3 show that the threonine and proline residues are buried by other atoms in the anxa6-002

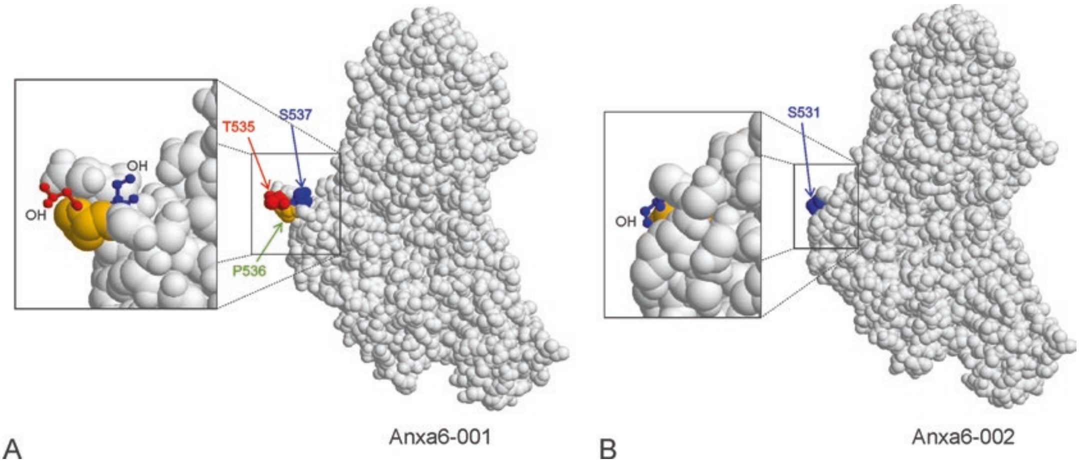


Fig. 3 The I-TASSER models for two splice isoforms of Annexin 6. (**a**) The "TPS" residues in anxa6-001 are exposed to solvent which helps increase the likelihood of phosphorylation as a post-translational modification. The hydroxyl groups, which are the target of kinases for phosphorylation, are highlighted in the inset. (**b**) Due to the absence of "VAAEIL" residues (aa 525–530 in anxa6-001) in the anxa6-002 variant, the "TPS" residues are either partially or completely buried by other atoms which significantly reduce the possibility of the protein for phosphorylation [Modified from reference [1]]

variant, whereas the hydroxyl group of serine (S531 in anxa6-002) seems quite accessible for phosphorylation (see inset of Fig. 3b). In order to search for phosphopeptides from the spliced region of anxa6, we performed a fresh analysis of the mass spectrometric data with our custom database using X! Tandem, specifying phosphorylation on serine or threonine (phospho(S) and phospho(T)) as potential residue modifications. Because phosphorylation is usually present at low stoichiometry and our dataset was not enriched for phosphopeptides, it was striking that we identified a spectrum from the normal sample that matched to the peptide "DQAQED AQVAAEILEIADTPSGDKTSLETR" with 3281.5 daltons as the calculated peptide mass plus a proton (mh). The unmodified mh of this peptide is 3201.5 daltons; the additional 80 daltons can be accounted for precisely by phosphorylation of either the threonine or serine residue in the peptide. We did not find such a phosphopeptide from the tumor sample. However, we did find multiple high-quality spectra from the tumor sample that identified the sequence "DQAQEDAQEIADTPSGDKTSLETR," the unique peptide that matches the anxa6 short variant (with residues "VAAEIL" missing). None of these spectra revealed modification by phosphorylation. Even though only a single spectrum identified the phosphopeptide from the unique region of the long anxa6 variant in the X! Tandem search, these observations are consistent with our functional inference from the structural comparison of the anxa6 variants [5] that the longer anxa-001 variant is more prone to undergo phosphorylation at Thr-535 or Ser-537 than in the anxa-002 variant at the Thr-529 or Ser-531 sites. Post-translational phosphorylation of anxa6 has been reported to be associated with cell growth in 3T3 fibroblasts and human T-lymphoblasts [52]; we previously predicted that the critical phosphorylation may occur at Thr-535 and/or Ser-537 in the loop region. We have now strikingly refined this prediction, which we hope experimentalists will test.

3.2 Methods for Annotation of Protein Isoform Structure and Predicted Functions, Using Support Vector Machine Multiple-Instance Learning to Predict Functions of Isoforms and Isoform-Level Networks

The prediction of isoform functions and networks can be formulated as a problem that can be addressed by multiple-instance learning (MIL) algorithms [53]. MIL works mainly in three steps (Fig. 4):

1. Initialization: label all isoforms in positive genes as Class 1 and all isoforms in negative genes as 0.
2. Model building: build a classifier using a given method such as support vector machines and Bayesian networks followed by using this classifier to score all isoforms.
3. Witness updating: reselect the highest scored isoform from positive genes as "witness," and label them as Class 1. All other isoforms are labeled as Class 0. If results are not converged, go to **step 2**. Otherwise, the converged model is stored for predicting isoform functions.

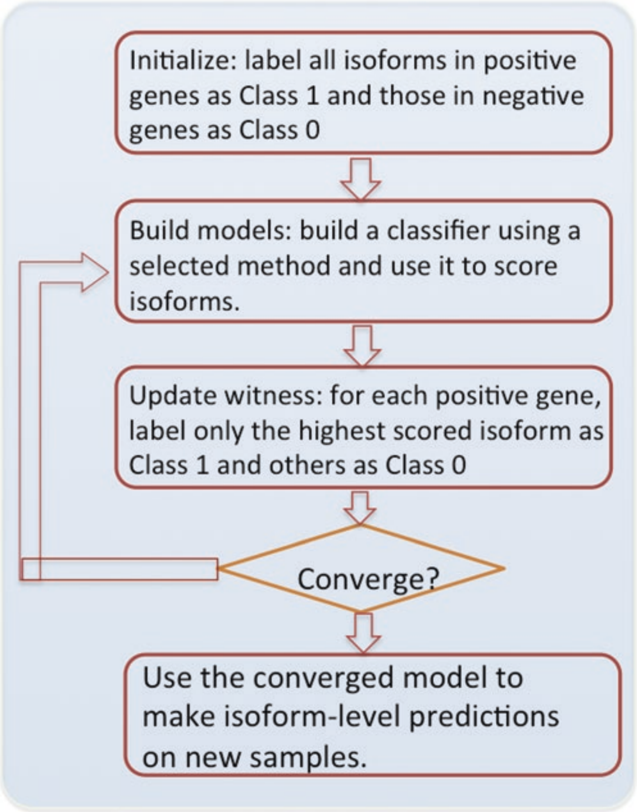
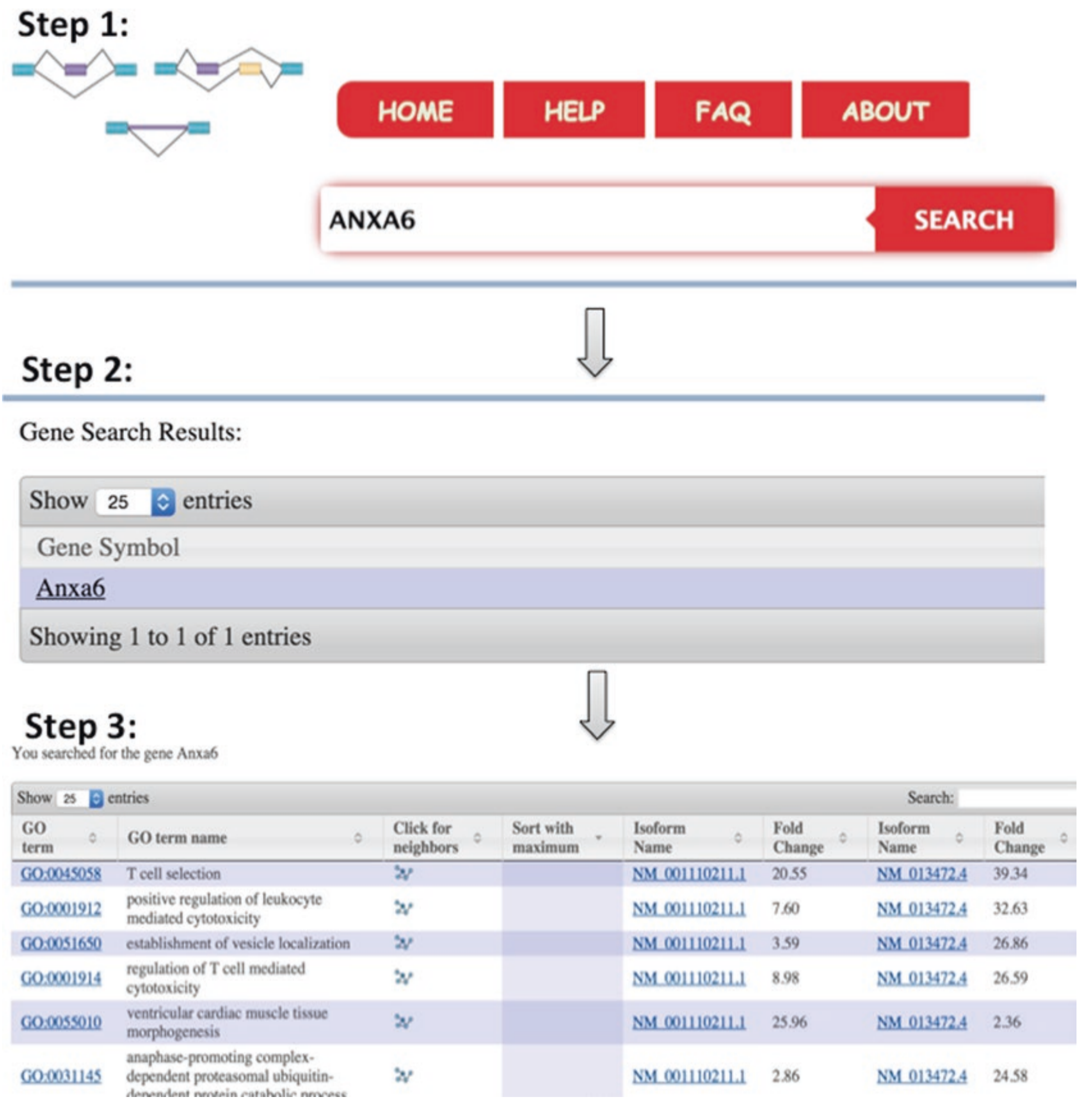

Initialize: label all isoforms in positive genes as Class 1 and those in negative genes as Class 0
Build models: build a classifier using a selected method and use it to score isoforms.
Update witness: for each positive gene, label only the highest scored isoform as Class 1 and others as Class 0
Converge?
Use the converged model to make isoform-level predictions on new samples.

Fig. 4 The schematic of iterative multiple-instance learning (MIL) for isoform-level function prediction

3.2.1 Genome-Wide Isoform Functions and Networks in the Mouse

Eksi et al. performed the first genome-wide isoform function predictions using the MIL algorithm for the mouse [53]. The method is described in the following steps:

1. Collect isoform feature data. A total of 811 RNA-Seq experiments (samples) were downloaded from the NCBI Sequence Read Archive (SRA) database (http://www.ncbi.nlm.nih.gov/sra), followed by estimating isoform expression using the TopHat and Cufflinks suites (v2.0.051). Quality control was conducted [53], and 365 samples were kept for subsequent analysis. All isoform expression from these 365 samples was collected into a data matrix.
2. Select "gold standard" of gene functions. Biological process terms in GO were used as "gold standard" for functional annotation.
3. Learn a support vector machine (SVM) model from the above collected data using the MIL algorithm (Fig. 4).
4. Use the learned SVM model to predict functions for all mouse isoforms.
5. Build a web server to make isoform functions publicly available and searchable (http://guanlab.ccmb.med.umich.edu/isoPred).



Gene Search Results:

Show 25 entries

Gene Symbol
Anxa6

Showing 1 to 1 of 1 entries

You searched for the gene Anxa6

Show 25 entries Search:

GO term	GO term name	Click for neighbors	Sort with maximum	Isoform Name	Fold Change	Isoform Name	Fold Change
GO:0045058	T cell selection			NM_001110211.1	20.55	NM_013472.4	39.34
GO:0001912	positive regulation of leukocyte mediated cytotoxicity			NM_001110211.1	7.60	NM_013472.4	32.63
GO:0051650	establishment of vesicle localization			NM_001110211.1	3.59	NM_013472.4	26.86
GO:0001914	regulation of T cell mediated cytotoxicity			NM_001110211.1	8.98	NM_013472.4	26.59
GO:0055010	ventricular cardiac muscle tissue morphogenesis			NM_001110211.1	25.96	NM_013472.4	2.36
GO:0031145	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process			NM_001110211.1	2.86	NM_013472.4	24.58

Fig. 5 An example query of *Anxa6* gene from the mouse isoform function database

Taking *Anxa6* (Annexin A6) as an example, the steps for querying the functions of its isoforms are described below (also shown in Fig. 5).

1. Go to website http://guanlab.ccmb.med.umich.edu/isoPred, type in "Anxa6" in the input box, and click the "Search" button (step 1, Fig. 5).
2. Users will be guided to http://guanlab.ccmb.med.umich.edu/isoPred/search_result.php. Locate "Anxa6" in the form on this page, and click it (step 2, Fig. 5).

3. Users will be guided to http://guanlab.ccmb.med.umich.edu/isoPred/gene_result.php?gene=Anxa6. On this page, see function predictions for each isoform presented in the table (step 3, Fig. 5).

Further, Li et al. constructed a genome-wide functional relationship network for the mouse [54, 55] with the following steps:

1. Collect isoform-pair feature data: RNA-Seq, exon array, pseudo-amino acid composition (pseudoAAC), and protein-docking data. For RNA-Seq and exon array, isoform expression was estimated followed by calculating isoform correlation as feature data. Correlation was also calculated between isoforms using their pseudoAAC profile. Protein-docking itself is an isoform-pair feature.
2. Construct "gold standard" of functionally related gene pairs. GO biological processes and KEGG and BioCyc pathways were used to construct a "gold standard" of functionally related gene pairs. Genes annotated to the same biological process or GO term are assumed to have a functional relationship.
3. Learn a model using MIL with Bayesian network as the base learner.
4. Use the learned model to make genome-wide predictions of functional relationship between any two isoforms.
5. Build a web server to allow users to search isoform networks. It is publicly available at http://guanlab.ccmb.med.umich.edu/isoformnetwork.

Users can go to this server, input their gene(s) of interest, click the "Search" button, and see isoform networks along with GO enrichment results.

Based on the mouse isoform network, Li et al. catalogued the highest connected isoforms (HCIs) as a predicted "canonical isoform" using the following approach [55].

1. Calculate an average functional relationship (AFR) score for each isoform of multi-isoform genes.
2. For each multi-isoform gene, choose the isoform with the highest AFR score as HCI. The remaining isoforms are considered as NCI (non-highest connected isoforms).
3. Use independent RNA-Seq and proteomic data to investigate the expression of HCI at both transcript and protein levels.
4. Identify a set of genes whose HCIs are most expressed at transcript level and are also expressed at protein level.

Further, the MIsoMine database was developed to provide an integrated platform for analyzing isoform expression, functions, and networks for the mouse [56]. Users can go to the website (http://guanlab.ccmb.med.umich.edu/misomine/) to perform isoform-level analyses.

3.2.2 Genome-Wide Isoform Functions and Networks in the Human

Panwar et al. predicted functions for splice isoforms in humans [57]. The approach mainly consists of the following steps:

1. Download the human RNA-Seq data from the ENCODE study [58]; 127 samples were used. The TopHat and Cufflinks suites were used to estimate isoform expression in terms of FPKM [59] using the Ensembl gene annotation (version74, available at http://www.ensembl.org/).
2. Use gene ontology biological processes to construct "gold standard" functional annotations.
3. Build an SVM model using the MIL algorithm, and use the model to predict the functions of all human isoforms.
4. Build a web server to store all the predictions and to make the predictions searchable (http://guanlab.ccmb.med.umich.edu/isofunc/).

In addition to isoform functions, Li et al. built a genome-wide function relationship network at the isoform level for the human [60], an effort from the chromosome 17 Human Proteome Project [61]. The pipeline is described below:

1. Collect four types of isoform-level feature data, including RNA-Seq expression, pseudo-amino acid composition, protein-docking score, and conserved domains.
2. Construct a "gold standard" of functionally related gene pairs using GO biological processes and KEGG pathways.
3. Build a Bayesian network classifier using the MIL algorithm, and use the model to make genome-wide predictions of function relationships between isoforms.
4. Build a web server to store the human isoform network (http://guanlab.ccmb.med.umich.edu/hisonet). Users are able to query isoform networks of their genes of interest.

In addition, using the same method described in Subheading 3.2.1, Li et al. catalogued a set of HCIs for the human.

3.2.3 Comparison of Isoform Functions and Networks in Mice and Humans

Li et al. compared the HCIs between mouse and human to investigate whether they are conserved:

1. Choose a set of 306 multi-isoform homologous genes between mouse and human. For each of these genes, denote its mouse and human HCI as HCI_m and HCI_h.
2. Identify 61 of the 306 genes whose HCI_m and HCI_h are homologs based on the HomoloGene database in NCBI (http://ncbi.nlm.nih.gov/homologene).

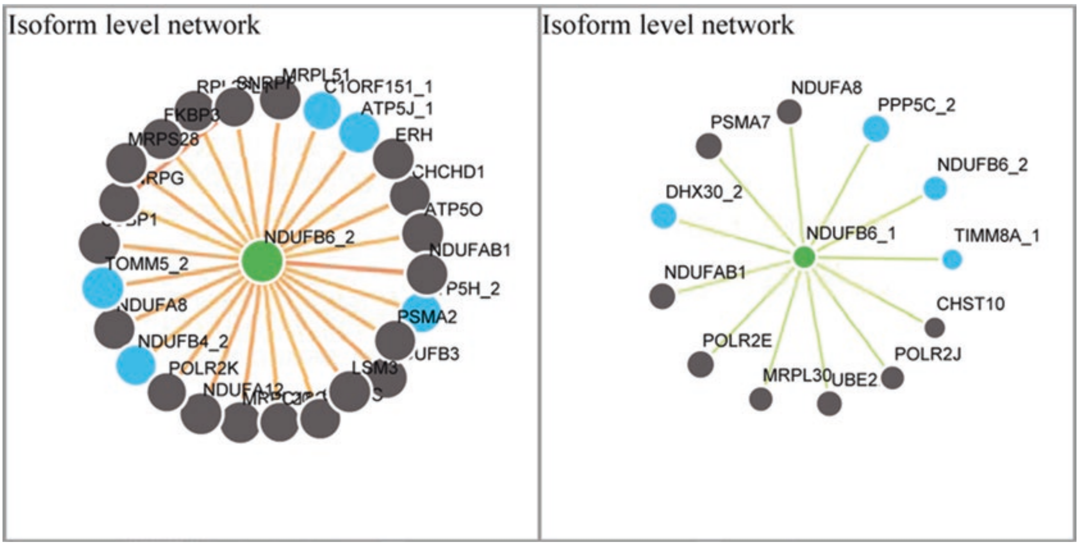


Fig. 6 The functional networks of the highest connected isoform (HCI) (NM_002493.4, NDUFB6_2) and non-highest connected isoform (NCI) (NM_182739.2, NDUFB6_1) of the NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 6 (NDUFB6) gene

3. Computationally generate a null distribution of the number of genes whose HCI_m and HCI_h are homologs by chance (41 ± 6).
4. Calculate a p-value by comparing the observed value [61] to the null distribution.

The overlap between mouse and human HCI is significant (p = 0.0003), providing additional evidence supporting the "canonical" features of HCIs. As an example, *NUDFB6* is a gene encoding two splice isoforms: HCI (NM_002493.4, NDUFB6_2) and NCI (NM_182739.2, NDUFB6_1); their networks are shown in Fig. 6. The HCI, but not the NCI, was reported to be expressed at the protein level in normal human retina [60], further supporting the "canonical" characteristics of HCIs.

3.2.4 Examples from the ERBB2 Amplicon and Pathways

The ERBB2 (HER2) gene is an epidermal growth factor (EGF) receptor of the receptor tyrosine kinase family. The protein encoded by this gene functions by forming a heterodimer through binding with ERBB1, ERBB3, or ERBB4 receptor proteins. Amplification or overexpression of this gene has been shown to be strongly associated with a major subset of human breast cancers. One can explore the isoform functions and networks for this gene:

1. Isoform functions. From the IsoFunc database (http://guanlab.ccmb.med.umich.edu/isofunc/), identify the functions of the five protein-coding splice isoforms of ERBB2 based on Ensembl annotation (version 74). For example, ENST0000\0541774 was

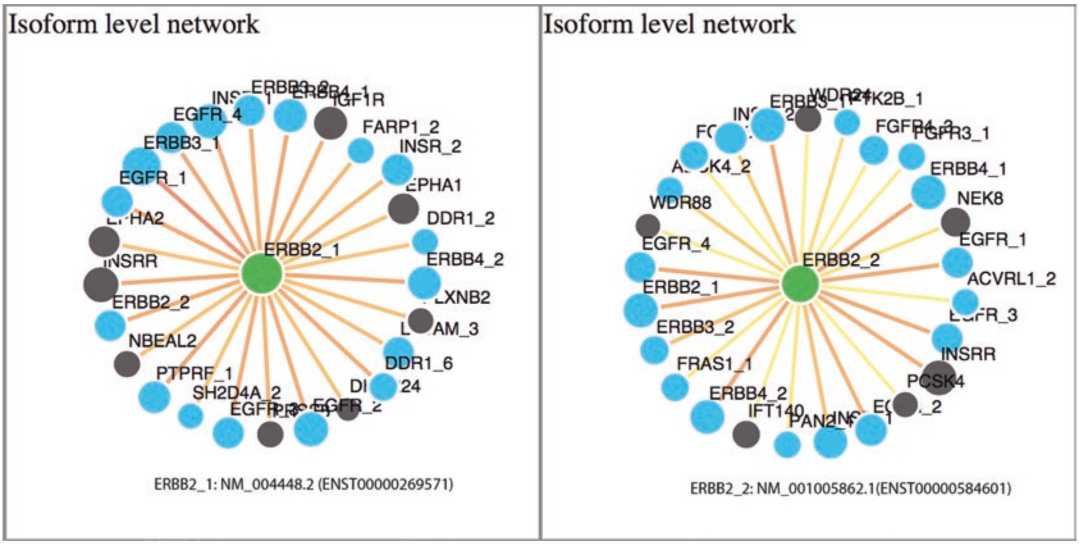


Fig. 7 Isoform networks of ERBB2_1 (NM_004448.2) and ERBB2_2 (NM_001005862) of the human ERBB2 gene

predicted to carry the "canonical Wnt signaling pathway" with a fold change = 37. In contrast, the likelihood for the other isoforms to have this function is much smaller (fold change ranging from 1.7 to 3.3). ENST00000445658, ENST00000269571, ENST00000584601, and ENST00000406381 are predicted to most likely function in "sterol biosynthetic process," "extracellular matrix disassembly," "extracellular matrix disassembly," and "cell-substrate junction assembly," respectively. These predictions suggest the potential functional differences between isoforms, though they need to be experimentally validated.

2. Isoform networks. To explore further the functional interactions, obtain the isoform networks of ERBB2 isoforms from http://guanlab.ccmb.med.umich.edu/hisonet/. This server provides networks for two isoforms, NM_004448.2(ENST00000269571) and NM_001005862.1(ENST00000584601), which are shown in Fig. 7. Both networks are enriched for GO biological processes such as "phosphoinositide 3-kinase cascade" and "protein amino acid autophosphorylation," suggesting their functional similarities. Functional differences were also predicted. For instance, GO term "regulation of MAP kinase activity" was enriched in the network of NM_004448.2 but not NM_001005862.1, suggesting possible different roles/extent for ERBB2 isoforms to be involved in the MAP kinase signaling pathway.

3.3 Concluding Remark

The combination of proteomics and transcriptomics with the bioinformatics algorithms and methods of structural biology and functional relationship networks can generate many new insights and provide testable hypotheses for experimental studies.

4 Notes

There are a variety of computational tools that have been developed in the field, including, e.g., Rosetta [62], HHsearch [63], and Modeller [64] for protein structure prediction and Concavity [48], FINDSITE [47], and ProFunc [65] for structure-based function annotations. While the I-TASSER family tools represent one of the most efficient sets of methods as demonstrated in various community-wide structure and functional modeling experiments [29, 34, 46, 66], it is important to remember that the results are predictions from automated computational programs. The accuracy and confidence of the models vary among different proteins, depending on the availability of homologous templates and size of the target sequences. We have developed two confidence scores to guide their use by biologist users.

First, C-score [27] is a measurement of confidence of protein structure models built by I-TASSER [11] and QUARK [12] programs. It was defined based on the significance score of structure templates identified by threading alignments and the structural density of Monte Carlo-based conformational search. A large-scale benchmark experiment based on 500 nonredundant proteins showed that there is a high correlation between the C-score and TM-score of the predicted models, with a Pearson correlation coefficient = 0.91 [27].

Second, we proposed an F-score [38] to estimate the accuracy of structure-based function predictions by COFACTOR [37] and COACH [39]. The F-score was defined based on the C-score of protein structure predictions and the structural and sequence similarities between the target and template proteins. A positive correlation between F-score and the accuracy of the predicted models was found in both COFACTOR and COACH predictions.

The I-TASSER family tools have been designed to predict protein structure and functions from the primary sequences. However, information from experimental data or human-based functional analyses can be of critical importance to improve the accuracy of the modeling. The on-line servers and downloadable packages of the I-TASSER family tools have provided entries that allow users to conveniently introduce experimental constraints, including contact and distance maps and specific template alignments, to the modeling systems.

There are multiple factors that would affect the prediction of functions and networks and thus subsequent comparisons, such as choice of gene annotation software for estimating isoform expression. For example, the predictions of human isoform functions [57] and networks [60] are based on RefSeq (version 37.2) and Ensembl (version 74) gene annotations, respectively, so preliminary interpretation of comparative results should be viewed with caution. RefSeq annotation is of high quality but is much less complete compared to Ensembl, which contains many more (predicted) genes and isoforms. This annotation difference will affect the estimation of splice isoform expression and the subsequent prediction of functions and networks. Also, note that Hisonet provides functions based on GO enrichment, which is different from the directly predicted isoform functions in IsoFunc.

References

1. Omenn GS, Menon R, Zhang Y (2013) Innovations in proteomic profiling of cancers: alternative splice variants as a new class of cancer biomarker candidates and bridging of proteomics with structural biology. J Proteomics 90:28–37
2. Menon R, Panwar B, Eksi R, Kleer C, Guan Y, Omenn GS (2015) Computational inferences of the functions of alternative/noncanonical splice isoforms specific to HER2+/ER-/PR- breast cancers, a chromosome 17 C-HPP study. J Proteome Res 14(9):3519–3529
3. Menon R, Omenn GS (2010) Proteomic characterization of novel alternative splice variant proteins in human epidermal growth factor receptor 2/neu-induced breast cancers. Cancer Res 70(9):3440–3449
4. Patro R, Mount SM, Kingsford C (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol 32(5): 462–464
5. Menon R, Roy A, Mukherjee S, Belkin S, Zhang Y, Omenn GS (2011) Functional implications of structural predictions for alternative splice proteins expressed in Her2/neu-induced breast cancers. J Proteome Res 10(12):5503–5511
6. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5(4):725–738
7. Wu S, Skolnick J, Zhang Y (2007) Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol 5:17
8. Zhang Y (2007) Template-based modeling and free modeling by I-TASSER in CASP7. Proteins 69(S8):108–117
9. Zhang Y (2009) I-TASSER: Fully automated protein structure prediction in CASP8. Proteins 77(S9):100–113
10. Zhang Y (2014) Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. Proteins 82(Suppl 2):175–187. doi:10.1002/prot.24341.
11. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y (2015) The I-TASSER Suite: protein structure and function prediction. Nat Methods 12(1):7–8
12. Xu D, Zhang Y (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 80(7):1715–1735
13. Xu, D, Zhang, Y (2012) Towards optimal fragment generations for ab initio protein structure assembly. Proteins. 10.1002/prot.24179.
14. Xu D, Zhang J, Roy A, Zhang Y (2011) Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. Proteins 79(Suppl 10):147–160
15. Wu S, Zhang Y (2007) LOMETS: a local meta-threading-server for protein structure prediction. Nucl Acids Res 35:3375–3382
16. Zhang Y, Kolinski A, Skolnick J (2003) TOUCHSTONE II: a new approach to ab initio protein structure prediction. Biophys J 85:1145–1164
17. Zhang Y, Skolnick J (2004) SPICKER: a clustering approach to identify near-native protein folds. J Comput Chem 25(6):865–871
18. Swendsen RH, Wang JS (1986) Replica Monte Carlo simulation of spin glasses. Phys Rev Lett 57(21):2607–2609
19. 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(3):665–676
20. Zhang Y (2014) Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. Proteins 82(Suppl 2):175–187
21. Wu S, Zhang Y (2008) A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. Bioinformatics 24(7):924–931
22. Wu S, Szilagyi A, Zhang Y (2011) Improving protein structure prediction using multiple sequence-based contact predictions. Structure 19(8):1182–1191

23. Wu S, Zhang Y (2010) Recognizing protein substructure similarity using segmental threading. Structure 18(7):858–867
24. Zhang J, Liang Y, Zhang Y (2011) Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. Structure 19(12):1784–1795
25. Xu D, Zhang Y (2013) Toward optimal fragment generations for ab initio protein structure assembly. Proteins 81(2):229–239
26. Xu D, Zhang Y (2011) Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. Biophys J 101(10):2525–2534
27. Zhang Y (2008) I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9:40
28. Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T (2007) Assessment of CASP7 predictions for template-based modeling targets. Proteins 69(S8):38–56
29. Battey JN, Kopp J, Bordoli L, Read RJ, Clarke ND, Schwede T (2007) Automated server predictions in CASP7. Proteins 69(S8):68–82
30. Cozzetto D, Kryshtafovych A, Fidelis K, Moult J, Rost B, Tramontano A (2009) Evaluation of template-based models in CASP8 with standard measures. Proteins 77(Suppl 9):18–28
31. Moult J, Pedersen JT, Judson R, Fidelis K (1995) A large-scale experiment to assess protein structure prediction methods. Proteins 23(3):ii–iv
32. Moult J (2005) A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 15(3):285–289
33. Mariani V, Kiefer F, Schmidt T, Haas J, Schwede T (2011) Assessment of template based protein structure predictions in CASP9. Proteins 79(Suppl 10):37–58
34. Montelione GT (2012) Template based modeling assessment in CASP10. Paper presented at the 10th community wide experiment on the critical assessment of techniques for protein structure prediction, Gaeta, Italy, 9–12 Dec 2012
35. Kinch LN, Li W, Monastyrskyy B, Kryshtafovych A, Grishin NV (2016) Evaluation of free modeling targets in CASP11 and ROLL. Proteins 84(Suppl 1):51–66. doi:10.1002/prot.24973.
36. Yang J, Roy A, Zhang Y (2013) BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. Nucleic Acids Res 41(D1):D1096–D1103
37. Roy A, Zhang Y (2012) Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. Structure 20(6):987–997
38. Roy A, Yang J, Zhang Y (2012) COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucleic Acids Res 40(Web Server issue):W471–W477
39. Yang J, Roy A, Zhang Y (2013) Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinformatics 29(20):2588–2595
40. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, Prlic A, Quesada M, Quinn GB, Westbrook JD, Young J, Yukich B, Zardecki C, Berman HM, Bourne PE (2011) The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res 39(Database issue):D392–D401
41. Benson ML, Smith RD, Khazanov NA, Dimcheff B, Beaver J, Dresslar P, Nerothin J, Carlson HA (2008) Binding MOAD, a high-quality protein-ligand database. Nucleic Acids Res 36(Database issue):D674–D678
42. Cheng T, Li X, Li Y, Liu Z, Wang R (2009) Comparative assessment of scoring functions on a diverse test set. J Chem Inf Model 49(4):1079–1093
43. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res 35(Database issue):D198–D201
44. Barrett AJ (1997) Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme Nomenclature. Recommendations 1992. Supplement 4: corrections and additions (1997). Eur J Biochem 250(1):1–6
45. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1):25–29
46. Schmidt T, Haas J, Gallo Cassarino T, Schwede T (2011) Assessment of ligand-binding residue predictions in CASP9. Proteins 79(Suppl 10):126–136
47. Brylinski M, Skolnick J (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. Proc Natl Acad Sci U S A 105(1):129–134
48. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput Biol 5(12):e1000585

49. Schwede T (2015) Montly summary of ligand binding prediction results in CAMEO is at http://www.cameo3d.org/lb.
50. Whiteaker JR, Zhang H, Zhao L, Wang P, Kelly-Spratt KS, Ivey RG, Piening BD, Feng LC, Kasarda E, Gurley KE, Eng JK, Chodosh LA, Kemp CJ, McIntosh MW, Paulovich AG (2007) Integrated pipeline for mass spectrometry-based discovery and confirmation of biomarkers demonstrated in a mouse model of breast cancer. J Proteome Res 6(10):3962–3975
51. Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33(7):2302–2309
52. Moss SE, Jacob SM, Davies AA, Crumpton MJ (1992) A growth-dependent post-translational modification of annexin VI. Biochim Biophys Acta 1160(1):120–126
53. Eksi R, Li H-D, Menon R, Wen Y, Omenn GS, Kretzler MK, Guan Y (2013) Systematically differentiating functions for alternatively spliced isoforms through integrating RNA-seq data. PLoS Comput Biol 9(11):e1003314
54. Li H-D, Menon R, Eksi R, Guerler A, Zhang Y, Omenn GS, Guan Y (2013) Modeling the functional relationship network at the splice isoform level through heterogeneous data integration. bioRxiv:doi: 10.1101/001719.
55. Li H-D, Menon R, Omenn GS, Guan Y (2014) Revisiting the identification of canonical splice isoforms through integration of functional genomics and proteomics evidence. Proteomics 14(23–24):2709–2718
56. Li H-D, Omenn GS, Guan Y (2015) MIsoMine: a genome-scale high-resolution data portal of expression, function and networks at the splice isoform level in the mouse. Database 2015. doi: 10.1093/database/bav1045.
57. Panwar B, Menon R, Eksi R, Li H-D, Omenn GS, Guan Y (2015) Genome-wide functional annotation of human protein-coding splice variants using multiple instance learning under revision
58. Consortium EP (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306:636–640
59. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7(3):562–578
60. Li H-D, Menon R, Govindarajoo B, Panwar B, Zhang Y, Omenn GS, Guan Y (2015) Functional networks of highest-connected splice isoforms: from the Chromosome 17 Human Proteome Project. J Proteome Res 14(9):3484–3491
61. Liu SL, Im H, Bairoch A, Cristofanilli M, Chen R, Deutsch EW, Dalton S, Fenyo D, Fanayan S, Gates C, Gaudet P, Hincapie M, Hanash S, Kim H, Jeong SK, Lundberg E, Mias G, Menon R, Mu ZM, Nice E, Paik YK, Uhlen M, Wells L, Wu SL, Yan FF, Zhang F, Zhang Y, Snyder M, Omenn GS, Beavis RC, Hancock WS (2012) A chromosome-centric Human Proteome Project (C-HPP) to characterize the sets of proteins encoded in Chromosome 17. J Proteome Res 12(1):45–57
62. Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 268(1):209–225
63. Soding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21(7):951–960
64. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234(3):779–815
65. Laskowski RA, Watson JD, Thornton JM (2005) ProFunc: a server for predicting protein function from 3D structure. Nucl Acids Res 33(Web Server issue):W89–W93
66. Haas J, Roth S, Arnold K, Kiefer F, Schmidt T, Bordoli L, Schwede T (2013) The Protein Model Portal—a comprehensive resource for protein structure and model information. Database (Oxford) 2013:bat031